Record Linkage Techniques— 1985

Proceedings of the Workshop on
Exact Matching Methodologies
Arlington, Virginia
May 9-10, 1985

Co-Sponsored with the
Washington Statistical Society
and the
Federal Committee on Statistical Methodology

Compiled and Edited by
Beth Kilss and Wendy Alvey

Department of the Treasury
Internal Revenue Service
Statistics of Income Division
December 1985

PREFACE

The Workshop on Exact Matching Methodologies was held on May 9-10, 1985, at the Rosslyn Westpark Hotel in Arlington, Virginia. The conference grew out of the efforts of the Matching Group, Administrative Records Subcommittee, of the Federal Committee on Statistical Methodology. It was co-sponsored with the Washington Statistical Society. This volume contains the papers from that event.

The current volume, Record Linkage Techniques -- 1985, is more than just a proceedings of the May conference. It is intended to serve as a handbook on modern matching theory, as well as to report on the current state of the art. For this reason, not only were the papers from the Workshop included here, but extensive background material and bibliographic citations have also been added.

Contents. -- The format for this volume essentially follows that of the Workshop agenda, with several sections added to help round out the actual presentations. The collection begins with an Introduction, which summarizes the objectives of the Matching Group in conducting a Workshop of this sort. It also proposes some recommendations for the statistical community to consider with regard to the future of exact matching. (This latter portion is based on comments made by the participants during and after the Workshop.)

The rest of the volume is set up as follows:

- o Section I provides selected background material, which lays the historical groundwork for current methodological thought. Some of these papers were distributed at the conference, but they were not presented as part of the agenda.

- o Section II begins the program for the Workshop. This contains three papers presented at the Opening Session, to introduce the theory and provide an overview of matching applications. A fourth (contributed) paper, describing the present state of general methodological issues, is also included.

- o Section III focuses on current theory and practice. It is comprised of three invited papers and their resulting discussions, as well as two related contributed papers.

- o Sections IV and V follow with papers which describe recent application case studies. Once again, two relevant papers have been added to the six invited papers and their discussions presented at the conference.

- o Finally, Section VI deals with computer software for exact matching. It provides the papers presented during the last portion of the Workshop.

The volume also contains two appendices:

- o Appendix A consists of selected bibliographies of exact matching methodologies and applications. Five separate collections of references are provided, each with a slightly different orientation.

- o Appendix B concludes the volume with information specific to the workshop, itself -- the agenda, the list of attendees, and the list of sponsors.

Copy Preparation. -- The contents of the papers included here are the responsibility of the authors. With the exception of previously published background papers, which were simply reproduced as is, all of the papers in this volume underwent only a limited peer review process. Each paper was read by at least one person familiar with the subject matter. It should be noted, however, that reviewers were instructed to focus on editorial concerns and gross factual problems. Since this did not constitute a formal referee process, authors were also encouraged to obtain their own technical review. Corrections and changes were either made by the authors themselves or cleared through them by the editors. Final layout of the papers was done by the editorial staff, with minor changes of a cosmetic nature considered the prerogative of the editors.

Acknowledgments. -- First of all, I would like to thank the Steering Committee -- Maria Gonzalez, Thomas B. Jabine, Matthew Jaro, Nancy Kirkendall, and Carol Utter -- and the Workshop Coordinators -- Beth Kilss and Wendy Alvey -- for their outstanding job in organizing such a successful conference.

Next, I join the editors of this volume in thanking the Workshop participants for their insightful remarks and helpful suggestions regarding both bibliographic references and recommendations for future research. Along those lines, a special thanks goes to Thomas B. Jabine and Nancy Kirkendall for their work in drafting the Matching Group Pecommendations, which appear following the Introduction. Appreciation is also due to all of the reviewers, for providing their thoughtful comments in such a timely manner. Finally, many thanks go out to the typists -- Michelle Cobb, Denise Herbert, Dawn Nester, Denise Reeder, and Susan Rhodes -- and especially, the editorial staff -- Clementine Brittain, Nancy Dutton, and Bettye Jamerson -- for their extensive contributions to the copy preparation process.

December 1985

Fritz Scheuren
Director
Statistics of Income Division

TABLE OF CONTENTS

Selected Background Papers (1959 - 1983)

Page

Overview of Applications and Introduction to Theory

Current Theory and Practice

Page

Application Case Studies I

Application Case Studies II

Page

Appendix B: Workshop Particulars

INTRODUCTION

In June 1984, the Administrative Records Subcommittee recommended to the Federal Committee on Statistical Methodology that a subcommittee be set up to explore integration of surveys and administrative records. The result was the creation of a Matching Group, whose initial (ambitious) goals were to examine policy strategies in conducting data linkages; look at such methodological issues as measurability of matching and analysis of statistical techniques in view of matching errors; and study previous linkage studies for suggestion of possible alternative approaches to matching problems. Both population-based linkages and establishment matches were of interest.

The Matching Group began by reviewing the available literature on exact matching. It soon became apparent that some gaps in knowledge existed that perhaps could be addressed by a workshop conducted by experts currently working in that field. Thus was born the Workshop on Exact Matching Methodologies.

The Workshop was designed to balance the disparate interests of the many different people involved in exact matching: statisticians, research analysts, and computer programmers. Subject matter interests ranged from the epidemiologists' concerns about person-matches to occupation and mortality data, to economists' desires to create estimates based on establishment linkages. As such, the Workshop was viewed as a means of summarizing the work done on matching over the past ten to fifteen years, filling in some of the holes we had discovered, and drawing this more current information together in one place -- this volume -- for use as a ready resource aid by the statistical community and its users. The conference was also seen as a means of building a network of people interested in matching, with a view towards establishing a more coordinated approach to future policy and research efforts.

The Workshop

The Workshop drew 140 registrants from both the U.S. and Canada, representing 47 different agencies, universities, and businesses -- a very sizable segment of the major contributors to the field of exact matching today. Not surprisingly, well over half of those who attended represented Federal agencies; most notably, the Bureau of the Census, Internal Revenue Service, Social Security Administration, Energy Information Administration, and the Bureau of Labor Statistics. Furthermore, about half of those who came expressed primary interest in application issues. The remainder were about equally divided between statistical theory and computational developments.

Workshop Results

In addition to the interactions which took place at the Workshop, there were also several important tangible products which resulted from that effort. First, based on discussions at the Workshop and subsequent correspondence, some recommendations for next steps in exact matching were developed. These were summarized by the Matching Group and appear following this Introduction. Next, selected papers representing the historical development of modern matching methodological thought were assembled. These have been represented here in Section I. Then, in Sections II through VI, presentations from the Workshop and a few additional related papers are provided. Their inclusion is intended to document the current state of exact matching methodology, application and computer software development.

Finally, extensive efforts were made to develop a comprehensive bibliography of exact matching literature. What resulted was a collection of five separate reference lists, each slightly different in orientation. These are provided in Appendix A. Also, along similar lines, the Matching Group developed a special software package containing a menu-prompt library of information on recent exact matching studies. This effort was dubbed Project LINK-LINK and is described in Section VI of this volume.

One of the most important outgrowths of the Workshop, however, was that it provided a long overdue forum for persons working in the area of exact data linkage. It not only sparked new interest in matching, but provided an atmosphere where participants could interact on a more personal basis -- a very important factor which, in the past, has been lacking, resulting in unnecessary duplication of effort in some cases. If nothing else, the Workshop has served its purpose if it acts as a catalyst for initiating more concerted efforts with regard to matching policy, methodological development and application. It was with this aim in mind that the Matching Group assembled the Recommendations which follow.

RECOMMENDATIONS

rive recommendations are provided here from the Matching Group, Administrative Records Subcommittee, Federal Committee on Statistical Methodology. The first recommendation calls for the establishment of a continuing interagency working group on record linkage systems and techniques: such a working group would be expected to play a significant role in implementing recommendations 2 through 5. The second recommendation calls for careful monitoring of external developments that might affect the prospects for undertaking record linkages for statistical purposes. Recommendations 3, 4 and 5 identify specific aspects of record linkage systems and techniques that deserve special emphasis in future research, development and evaluation activities. The five recommendations are:

1. Documentation should be improved and information on record linkage systems and techniques should be shared.

   It is recommended that the Matching Group of the Administrative Records Subcommittee be reconstituted as a Technical Working Group on Record Linkage Systems and Techniques, continuing to function under the auspices of the Federal Committee on Statistical Methodology. The main goal of the Working Group would be to promote the effective use of record linkage techniques for statistical purposes by encouraging the documentation of individual record linkage systems and techniques and the sharing of relevant technical information. A primary activity would be sponsorship and organization of workshops and meetings of professional societies to discuss relevant new developments and research, and to disseminate information on existing systems and techniques. In addition, the reconstituted working group would contribute, in appropriate ways, to the implementation of recommendations 2 through 5 below.

2. Changes in the external environment for record linkages should be monitored.

   Statistical users of record linkage techniques should track external developments that may influence their ability to perform record linkages. Such developments include changes in laws, regulations and policies affecting access to records and changes in the content of data files used in record linkages. Examples of the latter would include increased use of four-digit ZIP code add-ons ("ZIP + 4") and steps taken to promote the use of unique addresses in rural areas. In so far as possible, statistical users of record linkage techniques, working through the reconstituted Working Group (see recommendation 1), should attempt to influence the course of these developments in ways that will facilitate statistical applications. For example, the Working Group might try to promote the development of standards for reporting names and addresses of both businesses and individuals.

3. Comparative evaluation studies of record linkage systems should be undertaken.

   Several agencies of the United States and Canadian governments have invested substantial resources in the development of automated record linkage systems for use in a variety of statistical programs. For many new applications, use of an existing system is likely to be more cost-effective than development of a new one. To aid potential users of record linkage systems, it is recommended that resources be sought for comparative evaluations of existing systems and some of their components, such as name and address standardizers and blocking rules. The evaluation design should recognize that record linkage systems vary in their objectives, especially with respect to the kinds of units for which records are to be matched: persons or businesses. A much-needed first step is the development of a detailed evaluation plan that specifies the measures of quality and cost to be used in the evaluation and the nature of the files to be matched. Such evaluations may require data sets for which true match status is known. One possibility would be to create such data sets by simulation.

4. Research and development aimed at the improvement of record linkage systems and techniques should give priority to selected aspects.

   Recognizing that resources for the development of improved record linkage systems are limited, it is recommended that priority be given to the following aspects: (1) systems for linking business records, (2) name and address standardizers, (3) string comparators, (4) the choice of blocking strategies, (5) the development of "learning" systems, and (6) the role of manual intervention.

5. Errors associated with record linkages and their effects on analyses should be measured.

   It is recommended that more research be carried out on the error characteristics of record linkage systems and on the effects of errors on analyses performed with the linked data sets. To enhance the value of such research, consensus is desirable on standard measures of record linkage errors and on methods of measuring them. Promising error measurement methods include multiple matching techniques and direct contacts with samples of linked pairs to determine their true match status.

By design, the principal focus of the Workshop discussions and followup comments by participants was on methodological aspects of record linkages for statistical purposes. Legal and ethical considerations in such linkages were not part of the main agenda.

Nevertheless, the Matching Group of the Administrative Records Subcommittee recognizes that legal and ethical considerations must be weighed carefully by any organization that links records from different sources and that public perceptions of the appropriateness of various kinds of record linkages are also of critical importance. More research in these areas would also be desirable, addressing, in particular: (1) public understanding of and attitudes toward linkages performed for statistical and other purposes; (2) survey respondents' comprehension of informed consent statements currently being used, especially when such statements cover linkages of survey data and administrative records; and (3) the effects on survey response of varying the amount and kinds of information included in informed consent statements to respondents.

Members of the

MATCHING GROUP: ADMINISTRATIVE RECORDS SUBCOMMITTEE

(as of May 1985)

Mary Bentz
Internal Revenue Service

David Cordray
General Accounting Office

Jane Crane
National Center for Education Statistics

Maria Elena Gonzalez
Office of Management and Budget

Nick Greenia
Internal Revenue Service

David Hirschberg
Small Business Administration

Howard Hogan
Bureau of the Census

Thomas B. Jabine
Consultant, Committee on National Statistics

Matt Jaro
Bureau of the Census

Patrick Kelley
Bureau of the Census

Beth Kilss
Internal Revenue Service

Nancy Kirkendall
Energy Information Administration

Douglas Kleweno
Department of Agriculture

Thomas Reilly
Bureau of the Census

Douglas Sater
Bureau of the Census

Fritz Scheuren (Chair)
Internal Revenue Service

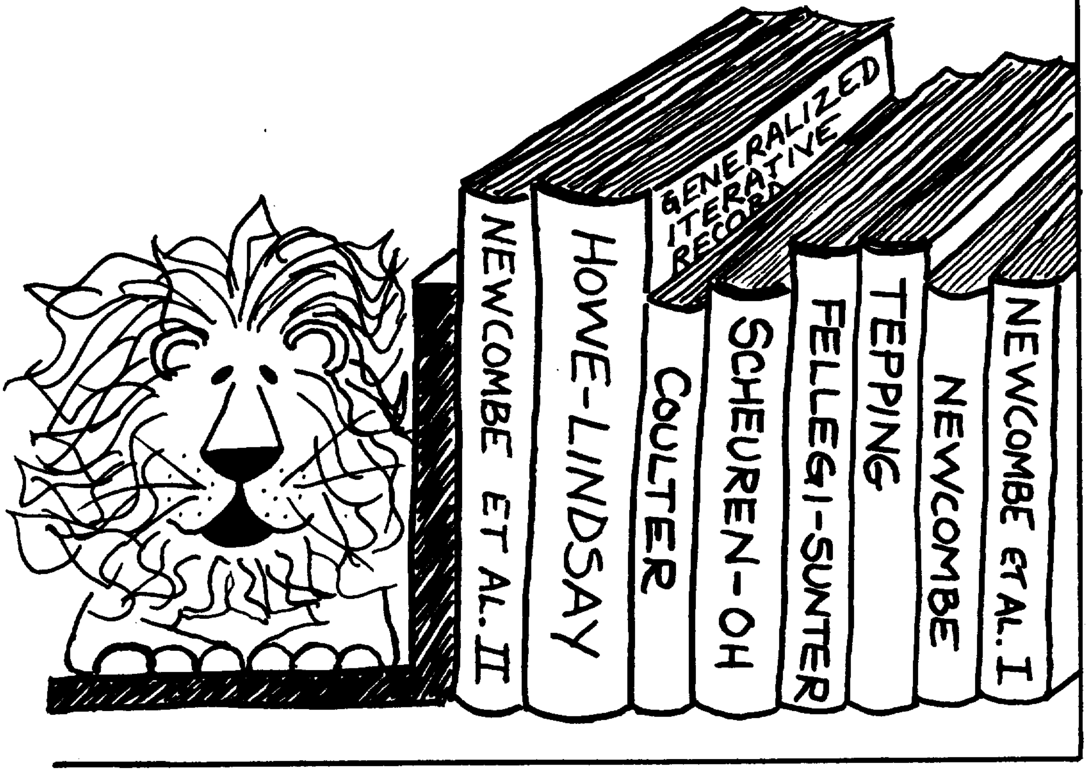
Wray Smith
Mathematica Policy Research

Wendel Thompson
Department of Energy

Carol Utter
Bureau of Labor Statistics

William Winkler
Energy Information Administration

Section I:
Selected Background Papers (1959 – 1983)



Automatic Linkage of Vital Records*

Computers can be used to extract "follow-up" statistics of families from files of routine records.

H. B. Newcombe, J. M. Kennedy, S. J. Axford, A. P. James

The term *record linkage* has been used to indicate the bringing together of two or more separately recorded pieces of information concerning a particular individual or family (*1*). Defined in this broad manner, it includes almost any use of a file of records to determine what has subsequently happened to people about whom one has some prior information.

The various facts concerning an individual which in any modern society are recorded routinely would, if brought together, form an extensively documented history of his life. In theory at least, an understanding might be derived from such collective histories concerning many of the factors which operate to influence the welfare of human populations, factors about which we are at present almost entirely in ignorance. Of course, much of the recorded information is in a relatively inaccessible form; but, even when circumstances have been most favorable, as in the registrations of births, deaths, and marriages, and in the census, there has been little recognition of the special value of the records as a source of statistics when they are brought together so as to relate the successive events in the lives of particular individuals and families. The chief reason for this lies in the high cost of searching manually for large numbers of single documents among vast accumulations of files. It is obvious that the searching could be mechanized, but as yet there has been no clear demonstration that machines can carry out the record linkages rapidly enough, cheaply enough, and with sufficient accuracy to make this practicable.

The need for various follow-up studies such as might be carried out with the aid of record linkage have been discussed in detail elsewhere (*1, 2*), and there are numerous examples of important surveys which could be greatly extended in scope if existing record files were more readily linkable (*3*). Our special interest in the techniques of record linkage relates to their possible use (i) for keeping track of large groups of individuals who have been exposed to low levels of radiation, in order to determine the causes of their eventual deaths (see *4*, chap. 8, para. 48; *5*), and (ii) for assessing the relative importance of repeated natural mutations on the one hand, and of fertility differentials on the other, in maintaining the frequency of genetic defects in human populations (see *4*, chap. 6, para. 36c).

Our own studies (*6*) were started as part of a plan to look for possible differentials of family fertility in relation to the presence or absence of hereditary disease (through the use of vital records and a register of handicapped children). The first step has been the development of a method for linking birth records to marriage records automatically with a Datatron 205 computer. For this purpose use has been made of the records of births which occurred in the Canadian province of British Columbia during the year 1955 (34,138 births) and of the marriages which took place in the same province over the 10-year period 1946-55 (114,471 marriages). Fortunately, these records were already in punch-card form as a part of Canada's National Index, and from them could be extracted most of the necessary information on names and other identifying particulars. An intensive study of the various sources of error in the automatic-linkage procedure has now been carried out on approximately one-fifth of these files.

Technical Problems

One of the chief difficulties arises from the unreliability of the identifying information contained in successive records which have to do with the same individual or married pair. The spellings of the surnames may be altered, the first Christian name on one record may become the second on another, and the birthplaces and ages may not be correctly stated. Much of the design effort must be directed toward ensuring that records can be linked in spite of such discrepancies, which in our files occurred with frequencies of about 10 percent of all record linkages involving live births and 25 percent of all linkages involving stillbirths.

A second problem relates to ambiguous linkage, in which it is uncertain whether or not a birth has arisen out of a particular marriage, or where there are two or more marriages any one of which might be that of the parents. These problems tend to occur when the husband's surname and the wife's maiden name are both common in the region studied, but they can also be associated with rarer family names, as in the marriage of two brothers to two sisters, and in certain racial minority groups. The difficulty increases with the size of the population under study.

At first sight these considerations might seem to preclude any extensive use of automatic record linkage as a source of statistics, since it is not at all obvious that the rules of judgment as exercised by a human being can be adapted to machine use. Also, partially mechanized record-linkage operations have proved laborious in the past (*7*).

Nevertheless, satisfactory procedures were eventually developed. These began with a series of small-scale attempts to link records visually, and thus to gain insight into the causes of any failures. The first of these studies was carried out at the Bureau of Statistics by one of us (S.J.A.) and made use of one of the standard phonetic name-coding systems to reduce the undesirable consequences of spelling discrepancies in linking records of sibling stillbirths. The gradual evolution of the method since that time has served to make it evident that further refinements can undoubted-

*Reprinted with permission from Science, Copyright 1959, by the American Association for the Advancement of Science, Vol. 130, No. 3381, October 16, 1959, pp. 954-959.

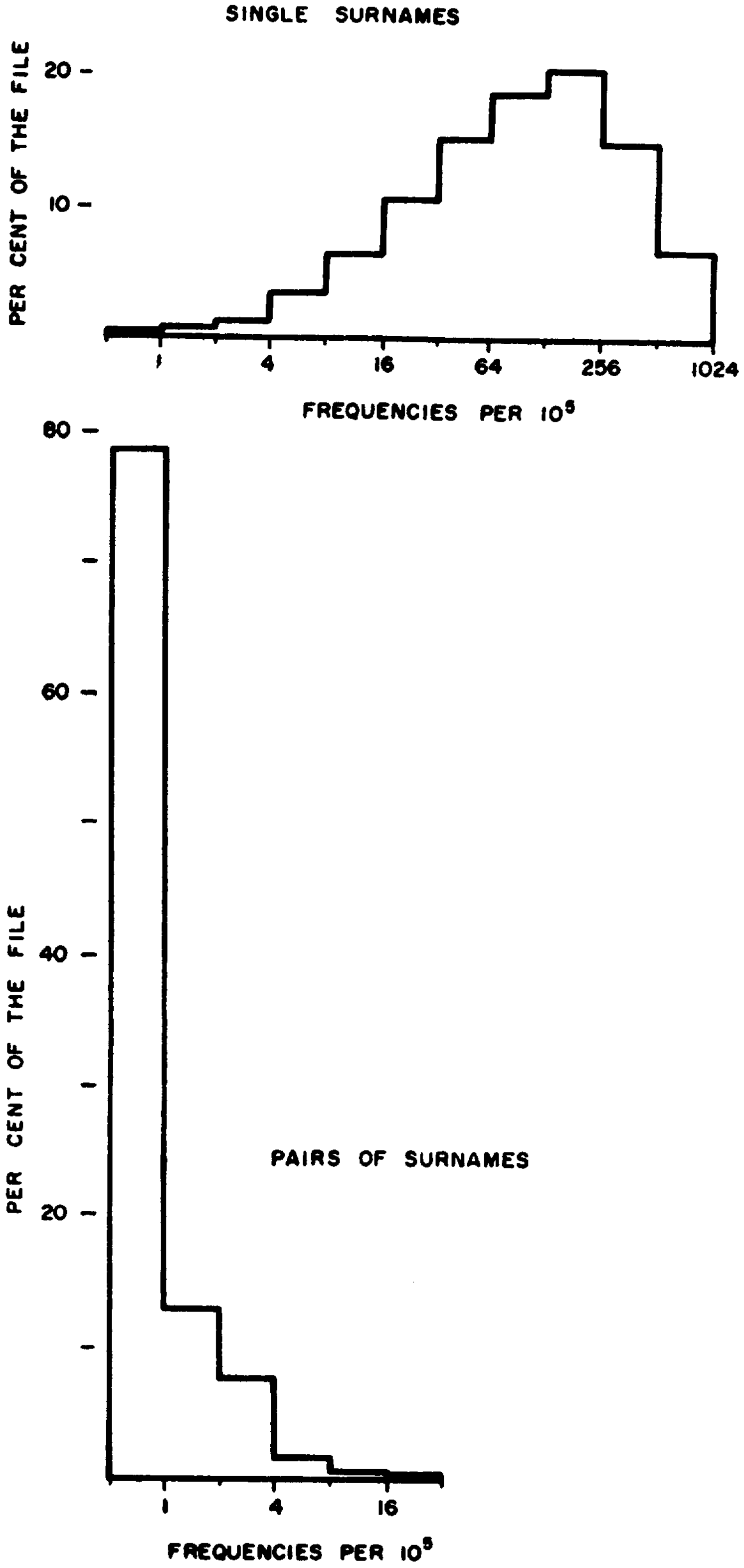


Fig. 1. (Top) Frequency distribution of brides' maiden names, in Soundex coded form, from records of 114,471 marriages in British Columbia for 1946–55. (Bottom) Frequency distribution of family-name pairs for married couples, in Soundex coded form, from the same records. Two East Indian names, of which one is customarily passed from mother to daughter and the other from father to son, were omitted. These occurred together in the same combination in approximately 100 marriages.

ly be developed and that no limit to the possible reliability of the linkages is yet in sight.

Methods

Of primary interest was the development of a procedure which would be fully automatic and free from piecemeal operations which might later limit the usefulness of the approach. This aim was achieved, chiefly because the use of a computer made it possible to compare each birth record in turn with all of the marriage records in appropriate sections of the marriage file. Since groups of marriages were sometimes scanned a number of times, it is apparent that this operation could not have been carried out with conventional card-handling equipment. Thus, without the computer, a visual search through printed lists would have been required to achieve some of the linkages.

To reduce the number of marriage records with which the computer must compare a birth record, it was decided to make use of both the husband's surname and the wife's maiden name, these being present on both the marriage and the birth cards. The surnames were first reduced to phonetic codes, consisting in each case of the first letter of the name followed by three numeric digits and known as the Russell Soundex Code (*8*), the computer being used for the coding operation. The codes served two purposes: They were designed to remain unchanged with many of the common spelling variations and in the present application were thus expected to bring together linkable records which would have been widely separated if arranged in a strictly alphabetic sequence. The coding also simplified the subsequent use of the Datatron computer, which is essentially a mathematical instrument and works more readily with numbers than it does with letters.

The extent to which two surnames are more efficient than one for identifying a family group has probably not been generally recognized. Thus, of the various brides' maiden names encountered in the marriage file, more than half recurred (in their coded forms) with frequencies in the range from 64 up to 1024 per 10^5. In contrast to this, nearly 80 percent of the pairs of family names (in their coded forms) were unique; that is, they occurred only once in our file in that particular combination, and extremely few had frequencies exceeding 4 per 10^5 (see Fig. 1). This meant that we could mechanically compare each birth for the entire year with all of the marriages, using the same pair of surname codes, and that only rarely would the number of code matchings exceed one or two per birth.

To enable the computer to decide whether or not a birth and a marriage relate to the same married pair, use must be made of other identifying particulars. We relied chiefly on six items: the full alphabetic family names of the husband and wife (limited to nine letters each), their provinces or countries of birth (each coded as a two-digit number), and their first initials. In addition, the ages of the married pair were available on our cards for all of the birth records and for about half of the marriage records (that is, for marriages in the period 1951-56); the second initials were present in the case of the birth file; and the name of the city or place of the event (restricted to six letters) was available throughout both files.

As mentioned earlier, no one piece of information was entirely reliable. Usually it was obvious on inspection that the two events did, or did not, relate to the same married pair, but occasionally the decision was difficult. For this reason the computer had to calculate a probability that the couples were the same, or were different. The operation was performed automatically when the files were first matched.

The principle on which such a probability was based is fairly simple. If, for example, the province or country of birth of both the husband and wife agree on the two records, these facts may influence somewhat our belief that these records relate to the same married pair. Of course, the weight which one attaches to the information will be small if both have been born in the home province of British Columbia, but it will be large if they happen to have been born in, let us say, Switzerland and New Zealand, respectively. To give this a mathematical form it is necessary to know the frequencies for the various birthplaces of brides and grooms, and these can be determined quite readily either from published statistics or from the files themselves.

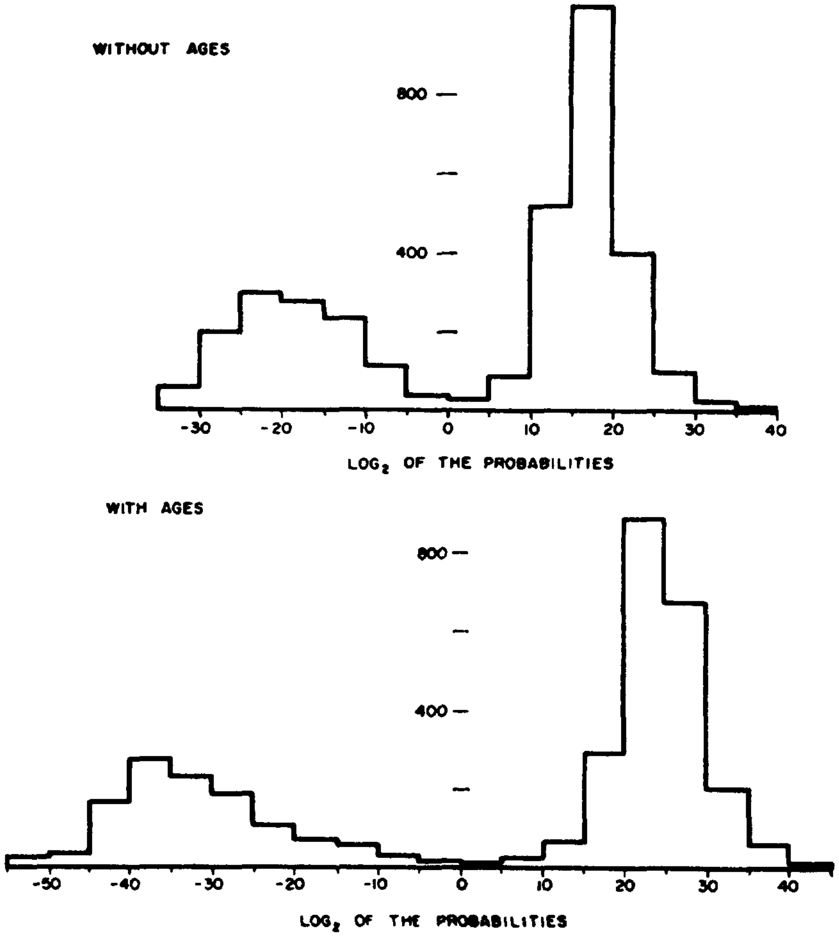


Fig. 2. (Top) Frequency distribution of the probabilities (in binits) obtained on comparing birth and marriage records having identical Soundex code pairs (calculated without using ages), based on records contained in the first fifth of the birth and marriage files (husband's surname beginning with *A*, *B*, or *C*). For this comparison only legitimate live births and marriages recorded in 1951-55 (a period for which ages are available) were considered. There were 2174 cases of genuine linkage and 1232 cases of accidental Soundex agreement. (Bottom) Same as above, except that the ages were used in calculating the probabilities.

Similar reasoning can be applied to any item of identifying information, and to both agreements and disagreements. In order that the probabilities may be added together they must be converted to logarithms, and it is conventional practice in information theory to use logarithms to the base 2 of the probabilities expressed in the form of the "odds," for or against. The units are known as "binits." Thus, if the odds were 16 to 1 in favor of a genuine linkage, this would be represented as plus 4 binits, and odds of 16 to 1 against would be minus 4 binits. It is convenient to remember that a value of 10 binits is equivalent to odds of approximately 1000 to 1.

For present purposes, the probability or odds associated with a given agreement or disagreement may be obtained in binit units from the expression:

$$\log_2 p_L - \log_2 p_F \qquad (1)$$

where p_L and p_F are the frequencies with which the agreement or disagreement occurs, respectively, in the linked pairs of records and in pairs which have been brought together by accident. The expression will have a positive value in the case of agreement and a negative value in the case of disagreement.

As applied to agreements of initials and birthplaces, the expression can usually be simplified without any great loss of accuracy, since the particular letter or place should agree in the linked records almost as often as it appears in the individual records, and the chance of a fortuitous agreement will in most cases be approximately the square of this frequency. By substitution, expression 1 thus becomes:

$$\log_2 p_R - \log_2 (p_R)^2 = -\log_2 p_R \quad (2)$$

where p_R is the frequency of the particular initial or birthplace in the individual records.

Table 1. Surname spelling discrepancies*.

Name	Number of linkages in sample	Total spelling discrepancies		Discrepancies affecting the phonetic codes	
		No.	Percentage	No.	Percentage
Husband's surname	3622	41	1.1	15	0.4
Wife's maiden name	3501	115	3.3	42	1.2
Combined			4.4		1.6

* Based on visual linkages of births with marriages. To detect spelling discrepancies in a random assortment of the family names of one partner, use was made of the parts of the files in which the family name of the spouse began with *A*, *B*, or *C*. Thus, the two samples of records each represented approximately 19 percent of the total files.

Table 2. Discrepancies in birthplaces and first initials*.

Category	Number of linkages in sample	Discrepancies	
		No.	Percentage
Birthplace of husband	2174	22	1.0
Birthplace of wife	2174	21	1.0
First initial of husband	2174	60	2.8
First initial of wife	2174	83	3.8
Total			8.6
Total, including surnames			11.4
Linkages having discrepancies in one or more of the six items			10.3

* Discrepancies in computer linkages of records contained in the first fifth of the birth and marriage files (husbands' surnames beginning with *A*, *B*, or *C*); only linkages of legitimate live births with marriages in the period 1951-56 (for which ages were available) were used. For the "total, including surnames," use was made of the data from Table 1.

The approach also lends itself to comparisons of the ages as stated on the two records, the lapse of time between the two events, and whether a discrepancy, if present, is slight or large, being taken into account. Even such an unlikely item as the place of the event can be used; if the marriage and the birth occurred in different places the fact carries little weight, but if they occurred in the same place (provided it was not the largest city in the province) the fact is important.

The items from which the probabilities were calculated in our study were the two alphabetic surnames, the two birthplaces, the two first initials, the two ages (where these were given on the cards), and the place of the event. For possible future use the computer also compared the birth order with the apparent duration of the marriage at the time of the birth, and wherever a first initial failed to agree, the computer looked for agreement between the first initial on the marriage record and the corresponding second initial on the birth record.

This sort of treatment can be adapted to linking almost any types of records where the information in common is sufficient for the purpose. Although tables of probabilities (in binits) containing over 300 items were used in the present study, they did not exhaust the capacity of the computer's memory unit. The limiting factor is the discriminating power inherent in the information supplied, and it is apparent that additional items of information can be of use even where they are of limited reliability.

The extent to which ages, for example, enable the computer to separate the genuine linkages from the fortuitous Soundex agreements can be seen from the data of Fig. 2. In this case, the number of record comparisons falling in the region from minus 10 to plus 10 binits, where the degree of certainty is less than 1000 to 1, is reduced by a factor of 3 when use is made of the additional information.

Reliability of the Linkages

Studies of the accuracy of the present computer-handling procedures indicate that about 98.3 percent of the potential linkages are detected in the existing record files, and that contamination with spurious linkages is 0.7 percent [see (*9*)]. This degree of accuracy is considered adequate for the statistical studies which have been planned, since the loss of such a small amount of data cannot in itself constitute a source of bias. Further, both the losses and the contaminations can be detected in the majority of cases by means of a subsequent check on the continuity of birth orders within families.

Variations in the spelling of the family names occur in about 4 to 5 percent of all linkages, but the losses from this source are reduced by the use of the phonetic codings to approximately a third of that value (see Table 1). The detection of such losses was accomplished by the simple expedient of resorting the files in a sequence which ignored the suspect code but trusted other identifying items, the files then being listed and examined visually. This operation could have been performed by the computer, and since the six main identifying items all agree in about 90 percent of the linked pairs of records (see Table 2), two additional arrangements of the files, each of which ignored one of the two Soundex codes, would be sufficient to reduce losses of this kind from the present 1.6 percent to about 0.16 percent. For the projected statistical studies such a procedure would hardly be worth while, the computer time being the limiting factor. It might become of value for other purposes, however, as computer speeds increase, especially as it is customary for central registry offices to keep two separate listings of marriages for searching purposes, arranged under grooms' surnames and brides' maiden names, respectively.

Failure of the calculated probabilities to make a correct distinction contributed a few additional losses and a few spurious linkages. These were detected by comparing the full Christian names as given on the original registration forms wherever the calculated probability fell within the range from minus 10 to plus 10 binits. Where age was used in calculating the probabilities there were only one loss and four spurious linkages from this source in a sample of over 2000 linkages (see Table 3). Although this degree of accuracy is adequate for almost any purpose, to make a further reduction in the number of spurious linkages would not be difficult.

Table 3. Losses and spurious linkages due to lack of sufficient identifying information, which occurred in the linkage reported in Table 2 (*9*).

Item	No. of linkages in sample	Losses		Spurious linkages	
		No.	Percentage	No.	Percentage
Age data used	2174	1	0.05	4	0.23
Age data not used	2174	5	0.2	26	1.2

The contamination with spurious linkages will tend, however, to vary in direct proportion to the size of the marriage file with which the births are compared. Thus, in any future studies of larger populations it might be desirable to make use of additional identifying information. Christian names (perhaps restricted to four letters each), the city of birth of the husband and of the wife, respectively (likewise restricted to a few letters), and the province and year of marriage (not shown at present on the birth registration form) would all be suitable data for this purpose. The last of these three groups of items, however, would be of special value in effectively reducing the size of the marriage file with which any one birth would have to be compared, and in this manner reducing the false linkages. Occasional inaccuracies in the additional information would not greatly alter its usefulness in view of the nature of the handling procedures.

It is doubtful whether the present accuracy of the procedure can be matched by that of conventional survey and interview techniques, and its potential accuracy is certainly much greater than that of conventional techniques.

Speed of Record Linkage

By far the largest part of the effort in this undertaking has gone into the preparation of the card files. This has included, in the case of the marriage cards, a mechanical reproduction of the information contained in the existing National Index marriage cards for brides and for grooms, respectively, on a single card of our own format. Likewise, a part of the contents of our birth cards was obtained by reproduction from existing National Index birth cards, but in this case the maiden name of the mother and a number of other items were then added from cards which had been especially key-punched for the purpose. The family names on all cards in both files were Soundex coded by means of the computer, and the files were sorted into a Soundex sequence by pairs of codes, and listed. For the purpose of the initial record-linkage study the part of the marriage file for married pairs in which the groom's surname began with *A*, *B*, or *C* (approximately one-fifth of the total file) was transferred to magnetic tape.

This done, the computer made the necessary birth-to-marriage comparisons when presented with the birth cards, matchings with respect to the pairs of name codes being achieved at a rate of approximately one comparison every 3 seconds. About half of these code agreements represented genuine linkages (*10*). (Subsequently the whole of the birth and marriage files were put on magnetic tape and linked automatically by the computer.)

The initial steps would be largely eliminated were the format of the cards which are prepared routinely designed with a view to their possible use for record-linkage purposes. Also, an improvement in the rate at which the computer makes the comparisons can be gained in later operations by limiting the longer computations to the relatively small number of comparisons where simpler tests are inadequate. Some other short cuts might well be effected in the program if it were used sufficiently to justify the time involved. Such improvements can be thought of as reducing the cost of record linkage, in which computer rentals may be a major item, and of increasing the ease with which statistics can be derived from the linkage process.

The use of a computer especially designed to handle alphabetic information would further reduce the time required for the linkages by virtue of this special design alone, and there are larger computers in which the basic logical steps are more rapid by an order of magnitude. Thus, the present rate of something like one linkage every 6 seconds might be increased perhaps 20- or 30-fold—that is, to 200 or 300 linkages per minute, with existing equipment.

It is difficult to guess to what extent these speeds will be exceeded in the next 10 years or so. However, circuits have been described in the literature in which the basic logical steps take much less time than those in any equipment at present on the market (*11*). Research with the more novel kinds of electrical switching devices, some of which are not only fast but extremely compact, may extend the present limit by at least another order of magnitude (*12*).

Well before such equipment becomes available, however, it should be possible to develop the data-processing methods by which record linkages are achieved to the point at which the extraction of a wide variety of family and follow-up statistics becomes practicable from any records which are in an accessible form.

References and Notes

1. H. L. Dunn, *Am. J. Public Health* **36** (Dec. 1946); J. T. Marshall, *Population Studies* **1**, 204 (1947).
2. H. L. Dunn and M. Gilbert, *Public Health Repts. (U.S.)* **71**, 1002 (1956); H. B. Newcombe, in *Effect of Radiation on Human Heredity* (World Health Organization, Geneva, 1957); ———, A. P. James, S. J. Axford, "Family Linkage of Vital and Health Records," *Atomic Energy Can. Rept. No. 470* (Chalk River, 1957); H. B. Newcombe, S. J. Axford, A. P. James, "A Plan for the Study of Fertility of Relatives of Children Suffering from Hereditary and Other Defects," *Atomic Energy Can. Rept. No. 551* (Chalk River, 1957); H. B. Newcombe, A. P. James, S. J. Axford, "Genetic hazards and vital statisitcs," *Proc. Intern. Congr. Genet. 10th Congr., Montreal* (1958), vol. 2, p. 205.
3. S. C. Reed and J. D. Palm, *Science* **113**, 294 (1951); S. C. Reed, E. W. Reed, J. D. Palm, *Eugenics Quart.* **1**, 44 (1954); T. E. Reed, *Japan. J. Human Genet.* **2**, suppl., 48 (1957); ——— and E. L. Kelly, *Ann. Human Genet.* **22**, part 2, 165 (1958); A. B. Hill, R. Doll, T. M. Galloway, J. P. W. Hughes, *Brit. J. Prevent. & Social Med.* **12**, 1 (1958).
4. *Report of the United Nations Scientific Committee on the Effects of Atomic Radiation, Suppl. No. 17 (A/3838)* (United Nations, New York, 1958).
5. H. B. Newcombe, *Science* **126**, 549 (1957).
6. We are indebted to John H. Doughty for his encouragement and constructive criticism in the course of this work, to Robert J. Montgomery for making available facilities for the preparation of the marriage file, and to George Selby for his help in this initial operation. We would also like to thank Elizabeth Kinsey for collaborating in the preparation of the record files and in the analysis of the results, and Arden Okasaki for her work in programming the computer. Permission to use the vital records in this study was obtained through the Dominion Bureau of Statistics, from the Health Branch, Department of Health and Welfare, Province of British Columbia. The permission was conditional upon strict observance of the oath of secrecy respecting the nonstatistical information contained in the records.
7. S. Shapiro and J. Schachter, *Estadistica* **10**, 688 (1952).
8. The rules of Soundex coding are as follows. (i) The first letter of a surname is uncoded and serves as the prefix letter. (ii) W and H are ignored completely. (iii) A, E, I, O, U, and Y are not coded but serve as separators (see v below). (iv) Other letters are coded as follows, until three digits have been used up (the remaining letters are ignored): B, F, P, V, coded 1; D, T, coded 3; L, coded 4; M, N, coded 5; R, coded 6; all other consonants (C, G, J, K, Q, S, X, Z), coded 2. (v) Exceptions are letters which follow letters having the same code, or prefix letters which would, if coded, have the same code. These are ignored in all cases unless a separator (see iii above) precedes them.
9. Since ages were available on only about half of the marriage cards, the average losses from this cause were 0.12 percent of all linkages, and the average spurious linkages were 0.7 percent. When these are added to the losses resulting from the Soundex discrepancies, as shown in Table 1, the total loss is 1.72 percent.
10. It is known that approximately 19 per cent of the surnames in the marriage file begin with A, B, or C, as determined from studies of the frequencies of brides' Soundex codes. Thus, of the 114,471 marriage records and 34,138 birth records, approximately 21,750 and 6500 records, respectively, were used in the initial linkage study. In all, 6375 comparisons (3484 with positive binit values and 2891 with negative) between birth records and marriage records having identical pairs of Soundex codes were made by the computer. Of these, 418 (20 positive and 398 negative) related to illegitimate births, 2549 (1285 positive and 1264 negative) related to legitimate births and to 1946-50 marriages, and 3408 (2179 positive and 1229 negative as determined by means of ages) related to legitimate births and to 1951-55 marriages. Since age records were available in the case of the 1951-55 marriages,

this latter group of 3408 comparisons was used for a detailed study of the reliability of the machine linkage process. (Revised tables of binit values were also derived from these comparisons.) Two of the 3408 comparison cards were removed because in each case one of the ages was missing. Of the remaining 3406 cards, 2174 represented genuine linkage (2173 positive cards plus one negative card) and 1232 represented accidental Soundex agreements (4 positive plus 1228 negative cards), as judged by comparisons of the full Christian names in all cases where the binit values fell within the range from minus 10 to plus 10. It will be noted that of the 6500 births of 1955 which were studied, 3484 (54 percent) were from marriages contracted in British Columbia during the 10-year period 1946-55. For a description of the manner in which visual record linkages (as distinct from computer linkages) were used to assess the losses due to spelling discrepancies, see footnote to Table 1.

11. R. M. Walker, D. E. Rosenheim, P. A. Lewis, A. G. Anderson, *IBM J. Research and Develop.* 1, 257 (1957).
12. R. F. Rutz, *ibid*, 1, 212 (1957); D. A. Buck, *Proc. I.R.E. (Inst. Radio Engrs.)* 44, 482 (1956); J. W. Crowe, *IBM J. Research and Develop.* 1, 295 (1957).

Editors' Note: In 1959 Dr. Newcombe and Dr. James were affiliated with the biology branch of Atomic Energy of Canada, Ltd., Chalk River, Ontario. Dr. Kennedy was affiliated with the theoretical physics branch of Atomic Energy of Canada. Dr. Axford was affiliated with the health and welfare division of the Dominion Bureau of Statistics, Ottawa.

Record Linking: The Design of Efficient Systems for Linking Records into Individual and Family Histories*

HOWARD B. NEWCOMBE

Biology Branch,
Chalk River Nuclear Laboratories,
Chalk River, Ontario.

INTRODUCTION

THE APPLICATIONS of computer technology to genetic problems discussed so far in this Supplement make use, primarily, of the ability of the machines to carry out involved mathematical procedures. In contrast, the application which I shall describe uses the computer as a kind of filing clerk. The task given it is that of building family histories of births, marriages, procreations, deaths, and ill health from the individual registrations of these events, and of doing so on a substantial scale.

Although the computer is at no point asked to carry out any mathematical operation more complicated than simple addition and subtraction, it must nevertheless perform a function that is much more unconventional for machines. It is required to simulate the judgment of a human clerk who attempts to file correctly the incoming correspondence from people who are careless about the way they spell their family names, who may sometimes use their middle names as if these were their first, and who may be writing from places that are not their usual addresses.

Provided that a computer can be instructed to carry out an operation of this kind with a degree of accuracy similar to that of a human filing clerk, the special talent which it may be expected to apply to the task is its speed. Current experience with this sort of computer application is particularly encouraging, in terms of accuracy, speed, and cost, and the capabilities of the machines will undoubtedly increase as time goes on. Thus, it is not unrealistic to think of integrating, in due course, some major fraction of the routine personal documentation dealing with reproduction and health into the form of individual and family histories.

CONCEPTS

A number of concepts will be discussed that are inherently simple, but the implications of these concepts will not necessarily be self evident.

The idea of linking records, for example, is particularly simple—the phrase *record linking* just means bringing together information from two independent sources about the same person—but with successive linkings the information may take on the characteristics of a collection of personal or family histories.

*Reprinted with permission from American Journal of Human Genetics, University of Chicago Press, Vol. 19, No. 3, Part I (May), 1967.

Even such familiar file upkeep operations as the insertion of address changes into a mailing list are elementary forms of record linking. However, the process as applied to human genetics will involve successive linkings of routinely collected records of procreative and health events to derive, eventually, multigeneration pedigrees for whole populations.

The two principal steps in any linking operation, namely, those of searching out the potentially linkable pairs of records for detailed comparison and of deciding whether or not a given pair is correctly matched, are commonplace in almost any operation by which a file is kept up-to-date. However, both of these steps, if they are to be carried out efficiently by machines, involve the use of stratagems of kinds that are employed almost unconsciously by a human filing clerk. For the *searching step,* the aim must be to reduce the number of failures to bring potentially linkable records together for comparison, such as may occur as a result of discrepancies in the file sequencing information, but this must be done without resorting to excessive amounts of additional searching. For the *matching step,* the problem is that of enabling the machine to apply in numerical form the rules of judgment by which a human clerk would decide whether or not a pair of records relates to the same person when some of the identifying information agrees and some disagrees.

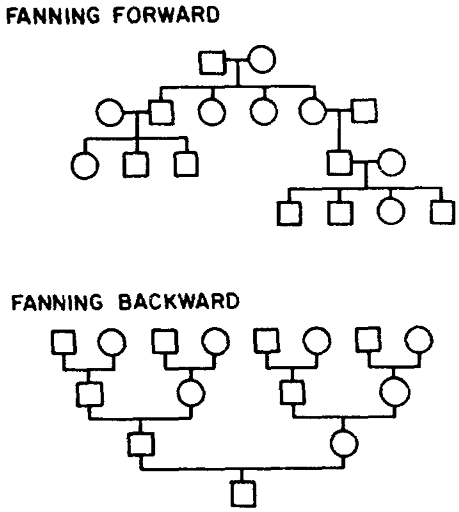
Similarly, the idea of arraying pedigree information in linear fashion to facilitate storage, updating, and retrieval by machines using magnetic tapes as the storage medium is simple and by no means new. Nevertheless, the forms which such linear arrays may take bear little resemblance to the conventional pedigree charts with which geneticists are most familiar. The great flexibility of the *linear pedigrees* and the ease with which family relationships of unlimited complexity may be represented in such a fashion are, for this reason, not generally appreciated. In comparison, however, the usual two-dimensional representations are exceedingly cumbersome (Fig. 1).

Finally, it has not been uncommon in the past to derive partial histories of individuals and families from the *routine vital and health records,* on a small scale, by manual means. However, the idea that some substantial fraction of these enormous files might be so organized and that we are at the point now where this would be technically feasible and not too expensive is one that has been slow in gaining acceptance. Nevertheless, the inherent possibilities are beginning to be recognized. A colleague of mine is reported to have remarked recently that we are still using old data on hemophilia, that there are many hemophiliacs in Canada, almost all of whom will wind up in a computer sooner or later, and "what a shame if it is only opposite a dollar sign."

The concepts may not be new, but such implications are.

METHODS OF RECORD LINKING

The two essential steps in the linking of records by computer, that is, the *searching* step and the *matching* step, have precise counterparts in many manual filing operations. Although the accuracies of such operations and the times required are generally regarded as important, it is unusual to judge the efficiencies in numerical terms or to set down the conditions under which



FANNING FORWARD

FANNING BACKWARD

Fig. 1. Conventional pedigree charts. Note the difficulty of representing in a single chart the ancestors, descendants, cousins, and in-laws.

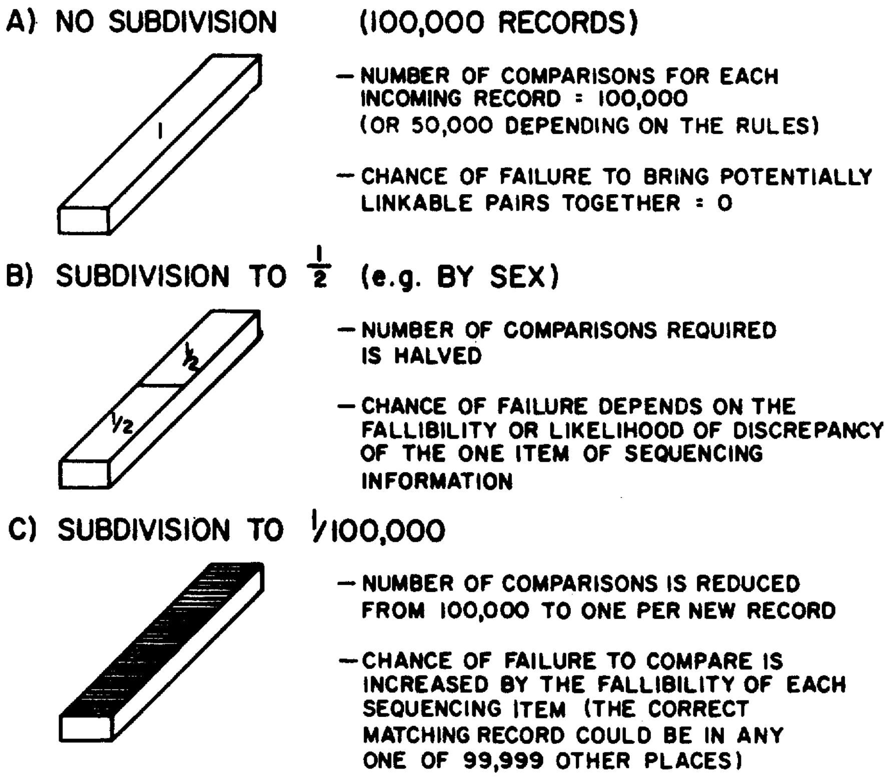
an optimum balance may be achieved between the level of accuracy and its cost as indicated by time required to achieve that level. Where such an undertaking is to be carried out on a very large scale by a computer, however, some thought may profitably be given to the efficiency of the operation in these terms.

1. *Optimizing the Searching Step*

In the case of the searching step, errors in the form of failures to bring potentially linkable pairs of records together for comparison could be reduced to zero simply by comparing each incoming record with all of the records already present in the master file. Where the files are large, however, such a procedure would generally be regarded as excessively costly in terms of the enormous numbers of wasted comparisons of pairs of records that are unlinkable.

For this reason, it is usual to arrange the file in some orderly sequence, using identifying information that is common to both the incoming records and those already present in the master file. Detailed comparisons then only need to be carried out within the small portions of the master file for which the sequencing information is the same as that on the incoming records (Fig. 2). For many purposes, it is common practice to use the alphabetic surnames and first given names for sequencing a file of personal records. The price that must be paid for the saving of time is an increase in the failures to bring potentially linkable pairs of records together for comparison, owing to discrepancies in the sequencing information on pairs that in fact relate to the same person. However, different kinds of information that might be used for the sequencing differ widely, both in their reliability and in the extents to which they subdivide a file.

Although alphabetic surnames are commonly employed, they are not particu-



A) NO SUBDIVISION (100,000 RECORDS)

— NUMBER OF COMPARISONS FOR EACH INCOMING RECORD = 100,000 (OR 50,000 DEPENDING ON THE RULES)

— CHANCE OF FAILURE TO BRING POTENTIALLY LINKABLE PAIRS TOGETHER = 0

B) SUBDIVISION TO $\frac{1}{2}$ (e.g. BY SEX)

— NUMBER OF COMPARISONS REQUIRED IS HALVED

— CHANCE OF FAILURE DEPENDS ON THE FALLIBILITY OR LIKELIHOOD OF DISCREPANCY OF THE ONE ITEM OF SEQUENCING INFORMATION

C) SUBDIVISION TO 1/100,000

— NUMBER OF COMPARISONS IS REDUCED FROM 100,000 TO ONE PER NEW RECORD

— CHANCE OF FAILURE TO COMPARE IS INCREASED BY THE FALLIBILITY OF EACH SEQUENCING ITEM (THE CORRECT MATCHING RECORD COULD BE IN ANY ONE OF 99,999 OTHER PLACES)

FIG. 2. Optimizing a single sequence search. Subdivision must be based on items of identifying information with the highest efficiency ratios and must be adjusted to an acceptable low level of losses or of wasted comparisons.

larly efficient for sequencing, because of the high frequency with which they are misspelled or altered. Considerable improvement can be achieved by setting aside temporarily the more fallible or labile parts of the information which the surnames contain, while retaining as much as possible of the inherent discriminating power. There are a number of systems for doing this, the most common of which is known as the Russell Soundex code. This is essentially a phonetic coding, based on the assignment of code digits which are the same for any of a phonetically similar group of consonants. (Details of a number of such surname coding systems are given in the Appendix.)

In practice, we have found that the Soundex code remains unchanged with about two-thirds of the spelling variations observed in linked pairs of vital records, and that it sets aside only a small part of the total discriminating power of the full alphabetic surname. The system is designed primarily for Caucasian surnames, but works well for files containing names of many different origins (such as those appearing on the records of the U. S. Immigration and Naturalization Service). This particular code is less satisfactory, however, where the files contain names of predominantly Oriental origin, because much of the discriminating power of these resides in the vowel sounds which the code ignores.

Any kind of identifying information that is available on all of the records may, of course, be used for sequencing the files, and it should not be assumed that surnames necessarily possess special merit for this purpose. The qualities required are reliability and discriminating power, both of which may be measured numerically. Usually, where the discriminating power of any one kind of information alone is insufficient to divide the file finely enough, two or more kinds of information may be used together to achieve a required degree of subdivision. However, each additional kind of information carries its own likelihood of discrepancy and thus contributes to the over-all tendency for the sequencing information to be reported differently on successive records relating to the same person, with a resulting increase in the frequency with which potentially linkable records will fail to be brought together for comparison. It is important, therefore, to choose the most appropriate kinds of information from among those that are available.

Fortunately, there are numerical tests which will indicate the relative merits of the different items of identifying information for the purpose of sequencing the files. Three values will be discussed, the *coefficient of specificity,* the *discriminating power,* which is simply another way of describing the specificity, and a so-called *merit ratio,* which may be used to indicate the amount of discriminating power per unit likelihood of discrepancy. This latter value can be used in selecting the most appropriate information to be employed in sequencing a file.

The fineness with which a file will be divided by a particular kind of identifying information may be represented by a single number, the *coefficient of specificity,*

$$C_s = \Sigma P_x^2 \tag{1}$$

where P_x is the fraction of the file falling in the xth block (see Fig. 3). C_s may be thought of as the fraction of the file falling within a block of strictly representative size. Since most identifying information divides a file unevenly into a mixture of small and large blocks, it is convenient to be able to indicate the effective degree of division of the file in this simple manner.

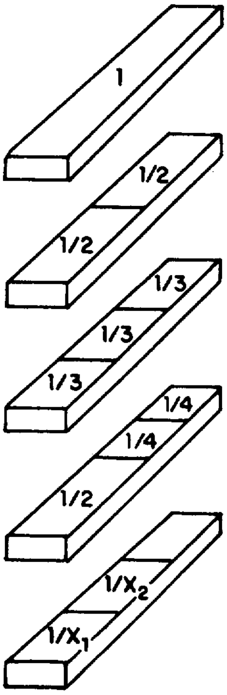
Unlike the coefficient of specificity, which gets smaller as a file becomes more finely divided, the *discriminating power* increases with the extent of the subdivision. Furthermore, it is usually regarded as an "addable" quantity. Thus, the discriminating power may be taken as the logarithm of the inverse of the coefficient of specificity, and in practice we have found it convenient to use logarithms to the base two (see Table 1):

$$D_p = \log_2 (1/C_s) \tag{2}$$

Finally, the merit of any particular kind of identifying information for sequencing the files may be taken as the ratio of the discriminating power to the likelihood of discrepancy or inconsistency of such information in linkable pairs of records:

$$M_t = D_p/I \tag{3}$$

In calculating this so-called *merit ratio,* we normally use the percentage likelihood of inconsistency as the numerical value of I.



$$C_s = ({}^1/_1)^2 = {}^1/_1 = 1$$

$$C_s = ({}^1/_2)^2 + ({}^1/_2)^2 = {}^2/_4 = {}^1/_2$$

$$C_s = ({}^1/_3)^2 + ({}^1/_3)^2 + ({}^1/_3)^2 = {}^3/_9 = {}^1/_3$$

$$C_s = ({}^1/_2)^2 + ({}^1/_4)^2 + ({}^1/_4)^2 = {}^6/_{16} = {}^1/_{2.7}$$

$$C_s = ({}^1/_{x_1})^2 + ({}^1/_{x_2})^2 + \ldots\ldots = \Sigma P_x{}^2$$

(where P_x is the proportion in the Xth block)

FIG. 3. Examples of coefficients of specificity.

TABLE 1. RELATIONSHIP OF COEFFICIENT OF SPECIFICITY AND DISCRIMINATING POWER

Coefficient of specificity $C_s = \Sigma P_x^2$	Discriminating power $\log_2 (1/C_s)$	Equivalent number of blocks if file equally divided
1	0	$2^0 = 1$
1/2	1	$2^1 = 2$
1/4	2	$2^2 = 4$
1/8	3	$2^3 = 8$
1/16	4	$2^4 = 16$
1/1024	10	$2^{10} = 1024$
$1/10^6$	20	$2^{20} = 10^6$

The most efficient sequencing of a file will be based on the items of identifying information that have the highest merit ratios, using enough different items to achieve a combined discriminating power that will subdivide the file to the required degree of fineness. In this manner, the minimum total likelihood of discrepancy or inconsistency will have been introduced into the sequencing items for any required degree of subdivision.

By means of such numerical values, the usefulness of surname information in its Soundex coded form can be shown to be considerably greater than

TABLE 2. RELATIVE MERITS OF ALPHABETIC VERSUS SOUNDEX CODED SURNAMES FOR SEQUENCING FILES

Surname information	Discriminating power D_p	Equivalent number of blocks of equal size $1/C_s$	Percentage likelihood of discrepancy* I	Merit ratio $Mt = D_p/I$
Alphabetic	+9	512	2.2	4.1
Soundex	+8	256	0.8	10.0
Residual	+1	2	1.4	0.7

*Average for husbands' and wives' birth surnames.

that of the full alphabetic surnames for the purpose of sequencing the files, the merit ratio being about two or three times as large (Table 2). The residual information that is omitted from the Soundex codes is of very low quality indeed, having a merit ratio that is less than one-tenth that of the Soundex codes.

The approach permits the searching step of a linkage operation to be optimized, in terms of the numbers of (1) wasted comparisons to which an incoming record must be subjected in order to be brought together with a potentially linkable counterpart from the master file, and (2) failures to bring such records together. A tolerable level may be set for either the wasted comparisons or the failures, and the other value may then be minimized. Adjustment is achieved by adding or deleting an item from the sequencing information, thus increasing or decreasing the fineness of subdivision and the errors simultaneously until the required balance is struck. At no time should the sequencing information include an item with a lower merit ratio where one with a higher ratio is available. The cost of the searching step is thus balanced against its precision with a view to getting the best possible bargain.

In practice, we have found that by sequencing a master file of 114,000 marriage records in order of the pairs of surname codes for the grooms and brides, the number of wasted comparisons was kept at a very low level, i.e., 0.6 per incoming birth record where the births had arisen from marriages represented in the master file and 1.6 for all other incoming birth records. The number of failures to bring potentially linkable records together for comparison due to spelling discrepancies that altered one or other of the Soundex codes amounted to 1.6% of the potentially possible linkages.

The discussion so far has assumed that all of the linkings will be carried out using files arranged in a single sequence. However, the cost of sorting by computer is rapidly diminishing. Where more than one sequence is permitted, an even better bargain may be struck in terms of the precision that can be achieved for any given number of wasted comparisons. Linkings may then be carried out using very fine subdivisions of the file sequences, based on information of quite limited reliability, with the assurance that potentially linkable pairs of records which are not brought together on the first search will be compared in one of the alternative sequences based on other identifying information.

One quite large manual test of such a procedure has been carried out in

Table 3. Identifying Information on Vital Records

Event and individual	Birth name	Birth-place*	Birth date (or age)
Marriage			
Groom	+	+	(+)
Bride	+	+	(+)
Father of groom	+	+	
Mother of groom	+	+	
Father of bride	+	+	
Mother of bride	+	+	
Birth			
Child	+	+	+
Father	+	+	(+)
Mother	+	+	(+)
Death			
Deceased	+	+	+
Spouse	+		
Father	+	+	
Mother	+	+	

*i.e., city or place, and province or country.

which initials and provinces of birth were substituted in the secondary sequences for one or other of the two surname codes. This test showed that a reduction in errors by more than tenfold could be achieved at the price of a two- to three-fold increase in wasted comparisons.

Where the avoidance of "lost" linkages is of special importance, the use of multiple alternative sequences represents an ultimate in refinement.

2. *Optimizing the Matching Step*

When pairs of records are brought together for comparison, decisions must be made as to whether these are to be regarded as linked, not linked, or possibly linked, depending upon the various agreements and disagreements of items of identifying information. It is also desirable that such decisions be based on numerical estimates of the degrees of assurance that the records do or do not relate to the same persons. The computer is asked, in effect, to simulate the processes of human judgment and to make the best use it can of the items of identifying information that are individually unreliable but collectively of considerable discriminating power.

The extent of the personal information that is usually entered in the vital registration makes the potential accuracy of the linkings of these records high indeed. Newborn children, grooms and brides, and deceased persons are commonly identified by their full birth names, their birth dates or ages, and their birthplaces. Together with this personal identification, there is a substantial amount of family information. The full names of the parents, including the maiden surname of the mother, are usually given, as well as their birthplaces. In addition, the ages of married couples are entered in the records of their marriages and the records of the births of their children (Table 3).

Thus, there is an abundance of overlapping information that may be used to link (1) deaths to births, (2) births to the parental marriages and to the births of older siblings, and (3) marriage records of brides and grooms to their birth records, to the marriage records of their parents, and to the birth and marriage records of their siblings (Table 4). Even where some of the items fail to agree, the combined discriminating power of such information is almost always large.

A human filing clerk attempting to carry out such a grouping operation would intuitively attach greater positive weight to some of the agreements than to others and greater negative weight to some of the disagreements than to others. In each instance, the question that is asked, almost unconsciously, is, "Would such an agreement be likely to have occurred by chance if the pair of records *did not* relate to the same person?" or "Would such a disagreement be likely to have occurred by chance if the pair of records *did* in fact relate to the same person?" The answer in each case will depend upon prior knowledge gained from experience. An initial known to be rare, such as "Z," will be regarded as less likely to agree by chance on a pair of records than would a commonly occurring initial such as "J." Similarly, a highly reliable and stable item of identification, such as sex, when it fails to agree, will argue more strongly that the people referred to are *not* the same than would, for example, disagreement of province of birth, which is known from our own experience to be discordant in about one per cent of genuinely linked pairs of records.

The mathematical basis of such intuitive assessments is really quite simple. In general, agreements of initials, birth dates, and such will be more common in genuinely linked pairs of records than in pairs brought together for comparison and rejected as unlinkable. The greater the ratio of these two frequencies, the greater will be the weight attached to the particular kind of agreement.

If we wish to obtain numerical weights that can be added to other such weights, the above ratio may simply be converted to a logarithm. In practice, the logarithm to the base two has proved particularly convenient. These so-called *binit weights* are simply

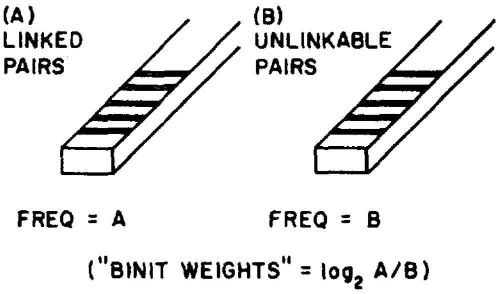
$$W_t = \log_2 (A/B) \qquad (4)$$

where A and B are the frequencies of the particular agreement, defined as specifically as one wishes, among linked pairs of records and among pairs that are rejected as unlinkable. The binit weights for agreements will have positive values because A in such circumstances is always greater than B (Fig. 4), and these weights may be regarded as strictly analogous to the discriminating powers discussed earlier except that they relate to particular values of the various items of identifying information.

There is no need to alter this formula when deriving the weights for disagreements. A and B may be regarded simply as the frequencies of the particular disagreement, defined in any way, among linked and unlinked pairs of records. Usually the weights will then be negative in sign, because disagree-

Table 4. Examples of Kinds of Linkage

Event		Parental information (husband × wife)				Individual information	
Kind	Year	Surnames	Initials	Birthplace codes	Ages	Name	Birth date (or age)
Death to birth							
Birth	1950	Doe × Cox	JA MB	09 09	30 25	Fred	15.6.50
Death	1955	Doe × Cox	JA MB	09 09	— —	Fred	15.6.50
Birth to parental marriage							
Parental marriage	1945	Doe × Cox	JA MB	09 09	25 20	—	—
Birth	1950	Doe × Cox	JA MB	09 09	30 25	Fred	15.6.50
Marriage of a groom, to own birth and own parents' marriage							
Parental marriage	1945	Doe × Cox	JA MB	09 09	25 20	—	—
Birth	1946	Doe × Cox	JA MB	09 09	26 21	Andy	18.5.46
Own marriage	1966	Doe × Cox	JA MB	09 09	— —	Andy	(age 20)



Examples

Kinds of agreements or disagreements	Frequency in linked pairs *A*	Frequency in unlinkable pairs *B*	Ratio *A/B*	Binit weight $\log_2 A/B$
Agreements				
Male sex	1/2	1/4	2	+1
Initial "J"	1/16	1/256	16	+4
Initial "Z"	1/1000	1/1,000,000	1000	+10
Disagreements				
City of residence	1/3	2/3	1/2	−1
Initial (any)	1/40	32/40	1/32	−5
Sex	1/8000	1/2	1/4000	−12

Fig. 4. Calculating "binit weights."

ments are, in most instances, less common among the linked than among the unlinked pairs; i.e., *A* will be less than *B*, and the logarithm of *A*/*B* will be negative.

Exceptions will occur in which an apparent disagreement is in reality a partial agreement. For example, a discrepancy of one year of age, after allowance is made for the interval of time between the two registered events, will frequently be a reflection of an underlying genuine agreement. Fortunately, however, it is not necessary to prejudge the issue. If the apparent discrepancy is predominantly a reflection of a partial agreement, the calculated weight will automatically turn out to be positive.

In practice, the formula is used to derive from the actual files a set of look-up tables of weights for agreements and disagreements of various items of information, broken down by the natures of these agreements and disagreements to whatever extent is necessary to make nearly full use of the discriminating powers. Such tables are stored in the memory of the computer. For each detailed comparison of a pair of records, the positive and negative weights appropriate for the different agreements and disagreements are added together, and the total weight is used to indicate the degree of assurance that the pair do, or do not, relate to the same person. The procedure assumes as a tolerable approximation that the weight for the individual agreements or disagreements are uncorrelated with each other; corrections are possible where this is not strictly true, but in our own experience these have been too small to be worth applying.

The derivation and use of the binit weighting factors have been described in greater detail elsewhere (Newcombe *et al.*, 1959; Newcombe and Kennedy, 1962). For present purposes, it is sufficient to indicate that there is great flexibility in the manner in which the weights can be employed and that they permit the introduction of numerous refinements so as to make nearly full use of the discriminating power inherent in the identifying information. For anyone planning an actual application, I would recommend that a number of small linking studies be carried out by hand to provide an opportunity to experiment with the system and become familiar with its characteristics.

The total binit weight represents the extent to which assurance of a genuine linkage is increased, or decreased, as a result of the comparisons made. Such weights are, in fact, logarithms to the base two of the factors by which the odds in favor of a linkage are increased over and above what they would have been in the absence of the comparisons.

In our own operation, the linkages are carried out within the very small "double surname pockets" of the master file, which contain on the average between one and two records apiece. Furthermore, an incoming record is quite likely to find a linkable counterpart there. Thus, even in the absence of the detailed comparisons, the probability of a match with a record drawn at random from the correct pocket of the master file will not be so very much less than 50% (i.e., odds of 1:1). In this situation, the total binit weight will closely approximate the $\log_2$ of the odds in favor of a linkage. Weights of +10 and of +20, for example, may in this situation be regarded as indicating favorable odds of approximately 1,000 to 1 and 1,000,000 to 1, respectively.

Using the double-surname sequenced files in this manner, no weights are attached to agreements of the items of sequencing information, i.e., to agreements of the surname codes. The reason is that the discriminating powers of these have already been taken into account automatically, since it is this information which determines the sizes of the pockets in the master file.

If binit weights were attached to agreements and disagreements of the sequencing information, incoming records would then have to be thought of as linking within a population of records consisting of the whole of the master file. Suppose, for example, that this contained 10^6 records and was known to include one which matched each of the incoming records. Under these conditions, the chance of an incoming record linking with a randomly chosen record from the master file would be $1/10^6$ ($= 2^{-20}$). However, if the detailed comparisons yielded a weight of +24, this would raise the odds from 2^{-20} up to 2^4, i.e., to 16:1 in favor of a genuine linkage.

Thus, to derive from the total binit weights the odds in favor of a linkage, allowance must be made for the size of the population of records within which the linkage is carried out by subtracting $\log_2$ of this population size. Similarly, allowance must also be made for the limited probability that there is, in fact, a matching record within that particular population. The $\log_2$ of this probability will be negative in sign and when added to the total binit weight will further reduce its value.

In practice, thresholds must be set which specify the ranges of binit weights

TABLE 5. TYPICAL MAGNETIC TAPE FORMAT FOR A VITAL RECORD

Information	Word*
Soundex pair	1
List word	2
Event (date, etc.)	3–6
Husband (name, etc.)	7–9
Wife	10–12
Offspring	13–14
Record linkage cross reference	15–17
Sibship cross reference	18–19
Statistics	20–24
Other cross reference	25

*One word equals ten octal digits or five alphanumeric characters.

which are to be regarded as representing linkage, no linkage, and possible linkage. Initially, these thresholds may be set to what seem intuitively to be reasonable values, but empirical tests are needed to ensure that false linkages, failures to link, and tentative linkages are balanced in a reasonable fashion.

In an actual operation, the total weights for linked pairs should be recorded permanently as evidence of the degree of assurance on which the linking was based. Similarly, for pairs of records that are judged to be neither positively linkable nor positively nonlinkable but which represent the most likely linkage available, it is prudent to retain permanently information about each such doubtful link and the weight associated with it. As more information accumulates about the family groupings, such as the sequences of birth orders in the families and the intervals between the births, this further knowledge may assist with the resolution of some of these doubtful linkings, provided that the information about them is retained on the files.

3. *Factors Affecting the Speed of the Record Linking Operation*

A number of practical considerations will influence the speed of a record linking operation.

The individual magnetic tape records should not be unnecessarily large, as this will increase the times required for input and output and for sorting the records. It will also limit the number of records that can be manipulated within the available core memory at any one time. The record format chosen for our own linking operation, using the vital registrations, consists of 25 words of 30 or 32 bits each (depending upon the magnetic tape units used). Each word may contain ten octal digits or five alphanumeric characters. This size of record was found to be sufficient for the storage of the individual and family identifying information, the statistics, and the cross-referencing information pertaining to a vital registration (Table 5).

Speeds are also affected by the amount of unused space on the magnetic tapes between records or between "blocks" of records. On the tapes used with the Control Data G20 computer, on which most of the recent work was done, records are stored in addressable blocks of 800 words each, i.e., con-

Table 6. Example of List Processing

New record	Position	Record	Links	
			Forward	Back
G	(1)	G*	0	0
B	(1)	G	0	2
	(2)	B*	1	0
D	(1)	G	0	3
	(2)	B*	3	0
	(3)	D	1	2
F	(1)	G	0	4
	(2)	B*	3	0
	(3)	D	4	2
	(4)	F	1	3
A	(1)	G	0	4
	(2)	B	3	5
	(3)	D	4	2
	(4)	F	1	3
	(5)	A*	2	0

*Indicates "flag" for head of list.

taining 32 records per block. If records are read singly onto tape rather than in blocks, a substantial fraction of the tape is used up in the inter-record gaps.

A special time-saving feature in our own linking operation has been the use of a so-called "list processing" method. Records entering a husband-wife double surname pocket in the master file are arranged, physically, simply in order of their entry or acquisition, regardless of the appropriate logical sequence in the family groups. The logical position of each record is indicated by the inclusion on it of the "entry number" (i.e., acquisition number) of the record that logically preceeds it and that of the record that logically succeeds it. These numbers are known respectively as the backward and forward links.

When a new record enters the double surname pocket, known as a "super-family," it is placed physically at the end; backward and forward links are then entered in the incoming record, and the existing links on the records that immediately precede and succeed it in the logical sequences are updated (Table 6). The saving of time occurs because with this procedure there is no need to alter the physical positions of the records already in a pocket to make room for a new record each time one is to be interfiled. The list processing method used has been described in detail by Kennedy *et al.* (1964).

Another factor that affects the speed of a linking operation has been mentioned earlier, namely, the size of the units into which the file is broken by the sequencing information. In our own experience, the use of two phonetically coded surnames relating to the husband-wife pair has divided a master file of 114,000 marriage records into units containing on the average about 1.6 records each. For approximately 80% of the file the pairs of surname codes are unique, i.e., they occur only once in that combination throughout the whole file.

Under the various conditions described above as pertaining to our own operation, incoming birth records have been merged and linked with a master file of parental marriages and earlier births at a rate of 2,300 per minute. Thus for the British Columbia population of 1.6 million people, with which this study is concerned, a year's crop of 35,000 birth records can be merged and linked with the master family file of ten years of marriages in somewhat less than 30 minutes of machine time, once the magnetic tape records have been prepared in the proper format and appropriately sequenced. At a machine rental of two dollars per minute this is equivalent to a cost of 0.1 cents per record, i.e., it is minute in comparison with the cost of producing the punchcards in the first place, as is done routinely for administrative and statistical purposes.

The ways in which these various time-saving devices have been employed are described in greater detail by Kennedy *et al.* (1965).

STORAGE AND RETRIEVAL

In the sections that follow, we will consider the manner in which records relating to sibship groups may be stored together, certain extensions of the procedures to permit the inclusion of pedigree information covering an indefinite number of generations, and methods of retrieving information from the sibship grouping and multigeneration pedigrees. The records pertaining to the sibships, of course, fall within the main file sequence based on the surname pairs in their phonetically coded forms (Table 7).

1. *Storage of Sibship Groupings of Records*

There is a natural sequence in which the vital and health records pertaining to a sibship group may be linked and stored. Starting with the parental marriage registration, which may be regarded as a "head-of-family" record, birth records are linked to the marriage record in chronological order, and records of the various events of ill health, including death, are linked to the birth records of the children to whom they relate, those for a particular child falling likewise in chronological order after his or her birth record (Table 8).

The experience which we have had with this kind of file organization relates to records of marriages, livebirths, stillbirths, and deaths, together with those from a special register of handicapping conditions of children and adults. In addition, detailed plans have been worked out for the possible future inclusion of substantial numbers of records from a universal scheme of hospital insurance. Off-line linkings with the birth registration records are needed in the case of the handicap and hospital records in order to pick up the mother's maiden name which is lacking on the original form. Only after this has been done can the handicap and hospital records be merged and linked with the master family file, which is arranged in order of the two parental surname codes.

Incompleteness of a sibship grouping of records poses no special problem. In the absence of the parental marriage record, for example, the birth record of the oldest child represented in the file may serve as the head-of-family record, and records of the births of younger siblings will be linked to it. A

TABLE 7. EXAMPLE OF DOUBLE SOUNDEX FILE SEQUENCE[*]

Adams × Adair	A 352	A 360
Adams × Baron	A 352	B 650
Adams × Caird	A 352	C 630
Adams × Danys	A 352	D 520
↓		
Baker × Allen	B 260	A 450
Baker × Barks	B 260	B 620
Baker × Caron	B 260	C 650
Baker × Duffy	B 260	D 200
↓		
Baird × Aubry	B 630	A 160
Baird × Baker	B 630	B 260
(and so on)		

[*] i.e., by husband's surname code followed by the wife's maiden surname code.

TABLE 8. EXAMPLE OF A SIBSHIP GROUP OF RECORDS

Record	Parental couple	Child
Parental marriage	Doe × Cox	—
Birth 1	Doe × Cox	Alan
Birth 2	Doe × Cox	Carl
Ill health	Doe × Cox	Carl
Death	Doe × Cox	Carl
Birth 3	Doe × Cox	Edna

death record may serve likewise as a head-of-family record where it relates to the oldest child represented in the family group and the birth record for this child is missing. Thus, all of the available records of vital and health events may be merged and linked into sibship arrays, regardless of the degree of completeness or incompleteness of these groupings, and the master file may be updated periodically by the introduction into it of successive crops of current records.

The times required to merge and link the death and handicap records to the master file are somewhat greater than those for the corresponding operation as applied to birth records. There are two reasons for this. First, an ill health or death record must scan all of the birth records present in the appropriate double surname pocket of the master file, and these will tend to be more numerous than the head-of-family records which the incoming births must scan. Second, where an incoming ill health or death record fails to find a matching birth record, it must scan the double surname pocket a second time in an attempt to find a head-of-family record with which to link.

In our own operation, handicap and death records were merged and linked with the master file at a rate of approximately 1,100 per minute, i.e., at about one-half of the speed for the merging and linking of birth records.

2. *Storage of Multigeneration Pedigrees*

The modifications of the above procedures needed to permit the linking and storage of the vital and health records in the form of multigeneration pedigrees are surprisingly simple. For most registration areas, the marriage records contain sufficient information to serve as bridges between the generations and between the in-law sibships.

Information from a marriage record may be treated in two ways. We have discussed already how it can be arranged into the form of a head-of-family record representing the marriage of a parental couple. Similarly, information from the registration form may also be fitted into the format of a record such as is used to describe an event in the life of an individual. The part of this latter kind of record entry that is assigned to family information would then contain the names and other identifying particulars of the parents of the newly married person, and the part of the record assigned to personal identification would contain his or her own name, age, and birthplace. This kind of entry of the marriage information is almost precisely analogous to a death record, since both relate to events in the lives of members of a sibship group. In the master file, the three entries pertaining to a particular event of marriage (i.e., the groom's entry, the bride's entry, and the head-of-family entry) will each become part of a different sibship group of records.

The only special requirement for the three marriage entry records is that each of them, before being placed in these various locations on the master tape, be cross-referenced to the other two. This is done by inserting in the cross-reference field of each record entry the double surname codes for the other two. These codes, together with the marriage registration number which is common to all three entries, provide both a means of access within the master file from one of the double surname pockets to the other two and a positive identification of the alternative entries when the pockets in which they occur have been located. The cross-referencing is illustrated in Tables 9 and 10.

The simplicity of the procedure resides in the use of essentially the same format for the marriage entries of grooms or brides as for their death records. In our own operation, the same programs that are used to build the sibship groupings of records will also be employed to insert into these groupings the grooms' and brides' marriage entries, just as they would the records of any other kinds of events in the lives of the same individuals.

The idea of thus putting family groups of records into a single linear array and of using cross references to indicate the relationships between the groupings that are filed as units is basic to any system by which computers may be employed to store and retrieve large quantities of pedigree information of unlimited complexity. The special features of the system described are merely matters of convenience. The choice of the sibship group as the unit of storage and of the surname pair as the sequencing information may have fairly wide application, but the details of the use of identifying particulars have been dictated largely by the nature of the vital records.

It would, of course, be feasible to store the same pedigree information more compactly if the family relationships were worked out in advance so that every individual could be assigned an identifying number containing as few

TABLE 9. EXAMPLE OF A MARRIAGE REGISTRATION AND OF THE MARRIAGE ENTRY RECORDS DERIVED FROM IT

Marriage registration	
Groom	Dunn, Alex
Bride	Rowe, Anna
Groom's father	Dunn, Carl
Groom's mother	Bell, Edna
Bride's father	Rowe, Paul
Bride's mother	Hill, Jean

Marriage entry records	Parental couple	Offspring
1. Head of family entry	Dunn × Rowe (Alex) (Anna)	—
2. Groom's entry	Dunn × Bell (Carl) (Edna)	Alex
3. Bride's entry	Rowe × Hill (Paul) (Jean)	Anna

TABLE 10. EXAMPLE OF CROSS-REFERENCING A SIBSHIP TO THE RELATED SIBSHIPS

Record	Parental couple	Offspring	Cross references
Parental marriage	Dunn × Bell		Dunn × Nash—father's sibship Bell × Mann—mother's sibship
Birth 1	Dunn × Bell	Alex	
Groom's entry	Dunn × Bell	Alex	Dunn × Rowe—new family Rowe × Hill—bride's sibship
Birth 2	Dunn × Bell	Stan	
Groom's entry	Dunn × Bell	Stan	Dunn × Knox—new family Knox × Fynn—bride's sibship

digits as possible, but the disadvantages of this approach where large populations are involved should perhaps be mentioned. A main objective of the present handling procedures has been to avoid entirely all manual manipulations so that full use can be made of the speeds of electronic computers. If this feature is to be preserved, the present kind of linking operation would have to be carried out anyway. A more important problem would be what to do with the borderline linkings when condensing the pedigree information into its more compact form, since both the extents of the uncertainties and the means for their later resolution would tend to be lost in the process. It might also be difficult to keep open the possibility, as the present system does, of merging at some future time the pedigrees drawn from a limited region, such as a province or a state, with those for a wider region such as the country as a whole.

3. *Retrieval of Pedigree Information*

The need for writing detailed programs does not end with the establishment of a master family file containing the required pedigree information. For almost any kind of genetic study, the extraction of the required tabular information from a printed listing of the master file would be almost unthinkably laborious and expensive.

In general, it is necessary first to prepare programs that will summarize in a single record whatever information is required about a particular family. A further program is then written to extract information in tabular form from the resulting file of these summary records. Two examples of such procedures will be described, relating to sibship groups and to multigeneration pedigrees, respectively.

Where the family units under study are restricted to the sibships, summaries of the events of birth, ill health, and death in the lives of the various members of a sibship will usually be derived in two steps. First, individual histories will be condensed so that there is just a single summary record for each child replacing the separate records for the various events. The resulting magnetic tape file of individual or personal summaries can be used repeatedly to prepare the much more compact family summary records, which may be of a variety of kinds depending upon the natures of the studies for which they are to be used (Table 11).

To facilitate subsequent tabulations, the family summary records will have a different fixed field for each of the siblings. There must also be provision for large families, which will sometimes overrun a family summary record of modest size. This is best taken care of by arranging for trailing records to act as extensions where needed.

In one study which we have done using this procedure, the coded causes of stillbirths, handicaps, and deaths were entered into the fields of the family summary record assigned to the particular siblings who were affected, and for the unaffected siblings just the fact of birth, the birth order, and the sex of the child were entered.

In this particular study, use was made of the family summaries to derive information about the magnitudes of the risks to the later-born siblings of children who had been stillborn, handicapped, or had died, as the result of diseases of various kinds. The tabulations contained, typically, the number of index cases of a disease, the numbers of earlier and later siblings of the index cases, and the number of later-born siblings suffering from the same condition (Table 12). For details of the computer programs by which the different steps in the extraction were carried out, the reader is referred to Smith *et al.* (1965).

A more elaborate procedure is required where multigeneration pedigrees are to be summarized, because as an initial step the sibship groupings of records relating to a particular family must be brought together from different parts of the master file. Before starting this step, certain sibships whose relatives one wishes to ascertain will have been extracted from the master file. These may be called "index sibships," and they will in most instances have been chosen because they include individuals who are affected by some disease of special interest.

TABLE 11. EXAMPLES OF INDIVIDUAL AND FAMILY SUMMARY RECORDS

		Event records for a sibship (one per event)		
Event code	Birth order	Family	Child	Disease code
J (birth)	1	Fox × Dow	Alan	—
J (birth)	2	Fox × Dow	John	—
J (birth)	3	Fox × Dow	Vera	—
Q (handicap)		Fox × Dow	Vera	123
J (birth)	4	Fox × Dow	Leon	—
R (death)		Fox × Dow	Leon	456
		Individual summary records (one per child)		
(J)	1	Fox × Dow	Alan	—
(J)	2	Fox × Dow	John	—
(Q)	3	Fox × Dow	Vera	123
(R)	4	Fox × Dow	Leon	456
		Family summary record (one per sibship)		
(Fox × Dow)	1 (J)---, 2 (J)---, 3 (Q) 123, 4 (R) 456.			

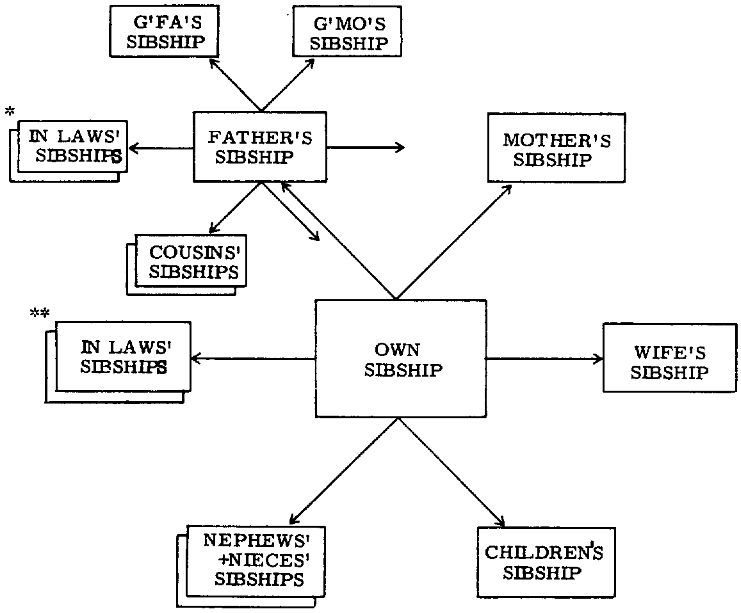
TABLE 12. EXAMPLE OF A TABULATION FROM FAMILY SUMMARY RECORDS

	Disease code 325 (*mental deficiency*)				
	Normal (J)	Stillborn (K)	Handicapped (Q)	Dead (R)	Handicapped and dead (S)
Index cases	0	0	506	9	58
Earlier sibs	208	2	6	16	0
Later sibs, same cause	0	0	11	0	1
Other later sibs	286	2	11	14	0

The records of the index sibships may contain cross-referencing information (in the form of double-surname codings and marriage registration numbers) indicating links with as many as six different kinds of related sibships, i.e.,

1. From the parental marriage (head-of-family) records to
   (a) the fathers' sibships and
   (b) the mothers' sibships.
2. From the marriage records of the "affected" individuals who got married (i.e., from the grooms' and brides' entries) to
   (c) their offspring's sibships and
   (d) their spouses' sibships.
3. From the marriage records of the brothers and sisters who got married to
   (e) the sibships of the nephews and nieces of the affected individuals and
   (f) the sibships of the spouses of the brothers and sisters who got married.

These six different kinds of cross references may be used in a single scan to draw from the master family file all of the groups of records pertaining to sibships that are removed by *one* degree of relationships from those in which the affected individuals occurred, including the in-law groups (Fig. 5).



*i.e., those of the paternal uncles and aunts by marriage.
**i.e., those of brothers' wives and sisters' husbands.

FIG. 5. Scanning the master file for related sibships.

Similarly, in a second scan of the master tape, use may be made of the further cross-referencing information contained in the sibship groups of these six different kinds to extract the sibships that are removed by *two* degrees of relationship from those in which the affected individuals occurred. Again, the in-law sibships may be extracted in the same way as those of the blood relatives. And so, with each successive scan, an expanding circle of more distant relatives may be identified and retrieved from the master file.

Each such scan will be exceedingly rapid even where large numbers of sibships groups are extracted. Thus, it is feasible to carry out the retrieval of multigeneration pedigrees on a truly massive scale.

From this point on, the making of summaries would follow much the same pattern as described earlier, except that the family summary record might be more complex than the sibship summary record.

The chief limiting factor in work of this kind is not the speed of the computer but the time required to develop the appropriate programs.

THE LIKELIHOOD OF FUTURE "TOTAL UTILIZATION" OF PEDIGREE INFORMATION

Geneticists will at first tend to think of the possible uses of record linking as applied simply to the familiar kinds of *ad hoc* studies of limited size and duration. The question arises whether it is realistic to go beyond this and to consider using for scientific purposes all of the pedigree information gathered routinely for whole populations through the vital registration systems, of doing so on a continuing basis, and of adding an increasing amount of medical documentation as time goes on.

Clearly, the cost would appear large if it were paid wholly from budgets for scientific research. But this would not necessarily be the case, because the information that is unlocked by linking and integrating the files into individual and family histories has many statistical and administrative uses, as well as other scientific uses beyond those of the geneticist.

Those geneticists who attempt to apply the methods of record linking will be in a particularly good position to see a variety of possible uses for the linked files and to develop procedures that will serve more than one purpose. Their own long-term interest may be furthered most where they exploit the fact that there arc other potential users.

Of course, with time the various files of routine records will, to an increasing extent, be linked and integrated anyway for administrative purposes, whether or not scientists take an interest in the matter. But the only way to ensure that scientific by-products will come out of this trend is for the scientists themselves to participate actively while the administrative procedures are being established.

APPENDIX
Surname Coding

Surnames may be converted into coded forms for either of two reasons: to set aside temporarily some unreliable component of the information that may vary on successive records relating to the same person, or for the sake of compactness. A number of systems have been designed to achieve one or other of these purposes, or both simultaneously. Some of the more useful of these codes will be described.

THE RUSSELL SOUNDEX CODE

This code is particularly efficient at setting aside unreliable components of the alphabetic surname information without losing more than a very small part of the total discriminating power. It is the method of choice for almost all populations, except where the names are predominantly of Oriental origin.

Rules:

1. The first letter of the surname is used in its uncoded form and serves as the prefix letter.
2. W and H are ignored entirely.
3. A, E, I, O, U, Y are not coded but serve as separators (see item 5 below).
4. Other letters are coded as follows until three digits are used up (the remaining letters are ignored):

B, P, F, V	coded 1
D, T	coded 3
L	coded 4
M, N	coded 5

R coded 6
All other consonants coded 2
(C, G, J, K, Q, S, X, Z)

5. Exceptions are letters which follow prefix letters which would, if coded, have the same code. These are ignored in all cases unless a separator (see item 3 above) precedes them.

Examples:

Anderson	= A 536
Bergmans, Brigham	= B 625
Birk, Berque, Birck	= B 620
Fisher, Fischer	= F 260
Lavoie	= L 100
Llwellyn	= L 450

NAME COMPRESSION

As indicated by its name, this form of coding is designed mainly to condense surnames, given names, and place names. However, the code does remain unchanged with some of the common spelling variations, although it is less efficient in this respect than the Soundex code.

Rules:

1. Delete the second of any pair of identical consonants.
2. Delete A, E. I, O, U, Y, except when the first letter of the name.

Examples:

BENNETT	= BNT
FISHER	= FSHR

ILL-SPELLED NAME ROUTINE

Where the insertion, deletion, or substitution of a single letter of a surname alters the coded form, recognition that a pair of names are the same necessarily depends upon residual similarities in the sequences of the letters in the two, despite any interruptions in these sequences. The "ill-spelled name routine" is not, strictly speaking, a system of coding but rather a system of comparison which employs the coded forms of the names as derived by "name compression." The system was designed for use with airline bookings (Davidson, 1962).

Rules:

1. Use "name compression" procedure, up to a total of four letters.
2. Search for and count the numbers of letters or blanks, up to a total of four in all, that agree without altering the sequence.
3. Where the agreements equal 3 or 4 in a pair of names, compare other identifying information.

Examples:

```
                                  Score
BOWMANN    = B M N -
             | | | |
BAUMAN     = B M N -                4
MCGONE     = M C G N
             | |   |
McKONE     = M C K N                3
ANGREIFF   = A N G R
               | | |
SINGER     = S N G R                3
MCGINESS   = M C G N
             |   / /
MAGINNES   = M G N S                3
LU         = L - - -
               | | |
ROO        = R - - -                3
```

ALPHANUMERIC CONVERSION

This is a highly specific numeric coding for all surnames. It is not designed to set aside the less stable parts of the information but rather to retain virtually all of the original specificity of the alphabetic form. The numeric form of the surname is compact, is more readily sorted on an electromechanical card sorter than the alphabetic form, and is nonrevealing to anyone who lacks the relevant look-up table. Furthermore, when sorted in numerical sequence the names fall in alphabetic order or a close approximation to it.

The coding is done by computer using a look-up table containing over 8,000 different entries. (See International Business Machines, 1960.)

Examples:

ABBIT	=	0008
ADLER	=	0105
BORNE	=	1058
BRYAN	=	1070
CLARK	=	1646
COX	=	1721
	↓	
ZZINA	=	9776

HOGBEN SURNAME CODE

This is a simple two-digit code for surnames based on a division of the names in a large telephone directory into 100 approximately equal parts. Although compact, it loses much of the discriminating power inherent in the full name and is therefore chiefly of historical interest. (Originally this was just a part of a much longer numeric code derived from the surname, first given name, sex, and birth date. See Hogben *et al.*, 1948.)

Examples:

00	=	AA	–	AK
01	=	AL		
02	=	AM	–	AR
03	=	AS	–	AZ
04	=	BAA	–	BAJ
05	=	BAK	–	BAQ
06	=	BAR		

(and so on)

REFERENCES

Davidson, L. 1962. Retrieval of misspelled names in an airlines passenger record system. *Commun. Assoc. Computing Machinery* 5: 169–171.

Hogben, L., Johnstone, M. M., and Cross, K. W. 1948. Identification of medical documents. *Brit. Med. J.* 1: 625–635.

International Business Machines. 1960. General Information Manual. *A Unique Computable Name Code for Alphabetic Account Numbering.*

Kennedy, J. M., Newcombe, H. B., Okazaki, E. A., and Smith, M. E. 1964. *List Processing Methods for Organizing Files of Linked Records.* Chalk River, Ontario: Atomic Energy of Canada, Ltd. Document AECL-2078.

Kennedy, J. M., Newcombe, H. B., Okazaki, E. A., and Smith, M. E. 1965. *Computer Methods for Family Linkage of Vital and Health Records.* Chalk River, Ontario: Atomic Energy of Canada, Ltd. Document AECL-2222.

Newcombe, H. B., and Kennedy, J. M. 1962. Record linkage: Making maximum use of the discriminating power of identifying information. *Commun. Assoc. Computing Machinery* 5: 563–566.

Newcombe, H. B., Kennedy, J. M., Axford, S. J., and James, A. P. 1959. Automatic linkage of vital and health records. *Science* 130: 954–959.

Smith, M. E., Schwartz, R. R., and Newcombe, H. B. 1965. *Computer Methods for Extracting Sibship Data from Family Groupings of Records.* Chalk River, Ontario: Atomic Energy of Canada, Ltd. Document AECL-2530.

A MODEL FOR OPTIMUM LINKAGE OF RECORDS*

BENJAMIN J. TEPPING
Bureau of the Census

A model is presented for the frequently recurring problem of linking records from two lists. The criterion for an optimum decision rule is taken to be the minimization of the expected total costs associated with the various actions that may be taken for each pair of records that may be compared. A procedure is described for estimating parameters of the model and for successively improving the decision rule. Illustrative results for an application to a file maintenance problem are given.

1. INTRODUCTION

THE problem of record linkage arises in many contexts. A typical example is that of file maintenance. In this example there is a file, which we shall call the master file, whose constitution is to be changed from time to time, by adding or deleting records or by altering specific records. Notice of these required changes is given by means of another file of records, which we shall call the transaction file. Presumably, each transaction record specifies the addition of a new master file record, or the deletion of an existing master file record, or the alteration of an existing master file record. It may not be known whether there exists a master file record that corresponds to a given transaction record so that the determination of whether a master file record is to be changed or a new master file record added must wait until it is found whether a corresponding master file record exists. Thus, the fundamental problem is to determine, for each transaction record, which master file record corresponds to it or that no master file record corresponds to it.

If each master file record and each transaction record carried a unique and error-free identification code, the problem would reduce to one of finding an optimum search sequence that would minimize the total number of comparisons. In most cases encountered in practice, the identification of the record is neither unique nor error-free. Thus it becomes necessary to make a decision as to whether or not a given transaction record ought to be treated as though it corresponded to a given master file record. The evidence presented by the identification codes of the two records in question may possibly be quite clear that the records correspond or that they do not correspond. On the other hand, the evidence may not clearly point to one or the other of these two decisions. Thus it may be reasonable to treat the records temporarily as if they corresponded or to treat them temporarily as if they did not correspond, but to seek further information. Or it may be reasonable in a particular case to take no overt action until further information has been obtained. The amount of effort that it is reasonable to expend in resolving a particular problem is also a variable. Thus it is clear that in making the decision on the correspondence between a transaction record and a master file record, there are available at least two and perhaps more possible decisions. If one considers now the costs of the various actions that might be taken and the utilities associated with their pos-

*Reprinted with permission from the Journal of the American Statistical Association, American Statistical Association, December 1968, Vol. 63, pp. 1321-1332.

sible outcomes, it appears to be desirable to choose decision rules that will in some sense minimize the costs of the operation.

There are many other contexts in which record linkage takes place. One example is that in which two files are to be consolidated. Information about some individuals may be contained in one or another of the two files, while for other individuals some information may be in one file and some in the other. Another example is that of multi-frame sample surveys in which it may be necessary to determine which of the sampling units in one frame are also included in the other frame. A third example is that of geographic coding in which the master file consists of a street address guide and the transaction records are particular addresses; the problem here is to assign to each address a geographic code as given by the street address guide. The reader can doubtless supply many other examples.

The literature on this subject is replete with descriptions of actual matching operations ([2], [3], [4], [7], [8], [10], [11], [12], [13], [17], [18]). Several also deal with principles for the design of matching operations ([4], [7], [8], [9], [11], [12]). Some formulate mathematical models to serve as a basis for the design of a matching process that will be optimum in some sense. Thus, in analogy to the Neyman-Pearson theory of testing statistical hypotheses, Sunter and Fellegi [14][1] fix the probabilities of erroneous matches and erroneous non-matches and minimize the probability of cases for which no decision is made. Nathan ([5], [6]) proposes a model that involves minimization of a cost function, but restricts detailed discussion to cases in which the information used for matching appears in precisely the same form whenever the item exists in either list. Du Bois' [1] approach is to attempt to maximize the set of correct matches while minimizing the set of erroneous matches.

This paper proposes a mathematical model of the record linkage problem and a decision rule which minimizes the cost. The implementation of this model in practice depends upon the estimation of the parameters of the model. These parameters are costs and certain probabilities. The parameters may be difficult to determine. Also, it will be seen, the mathematical model (as usual) is not an exact representation of the real world. Nevertheless, the model provides useful guides for the construction of efficient linkage rules, as will be illustrated in the sequel.

2. A MATHEMATICAL MODEL

There are given two lists: a list A (the master file, say) which consists of a set of labels $\{\alpha\}$ and a list B (the transaction file, say) consisting of a set of labels $\{\beta\}$. (See Section 6 for a simple example.) Each label α is to be compared with each label β and an action taken on the basis of that comparison. The action taken must be one of a list of possible actions exemplified by, but not confined to, the following:

1. Treat the labels α and β as if they designated the same individual of some population. We shall say that the pair (α, β) is a "link".

[1] The notation and terminology used here follow, generally, those of the Sunter-Fellegi paper.

2. Temporarily treat the labels α and β as a link but obtain additional information before classifying the pair as a link or a non-link.
3. Take no action immediately but obtain additional information before classifying the pair as a link or non-link.
4. Temporarily treat the labels α and β as if they were associated with different individuals of the population, but obtain additional information before classifying the pair as link or non-link.
5. Treat the labels α and β as if they were associated with different individuals of the population (non-link).

Other actions may be added to the list, including for example the use of a randomizing device to determine the treatment of the pair (α, β). Each pair (α, β) will be called a "comparison pair." It is assumed that each pair (α, β) is either a "match" (the labels α and β are associated with the same individual of the population) or a "nonmatch" (the labels α and β are associated with different individuals of the population). Thus the set of all comparison pairs is the sum of mutually exclusive sets M (the "match" pairs) and U (the "nonmatch" pairs).

It should be noted that the labels α and β are, in general, vector-valued. Thus a label may contain, for example, a name, address, age, and other characteristics of a person.

Theoretically, any comparison of the label α with the label β consists of constructing a vector-valued function γ of the comparison pair (α, β). (See Section 6 for a simple example of a comparison function.) The comparison function γ serves to classify all pairs into classes: (α_1, β_1) and (α_2, β_2) are members of the same class if and only if $\gamma(\alpha_1, \beta_1) = \gamma(\alpha_2, \beta_2)$. The comparison pairs in each given class are to be subjected to exactly one of s possible "actions" $a_1, a_2, \cdots, a_s$. (Examples of five possible actions were given above.) A "linkage rule" consists of the assignment of an action to each class.

Let a label α be selected at random from list A and a label β from list B, and let a non-negative loss $g(a_i; \alpha, \beta)$ be associated with taking action a_i on a pair (α, β). Let

$$P[M \mid \gamma] \equiv \text{Prob}[(\alpha, \beta) \epsilon M \mid \gamma(\alpha, \beta)]$$

denote the conditional probability that the pair (α, β) is a match, given the value of γ.

We assume here that G, the expected value of $g(a_i; \alpha, \beta)$, is a function only of a_i and $P[M|\gamma]$. (This assumption is discussed below, in Section 4.) Thus

$$G = \mathcal{E}\{g(a_i; \alpha, \beta) \mid a_i, P[M \mid \gamma]\} = G(a_i, P[M \mid \gamma]).$$

Given a linkage rule, the total expected loss of the rule is

$$\sum P(\gamma) \times G(a_i, P[M \mid \gamma])$$

where a_i is the action specified for γ by the linkage rule, and the summation extends over all γ. To minimize the total loss, we need only minimize each term of the sum, each term being non-negative.

A special case of the above is that in which there is a loss G_{i1} associated with taking action a_i on a pair (α, β) when in fact that pair is a match, and a loss G_{i2} when in fact the pair is a nonmatch. In this case G, the expected value of the loss, can easily be seen to be a linear function of the conditional probability that the comparison pair is a match, given γ, for each action a_i.

If the functions G are linear in $P(M|\gamma)$, the interval (0, 1) for the probability of a match is divided into at most s "action intervals" each of which corresponds to one of the possible s actions. The action interval for a given action is the interval in which the cost function G for that action is less than the cost function for any other action.

Figure 1 illustrates a case in which $G(a_i, P[M|\gamma])$ is a linear function of $P[M|\gamma]$ for each a_i. In this illustration, the optimum linkage rule specifies:

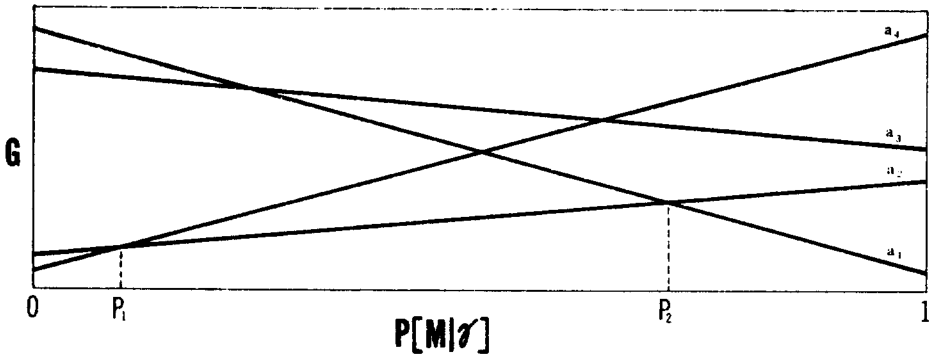


Fig. 1.

$$\begin{aligned}
&\text{Take action } a_4 \quad \text{if} \quad 0 \leqq P[M|\gamma] \leqq P_1\\
&\text{Take action } a_2 \quad \text{if} \quad P_1 < P[M|\gamma] \leqq P_2\\
&\text{Take action } a_1 \quad \text{if} \quad P_2 < P[M|\gamma] \leqq 1
\end{aligned}$$

If the functions G are not linear in $P[M|\gamma]$, an "action set" of points of the interval (0, 1) that correspond to one of the possible actions will not be an interval in general. The treatment of the nonlinear case, however, proceeds along the same lines.

The conditional probability that a comparison pair is a match, given that the comparison function γ has a stated value depends upon the prior definition of the comparison function γ or, equivalently, upon the definition of the corresponding classification of comparison pairs.

As noted above, any comparison function γ defines a classification of the pairs (α, β). Let γ' be any other comparison function, which therefore defines another classification. It is possible to pass from the classification γ to the classification γ' by a sequence of steps, each of which consists either of splitting a class into two classes or of combining two classes into a single class. Therefore, if we begin with a tentative comparison function γ, we may seek ways of splitting some classes or combining some classes in such a way as to reduce the contribution of the classes involved to the loss function.

Consider the case of splitting a class γ into two classes γ_1 and γ_2. Without loss of generality, we may assume that

$$P(M \mid \gamma_1) \leqq P(M \mid \gamma_2).$$

But then, clearly,

$$P(M \mid \gamma_1) \leqq P(M \mid \gamma) \leqq P(M \mid \gamma_2).$$

If $P(M|\gamma_1)$ and $P(M|\gamma_2)$ are in the same action set as $P(M|\gamma)$, there is no gain in making the split. But if either $P(M|\gamma_1)$ or $P(M|\gamma_2)$ falls into a different action set, the loss is necessarily (and sometimes materially) reduced.

To determine for which classes splits should be considered, one may first calculate the expected loss contribution for each class. It is evident that if the expected loss for a class is a small proportion of the total, little can be gained by splitting that class. Therefore, attention should be given first to classes whose expected loss contribution is a substantial proportion of the total. The illustration given subsequently shows that large reductions in the total expected cost can be attained by this technique.

With regard to the combining of classes, it is clear that this cannot result in reducing the expected cost. But if the classes to be combined are in the same action set, no increase in the cost will be sustained while the combination may reduce somewhat the operational costs of implementing the linkage rule. The combining of classes is useful also as an initial step, for the purpose of reducing the number of classes for which estimates need to be made, as detailed in Section 3, below.

3. ESTIMATION PROBLEMS

The application of the mathematical model involves estimating the cost function for each action as a function of the probability of a match, and estimating the probability that a comparison pair is a match.

The estimation of the cost function is often extremely difficult. Usually the cost consists of two classes of components, one class consisting of the cost of actual operations that may be involved and the other of the less tangible losses associated with the occurrence of errors of matching. The former can often be estimated very well, but estimates of the latter may depend upon judgment in large part. Despite the possible dependence on judgment, in the framework of the mathematical model even rough guesses at the cost function are extremely useful.

It may be noted that the first class of components of the cost function usually contains some components that are functions of the linkage rule (specifically, of the classification imposed). This is not reflected in the model, which only defines an optimum linkage rule for a fixed classification or comparison function.

It should be noted in connection with the estimation of the probabilities that it is necessary only to determine in which of the action sets a given probability falls. Ordinarily the probabilities will be estimated by selecting a sample in each comparison class. The sampling designs used should be chosen with the whole problem in mind, so that unnecessary sampling costs are avoided when, for example, the probability being estimated is near the center of an action interval or when an error in the estimate of the probability will have little effect on the total cost. The latter may occur if the frequency of the given comparison class is small or if the alternative actions in the neighborhood of a given probability lead to costs which are only slightly different.

The successive steps in the application of the mathematical model may be described as follows:

1. The possible actions that may be taken on a comparison pair are listed.
2. For each action, the mathematical expectation of the cost as a function of the probability of a match is estimated.
3. An initial comparison function, i.e., an initial classification of comparison pairs into comparison classes, is determined on the basis of judgment or past experience (see, for example, [2], [3], [4], [7], [8], [9], [10], [11], [12], [15], [17], [18]), or on the basis of mathematical conclusions following from specified assumptions[2] about the interaction of the components of the labels α and β. The more nearly the initial classification resembles the optimum classification, the less is the amount of subsequent work required to attain the classification that will finally be used.
4. Samples are selected from each comparison class and the probability of a match estimated for each comparison class. This determines the optimum action pattern for the given classification.
5. The contributions of the several comparison classes to the total cost is now analyzed, and the classes that provide large contributions to that total cost are identified.
6. On the basis of that analysis, the classification is revised by splitting and recombining classes.
7. Steps 4 to 6 are repeated until step 6 indicates that no substantial additional reduction of cost can be made.

4. SOME COMMENTS ON THE MODEL

As is usually the case with a mathematical model, the model does not, in every respect, faithfully represent the real world that it is intended to describe.

The model assumes that every possible comparison pair will actually be examined. With large files, this would involve an inordinate number of comparisons. In practice, comparisons would be confined to specified subsets of the master file, and corresponding subsets of the transaction file. From the point of view of the mathematical model, the comparisons not actually made are being treated as non-links.

A limitation of the model is that it permits a given element of the transaction file to be treated as a link with more than one element of the master file. In many situations, this treatment may be intolerable. The difficulty can be handled by subjecting all such multiple-link cases to a subsequent stage in

[2] Thus Sunter and Fellegi [14] suggest that the components of the comparison vector may be grouped into sub-vectors which are statistically independent on each of the sets M and U. They then show how the value of a parameter equivalent to $P[M \mid \gamma]$ may be estimated on the basis of a knowledge of the frequency distribution of γ. This would serve to define an initial comparison function, even if the assumption of independence is not a satisfactory one.

which the transaction record is linked with at most one of the master file records associated with it in the first stage. If the cost or frequency of such cases is small, the mathematical model described in this paper remains a useful one for guiding the design of the linkage rule.

Similarly, there exist situations in which the linkage of a master file record with more than one transaction record is not tolerated.

There are some situations in which the cost is not only a function of the probability of a match but also of some other characteristic of the comparison pair. Thus, there may be two types of master file records, with the cost of an erroneous link being different for the two types. In such a situation, the comparison pairs may be classified in such a way that the characteristic is constant within each class and then the problem of optimum linkage may be treated as a separate problem in each of these classes.

The model is applicable also to cases in which the master file is not fixed but changes from one time period to another. Each transaction record is to be compared with the master file as it exists at the time period when the transaction record enters the system. We may consider the sequence of master files as constituting list A and a corresponding sequence of transaction files as constituting list B. The identity of the particular file becomes a component of the comparison vector γ, and we may define (α, β) to be a member of U if α and β are not from corresponding files. In this manner, this situation is covered by the model.

Some comments on the characteristics of useful comparison function are in order. Typically, the cost function

$$G(P) = \min_{a_i} G(a_i, P[M \mid \gamma])$$

is a concave function of P, with $G(0) = G(1) = 0$. Thus, the ideal comparison function is one for which $P[M|\gamma]$ is either 0 or 1 for every value of γ that may be observed. This ideal is usually not attained. However, one can usually find an initial comparison function such that the distribution of $P[M|\gamma]$ over the set of all comparison pairs is U-shaped, with low frequency where the cost function is high and high frequency where the cost function is low. Carrying through the steps given in Section 3 will often result in revising the comparison function γ so that the distribution of $P[M|\gamma]$ is shifted nearer the endpoints of the interval (0, 1).

Finally, it should be noted that the successive steps listed in Section 3 do not necessarily converge to the optimum decision rule. The procedure does provide an effective means of reducing the cost, as illustrated in Section 5.

5. AN ILLUSTRATION

The model described above was developed in connection with a file maintenance application, the master files being the lists of subscribers of two large magazine publishers ([15], [16]). In connection with the development of a system employing a large-scale electronic computer for the maintenance of the files of subscribers, it was necessary to develop explicit rules for matching the transaction file with the master file of subscribers. Initially, matching rules were developed on an intuitive basis, but the subsequent development of the mathematical model indicated ways in which the matching rules could be substantially improved. The illustration presented here is confined to transactions which are subscription orders. (Other types of transactions included changes of address, complaints of non-delivery, subscription cancellations, and so forth. Separate linkage rules should be established for each type.)

TABLE 1. TENTATIVE UNIT COSTS

Action	True Status	
	Match	Non-match
1	$0.00	$6.01
2	.41	1.13
3	.77	.77
4	.82	.41
5	2.59	.00

Table 1 shows tentative unit costs developed by the staff of one of the publishers on the basis of consideration of the character of the actions and the consequences of these actions. The actions listed are roughly the same as those given above as examples in the description of the model. Computation from these unit costs would indicate that the optimum action intervals are as follows:

Action	Probability of a Match
1	$P > .92$
2	$.64 < P < .92$
3	—
4	$.19 < P < .64$
5	$P < .19$

Figure 2 shows the cost function for each of the possible actions. Note that action 3 is never used, since its cost function lies everywhere above some other cost function.

A systematic sample of approximately 10,000 subscription orders during a period of four months was selected. The portion of the master file used for this study consisted of those records for which the post office and the first four letters of the surname were the same as some record in the sample of transactions. Thus, comparison pairs to be examined were confined to those in which the post office and the first four letters in the surname were the same in the two members of the pair. (This is consonant with the comment made above in Section 4 that, in practice, comparisons are usually confined to specified subsets of the master file and the transaction file. This procedure adds, to the cost of any of the alternative linkage rules considered, the contribution from linking errors made for pairs (α, β) that are not actually examined.) To reduce the size of the master file for the purpose of this study, a subsample of one in ten of the master file records not matching a transaction record was selected from those sets that contained 100 or more records, a set here being defined as

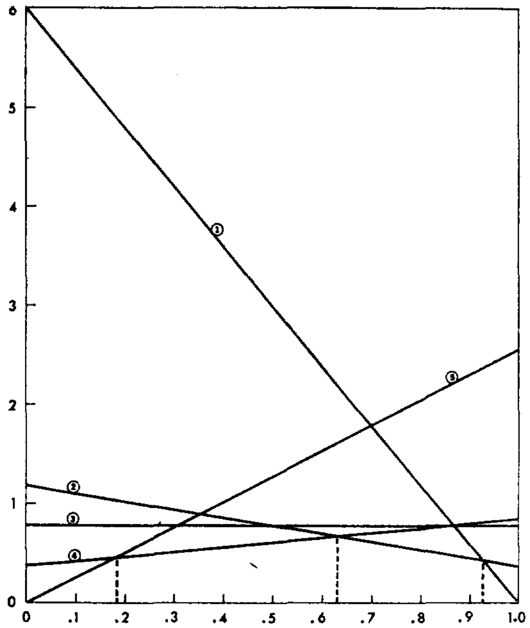


FIG. 2. Cost function for each of five actions, and the optimum action intervals.

a group of master file records having the same post office and first four letters of surname. The number of master file records in the final sample was about 83,000 and the number of comparison pairs about 192,000.

The comparison pairs in the sample were then classified into comparison classes that corresponded to the initial intuitive rule already being employed in the system. The probability of a match in each comparison class was estimated as the proportion of the comparison pairs in that class that were judged to correspond to each other. The determination as to whether a given comparison pair was or was not a match cannot be regarded as definitive since that determination was based upon judgment. However, there were at least two independent judgments for each case, and discrepancies between the judgments were resolved by further review and judgments. It was planned, but never carried out, that results should be refined by selecting a subsample of comparison pairs from the classes defined and then making more intensive investigations of each of the subsample pairs in an effort to determine definitively whether or not the pair was a match. However, it is suggestive to consider some of the consequences if the match status assigned is assumed to be correct. For example, it is interesting to consider the difference in the cost of the initial intuitive rule and the optimum rule based upon the assumed cost system.

Table 2 lists the 52 classes of comparison pairs with the size of each class and the estimated probability of a match in each class. For the initial intuitive rule and for the optimum rule, the table shows the action to be taken for each class, the expected cost for this sample, and the percentage of the total cost. Thus, it is estimated that the expected cost using the initial rule would have been $1,800 for this sample while the cost using the optimum rule was reduced

TABLE 2. COSTS FOR THE SAMPLE, FOR TWO MATCHING RULES, ASSUMING THE TENTATIVE UNIT COSTS

Comparison class	Total pairs	Estimated percent match	Estimated Expected Costs					
			Initial Rule			Optimum Rule		
			Act	$	% of total	Act	$	% of total
1	1,496	99.5	1	42.07	2.3	1	42.07	4.4
2	17	47.1	1	54.09	3.0	4	13.53	1.4
3	544	87.5	1	408.68	22.7	2	272.00	28.7
4	31	96.8	1	6.01	.3	1	6.01	.6
5	38	97.4	1	6.01	.3	1	6.01	.6
6	59	100.0	1	0.00	.0	1	0.00	.0
7	4	100.0	1	0.00	.0	1	0.00	.0
8	63	98.4	1	6.01	.3	1	6.01	.6
9	16	50.0	1	48.08	2.7	4	9.84	1.0
10	14	100.0	1	0.00	.0	1	0.00	.0
11	13	92.3	1	6.01	.3	1	6.01	.6
12	84	94.0	1	30.05	1.7	1	30.05	3.2
13	17	94.1	1	6.01	.3	1	6.01	.6
14	13	53.8	1	36.06	2.0	4	8.20	.9
15	10	70.0	1	18.03	1.0	2	6.26	.7
16	93	86.0	1	84.14	4.7	2	48.21	5.1
17	56	46.4	1	180.30	10.0	4	33.62	3.6
18	56	98.2	2	23.68	1.3	1	6.01	.6
19	26	0	2	29.38	1.6	5	0.00	.0
20	161	8.1	2	172.57	9.6	5	33.67	3.6
21	53	100.0	2	21.73	1.2	1	0.00	.0
22	17	0	2	19.21	1.1	5	0.00	.0
23	77	19.5	2	76.21	4.2	4	37.72	4.0
24	66	54.5	2	48.66	2.7	4	31.47	3.3
25	11	90.9	4	8.61	.5	2	5.23	.6
26	44	0	4	18.04	1.0	5	0.00	.0
27	97	3.1	4	41.00	2.3	5	7.77	.8
28	17	94.1	4	13.53	.8	1	6.01	.6
29	6	0	4	2.46	.1	5	0.00	.0
30	52	7.7	4	22.96	1.3	5	10.36	1.1
31	30	6.7	4	13.12	.7	5	4.10	.4
32	101	9.9	4	45.51	2.5	5	23.90	2.5
33	36	8.3	4	15.99	.9	5	7.77	.8
34	24	29.2	4	18.31	1.0	4	12.71	1.3
35	163	0	5	0.00	.0	5	0.00	.0
36	454	0.2	5	2.59	.1	5	2.59	.3
37	62	0	5	0.00	.0	5	0.00	.0
38	2,822	1.1	5	77.70	4.3	5	77.70	8.2
39	43,678	0	5	0.00	.0	5	0.00	.0
40	129,936	0.005	5	15.54	.9	5	15.54	1.6
41	265	2.3	5	15.54	.9	5	15.54	1.6
42	30	16.7	5	12.95	.7	5	12.95	1.4
43	646	0	5	0.00	.0	5	0.00	.0
44	1,709	0	5	0.00	.0	5	0.00	.0
45	74	0	5	0.00	.0	5	0.00	.0
46	62	0	5	0.00	.0	5	0.00	.0
47	25	8.0	5	5.18	.3	5	5.18	.5
48	8	37.5	5	7.77	.4	4	4.51	.5
49	491	1.2	5	15.54	.9	5	15.54	1.6
50	1	100.0	5	2.59	.1	1	0.00	.0
51	168	20.2	5	83.06	4.9	4	82.82	8.7
52	8,089	0.2	5	33.67	1.9	5	33.67	3.6
Totals	192,125			$1,799.65	99.8%		$946.59	99.6%

to about $950, or about one-half. The estimated standard error of the estimated percentage reduction in cost is approximately 2 percentage points. It is also suggestive to note that 4 of these comparison classes account for more than half of the expected cost of the optimum rule but involve fewer than 2 per cent of all comparison pairs. There is a distinct possibility that an intensive investigation of these 4 comparison classes could markedly reduce the cost of the optimum rule by subdividing these comparison classes.

6. A SIMPLE EXAMPLE OF A COMPARISON FUNCTION

To clarify the notion of a comparison function, the following simple example is given. The example is given for illustration only and bears no direct relationship to the numerical illustration given above, in which the comparison classes are defined in a more complex way.

Let each label α or β consist of the following components, a "blank" being an admissible entry for a component:

1. Surname
2. Given name
3. House number
4. Street name
5. Post office zip code

Then $\gamma(\alpha, \beta)$ may be defined as a vector $(\gamma_1, \gamma_2, \gamma_3, \gamma_4, \gamma_5)$ where

$\gamma_1=0$ if the surname is blank in either α or β.

1 if the surname is the same in α and β, and is a member of a specified list of common surnames.

2 if the surname is the same in α and β, and is not a member of the specified list of common surnames.

3 if the surname is different in α and β, and at least one of them is a member of the specified list of common surnames.

4 if the surname is different in α and β, and neither is a member of the specified list of common surnames.

$\gamma_2=0$ if the given name is blank in either α or β.

1 if the given name is the same in α and β.

2 if the given name is different in α and β.

$\gamma_3=0$ if the house number is blank in either α or β.

1 if the house number is the same in α and β.

2 if the house numbers are different in α and β, but one is a permutation of the other.

3 if the house numbers are different in α and β, and one is not a permutation of the other.

$\gamma_4=0$ if the street name is blank in either α or β.

1 if the street names are the same in α and β.

2 if the street names are different in α and β.

$\gamma_5=1$ if the zip codes are the same in α and β.

2 if the zip codes are different in α and β.

(It is assumed that the zip code is always present or can be supplied.) Thus the function γ may have up to 360 distinct values in this example.

It should be noted that the number of distinct values of the comparison function may be reduced by a process of combination. That is, we may define another comparison function γ' in terms of *sets* of values γ. Let the 360 possible values of γ be classified into sets S_i. Then $\gamma'(\alpha, \beta) = \gamma'_{(i)}$ if and only if $\gamma(\alpha, \beta) \epsilon S_i$.

I thank the referees for their helpful comments.

REFERENCES

[1] Du Bois, N. S. D'Andrea. "On the problem of matching documents with missing and inaccurately recorded items (Preliminary report)." *Annals of Mathematical Statistics*, 35 (1964), p. 1404 (Abstract).

[2] Fasteau, Herman H. and Minton, George. *Automated Geographic Coding System*. 1963 Economic Census: Research Report No. 1, U. S. Bureau of the Census (unpublished). (1965).

[3] Kennedy, J. M. *Linkage of Birth and Marriage Records Using a Digital Computer*. Document No. A.E.C.L.-1258, Atomic Energy of Canada Limited, Chalk River, Ontario. (1961).

[4] Kennedy, J. M. "The use of a digital computer for record linkage." *The Use of Vital and Health Statistics for Genetic and Radiation Studies*, United Nations, New York, (1962), pp. 155–60.

[5] Nathan, Gad. *On Optimal Matching Processes*. Doctoral Dissertation, Case Institute of Technology, Cleveland, Ohio (1964).

[6] Nathan, Gad. "Outcome probabilities for a record matching process with complete invariant information." *Journal of the American Statistical Association*, 62 (1967), pp. 454–69.

[7] Newcombe, H. B. "The study of mutation and selection in human populations." *The Genetics Review*, 57 (1965), pp. 109–25.

[8] Newcombe, H. B. and Kennedy, J. M. "Record linkage: Making maximum use of the discriminating power of identifying information." *Communications of the Association for Computing Machinery*, 5 (1962), pp. 563–66.

[9] Newcombe, H. B., Kennedy, J. M., Axford, S. J., and James, A. P. "Automatic linkage of vital records." *Science*, 130 (1959), pp. 954–9.

[10] Newcombe, H. B. and Rhynas, P. O. W. "Child spacing following stillbirth and infant death." *Eugenics Quarterly*, 9 (1962), pp. 25–35.

[11] Nitzberg, David M. and Sardy, Hyman. "The methodology of computer linkage of health and vital records." *Proceedings, Social Statistics Section, American Statistical Association*. (1965), pp. 100–6.

[12] Perkins, Walter M. and Jones, Charles D. "Matching for Census Coverage Checks." *Proceedings, Social Statistics Section, American Statistical Association*. (1965), pp. 122–39.

[13] Phillips, William and Bahn, Anita K. "Experience with matching of names." *Proceedings, Social Statistics Section, American Statistical Association*. (1963), pp. 26–9.

[14] Sunter, A. B. and Fellegi, I. P. *An Optimal Theory of Record Linkage*. Unpublished paper presented at the 36 Session of the International Statistical Institute, Sydney, Australia (1967).

[15] Tepping, Benjamin J. *Progress Report on the 1959 Matching Study*. National Analysts, Inc., Philadelphia, Pa. (1960).

[16] Tepping, Benjamin J. and Chu, John T. *A Report on Matching Rules*. National Analysts, Inc., Philadelphia, Pa. (1958).

[17] U.S. Bureau of the Census. *Evaluation and Research Program of the U. S. Censuses of Population and Housing, 1960: Record Check Studies of Population Coverage*. Series ER 60, No. 2. U. S. Government Printing Office, Washington, D. C. (1964).

[18] U.S. Bureau of the Census. *Evaluation and Research Program of the U.S. Censuses of Population and Housing, 1960: Accuracy of Data on Population Characteristics as Measured by CPS-Census Match*. Series ER 60, No. 5. U. S. Government Printing Office, Washington, D. C. (1964)

A THEORY FOR RECORD LINKAGE*

IVAN P. FELLEGI AND ALAN B. SUNTER
Dominion Bureau of Statistics

A mathematical model is developed to provide a theoretical framework for a computer-oriented solution to the problem of recognizing those records in two files which represent identical persons, objects or events (said to be *matched*).

A comparison is to be made between the recorded characteristics and values in two records (one from each file) and a decision made as to whether or not the members of the comparison-pair represent the same person or event, or whether there is insufficient evidence to justify either of these decisions at stipulated levels of error. These three decisions are referred to as *link* (A_1), a *non-link* (A_3), and a *possible link* (A_2). The first two decisions are called positive dispositions.

The two types of error are defined as the error of the decision A_1 when the members of the comparison pair are in fact unmatched, and the error of the decision A_3 when the members of the comparison pair are, in fact matched. The probabilities of these errors are defined as

$$\mu = \sum_{\gamma \epsilon \Gamma} u(\gamma) P(A_1 \mid \gamma)$$

and

$$\lambda = \sum_{\gamma \epsilon \Gamma} m(\gamma) P(A_3 \mid \gamma)$$

respectively where $u(\gamma)$, $m(\gamma)$ are the probabilities of realizing γ (a comparison vector whose components are the coded agreements and disagreements on each characteristic) for unmatched and matched record pairs respectively. The summation is over the whole comparison space Γ of possible realizations.

A *linkage rule* assigns probabilities $P(A_1|\gamma)$, and $P(A_2|\gamma)$, and $P(A_3|\gamma)$ to each possible realization of $\gamma \epsilon \Gamma$. An optimal linkage rule L (μ, λ, Γ) is defined for each value of (μ, λ) as the rule that minimizes $P(A_2)$ at those error levels. In other words, for fixed levels of error, the rule minimizes the probability of failing to make positive dispositions.

A theorem describing the construction and properties of the optimal linkage rule and two corollaries to the theorem which make it a practical working tool are given.

1. INTRODUCTION

THE necessity for comparing the records contained in a file L_A with those in a file L_B in an effort to determine which pairs of records relate to the same population unit is one which arises in many contexts, most of which can be categorized as either (a) the construction or maintenance of a master file for a population, or (b) merging two files in order to extend the amount of information available for population units represented in both files.

The expansion of interest in the problem in the last few years is explained by three main factors:

1) the creation, often as a by-product of administrative programmes, of large files which require maintenance over long periods of time and which often contain important statistical information whose value could be increased by linkage of individual records in different files;

*Reprinted with permission from the Journal of the American Statistical Association, American Statistical Association, December 1969, Vol. 64, No. 328, pp. 1183-1210.

2) increased awareness in many countries of the potential of record linkage for medical and genetic research;
3) advances in electronic data processing equipment and techniques which make it appear technically and economically feasible to carry out the huge amount of operational work in comparing records between even medium-sized files.

A number of computer-oriented record linkage operations have already been reported in the literature ([4], [5], [6], [7], [8], [11], [12], [13]) as well as at least two attempts to develop a theory for record linkage ([1], [3]). The present paper is, the authors hope, an improved version of their own earlier papers on the subject ([2], [9], [10]). The theory, developed along the lines of classical hypothesis testing, leads to a linkage rule which is quite similar to the intuitively appealing approach of Newcombe ([4], [5], [6]).

The approach of the present paper is to create a mathematical model within the framework of which a theory is developed to provide guidance for the handling of the linkage problem. Some simplifying assumptions are introduced and some practical problems are examined.

2. THEORY

There are two populations A and B whose elements will be denoted by a and b respectively. We assume that some elements are common to A and B. Consequently the set of ordered pairs

$$A \times B = \{(a, b); a \epsilon A, b \epsilon B)$$

is the union of two disjoint sets

$$M = \{(a, b); a = b, a \epsilon A, b \epsilon B\} \tag{1}$$

and

$$U = \{(a, b); a \neq b, a \epsilon A, b \epsilon B\} \tag{2}$$

which we call the *matched* and *unmatched* sets respectively.

Each unit in the population has a number of characteristics associated with it (e.g. name, age, sex, marital status, address at different points in time, place and date of birth, etc.). We assume now that there are two record generating processes, one for each of the two populations. The result of a record generating process is a record for each member of the population containing some selected characteristics (e.g. age at a certain date, address at a certain date, etc.). The record generating process also introduces some errors and some incompleteness into the resulting records (e.g. errors of reporting or failure to report, errors of coding, transcribing, keypunching, etc.). As a result two unmatched members of A and B may give rise to identical records (either due to errors or due to the fact that an insufficient number of characteristics are included in the record) and, conversely, two matched (identical) members of A and B may give rise to different records. We denote the records corresponding to members of A and B by $\alpha(a)$ and $\beta(b)$ respectively.

We also assume that simple random samples, denoted by A_s and B_s respectively, are selected from each of A and B. We do not, however, exclude the possibility that $A_s = A$ and $B_s = B$. The two given files, L_A and L_B, are considered to be the result of the application of the record generating process to A_s and B_s respectively. For simplicity of notation we will drop the subscript s.

The first step in attempting to link the records of the two files (i.e. identifying the records which correspond to matched members of A and B) is the comparison of records. The result of comparing two records, is a set of codes encoding such statements as "name is the same," "name is the same and it is Brown," "name disagrees," "name missing on one record," "agreement on city part of address, but not on street," etc. Formally we define the *comparison vector* as a vector function of the records $\alpha(a)$, $\beta(b)$:

$$\gamma[\alpha(a), \beta(b)] = \{\gamma^1[\alpha(a), \beta(b)], \cdots, \gamma^K[\alpha(a), \beta(b)]\} \tag{3}$$

It is seen that γ is a function on $A \times B$. We shall write $\gamma(a, b)$ or $\gamma(\alpha, \beta)$ or simply γ as it serves our purpose. The set of all possible realizations of γ is called the *comparison space* and denoted by Γ.

In the course of the linkage operation we observe $\gamma(a, b)$ and want to decide either that (a, b) is a matched pair $(a, b) \in M$ (call this decision, denoted by A_1, a *positive link*) or that (a, b) is an unmatched pair $(a, b) \in U$ (call this decision, denoted by A_3, a *positive non-link*). There will be however some cases in which we shall find ourselves unable to make either of these decisions at specified levels of error (as defined below) so that we allow a third decision, denoted A_2, a *possible link*.

A *linkage rule* L can now be defined as a mapping from Γ, the comparison space, onto a set of random decision functions $D = \{d(\gamma)\}$ where

$$d(\gamma) = \{P(A_1 \mid \gamma), P(A_2 \mid \gamma), P(A_3 \mid \gamma)\}; \quad \gamma \epsilon \Gamma \tag{4}$$

and

$$\sum_{i=1}^{3} P(A_i \mid \gamma) = 1. \tag{5}$$

In other words, corresponding to each observed value of γ, the linkage rule assigns the probabilities for taking each of the three possible actions. For some or even all of the possible values of γ the decision function may be a degenerate random variable, i.e. it may assign one of the actions with probability equal to 1.

We have to consider the levels of error associated with a linkage rule. We assume, for the time being, that a pair of records $[\alpha(a), \beta(b)]$ is selected for comparison according to some probability process from $L_A \times L_B$ (this is equivalent to selecting a pair of elements (a, b) at random from $A \times B$, due to the construction of L_A and L_B). The resulting comparison vector $\gamma[\alpha(a), \beta(b)]$ is a random variable. We denote the conditional probability of γ, given that $(a, b) \in M$ by $m(\gamma)$. Thus

$$\begin{aligned} m(\gamma) &= P\{\gamma[\alpha(a), \beta(b)] \mid (a, b) \in M\} \\ &= \sum_{(a,b) \epsilon M} P\{\gamma[\alpha(a), \beta(b)]\} \cdot P[(a, b) \mid M]. \end{aligned} \tag{6}$$

Similarly we denote the conditional probability of γ, given that $(a, b) \in U$ by $u(\gamma)$. Thus

$$\begin{aligned} u(\gamma) &= P\{\gamma[\alpha(a), \beta(b)] \mid (a, b) \in U\} \\ &= \sum_{(a,b) \in U} P\{\gamma[\alpha(a), \beta(b)]\} \cdot P[(a, b) \mid U]. \end{aligned} \tag{7}$$

There are two types of error associated with a linkage rule. The first occurs when an unmatched comparison is linked and has the probability

$$P(A_1 \mid U) = \sum_{\gamma \in \Gamma} u(\gamma) \cdot P(A_1 \mid \gamma). \tag{8}$$

The second occurs when a matched comparison is non-linked and has the probability

$$P(A_3 \mid M) = \sum_{\gamma \in \Gamma} m(\gamma) \cdot P(A_3 \mid \gamma). \tag{9}$$

A linkage rule on the space Γ will be said to be a linkage rule at the levels μ, λ ($0<\mu<1$ and $0<\lambda<1$) and denoted by $L(\mu, \lambda, \Gamma)$ if

$$P(A_1 \mid U) = \mu \tag{10}$$

and

$$P(A_3 \mid M) = \lambda. \tag{11}$$

Among the class of linkage rules on Γ which satisfy (10) and (11) the linkage rule $L(\mu, \lambda, \Gamma)$ will be said to be the *optimal linkage rule* if the relation

$$P(A_2 \mid L) \leqq P(A_2 \mid L') \tag{12}$$

holds for every $L'(\mu, \lambda, \Gamma)$ in the class.

In explanation of our definition we note that the optimal linkage rule maximizes the probabilities of positive dispositions of comparisons (i.e. decisions A_1 and A_3) subject to the fixed levels of error in (10) and (11) or, put differently, it minimizes the probability of failing to make a positive disposition. This seems a reasonable approach since in applications the decision A_2 will require expensive manual linkage operations; alternatively, if the probability of A_2 is not small, the linkage process is of doubtful utility.

It is not difficult to see that for certain combinations of μ and λ the class of linkage rules satisfying (10) and (11) is empty. We admit only those combinations of μ and λ for which it is possible to satisfy equations (10) and (11) simultaneously with some set D of decision functions as defined by (4) and (5). For a more detailed discussion of admissibility see Appendix 1. At this point it is sufficient to note that a pair of values (μ, λ) will be inadmissible only if one or both of the members are too large, and that in this case we would always be happy to reduce the error levels.

2.1. *A fundamental theorem*

We first define a linkage rule L_0 on Γ. We start by defining a unique ordering of the (finite) set of possible realizations of γ.

If any value of γ is such that both $m(\gamma)$ and $u(\gamma)$ are equal to zero, then the (unconditional) probability of realizing that value of γ is equal to zero. and hence it need not be included in Γ. We now assign an order arbitrarily to all γ for which $m(\gamma)>0$ but $u(\gamma)=0$.

Next we order all remaining γ in such a way that the corresponding sequence of

$$m(\gamma)/u(\gamma)$$

is monotone decreasing. When the value of $m(\gamma)/u(\gamma)$ is the same for more than one γ we order these γ arbitrarily.

We index the ordered set $\{\gamma\}$ by the subscript i; $(i=1, 2, \cdots, N_\Gamma)$; and write $u_i=u(\gamma_i)$; $m_i=m(\gamma_i)$.

Let (μ, λ) be an admissible pair of error levels and choose n and n' such that

$$\sum_{i=1}^{n-1} u_i < \mu \leqq \sum_{i=1}^{n} u_i \tag{13}$$

$$\sum_{i=n'}^{N_\Gamma} m_i \geqq \lambda > \sum_{i=n'+1}^{N_\Gamma} m_i \tag{14}$$

where N_Γ is the number of points in Γ.

We assume for the present that when (13) and (14) are satisfied we have $1<n\leq n'-1<N_\Gamma$. This will ensure that the levels (μ, λ) are admissible. Let $L_0(\mu, \lambda, \Gamma)$ denote the linkage rule defined as follows: having observed a comparison vector, γ_i, take action A_1 (positive link) if $i\leqq n-1$, action A_2 when $n<i\leqq n'-1$, and action A_3 (positive non-link) when $i\geqq n'+1$. When $i=n$ or $i=n'$ then a random decision is required to achieve the error levels μ and λ exactly. Formally,

$$d(\gamma_i) = \begin{cases} (1,0,0) & i \leq n-1 & \text{(a)} \\ (P_\mu, 1-P_\mu, 0) & i = n & \text{(b)} \\ (0,1,0) & n < i \leq n'-1 & \text{(c)} \\ (0, 1-P_\lambda, P_\lambda) & i = n' & \text{(d)} \\ (0,0,1) & i \geq n'+1 & \text{(e)} \end{cases} \tag{15}$$

where P_μ and P_λ are defined as the solutions to the equations

$$u_n \cdot P_\mu = \mu - \sum_{i=1}^{n-1} u_i \tag{16}$$

$$m_{n'} \cdot P_\lambda = \lambda - \sum_{i=n'+1}^{N_\Gamma} m_i. \tag{17}$$

THEOREM[1]: Let $L_0(\mu, \lambda, \Gamma)$ be the linkage rule defined by (15). Then L is a best linkage rule on Γ at the levels (μ, λ). The proof is given in Appendix 1.

The reader will have observed that the whole theory could have been formulated, although somewhat awkwardly, in terms of the classical theory of hypothesis testing. We can test first the null hypothesis that $(a, b)\in U$ against

[1] A slightly extended version of the theorm is given in Appendix 1.

the simple alternative that $(a, b) \in M$, the action A_1 being the rejection of the null hypothesis and μ the level of significance. Similarly the action A_3 is the rejection at the significance level λ of the null hypothesis that $(a, b) \in M$ in favour of the simple alternative that $(a, b) \in U$. The linkage rule L is equivalent to the likelihood ratio test and the theorem above asserts this to be the uniformly most powerful test for either hypothesis.

We state, without proof, two corollaries to the theorem. These corollaries, although mathematically trivial, are important in practice.

Corollary 1: If

$$\mu = \sum_{i=1}^{n} u_i, \quad \lambda = \sum_{i=n}^{N_\Gamma} m_i, \quad n < n',$$

the $L_0(u, \lambda, \Gamma)$, the best linkage rule at the levels (μ, λ) becomes

$$d(\gamma_i) = \begin{cases} (1, 0, 0) & \text{if} \quad 1 \leq i \leq n \\ (0, 1, 0) & \text{if} \quad n < i < n' \\ (0, 0, 1) & \text{if} \quad n' \leq i \leq N_\Gamma. \end{cases} \tag{18}$$

If we define

$$T_\mu = \frac{m(\gamma_n)}{u(\gamma_n)}$$

$$T_\lambda = \frac{m(\gamma_{n'})}{u(\gamma_{n'})}$$

then the linkage rule (18) can be written equivalently[2] as

$$d(\gamma) = \begin{cases} (1, 0, 0) & \text{if} \quad T_\mu \leq m(\gamma)/u(\gamma) \\ (0, 1, 0) & \text{if} \quad T_\lambda < m(\gamma)/u(\gamma) < T_\mu \\ (0, 0, 1) & \text{if} \quad m(\gamma)/u(\gamma) \leq T_\lambda. \end{cases} \tag{19}$$

Corollary 2: Let T_μ and T_λ be any two positive numbers such that

$$T_\mu > T_\lambda.$$

Then there exists an admissible pair of error levels (μ, λ) corresponding to T_μ and T_λ such that the linkage rule (19) is best at these levels. The levels (μ, λ) are given by

$$\mu = \sum_{\gamma \in \Gamma_\mu} u(\gamma) \tag{20}$$

$$\lambda = \sum_{\gamma \in \Gamma_\lambda} m(\gamma) \tag{21}$$

where

$$\Gamma_\mu = \{\gamma : T_\mu \leq m(\gamma)/u(\gamma)\} \tag{22}$$

$$\Gamma_\lambda \{\gamma : m(\gamma)/u(\gamma) \leq T_\lambda\} \tag{23}$$

[2] We are grateful to the referee for pointing out that (19) and (18) are exactly equivalent only if $m_n/u_n < m_{n+1}/u_{n+1}$ and $m_{n'-1}/u_{n'-1} < m_n/u_n$.

In many applications we may be willing to tolerate error levels sufficiently high to preclude the action A_2. In this case we choose n and n' or, alternatively, T_μ and T_λ so that the middle set of γ in (18) or (19) is empty. In other words every (a, b) is allocated either to M or to U. The theory for the allocation of observations to one of two mutually exclusive populations may thus be regarded as a special case of the theory given in this paper.

3. APPLICATIONS

3.1. *Some Practical Problems*

In attempting to implement the theory developed in the previous section several practical problems need to be solved. They are outlined briefly below and taken up in more detail in subsequent sections.

a) The large number of possible values of $m(\gamma)$ and $u(\gamma)$. Clearly the number of distinct realizations of γ may be so large as to make the computation and storage of the corresponding values of $m(\gamma)$ and $u(\gamma)$ impractical. The amount of computation and storage can be substantially reduced on the basis of some simplifying assumptions.
b) Methods to calculate the quantities $m(\gamma)$ and $u(\gamma)$. Two methods are proposed.
c) Blocking the files. Implicit in the development of the theory is the assumption that if two files are linked then all possible comparisons of all the records of both files will be attempted. It is clear that even for medium sized files the number of comparisons under this assumption would be very large, (e.g. 10^5 records in each file would imply 10^{10} comparisons). In practice the files have to be "blocked" in some fashion and comparisons made only within corresponding blocks. The impact of such blocking on the error levels will be examined.
d) Calculations of threshold values. It should be clear from Corollary 2 that we do not have to order explicitly the values of γ in order to apply the main theorem since for any particular γ the appropriate decision (A_1, A_2 or A_3) can be made by comparing $m(\gamma)/u(\gamma)$ with the threshold values T_μ and T_λ. We shall outline a method of establishing these threshold values corresponding to the required error levels μ and λ.
e) Choice of the comparison space. The main theorem provides an optimal linkage rule for a given comparison space. Some guidance will be provided on the choice of the comparison space.

3.2. *Some simplifying assumptions*

In practice the set of distinct (vector) values of γ may be so large that the estimation of the corresponding probabilities $m(\gamma)$ and $u(\gamma)$ becomes comletely impracticable. In order to make use of the theorem it will be necessary to make some simplifying assumptions about the distribution of γ.

We assume that the components of γ can be re-ordered and grouped in such a way that

$$\gamma = (\gamma^1, \gamma^2, \cdots, \gamma^K)$$

and that the (vector) components are mutually statistically independent with respect to each of the conditional distributions. Thus

$$m(\gamma) = m_1(\gamma^1) \cdot m_2(\gamma^2) \cdot \cdot \cdot m_k(\gamma^K) \tag{24}$$

$$u(\gamma) = u_1(\gamma^1) \cdot u_2(\gamma^2) \cdot \cdot \cdot u_k(\gamma^K) \tag{25}$$

where $m(\gamma)$ and $u(\gamma)$ are defined by (4) and (5) respectively and

$$m_i(\gamma^i) = P(\gamma^i \mid (a, b) \in M)$$

$$u_i(\gamma^i) = P(\gamma^i \mid (a, b) \in U).$$

For simplicity of notation we shall write $m(\gamma^i)$ and $u(\gamma^i)$ instead of the technically more precise $m_i(\gamma^i)$ and $u_i(\gamma^i)$. As an example, in a comparison of records relating to persons γ^1 might include all comparison components that relate to surnames, γ^2 all comparison components that relate to addresses. The components γ^1 and γ^2 are themselves vectors; the subcomponents of γ^2 for example might represent the coded results of comparing the different components of the address (city name, street name, house number, etc.). If two records are matched (i.e. when in fact they represent the same person or event), then a disagreement configuration could occur due to errors. Our assumption says that errors in names, for example, are independent of errors in addresses. If two records are unmatched (i.e. when in fact they represent different persons or events) then our assumption says that an accidental agreement on name, for example, is independent of an accidental agreement on address. In other words what we do assume is that $\gamma^1, \gamma^2, \cdot \cdot \cdot, \gamma^K$ are conditionally independently distributed. We emphasize that we do *not* assume anything about the unconditional distribution of γ.

It is clear that any monotone increasing function of $m(\gamma)/u(\gamma)$ could serve equally well as a test statistic for the purpose of our linkage rule. In particular it will be advantageous to use the logarithm of this ratio and define

$$w^k(\gamma^k) = \log m(\gamma^k) - \log u(\gamma^k). \tag{26}$$

We can then write

$$w(\gamma) = w^1 + w^2 + \cdot \cdot \cdot + w^K \tag{27}$$

and use $w(\gamma)$ as our test statistic with the understanding that if $u(\gamma) = 0$ or $m(\gamma) = 0$ then $w(\gamma) = +\infty$ (or $w(\gamma) = -\infty$) in the sense that $w(\gamma)$ is greater (or smaller) than any given finite number.

Suppose that γ^k can take on n_k different *configurations*, $\gamma_1^k, \gamma_2^k, \cdot \cdot \cdot, \gamma_{n_k}^k$. We define

$$w_j^k = \log m(\gamma_j^k) - \log u(\gamma_j^k). \tag{28}$$

It is a convenience for the intuitive interpretation of the linkage process that the weights so defined are positive for those configurations for which $m(\gamma_j^k) > u(\gamma_j^k)$, negative for those configurations for which $m(\gamma_j^k) < u(\gamma_j^k)$, and that this property is preserved by the weights associated with the total configuration γ.

The number of total configurations (i.e. the number of points $\gamma \in \Gamma$) is obviously $n_1 \cdot n_2 \cdot \cdot \cdot \cdot n_K$. However, because of the additive property of the weights defined for components it will be sufficient to determine $n_1+n_2+\cdots+n_K$ weights. We can then always determine the weight associated with any γ by employing this additivity.

3.3. *The Calculation of Weights*

An assumption made at the outset of this paper was that the files L_A and L_B represent samples A_s and B_s of the populations A and B. This assumption is often necessary in some applications when one wishes to use a set of values of $m(\gamma^k)$ and $u(\gamma^k)$, computed for some large populations A and B while the actually observed files L_A and L_B correspond to some subpopulations A_s and B_s. For example, in comparing a set of incoming records against a master file in order to update the file one may want to consider the master file and the incoming set of records as corresponding to samples A_s and B_s of some conceptual populations A and B. One might compute the weights for the full comparison space Γ corresponding to A and B and apply these weights repeatedly on different update runs; otherwise one would have to recompute the weights on each occasion.

Of course it seldom occurs in practice that the subpopulations represented by the files L_A and L_B are actually drawn at random from any real populations A and B. However it is clear that all the theory presented in this paper will still hold if the assumption is relaxed to the assumption that the condition of entry of the subpopulation into the files is uncorrelated with the distribution in the populations of the characteristics used for comparisons. This second assumption obviously holds if the first does, although the converse is not necessarily true.

In this paper we propose two methods for calculating weights. In the first of these we assume that prior information is available on the distribution in the populations A and B of the characteristics used in comparison as well as on the probabilities of different types of error introduced into the files by the record generating processes. The second method utilizes the information in the files L_A and L_B themselves to estimate the probabilities $m(\gamma^k)$ and $u(\gamma^k)$. The validity of these estimates is strongly predicated on the independence assumption of the previous section. Specifically it requires that the formal expression for that independence should hold almost exactly in the subpopulation $L_A \times L_B$, which, in turn, requires that the files L_A and L_B should be large and should satisfy at least the weaker of the assumptions of the previous paragraph.

Another procedure, proposed by Tepping ([11], [13]), is to draw a sample from $L_A \times L_B$, identify somehow (with negligible error) the matched and unmatched comparisons in this sample, and thus estimate $m(\gamma)$ and $u(\gamma)$ directly. The procedure seems to have some difficulties associated with it. If and when the identification of matched and unmatched records can in fact be carried out with reasonable accuracy and with reasonable economy (even if only at least occasionally) then it might provide a useful check or corroboration of the reasonableness of assumptions underlying the calculation of weights.

Finally, the weights $w(\gamma)$ or alternatively the probabilities $m(\gamma)$ and $u(\gamma)$, derived on one occasion for the linkage $L_A \times L_B$ can continue to be used on a subsequent occasion for the linkage, say $L_A' \times L_B'$, provided A_s and B_s can be regarded as samples from the same populations as A_s and B_s and provided the record generating processes are unaltered.

3.3.1. *Method I*

Suppose that one component of the records associated with each of the two populations A and B is the surname. The comparison of surnames on two records will result in a component of the comparison vector. This component may be a simple comparison component such as "name agrees" or "name disagrees" or "name missing on one or both records" (in this case γ^k is a scalar); or it may be a more complicated vector component such as for example "records agree on Soundex code, the Soundex code is B650; the first 5 characters of the name agree; the second 5 characters of the name agree; the surname is BROWNING."

In either of the two files the surname may be reported in error. Assume that we could list all error-free realizations of all surnames in the two populations and also the number of individuals in the respective populations corresponding to each of these surnames. Let the respective frequencies in A and B be

$$f_{A_1}, f_{A_2}, \cdots, f_{A_m}; \qquad \sum_{j=1}^{m} f_{A_j} = N_A$$

and

$$f_{B_1}, f_{B_2}, \cdots, f_{B_m}; \qquad \sum_{j=1}^{m} f_{B_j} = N_B.$$

Let the corresponding frequencies in $A \cap B$ be

$$f_1, f_2, \cdots, f_m; \qquad \sum_{j} f_j = N_{AB}.$$

The following additional notation is needed:

e_A or e_B the respective probabilities of a name being misreported in L_A or L_B (we assume that the probability of misreporting is independent of the particular name);

e_{A0} or e_{B0} the respective probabilities of a name not being reported in L_A or L_B (we assume that the probability of name not being reported is independent of the particular name);

e_T the probability the name of a person is differently (though correctly) reported in the two files (this might arise, for example, if L_A and L_B were generated at different times and the person changed his name).

Finally we assume that e_A and e_B are sufficiently small that the probability of an agreement on two identical, though erroneous, entries is negligible and that the probabilities of misreporting, not reporting and change are independent of one another.

We shall first give a few rules for the calculation of m and u corresponding to the following configurations of γ: name agrees and it is the jth listed name, name disagrees; name missing on either record.

m (name agrees and is the jth listed name)

$$\begin{aligned} &= \frac{f_j}{N_{AB}}(1 - e_A)(1 - e_B)(1 - e_T)(1 - e_{A0})(1 - e_{B0}) \\ &\doteq \frac{f_j}{N_{AB}}(1 - e_A - e_B - e_T - e_{A0} - e_{B0}) \end{aligned} \tag{29}$$

m (name disagrees)

$$\begin{aligned} &= [1 - (1 - e_A)(1 - e_B)(1 - e_T)](1 - e_{A0})(1 - e_{B0}) \\ &\doteq e_A + e_B + e_T \end{aligned} \tag{30}$$

m (name missing on either file)

$$= 1 - (1 - e_{A0})(1 - e_{B0}) \doteq e_{A0} + e_{B0} \tag{31}$$

u (name agrees and is the jth listed name)

$$\begin{aligned} &= \frac{f_{Aj}}{N_A}\frac{f_{Bj}}{N_B}(1 - e_A)(1 - e_T)(1 - e_{A0})(1 - e_{B0}) \\ &\doteq \frac{f_{Aj}}{N_A}\frac{f_{Bj}}{N_B}(1 - e_A - e_B - e_T - e_{A0} - e_{B0}) \end{aligned} \tag{32}$$

u (name disagrees)

$$\begin{aligned} &= \left[1 - (1 - e_A)(1 - e_B)(1 - e_T)\sum_j \frac{f_{Aj}}{N_A}\frac{f_{Bj}}{N_B}\right](1 - e_{A0})(1 - e_{B0}) \\ &\doteq \left[1 - (1 - e_A - e_B - e_T)\sum_j \frac{f_{Aj}}{N_A}\frac{f_{Bj}}{N_B}\right](1 - e_{A0} - e_{B0}) \end{aligned} \tag{33}$$

u (name missing on either file)

$$= 1 - (1 - e_{A0})(1 - e_{B0}) = e_{A0} + e_{B0}. \tag{34}$$

The proportions f_{Aj}/N_A, f_{Bj}/N_B, f_j/N may be taken, in many applications, to be the same. This would be the case, for example, if two large files can be assumed to be drawn from the same population. These frequencies may be estimated from the files themselves.

A second remark relates to the interpretation of weights. It will be recalled that according to (28) the contribution to the overall weight of the name component is equal to log (m/u) and that comparisons with a weight higher than a specified number will be considered linked, while those whose weight is below a specified number will be considered unlinked. It is clear from (29–34) that an agreement on name will produce a positive weight and in fact the rarer the name, the larger the weight; a disagreement on name will produce a negative weight which decreases with the errors e_A, e_B, e_T; if the name is missing on either record, the weight will be zero. These results seem intuitively appealing.

We should emphasize that it is not necessary to list all possible names for the validity of formulae (29) to (34). We might only list the more common names separately, grouping all the remaining names. In the case of groupings the appropriate formulae in (29) to (34) have to be summed over the corresponding values of the subscript j. The problem of how to group configurations is taken up in a later section.

Finally we should mention that formulae (29) to (34) relate to reasonably simple realizations of γ, such as a list of names, or list of ages, or lists of other possible identifiers. In more complex cases one may be able to make use of these results, with appropriate modifications, in conjunction with the elementary rules of probability calculus. Alternatively one may have recourse to the method given below.

3.3.2. *Method II*

The formulae presented in Appendix 2 can be used, under certain circumstances, to estimate the quantities $m(\gamma^k)$, $u(\gamma^k)$ and N, the number of matched records, simply by substituting into these formulae certain frequencies which can be directly (and automatically) counted by comparing the two files. Mathematically, the only condition for the validity of these formulae is that γ should have at least three components which are independent with respect to the probability measures m and u in the sense of (24) and (25). It should be kept in mind, however, that for agreement configurations $m(\gamma^k)$ is typically very close to one, $u(\gamma^k)$ is very close to zero, and conversely for diagreement configurations. Therefore the estimates of $u(\gamma^k)$ and $m(\gamma^k)$ can be subject to substantial sampling variability unless the two files represent censuses or large random samples of the populations A and B.

The detailed formulae and their proofs are included in the Appendix. At this point only an indication of the methods will be given. For simplicity we present the method in terms of three components. If, in fact, there are more than three components they can be grouped until there are only three left. Clearly this can be done without violating (24) and (25).

For each component vector of γ designate the set of configurations to be considered as "agreements" and denote this set (of vectors) for the hth component by S_h. The designation of specific configurations as "agreements" may be arbitrary but subject to some numerical considerations to be outlined in the Appendix.

The following notation refers to the frequencies of various configurations of γ. Since they are not conditional frequencies, they can be obtained as direct counts by comparing the files L_A and L_B:

M_h: the proportion of "agreement" in all components except the hth; any configuration in the kth component;

U_h: the proportion of "agreement" in the hth component; any configuration in the others;

M: the proportion of "agreement" in all components.

Denote also the respective conditional probabilities of "agreements" by

$$m_h = \sum_{\gamma \epsilon S_h} m(\gamma) \qquad (35)$$

$$u_h = \sum_{\gamma \epsilon S_h} u(\gamma). \qquad (36)$$

It follows from the assumptions (24) and (25) that the expected values of M_h, U_h, and M with respect to the sampling procedure (if any) and the record generating process through which the files L_A and L_B arose from the populations A and B can be expressed simply in terms of m_h and u_h as follows.

$$N_A N_B E(M_h) = E(N) \prod_{\substack{j=1 \\ j \neq h}}^{3} m_j + [N_A N_B - E(N)] \prod_{\substack{j=1 \\ j \neq h}}^{3} u_{\delta}; \qquad h = 1, 2, 3 \quad (37)$$

$$N_A N_B E(U_h) = E(N) m_h + [N_A N_B - E(N)] u_h \qquad (38)$$

$$N_A N_B E(M) = E(N) \prod_{j=1}^{3} m_j + [N_A N_B - E(N)] \prod_{j=1}^{3} u_j \qquad (39)$$

where N_A and N_B are the known number of records in the files L_A and L_B and N is the unknown number of matched records.

Dropping the expected values we obtain seven equations for the estimation of the seven unknown quantities N, m_h, $u_h (h=1, 2, 3)$. The solution of these equations is given in Appendix 2.

Having solved for m_h, u_h and N the quantities $m(\gamma^k)$ and $u(\gamma^k)$ are easily computed by substituting some additional directly observable frequencies into some other equations, also presented in Appendix 2. The frequency counts required for all the calculations can be obtained at the price of three sorts of the two files.

It is our duty to warn the reader again that although these equations provide statistically consistent estimates, the sampling variability of the estimates may be considerable if the number of records involved ($N_A N_B$) is not sufficiently large. One might get an impression of the sampling variabilities through the method of random replication, i.e., by splitting both of the files at random into at least two parts and by performing the estimation separately for each. Alternatively, one can at least get an impression of the sampling variabilities of M_h, U_h and M by assuming that they are estimated from a random sample of size $N_A N_B$.

Another word of caution may be in order. The estimates are computed on the basis of the independence assumptions of (24) and (25). In the case of departures from independence the estimates, *as estimates of the probabilities* $m(\gamma^k)$ *and* $u(\gamma^k)$, may be seriously affected and the resulting weights $m(\gamma^k)/u(\gamma^k)$ would lose their probabilistic interpretations. What is important, of course, is their effect on the resulting linkage operation. We believe that if sufficient identifying information is available in the two files to carry out the linkage operation in the first place, then the operation is quite robust against departures from independence. One can get an impression of the extent of the departures from independence by carrying out the calculations of Appendix 2 on the basis of alternative designations of the "agreement" configurations.

3.4. *Restriction of Explicit Comparisons to a Subspace*

In practice of course we do not select comparisons at random from $L_A \times L_B$. But then in practice we are not concerned with the *probability* of the event $(A_1 | U)$ or the event $(A_3 | M)$ for any particular comparison but rather with the *proportion* of occurrences of these two events in the long run. Clearly if our linkage procedure is to examine *every* comparison $(\alpha, \beta) \in L_A \times L_B$ then we could formally treat any particular comparison as if it had been drawn at random from $L_A \times L_B$. The only change in our theory in this case would be the replacement of *probabilities* with *proportions*. In particular the probabilities of error μ and λ would then have to be interpreted as proportions of errors. With this understanding we can continue to use the notation and concepts of probability calculus in this paper even though often we shall think of probabilities as proportions.

We have now made explicit a second point which needs to be examined. We would seldom be prepared to examine every $(\alpha, \beta) \in L_A \times L_B$ since it is clear that even for medium sized files (say 10^5 record each) the number of comparisons (10^{10}) would outstrip the economic capacity of even the largest and fastest computers.

Thus the number of comparisons we will examine explicitly will be restricted to a subspace, say Γ^*, of Γ. This might be achieved for example by partitioning or "blocking" the two files into Soundex-coded Surname "blocks" and making explicit comparisons only between records in corresponding blocks. The subspace Γ^* is then the set of γ for which the Soundex Surname component has the agreement status. All other γ are implicit positive non-links (the comparisons in $\Gamma - \Gamma^*$ will not even be actually compared hence they may not be either positive or possible links). We consider the effect that this procedure has on the error levels established for the all-comparison procedure.

Let Γ_μ and Γ_λ be established (as in Corollary 2) for the all-comparison procedure so as to satisfy

$$\Gamma_\mu = \{\gamma : T_\mu \leq m(\gamma)/u(\gamma)\}$$
$$\Gamma_\lambda = \{\gamma : m(\gamma)/u(\gamma) \leq T_\lambda\}$$

where

$$\mu = \sum_{\gamma \epsilon \Gamma_\mu} u(\gamma)$$
$$\lambda = \sum_{\gamma \epsilon \Gamma_\lambda} m(\gamma).$$

If we now regard all $\gamma \in (\Gamma - \Gamma^*)$ as implicit positive non-links we must adjust our error levels to

$$\mu^* = \mu - \sum_{\Gamma_\mu \cap \bar{\Gamma}^*} u(\gamma) \tag{40}$$

$$\lambda^* = \lambda + \sum_{\bar{\Gamma}_\lambda \cap \bar{\Gamma}^*} m(\gamma) \tag{41}$$

where $\bar{\Gamma}_\lambda$ and $\bar{\Gamma}^*$ denote complements taken with respect to Γ (i.e. $\Gamma - \Gamma_\lambda$ and $\Gamma - \Gamma^*$, respectively).

The first of these expressions indicates that the level of μ is reduced by the sum of the u-probabilities of those comparisons which would have been links under the all-comparison procedure but are implicit non-links under the blocking procedure. The second expression indicates that the actual level of λ is increased by the sum of the m-probabilities of the comparisons that would be links or possible links under the all-comparison procedure but are implicit non-links under the blocking procedure.

The probabilities of a failure to make a positive disposition under the blocking procedure are given by

$$P^*(A_2 \mid M) = \sum_{\gamma \epsilon \bar{\Gamma}_\mu \cap \bar{\Gamma}_\lambda} m(\gamma) - \sum_{\gamma \epsilon \bar{\Gamma}_\mu \cap \bar{\Gamma}_\lambda \cap \bar{\Gamma}^*} m(\gamma) \tag{42}$$

$$P^*(A_2 \mid U) = \sum_{\gamma \epsilon \bar{\Gamma}_\mu \cap \bar{\Gamma}_\lambda} u(\gamma) - \sum_{\gamma \epsilon \bar{\Gamma}_\mu \cap \bar{\Gamma}_\lambda \cap \bar{\Gamma}^*} u(\gamma) \tag{43}$$

the second term on the right in each case being the reduction due to the blocking procedure.

These expressions will be found to be useful when we consider the best way of blocking a file.

3.5. *Choice of Error Levels and Choice of Subspace*

In choosing the error levels (μ, λ) we may want to be guided by the consideration of losses incurred by the different actions.

Let $G_M(A_i)$ and $G_U(A_i)$ be non-negative loss functions which give the loss associated with the disposition A_i; $(i=1, 2, 3)$; for each type of comparison. Normally, we would set

$$G_M(A_1) = G_U(A_3) = 0$$

and we do so here. Reverting to the all-comparison procedure we set (μ, λ) so as to minimize the expected loss given by the expression

$$\begin{aligned} &P(M) \cdot E[G_M(A_i)] + P(U) \cdot E[G_U(A_i)] \\ &= P(M)[P(A_2 \mid M) \cdot G_M(A_2) + \lambda \cdot G_M(A_3)] \\ &\quad + P(U)[\mu \cdot G_U(A_1) + P(A_2 \mid U) \cdot G_U(A_2)] \end{aligned} \tag{44}$$

Note that $P(A_2 | M)$ and $P(A_2 | U)$ are functions of μ and λ. We give later a practical procedure for determining the values of (μ, λ) which minimize (44).

Suppose that (μ, λ) have been set so as to minimize (44). We now consider the effects of blocking the files and introduce an additional component in the loss function which expresses the costs of comparisons, $G_{\Gamma^*}(L_A \times L_B)$, under a blocking procedure equivalent to making implicit comparisons in a subspace Γ^*. We seek that subspace Γ^* which minimizes the total expected loss,

$$\begin{aligned} &c\{P(M) \cdot E[G_M(A_i)] + P(U) \cdot E[G_U(A_i)]\} \\ &\quad + G_{\Gamma^*}(L_A \times L_B) \\ &= c\{P(M)[P^*(A_2 \mid M) G_M(A_2) + \lambda^* G_M(A_3)] \\ &\quad + P(U)[\mu^* G_U(A_1) + P^*(A_2 \mid U) G_U(A_2)]\} \\ &\quad + G_{\Gamma^*}(L_A \times L_B) \end{aligned} \tag{45}$$

where P^* denotes probabilities under the blocking procedure given by (42) and (43) respectively and c denotes the number of comparisons in $L_A \times L_B$. Now if the processing cost of comparisons under any blocking Γ^* is simply proportional to the number of comparisons, c^*, i.e.

$$G_{\Gamma^*}(L_A \times L_B) = \alpha c^*$$

then we can minimize

$$\begin{aligned} &P(M)[P^*(A_2 \mid M)G_M(A_2)\lambda^* G_M(A_3)] \\ &+ P(U)[\mu^* G_U(A_1) + P^*(A_2 \mid U)G_U(A_2)] + \frac{\alpha c^*}{c} \cdot \qquad (46) \end{aligned}$$

The last term is the product of the cost, α, per comparison and the reduction ratio in the number of comparisons to be made explicitly.

No explicit solution of (46) seems possible under such general conditions. However, (46) can be used to compare two different choices of Γ^*. Once a choice of Γ^* has been made, the "theoretical" error levels μ, λ can be chosen, using (40) and (41), so that the actual error levels μ^*, λ^* meet the error specification. The threshold values T_μ, T_λ are then calculated from the "theoretical" error levels.

3.6. *Choice of comparison space*

Let Γ and Γ' be two comparison spaces, with conditional distributions $m(w)$, $u(w)$ and $m'(w)$, $u'(w)$ and threshold values T_μ, T_λ and T'_μ, T'_λ respectively (the threshold values being in both cases so determined that they lead to the same error levels μ, λ).

Now in a manner precisely analogous to our linkage criterion we might say that a comparison space Γ is better than a comparison space Γ' at the error levels (μ, λ) if

$$P(T_\lambda < w(\gamma) < T_\mu) < P(T'_\lambda < w'(\gamma') < T'_\mu) \qquad (47)$$

where it is assumed that the comparisons are made under the optimal linkage rule in each case. The linkage criterion developed for a given Γ is independent of (μ, λ) and $P(M)$. Clearly we cannot hope for this to be the case in general with a criterion for the choice of a comparison space.

Expanding the expression (47) we have as our criterion at the level (μ, λ)

$$\begin{aligned} &P(M) \cdot \sum_{T_\lambda < w < T_\mu} m(w) + P(U) \cdot \sum_{T_\lambda < w < T_\mu} u(w) \\ &\qquad < P(M) \cdot \sum_{T_\lambda < w < T_\mu} m(w') + P(U) \cdot \sum_{T_\lambda < w' < T_\mu} u(w') \qquad (48) \end{aligned}$$

In most practical cases of course $P(M)$ is very small and the two sides of (48) are dominated by the second term. However if a "blocking" procedure has reduced the number of unmatched comparisons greatly it would be more appropriate to use $P^*(M)$ and $P^*(U)$ appropriate to the subspace Γ^* (i.e. to the set of comparisons that will be made explicitly), than to use $P(M)$ and $P(U)$ provided the same "blocking" procedure is to be used for each choice of comparison space. $P(M)$ and $P(U)$, or alternatively $P^*(M)$ and $P^*(U)$, have to be guessed at for the application of (48). The difference between the right hand side and the left hand side of (48) is equal to the reduction of $P(A_2)$ due to the choice of the comparison space.

In practice the difference between two comparison spaces will often be the number of configurations of component vectors which are listed out in addition to the simple "agreement"—"disagreement" configurations (e.g. "agreement on name Jones," "agreement on name Smith," etc.). The formula (48) can be used to compare the loss or gain in dropping some special configurations or listing out explicitly some more.

3.7. *Calculation of threshold values*

Having specified all the relevant configurations γ_j^k and determined their associated weights w_j^k; $k=1, 2, \cdots, K; j=1, 2, \cdots, n_k$ it remains to set the threshold values T_μ and T_λ corresponding to given μ and λ and to estimate the number or proportion of failures to make positive dispositions of comparisons.

As shown before, the number of weights to be determined is equal to $n_1+n_2 \cdots +n_K$. The total number of different configurations is, however, $n_1n_2 \cdots n_K$. Since the number of total configurations will, in most practical situations, be too large for their complete listing and ordering to be feasible we have resorted to sampling the configurations in order to estimate T_μ and T_λ. Since we are primarily interested in the two ends of an ordered list of total configurations we sample with relatively high probabilities for configurations which have very high or very low weights $w(\gamma)$.

The problem is made considerably easier by the independence of the component vectors γ^k. Thus if we sample independently the component configurations $\gamma_{j_1}^1, \gamma_{j_2}^2, \cdots, \gamma_{j_K}^K$ with probabilities $z_{j_1}^1, z_{j_2}^2, \cdots, z_{j_K}^K$ respectively we will have sampled the total configuration $\gamma_j=(\gamma_{j_1}^1, \gamma_{j_2}^2, \cdots, \gamma_{j_K}^K)$ with probability $z_j=z_{j_1}^1, z_{j_2}^2 \cdots z_{j_K}^K$. Hence we do not need to list all configurations of γ for sampling purposes, only all configurations of γ^k for each k.

We speed up the sampling process and increase the efficiency of the sample by ordering the configurations listed for each component by decreasing values w^k, and sampling according to the following scheme:

1) Assign selection probabilities $z_1^k, z_2^k, \cdots, z_{n_k}^k$ roughly proportional to $|w_j^k|\cdot$
2) Choose a configuration from each component. If the configuration γ_j^k is chosen from the kth component (with probability z_j^k) choose also the configuration $\gamma_{n_k-j+1}^k$.
3) Combine the first members of the pairs chosen from each component to give one total configuration and the second members to give another.
4) Repeat the whole procedure $S/2$ times to give a with-replacement sample of S total configurations.

The sample is then ordered by decreasing values of

$$w = w_1 + w_2 + \cdots + w_K. \tag{49}$$

Let $\gamma_h(h=1, 2, \cdots, S)$ be the hth member of the ordered listing of the sample. (Note: If a configuration with the same value of w occurs twice in the sample, it is listed twice.) Then $P(w(\gamma)<w(\gamma_h)|\gamma\in M)$ is estimated by

$$\lambda_h = \sum_{h'=h}^{S} m(\gamma_{h'})/\pi(\gamma_{h'}) \tag{50}$$

where

$$\pi(\gamma_h) = \frac{S}{2}\cdot z'(\gamma_h) \tag{51}$$

and

$$z'(\gamma_h) = \overset{1}{z}_{h_1}\overset{2}{z}_{h_2}\cdots\overset{K}{z}_{h_K} + \overset{1}{z}_{n_1-h_1+1}\overset{2}{z}_{n_2-h_2+1}\cdots\overset{K}{z}_{n_K-h_K+1} \tag{52}$$

while

$$P'(w(\gamma) < w(\gamma_h) \mid \gamma \in U) \qquad \text{is estimated by}$$

$$\mu_h = \sum_{h'=1}^{h} u(\gamma_{h'})/\pi(\gamma_{h'}). \tag{53}$$

The threshold values $T(\lambda_{h'})$ and $T(\mu_{h'})$, are simply the weights $w(\gamma_{h'})$ and $w(\gamma_{h'})$.

We have written a computer program which, working from a list of configurations for each vector component and associated selection probabilities, selects a sample of total configurations, orders the sample according to (49), calculates the estimates (50) and (53) and finally prints out the whole list giving for each total configuration its associated λ_h, μ_h, $T(\lambda_h)$, and $T(\mu_h)$.

We can use the same program to examine alternative blocking procedures (see Section 3.4). Thus in the ordered listing of sampled configurations we can identify those which would be implicit positive non-links under a blocking procedure which restricts explicit comparisons to a subspace Γ^*. Thus corresponding to any values of T_μ and T_λ (or μ and λ) we can obtain the second terms in each of the expressions (40), (41), (42), and (43). Alternatively if the implicit positive non-links are passed over in the summations (40) and (41) we can read off the values of the left-hand sides of those expressions. If we arrange this for alternative blocking procedures we are able to use the output of the program to make a choice of blocking procedures according to (46).

4. ACKNOWLEDGMENTS

The authors would like to express their gratitude to the Dominion Bureau of Statistics for providing opportunties for this research and in particular to Dr. S. A. Goldberg for his continued support.

The authors would also like to express their appreciation to H. B. Newcombe for his pioneering work in the field of record linkage and for his generous encouragement of an approach which, in many respects, differs from his own. The contributions of J. N. Gauthier, the systems analyst in charge of programming for our pilot project, have been essential to whatever success we have enjoyed so far and will continue to be essential in what remains for us to do.

REFERENCES

[1] Du Bois, N. S. D., "A solution to the problem of linking multivariate documents, *Journal of the American Statistical Association,* 64 (1969) 163–174.

[2] Fellegi, I. P. and Sunter, A. B., "An optimal theory of record linkage," *Proceedings of the International Symposium on Automation of Population Register Systems, Volume 1*, Jerusalem, Israel, 1967.

[3] Nathan, G., "Outcome probabilities for a record matching process with complete invariant information," *Journal of the American Statistical Association*, 62 (1967) 454–69.

[4] Newcombe, H. B. and Kennedy, J. M., "Record linkage: Making maximum use of the discriminating power of identifying information," *Communications of the A.C.M.* 5 (1962) 563.

[5] Newcombe, H. B., Kennedy, J. M., Axford, S. L., and James, A. P., "Automatic linkage of vital records," *Science* 130, (1959) 954.

[6] Newcombe, H. B. and Rhynas, P. O. W., "Family linkage of population records," Proc. U.N. / W. H. O. Seminar on Use of Vital and Health Statistics for Genetic and Radiation Studies; United Nations Sales No: 61, XVII 8, New York, 1962.

[7] Nitzberg, David M. and Sardy, Hyman, "The methodology of computer linkage of health and vital records," *Proc. Soc. Statist. Section, American Statistical Association*, Philadelphia, 1965.

[8] Phillips, Jr., William and Bahn, Anita K., "Experience with computer matching of names," *Proc. Soc. Statist. Section, American Statistical Association*, Philadelphia, 1965.

[9] Sunter, A. B., "A statistical approach to record linkage; record linkage in medicine," *Proceedings of the International Symposium*, Oxford, July 1967; E. & S. Livingstone Ltd., London, 1968.

[10] Sunter, A. B., and Fellegi, I. P., "An optimal theory of record linkage," 36th Session of the International Statistical Institute, Sydney, Australia, 1967.

[11] Tepping, B. J., "Study of matching techniques for subscriptions fulfillment," National Analysts Inc., Philadelphia, August, 1955.

[12] Tepping, B. J., "A model for optimum linkage of records," *Journal of the American Statistical Association*, 63 (1968) 1321–1332.

[13] Tepping, B. J., and Chu, J. T., "A report on matching rules applied to readers digest data," National Analysts Inc., Philadelphia, August, 1958.

APPENDIX I

A FUNDAMENTAL THEOREM FOR RECORD LINKAGE

We stated that (μ, λ) is an admissible pair of error levels provided μ and λ are not both too large. We will make this statement more precise.

Let

$$U_n = \sum_{i=1}^{n} u_i; \qquad n = 1, 2, \cdots, N_\Gamma \tag{1}$$

$$U_0 = 0 \tag{2}$$

$$M_{n'} = \sum_{i=n'}^{N_\Gamma} m_i; \qquad n' = 1, 2, \cdots, N_\Gamma \tag{3}$$

$$M_{N_\Gamma+1} = 0 \tag{4}$$

and define $f(\mu)$, as shown in Figure 1, on the interval (0, 1) as the monotone decreasing polygon line passing through the points (U_n, M_{n+1}) for $n=0, 1, \cdots, N$. It is possible of course to state the definition more precisely, but unnecessary for our purposes.

The area contained by the axes and including the line $\lambda = f(\mu)$ defines the region of admissible pairs (μ, λ). In other words (μ, λ) is an admissible pair if

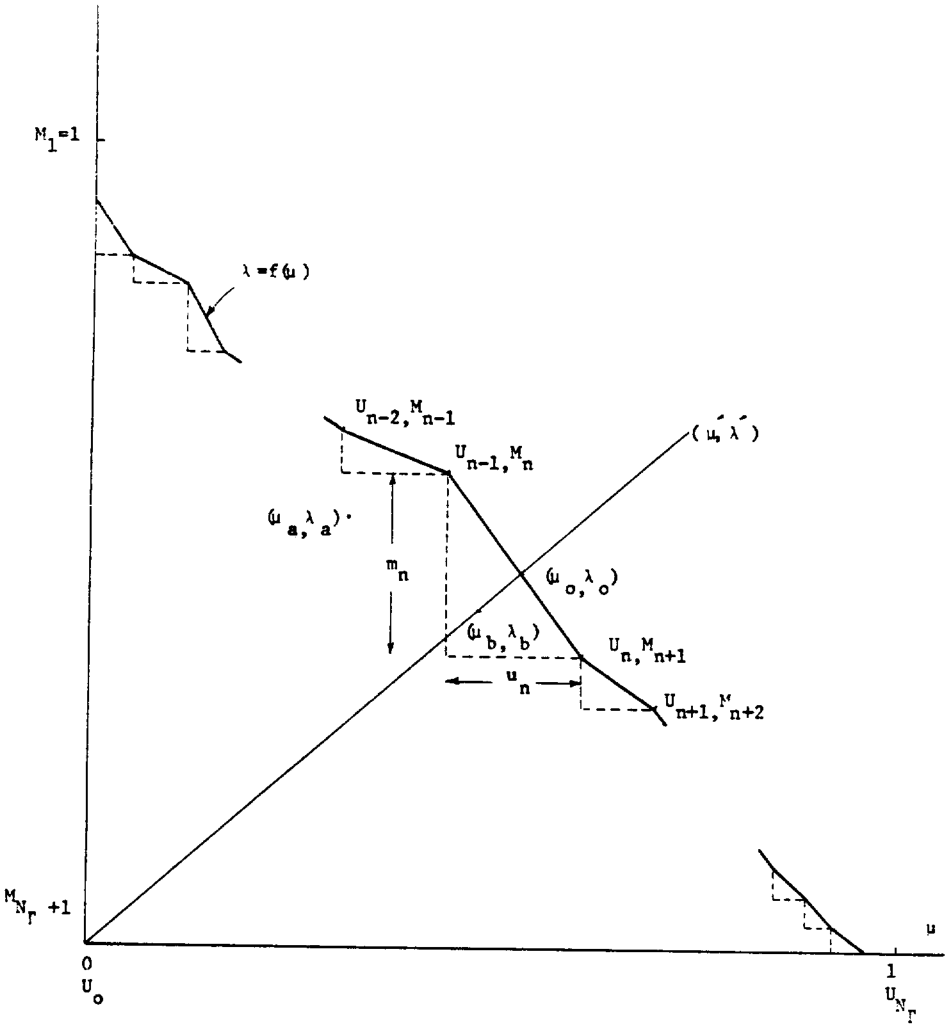


FIG. 1

$$
\begin{aligned}
&0 < \lambda \le f(\mu) \\
\text{and}\quad &0 < \mu.
\end{aligned}
\tag{5}
$$

Let $n(\mu)$ be the integer such that

$$U_{n(\mu)-1} < \mu \le U_{n(\mu)} \tag{6}$$

and $n'(\lambda)$ the integer such that

$$M_{n'(\lambda)} \ge \lambda > M_{n'(\lambda)+1}. \tag{7}$$

Define

$$P_\lambda = \frac{\lambda - M_{n'(\lambda)+1}}{m_{n'(\lambda)}} \tag{8}$$

and

$$P_\mu = \frac{\mu - U_{n(\mu)-1}}{u_{n(\mu)}}\,. \tag{9}$$

It follows from the way in which the configurations were ordered and the restrictions on μ and λ that the denominators of the expressions on the right of (8) and (9) are positive.

It is easy to see from Figure 1 that

$$0 < P_\lambda \leq 1 \quad \text{and} \quad 0 < P_\mu \leq 1. \tag{10}$$

It is also clear from Figure 1 that (μ, λ) are admissible if and only if

$$\begin{aligned} &\text{(a)} \quad n'(\lambda) \geq n(\mu) + 1 \\ &\qquad \text{(e.g. } (\mu_a, \lambda_a) \text{ in Figure 1)} \\ &\qquad\qquad\qquad \text{or} \\ &\text{(b)} \quad n'(\lambda) = n(\mu) \quad \text{and} \quad P_\lambda + P_\mu \leq 1 \\ &\qquad \text{(e.g. } (\mu_b, \lambda_b) \text{ in Figure 1).} \end{aligned} \tag{11}$$

Thus (a) and (b) simply divide the admissible region into two areas, one bounded by the axes and the broken lines in Figure 1, and the other bounded by the broken lines and the polygon line $\lambda = f(\mu)$.

Finally, from Figure 1 and the definitions of $n(\mu)$ and $n'(\lambda)$ we see that $\lambda = f(\mu)$ if and only if

$$\text{(a)} \quad n'(\lambda) = n(\mu) + 1 \quad \text{and} \quad P_\lambda = P_\mu \tag{12}$$

(i.e. the vertices of $\lambda = f(\mu)$).

or

$$\text{(b)} \quad n'(\lambda) = n(\mu) \quad \text{and} \quad P_\lambda + P_\mu = 1 \tag{13}$$

(i.e. points on $\lambda = f(\mu)$ other than vertices).

Let (μ, λ) be an admissible pair of error levels on Γ. We define a linkage rule $L_0(\mu, \lambda, \Gamma)$ as follows:

1) If $n'(\lambda) > n(\mu) + 1$ then

$$d_0(\gamma_i) = \begin{cases} (1, 0,) & \text{if } i \leq n(\mu) - 1 \\ (P_\mu, 1 - P_\mu, 0) & \text{if } i = n(\mu) \\ (0, 1, 0) & \text{if } n(\mu) + 1 \leq i \leq n'(\lambda) - 1 \\ (0, 1 - P_\lambda, P_\lambda) & \text{if } i = n'(\lambda) \\ (0, 0, 1) & \text{if } i \geq n'(\lambda) + 1 \end{cases}$$

2) If $n'(\lambda) = n(\mu)$ and $P_\lambda + P_\mu \leq 1$

$$d_0(\gamma_i) = \begin{cases} (1, 0, 0) & \text{if } i \leq n(\mu) - 1 \\ (P_\mu, 1 - P_\mu - P_\lambda, P_\lambda) & \text{if } i = n(\mu) = n'(\lambda) \\ (0, 0, 1) & \text{if } i \geq n'(\lambda) + 1. \end{cases}$$

(It is easy to see that (μ, λ) is admissible if and only if one of the two conditions above holds.)

We have now defined a linkage rule for an arbitrary pair of admissible levels (μ, λ). It follows immediately from the definition of $L_0(\mu, \lambda, \Gamma)$ that $P(A_2)=0$ if and only if $\lambda=f(\mu)$

Theorem: If (μ, λ) is an admissible pair of error levels on Γ then $L_0(\mu, \lambda, \Gamma)$ is the best linkage rule on Γ at the levels μ and λ. If (μ, λ) is not admissible on Γ then there are levels (μ_0, λ_0) with

$$\mu_0 \leqq \mu, \quad \text{and} \quad \lambda_0 \leqq \lambda \tag{14}$$

(with at least one of the inequalities in (14) being a definite inequality) such that $L_0^*(\mu_0, \lambda_0, \Gamma)$ is better than $L_0(\mu, \lambda, \Gamma)$ and for which

$$P_{L_0}(A_2) = 0. \tag{15}$$

This theorem explains the terminology "inadmissible." This simply means that we should not consider linkage rules at inadmissible error levels, since in this case L_0^* always provides a linkage rule at lower error levels for which we still have $P(A_2)=0$ (i.e. only the positive dispositions A_1 and A_3 occur).

Proof:

Let $L'(\mu, \lambda, \Gamma)$ be any linkage rule with admissible levels (μ, λ). Then $L'(\mu, \lambda, \Gamma)$ can be characterized by the set of decision functions

$$d'(\gamma_i) = (P'_{i1}, P'_{i2}, P'_{i3}), \qquad \sum_{j=1}^{3} P'_{ij} = 1 \qquad i = 1, 2, \cdots, N_\Gamma \tag{16}$$

where

$$P'_{ij} = P(A_j \mid \gamma_i), \qquad j = 1, 2, 3; \qquad i = 1, 2, \cdots, N_\Gamma. \tag{17}$$

Clearly

$$P_{L'}(A_1 \mid U) = \sum_{i=1}^{N_\Gamma} u_i P'_{i1} = \mu \tag{18}$$

$$P_{L'}(A_3 \mid M) = \sum_{i=1}^{N_\Gamma} m_i P'_{i3} = \lambda. \tag{19}$$

Consider the linkage rule $L_0(\mu, \lambda, \Gamma)$. It is characterized by equations analogous to (16) to (19) but P'_{ij} replaced by P_{ij} as defined above. We shall prove that

$$P(A_2 \mid L_0) \leqq P(A_2 \mid L') \tag{20}$$

According to the construction of L_0 the u_i which happen to be zero have the smallest subscripts, the m_i which happen to be zero have the largest subscripts. More rigorously, there are subscripts r and s such that

$$u_i = 0 \quad \text{if } i \leqq r-1, \qquad u_i > 0 \quad \text{if } i \geq r \tag{21}$$

$$m_i = 0 \quad \text{if } i \geqq s+1, \qquad m_i > 0 \quad \text{if } i \leq s \tag{22}$$

We have seen previously that

$$u_{n(\mu)} > 0$$

and

$$m_{n'(\lambda)} > 0$$

hence

$$n(\mu) \geqq r$$
$$n'(\lambda) \leqq s$$

hence

$$P_{i1} = 1 \qquad \text{for } i = 1, 2, \cdots, r - 1 \tag{23}$$

$$P_{i3} = 1 \qquad \text{for } i = s + 1, s + 2, \cdots, N_\Gamma \tag{24}$$

that is, whenever u_i is zero then $P_{i1}=1$ and whenever $m_i=0$ then $P_{i3}=1$.

By definition of μ, it follows that

$$\sum_{i=1}^{N_\Gamma} u_i P_{i1} = \sum_{i=1}^{N_\Gamma} u_i P'_{i1} = \mu. \tag{25}$$

Putting $n=n(\mu)$ and observing that $P_{i1}=1$ if $i \leqq n-1$ we can express (25) as follows:

$$\sum_{i=1}^{n-1} u_i + u_n P_\mu = \sum_{i=1}^{N_\Gamma} u_i P'_{i1}$$

or

$$\sum_{i=1}^{n-1} u_i(1 - P'_{i1}) + u_n(P_\mu - P'_{n,1}) = \sum_{i=n+1}^{N_\Gamma} u_i P'_{i1}. \tag{26}$$

With the possible exception of the last term on the left it is clear that every term in (26) is non-negative. We assume, without loss of generality, that the term in question *is* non-negative for, if it were negative, we would simply transfer it to the other side of the equality and all of the steps to follow would hold. It follows that if not every term in (26) is equal to zero then both sides are positive. Assume for the moment that this is the case.

It follows from the ordering of Γ that

$$u_i m_j \leqq u_j m_i \qquad \text{whenever } i < j. \tag{27}$$

It is now seen that

$$\begin{aligned} &\left[\sum_{j=n+1}^{N_\Gamma} m_j P'_{j1}\right]\left[\sum_{i=1}^{n-1} u_i(1 - P'_{i1}) + u_n(P_\mu - P'_{n,1})\right] \\ &\qquad \leqq \left[\sum_{i=1}^{n-1} m_i(1 - P'_{i1}) + m_n(P_\mu - P'_{n,1})\right]\left[\sum_{j=n+1}^{N_\Gamma} u_j P'_{j1}\right] \end{aligned} \tag{28}$$

since by (27) every term in the expansion of the left hand side is of the form

$$m_j u_i P'_{j1}(1 - P'_{i1}) \quad \text{or} \quad m_j u_n P'_j(P_\mu - P_{n,1}) \qquad (i \leqq n < j)$$

and corresponding to each there is a similar term on the right hand side but with $m_j u_i$ replaced by $m_i u_j$ and $m_j u_n$ replaced by $m_n u_j$. Dividing (28) by (26) we get

$$\sum_{j=n+1}^{N_\Gamma} m_j P'_{j1} \leqq \sum_{j=1}^{n-1} m_j(1 - P'_{j1}) + m_n(P_\mu - P'_{n,1})$$

or

$$\sum_{i=1}^{N_\Gamma} m_i P'_{i1} \leqq \sum_{i=1}^{N_\Gamma} m_i P_{i1}. \tag{29}$$

If every term in (26) was zero (29) would still hold since in that case we would have

$$P_{i1} = P'_{i1} \qquad \text{for } i \geqq r$$

i.e. whenever $u_i \neq 0$ and we would have

$$P_{i1} = 1 \geqq P'_{i1} \qquad \text{for } i \leqq r - 1$$

because of (23) and because $P'_{i1} \leqq 1$ for every i. Hence (29) would hold in this case as well.

By definition

$$\sum_{i=1}^{N_\Gamma} m_i P'_{i3} = \sum_{i=1}^{N_\Gamma} m_i P_{i3} = \lambda. \tag{30}$$

From (29) and (30) we get

$$\sum_{i=1}^{N_\Gamma} m_i(P'_{i1} + P'_{i3}) \leqq \sum_{i=1}^{N_\Gamma} (P_{i1} + P_{i3})$$

or

$$\sum_{i=1}^{N_\Gamma} m_i(1 - P'_{i2}) \leqq \sum_{i=1}^{N_\Gamma} m_i(1 - P_{i2}). \tag{31}$$

Because

$$\sum_{i=1}^{N_\Gamma} m_i = 1, \qquad \text{we get}$$

$$\sum_{i=1}^{N_\Gamma} m_i P_{i2} \leqq \sum_{i=1}^{N_\Gamma} m_i P'_{i2}$$

or

$$P_{L_0}(A_2 \mid M) \leqq P_{L'}(A_2 \mid M). \tag{32}$$

It can be shown similarly that

$$P_{L_0}(A_2 \mid U) \leqq P_{L'}(A_2 \mid U). \tag{33}$$

But (32) and (33) together state that

$$P(A_2 | L_0) \leqq P(A_2 | L') \tag{34}$$

which completes the proof of the first part of the theorem. Note that we have actually proved more than (34) since we have proved that L_0 is optimal separately under both the conditions M *and* the condition U. This also explains why the prior probabilities $P(M)$ and $P(U)$ do not enter either the statement or the proof of the theorem; our result is independent of these prior probabilities. The underlying reason, of course, lies in the fact that the error levels are concerned with conditional probabilities of misallocation. The situation would change if one tried to minimize the unconditional probability of misallocation or if one tried to minimize some general loss function.

As for the proof of the second part, let (μ', λ') be an inadmissible pair of error levels $(0<\mu<1,\ 0<\lambda<1)$. Since $f(\mu)$ is a strictly monotone decreasing continuous function in the range determined by

$$0 < \mu < 1$$
$$0 < f(\mu) < 1$$

it will intersect at a unique point the straight line drawn through (0, 0) and (μ', λ'). This is illustrated in Figure 1. Denote this point by (μ_0, λ_0). Then

$$0 < \mu_0 < \mu' < 1$$
$$0 < \lambda_0 < \lambda' < 1$$

and

$$\lambda_0 = f(\mu_0). \tag{35}$$

The linkage rule $L_0(\mu_0, \lambda_0, \Gamma)$ is, in light of (36), (12), and (13) such that

$$P(A_2 | L_0) = 0.$$

Hence $L_0(\mu_0, \lambda_0, \Gamma)$ is a better linkage rule than any other linkage rule at the level (μ', λ').

This completes the full proof of our theorem.

The form of the theorem given in the text is an immediate corollary of the theorem above and the expression (11).

APPENDIX II

METHOD II FOR THE CALCULATION OF WEIGHTS

Denoting

$$N_A N_B = c$$

the equations resulting from (37) to (39) by dropping expected values can be written as

$$M_k = \frac{N}{c} \prod_{j=1, j\neq k}^{3} m_j + \frac{c-N}{c} \prod_{j=1, j\neq k}^{3} u_j \qquad k = 1, 2, 3 \tag{1}$$

$$U_k = \frac{N}{c} m_k + \frac{c - N}{c} u_k \qquad k = 1, 2, 3 \tag{2}$$

$$M = \frac{N}{c} \prod_{j=1}^{3} m_j + \frac{c - N}{c} \prod_{j=1}^{3} u_j. \tag{3}$$

We introduce the transformation

$$\overset{*}{m_k} = m_k - U_k \tag{4}$$

$$\overset{*}{u_k} = u_k - U_k. \tag{5}$$

Substituting m_k and u_k from (4) and (5) into (2) we obtain

$$\frac{N}{c} \overset{*}{m_k} + \frac{c - N}{c} \overset{*}{u_k} = 0 \qquad k = 1, 2, 3. \tag{6}$$

Substituting (4) and (5) into (1) and then substituting in the resulting equations u_k^* from (6) we obtain

$$\prod_{j=1, j \neq k}^{3} \overset{*}{m_j} = \frac{c - N}{N} \left[M_k - \prod_{j=1, j \neq k}^{3} U_j \right] \qquad k = 1, 2, 3. \tag{7}$$

Denoting

$$R_k = M_k - \prod_{j=1, j \neq k}^{3} U_j \qquad k = 1, 2, 3 \tag{8}$$

we obtain by multiplying the three equations under (7) and by taking square roots

$$\prod_{j=1}^{3} \overset{*}{m_j} = \left(\frac{c - N}{N} \right)^{\frac{3}{2}} \left(\prod_{j=1}^{3} R_j \right)^{\frac{1}{2}} \tag{9}$$

Dividing (9) by (7) and putting

$$X = \sqrt{(c - N)/N} \tag{10}$$

$$B_k = \sqrt{\prod_{j=1, j \neq k}^{3} R_j / R_k} \qquad k = 1, 2, 3 \tag{11}$$

we get

$$\overset{*}{m_k} = B_k X \qquad k = 1, 2, 3 \tag{12}$$

and, from (4) to (6),

$$m_k = U_k + B_k X \qquad k = 1, 2, 3 \tag{13}$$

$$u_k = U_k - B_k / X \qquad k = 1, 2, 3. \tag{14}$$

We can now substitute into (3) m_k and u_k from (13) and (14) respectively and N as expressed from (10). We obtain

$$\frac{1}{X^2 + 1} \prod_{j=1}^{3} (U_j + B_j X) + \frac{X^2}{X^2 + 1} (U_j - B_j / X) = M. \tag{15}$$

After expanding (15), some cancellations and substitution of B_k from (11) we get the following quadratic equation in X:

$$\sqrt{\prod_{j=1}^{3} R_j}\,(X^2 - 1) + \left[\prod_{j=1}^{3} U_j + \sum_{j=1}^{3} R_jU_j - M\right] X = 0. \qquad (16)$$

The positive root of this equation is

$$X = \left\{M - \sum_{j=1}^{3} R_jU_j - \prod_{j=1}^{3} U_j + \sqrt{\left[M - \sum_{j=1}^{3} R_jU_j - \prod_{j=1}^{3} U_j\right]^2 + 4\prod_{j=1}^{3} R_j}\right\} \Big/ 2\sqrt{\prod_{j=1}^{3} R_j}. \qquad (17)$$

The estimates of m_k, u_k and N are now easily obtained from (10), (13) and (14).

Having solved these equations we can proceed to estimate the specific values of $m(\gamma)$ and $u(\gamma)$ which are required. We introduce some additional notation which, as before, refers to observable frequencies:

$M_k(\gamma_i^k)$ = the proportion of "agreement" in all components except the kth; the specific configuration γ_i^k in the kth component

$U_1(\gamma_i^2)$ = the proportion of "agreement" in the first, γ_i^2 in the second and any configuration in the third component

$U_1(\gamma_i^3)$ = the proportion of "agreement" in the first, γ_i^3 in the third and any configuration in the third component

$U_2(\gamma_i^1)$ = the proportion of γ_i^1 in the first, "agreement" in the second and any configuration in the third component.

The required values of $m(\gamma_i^k)$ and $u(\gamma_i^k)$ are estimated as

$$m(\gamma_i^1) = \frac{M_1(\gamma_i^1) - u_3U_2(\gamma_i^1)}{m_2(m_3 - u_3)}(X^2 + 1) \qquad (18)$$

$$m(\gamma_i^2) = \frac{M_2(\gamma_i^2) - u_3U_1(\gamma_i^2)}{m_1(m_3 - u_3)}(X^2 + 1) \qquad (19)$$

$$m(\gamma_i^3) = \frac{M_3(\gamma_i^3) - u_2U_1(\gamma_i^3)}{m_1(m_2 - u_2)}(X^2 + 1) \qquad (20)$$

$$u(\gamma_i^1) = \frac{m_3U_2(\gamma_i^1) - M_1(\gamma_i^1)}{u_2(m_3 - u_3)}\;\frac{X^2 + 1}{X^2} \qquad (21)$$

$$u(\gamma_i^2) = \frac{m_3U_1(\gamma_i^2) - M_2(\gamma_i^2)}{u_1(m_3 - u_3)}\;\frac{X^2 + 1}{X^2} \qquad (22)$$

$$u(\gamma_i^3) = \frac{m_2U_1(\gamma_i^3) - M_2(\gamma_i^3)}{u_1(m_2 - u_2)}\;\frac{X^2 + 1}{X^2} \qquad (23)$$

The formulae (18) to (23) are easily verified by expressing the expected values of the quantitites $M_k(\gamma_i^k)$, $U_1(\gamma_i^2)$, etc. in terms of m_k, u_k, $m(\gamma_i^k)$ and $u(\gamma_i^k)$, dropping the expected values and solving the resulting equations (there will be two equations for each pair $m(\gamma_i^k)$ and $u(\gamma_i^k)$).

The necessary and sufficient conditions for the mechanical validity of the formulae in this section are that

$$m_k \neq u_k \qquad k = 1, 2, 3$$

and

$$R_k > 0 \qquad k = 1, 2, 3$$

Since

$$m_k = m(S_k) = \Pr(S_k \mid M)$$
$$u_k = u(S_k) = \Pr(S_k \mid U)$$

clearly for sensible definitions of "agreement" $m_k > u_k$ should hold for $k = 1, 2, 3$. In this case $R_k > 0$ will hold as well. The latter statement can easily be verified by substituting (1) and (2) into (8).

FIDDLING AROUND WITH NONMATCHES AND MISMATCHES

Fritz Scheuren and H. Lock Oh, Social Security Administration

The necessity of linking records from two or more sources arises in many contexts. One good example would be merging files in order to extend the amount or improve the quality of information available for population units represented in both files. In developing procedures for linking records from two or more sources, tradeoffs exist between two types of mistakes: (1) the bringing together of records which are for different entities (mismatches), and (2) the failure to link records which are for the same entity (erroneous nonmatches). Whether or not one is able to utilize one's resources in an "optimal" way, it is almost certainly going to be true that in most situations of practical interest some mismatching and erroneous nonmatching will be unavoidable. How to deal with these problems depends, of course, to a great extent on the purposes for which the data linkage is being carried out. Because these reasons can be so diverse, no general strategy for handling mismatches and nonmatches will be offered here. Instead, we will examine the impact of these difficulties on the analysis of a specific study. The study chosen is a large-scale matching effort, now nearing completion, which had as its starting point the March 1973 Current Population Survey (CPS).

THE 1973 CENSUS - SOCIAL SECURITY EXACT MATCH STUDY

The primary identifying information in the 1973 Census-Social Security study was the social security number (SSN). The problems which arise when using the SSN to link Current Population Survey interview schedules to Social Security records differ in degree, but not in kind, from the problems faced by other "matchmakers."

In the 1973 study, as in prior CPS-SSA linkages, the major difficulty encountered was incompleteness in the identifying information [1]. Manual searches had to be carried out at SSA for over 22,000 individuals for whom no SSN had been reported by the survey respondent [2]. Another major problem was reporting errors in the social security number or other identifiers (name and date of birth, etc.). SSN's were manually searched for at SSA in cases where severe discrepancies between the CPS and SSA information were found after matching the two sources using the account number initially provided [3]. Because of scheduling and other operational constraints, an upper limit of 4,000 manual searches had to be set for this part of the project. Therefore, it was possible to look for account numbers only in the most "likely" instances of CPS misreporting of the SSN. The cases sent through this search procedure were those for which both name and date of birth were in substantial disagreement. For social security beneficiaries, computerized (machine) searches at SSA were also conducted for both missing and misreported SSN's. This was made possible through an administrative cross-reference system which links together persons who receive benefits on the same claim number. About 1,000 potentially usable SSN's were obtained in this way.

Operational Restrictions on the Matching.-- One of the concerns the 1973 work has in common with earlier Census-SSA linkage efforts is the great care that is being taken to ensure the confidentiality of the shared information. The laws and regulations under which the agencies operate impose very definite restrictions on such exchanges, and special procedures have been followed throughout, so as to adhere to these provisions--in particular, to ensure that the shared information is used only for statistical purposes and not for administrative ones.1/ Another major restriction on the study was, of course, that it had to be conducted using data systems which were developed and are used principally for other purposes. The CPS, for instance, lacks a number of pieces of information that would, if available, have materially increased the chances of finding the surveyed individual in SSA's files. Finally, the manual searching for over 26,000 account numbers at Social Security imposed a sizable addition to the normal administrative workload in certain parts of the agency. Therefore, in order to obtain a reasonable priority for the project, numerous operational compromises were made which precluded the employment of "optimal" matching techniques [e.g., 4, 5, 6, 7, 8]. One of the most serious of these was the decision basically not to "re-search" for the missing and misreported SSN's of individuals for whom no potentially usable number was found after just one search.

Basic Match Results.--There were 101,287 interviewed persons age 14 or older who were included in the 1973 Census-Social Security Exact Match Study. Of the total, about 2 percent had not yet been issued an SSN at the time of the interview and, hence, were not eligible for matching. In another 8 percent of the cases, no potentially usable social security numbers could be found even though one was believed to exist. For the remaining 90,815 sampled individuals, an SSN was available, and CPS and SSA data could be linked. Of these account numbers, 77,465 were supplied by CPS respondents initially. There were also 3,347 cases where the SSN provided originally was replaced with an account number obtained from the manual and machine searches of SSA's files which were described above. In a few of these cases--about 200--the SSN's used as replacements were taken from a supplementary Census source. Finally, there were 10,003 sampled individuals for whom no account number had been provided initially, but one was obtained subsequently by a search of SSA's files.

ALTERNATIVE COMPUTERIZED MATCH RULES

In general, aside from certain obvious errors (which have already been eliminated), it is not possible to determine whether the SSN we have for a particular individual is his own or has been erroneously ascribed to him. One can, however, estimate the likelihood that a potentially usable account number is incorrect. To do this, five confirmatory variables common to both data sets were used: surname (first six characters), age attained in 1972 (in years), race, sex, and month of birth. The pattern of agreements and disagreements that might be expected between the CPS and SSA reporting on these variables depends, of course, on whether the records brought together are "mismatches" or "truematches." (See figure 1 below for definitions.)

Figure 1 -- Match Definitions

<u>TRUEMATCH</u> -- A match between a Social Security Administration (SSA) record and a Current Population Survey (CPS) interview schedule where the two sets of documents were for the same individual.

<u>MISMATCH</u> -- The erroneous matching of data from the two sources when the information brought together was not for the same individual.

<u>TRUE NONMATCHES</u> -- Individuals in the Current Population Survey who have not yet been issued a social security number (SSN) and therefore do not have a Social Security Administrative record.

<u>ERRONEOUS NONMATCH</u> -- A case where <u>either</u> no SSN could be found even though it had been issued (making it impossible to match the sources together) <u>or</u> the two sources were brought together but because of the <u>rule</u> used to decide what would be called a "match" they were treated erroneously as nonmatches.

<u>Mismatches</u>.--If mismatches arise on a purely chance basis, then the probability of agreement on any one variable would depend just on the marginal distribution of that variable in the two data sets being linked. This is the assumption we have made here. The conditional probability given a mismatch of a particular combination of agreements (disagreements) on the confirmatory information, denoted by $\{P^{MM}\}$, was thus estimated as the product of the observed marginal proportions of agreement and disagreement for each variable separately.

Two separate mismatch models were fit: one for SSN's obtained in manual searching and one for all other SSN's. This was necessary because of the nature of SSA's manual searching procedures where, for a number to be returned from the search, there usually must be at least rough agreement on surname and age. (Hence, these two variables could not be used for evaluating mismatches among persons with SSN's obtained from manual searching.)

<u>Truematches</u>.-- Differences between the CPS and SSA variables can arise quite frequently even when the data is for the same person. The information in the two systems is collected at very different times; perhaps as long as 30 or more years separate the two observations. Furthermore, the respondent on the two occasions may very well be different. For the most part, the Social Security variables were obtained from the individual himself, while in the CPS, over half the information was obtained by proxy.

The extent of agreement for "truematches" has also been modelled by assuming independence among the confirmatory variables. However, the conditional probabilities of agreement, given a truematch, denoted by $\{P^{TM}\}$, cannot be estimated separately from the overall mismatch rate, "α," that exists among the 90,815 individuals with potentially usable SSN's. To obtain estimates an Information Theoretic approach was taken; the $\{P^{TM}\}$ and α were obtained by (iteratively) fitting the observed proportions $\{\pi\}$ for each of the combinations of agreement or disagreement on the confirmatory variables that were found in the sample. The estimating equation was of the form

$$(1) \qquad \pi = (1 - \alpha)\, P^{TM} + \alpha\, \hat{P}^{MM}$$

where the $\{\hat{P}^{MM}\}$ were calculated as described above, with α and the $\{P^{TM}\}$ being chosed such that

$$(2) \qquad I(\hat{\pi};\pi) = \Sigma\, \hat{\pi} \ln \frac{\hat{\pi}}{\pi}$$

was a minimum. The $\{\hat{\pi}\}$ are given by the expression

$$(3) \qquad \hat{\pi} = (1 - \hat{\alpha})\, \hat{P}^{TM} + \hat{\alpha}\, \hat{P}^{MM}$$

and were used in obtaining table 1.

These models were judged to be adequate except for cases where there was perfect or near perfect agreement on the confirmatory variables. For such individuals, research from other SSA studies indicated that the estimated number of mismatches was probably too small, and some upward adjustments were made to the fitted results.2/

<u>Alternate Match Rules</u>.--The match rules considered in the remainder of this paper all use the extent of agreement on age, race, sex, month of birth, and surname to determine whether CPS and SSA records linked by common SSN's should be treated as "matches" or "nonmatches." Four ad hoc rules were examined:

1. <u>"Perfect" Agreement Rule</u>.--For this rule all five confirmatory variables had to agree within tolerance. For surname, which

Table 1. -- Estimated Number of Mismatches and Erroneous Nonmatches by Match Rule for March 1973 CPS Interviewed Persons 14 Years of Age and Older

Item	Perfect Agreement Rule	Surname Agreement Rule	CPS-SER Agreement Rule	Potentially Usable Rule
Total	90,815	90,815	90,815	90,815
Matched, Total	76,294	85,293	86,910	90,815
Truematches............	76,276	84,784	86,537	88,962
Mismatches.............	18	509	373	1,853
Mismatches as a Percent of Total Matches........	0.02	0.60	0.43	2.04
Nonmatches, Total	14,521	5,522	3,905	-
True Nonmatches........	1,835	1,344	1,480	-
Erroneous Nonmatches...	12,686	4,178	2,425	-

Note: Based on an unweighted CPS sample of all individuals with potentially usable SSN's,including a small number of Armed Forces members excluded from the weighted figures in the remaining tables.

depends on a character-by-character agreement of the first six letters of the last name, a tolerance of two letters was allowed. Similarly, a difference of four years was permitted in defining agreement on age. For sex, race, and month of birth, no tolerance was allowed.

2. Surname Agreement Rule.--This rule requires at least four of the first six letters of the surname to be the same. (The other confirming variables were not considered.) The surname rule is based on a modified version of the administrative procedures now in use at IRS and SSA to verify the correctness of the social security number supplied.

3. CPS-SER Agreement Rule.--This rule basically requires that four out of the five confirmatory variables agree (within the tolerances mentioned in the first rule above). In selected cases (361 altogether), agreement on just three variables was enough to consider the individual a match. It was this rule, discussed in report no. 4 of SSA's Series on Studies from Interagency Data Linkages, which has been employed for the first public-use match file prepared from the project and described in reports nos. 5 and 6 of that Series.

4. Potentially Usable Rule.--This is the least stringent of the rules in that no restrictions are placed on what is to be called a "match."

IMPACT OF ALTERNATE MATCH RULES ON EARNINGS

In assessing the four match rules being considered, it is not enough simply to look at them in terms of their respective mismatch and erroneous nonmatch rates. What we need to do is to take account of the bias and variance implications of the matching error on some of the chief variables to be provided by the linkage. Among the most important of these data items are the 1972 earnings information reported to the Census Bureau and to Social Security. In this section, therefore, we will compare these earnings data under each match rule. First, we will examine the extent to which one's overall "level" estimators of the CPS or SSA earnings distribution are affected by the different match rules. The level estimates are of interest principally because a standard exists for these against which a comparison can be made. What is crucial to our evaluation, however, is the sensitivity of the relationships between CPS and SSA earnings amounts to the match rule chosen. Here, of course, no outside standard exists, since it was to examine these relationships that the study was mounted.

<u>Level Comparisons</u>.--Tables 2 and 3 below compare the percentage distributions of CPS and SSA earnings for each procedure with preliminary overall survey or administrative control figures. No correction has been made for erroneous nonmatches or mismatches, but the sample has been reweighted to make a rough adjustment for differences which arise because of survey undercoverage [9].

Sizable discrepancies among the various estimates can be observed in the tables. For example, from table 2, it can be seen that the difficulty of obtaining an SSN may have been relatively greater for individuals who were not identified in the CPS as having worked in 1972. Large differences (statistically significant at $\alpha = 0.01$) exist, in fact, between each of the match results and the control for the "no earnings" category of the CPS classifier. On the other hand, both tables 2 and 3 show that persons with CPS or SSA earnings of $9,000 or more are always proportionately overrepresented in the sample. For the SSA classifier the observed differences for the $9,000 or more class are all significant at the $\alpha = 0.01$ level.

<u>Relationship Comparisons</u>.--The relationships between CPS and SSA reported earnings can be investigated in a number of ways. One of the standard methods is to cross-classify the two amounts by the same dollar size-classes and count the fraction of cases which fall into the same interval or into a higher or lower interval [11]. Table 4 provides a summary of such cross-tabulations for each match rule where the dollar size-classes used are the same as those shown in tables 2 and 3.

Table 2. -- Unadjusted CPS Earnings Percentage Distributions Under Alternate Match Rules, as Compared to the Overall Survey Estimate: Civilians 14 or Older with SSN's

Size of CPS Earnings	Overall Survey Estimate	Match Rule			
		Perfect Agreement Rule	Surname Agreement Rule	CPS-SER Rule	Potentially Usable Rule
TOTAL..........	100.0	100.0	100.0	100.0	100.0
None	35.0	32.8	33.6	34.0	34.2
$1 to $999 or Loss..	10.9	10.5	10.6	10.7	10.6
$1,000 to $1,999....	5.8	5.9	5.9	6.0	6.0
$2,000 to $2,999....	4.4	4.5	4.5	4.5	4.5
$3,000 to $3,999...	4.4	4.5	4.6	4.6	4.6
$4,000 to $4,999...	4.4	4.5	4.5	4.5	4.5
$5,000 to $5,999...	4.5	4.7	4.7	4.7	4.7
$6,000 to $6,999...	4.1	4.3	4.3	4.2	4.2
$7,000 to $7,999...	4.2	4.3	4.3	4.2	4.2
$8,000 to $8,999...	3.5	3.6	3.5	3.5	3.5
$9,000 or More.....	18.9	20.4	19.5	19.2	19.0

Note: Based on weighted sample counts for civilians, adjusted as explained in the text. Detail may not add to totals because of rounding.

Table 3. -- Unadjusted SSA Earnings Percentage Distributions Under Alternate Match Rules, as Compared to the Administrative Controls: Civilians 14 or Older with SSN's

Size of SSA Earnings	Administrative Control	Match Rule			
		Perfect Agreement Rule	Surname Agreement Rule	CPS-SER Rule	Potentially Usable Rule
TOTAL.......		100.0	100.0	100.0	100.0
None.............	40.9	39.2	40.0	40.6	41.0
$1 to $999.......	10.2	9.7	9.8	9.9	9.8
$1,000 to $1,999.	6.5	6.3	6.3	6.2	6.2
$2,000 to $2,999.	4.7	4.6	4.7	4.7	4.6
$3,000 to $3,999.	4.4	4.4	4.4	4.4	4.4
$4,000 to $4,999.	4.3	4.5	4.4	4.4	4.4
$5,000 to $5,999.	4.1	4.2	4.1	4.1	4.0
$6,000 to $6,999.	3.7	3.9	3.9	3.8	3.8
$7,000 to $7,999.	3.3	3.6	3.5	3.5	3.5
$8,000 to $8,999.	3.1	3.0	3.0	2.9	2.9
$9,000 or More...	14.8	16.5	15.8	15.5	15.3

Note: Based on weighted sample counts for civilians, adjusted as explained in the next. Detail may not add to totals because of rounding.

As can be seen from table 4, marked differences exist among the procedures in the proportion of individuals whose CPS and SSA earnings class agree. The percentages vary from a high of 68 percent for the perfect agreement rule to a low of 66 percent for the potentially usable one, with the surname and CPS-SER rules having class agreements of around 67 percent. The standard errors for the four estimators of the extent of earnings class agreement average about 0.25 percentage points. The range of the agreement figures (at 2.0 percentage points) is thus eight times the standard error.

Since our focus is on the matching process itself, we will leave to others [12, 13] a detailed study of the relationships between the earnings distributions shown in table 4. Instead, we will proceed (in the next section) to examine the bias and variance impact of adjustments designed to lessen the effect of errors in the matching.

UTILITY OF POST-HOC ADJUSTMENT PROCEDURES

In this section a combination of procedures is examined which is designed to adjust for mismatching and erroneous nonmatches. Successive adjustments will be made to the data: first, by reweighting to account for the nonmatches; then, by "raking" the results to the overall survey and administrative controls shown in tables 2 and 3; and, finally, by "subtracting out" estimates of the effect of the mismatching. The utility of each step taken will be evaluated in terms of its bias and variance impact.

<u>Reweighting for Nonmatches</u>.--No matter which of the four match rules is used, important differences exist between those who are treated as "matches" and those believed to have SSN's but for whom no usable account number could be determined. This is evident not only from tables 2 and 3, but also from previous papers which have discussed the reporting of social security numbers in the March 1973 Current Population Survey [i.e., 1, 2, 3]. For example, large differences exist between the two groups by earnings, age, race, sex, and respondent status.3/

One way to "correct" for these differentials (the method adopted in this paper) is to consider the cases where SSN's were obtained through manual searching as a sample from the entire group of

Table 4. -- Percentage Distribution of Earnings Class Agreement Between CPS and SSA Reported Amounts Under Alternate Match Rules Before Adjustment: Civilians 14 or Older with SSN's

Extent of Earnings Class Agreement	Perfect Agreement Rule	Surname Agreement Rule	CPS-SER Agreement Rule	Potentially Usable Rule
Total.............	100.00	100.00	100.00	100.00
SSA Earnings in Higher Interval than CPS........	10.84	11.35	11.05	11.70
CPS and SSA Earnings Class Agree....................	68.08	67.13	67.42	66.05
CPS Earnings in Higher Interval than SSA........	21.08	21.52	21.53	22.25

Note: Based on weighted sample counts for civilians, adjusted as explained in the text. Detail may not add to totals because of rounding.

individuals who "should" have usable numbers but do not. The exact procedure followed was to subtract from the estimated total with SSN's, the weighted number of adults who had an acceptable SSN but who had not obtained it from the manual search. The weighted manual search cases were then ratioed up to this difference and added to the estimates obtained from the rest of the sample. These steps were carried out for each of the eight CPS rotation groups separately in order to be able to come up with an approximation to the variance.4/ The overall adjustment factors applied are shown below for each match rule along with the (weighted) fraction of sample cases with SSN's but for which no usable SSN could be found.

Match Rule	Percent with No Usable SSN Found	Weighting Factor for Manual Search Cases
Perfect agreement rule..	26.9	3.4
Surname agreement rule..	13.2	2.2
CPS-SER rule...........	10.9	2.0
Potentially usable rule.	5.9	1.5

The reweighting procedure just described, while crude in many respects, does have a certain logic to it since the great bulk of the cases for whom no SSN is available were searched for manually in SSA's files. It might also be noted in passing that such an approach is quite analogous to the classical method for utilizing follow-up samples of those persons who, in the survey's initial wave, were nonrespondents [14].

To help evaluate the impact of the reweighting scheme, table 5 is provided below. As can be seen, for all match rules, the reweighting reduces the amount of CPS-SSA earnings-class agreement. In fact, the average declined by about 0.8 percent, from 67.17 percent to 66.40 percent. From internal evidence in the CPS, there seems to be a definite tendency for persons who provide usable SSN's to be better respondents than those who do not. Thus, this reduction in earnings-class agreement (with accompanying increases elsewhere) probably reduces the overall nonmatch bias which exists for all of the estimators. There is, of course, no way of knowing whether the magnitude of the changes is appropriate, but it is encouraging to note that the net effect of the reweighting is to bring the estimates for the four rules closer together. (The range of the percentages for earnings-class agreement dropped from 2.0 percent to 1.1 percent.

For the probable reduction in the nonmatch bias, a price has been paid in increasing the standard error of nearly all the estimators shown in the table. These increases range from small to moderate for the potentially usable,surname,and CPS-SER rules. However, for the perfect agreement

Table 5. -- Percentage Distribution of Earnings Class Agreement Between CPS and SSA Reported Amounts Under Alternate Match Rules After Reweighting: Civilians 14 or Older with SSN's

Extent of Earnings Class Agreement	Perfect Agreement Rule	Surname Agreement Rule	CPS-SER Agreement Rule	Potentially Usable Rule
Total...............	100.00	100.00	100.00	100.00
SSA Earnings in Higher Interval than CPS..........	11.99	12.01	11.50	12.01
CPS and SSA Earnings Class Agree......................	66.74	66.34	66.81	65.70
CPS Earnings in Higher Interval than SSA..........	21.26	21.65	21.60	22.29

Note: Based on weighted sample counts for civilians, adjusted as explained in the text. Detail may not add to totals because of rounding.

rule, the increase is sizable; if such a rule were seriously being contemplated, some other method of adjustment would, in all likelihood, be desirable.

Raking Adjustment for Nonmatches.--The reweighting scheme just described tends to bring the matched CPS and SSA earnings distributions closer to the control totals shown in tables 2 and 3. However, the remaining discrepancies are still large. Unlike biases in the CPS-SSA interrelationships, which can only be adjusted indirectly and incompletely, it is possible to alter the sample earnings marginals so they conform simultaneously to both sets of controls more or less exactly. There are a number of well-known procedures for doing this. The approach employed here is due to Deming and Stephan [15], and we have referred to it, following the practice at the Census Bureau, as "raking." (Perhaps it is better known elsewhere as "the method of iterative proportions" [16].)

Table 6 provides a summary of the impact of the raking on the extent of agreement between CPS and SSA earnings. As will be seen, our estimators of the amount of agreement have declined still more as a result of this additional adjustment (from an average of 66.4 percent after reweighting to 66.2 percent after raking). The range in the extent of agreement has also narrowed further, from 1.1 percent to 0.9 percent, respectively, with the largest proportion on the main diagonal being 66.4 percent (CPS-SER) and the smallest, 65.5 percent (potentially usable rule). Again, we believe that this change represents a further reduction in the nonmatch bias. Not unexpectedly, the raking has also produced reductions in the standard errors, although not uniformly so. (For 8 of the 12 estimators in the table, there was some reduction. In the four instances where increases occurred, they were slight.)

Mismatch Adjustment.--If two linked records have been brought together just by chance, then it is highly unlikely for them to agree on earnings class. Thus, a "natural" consequence of the mismatching which exists under each rule is that the estimates of the extent of agreement, as shown in table 6, understate the true underlying amount of agreement. Some further adjustment, therefore, is necessary. There are a number of ways of taking account of the mismatches, depending on the assumptions one is willing to make about their affect on the relationship between the CPS and SSA classifiers. The model chosen here is a fairly simple one which may not be too unrealistic. Basically, it assumes that the mismatch rates do not depend on earnings levels and that, when a mismatch occurs, the matched CPS and SSA amounts are independently distributed. Put another way, the mismatches can be thought of as having the same row $\{P_{i.}\}$ and column $\{P_{.j}\}$ marginal proportions for CPS and SSA earnings, respectively, as the truematches; but such that the

Table 6. -- Percentage Distribution of Earnings Class Agreement Between CPS and SSA Reported Amounts Under Alternate Match Rules After Reweighting and Raking: Civilians 14 or Older with SSN's

Extent of Earnings Class Agreement	Perfect Agreement Rule	Surname Agreement Rule	CPS-SER Agreement Rule	Potentially Usable Rule
Total...............	100.00	100.00	100.00	100.00
SSA Earnings in Higher Interval than CPS..........	11.78	11.82	11.47	11.98
CPS and SSA Earnings Class Agree......................	66.01	65.89	66.36	65.45
CPS Earnings in Higher Interval than SSA..........	22.21	22.30	22.17	22.57

Note: Based on weighted sample counts for civilians, adjusted as explained in the text. Detail may not add to totals because of rounding.

proportion of mismatches for any particular combination ij of CPS and SSA earnings classes, denoted $\{P_{ij}^{MM}\}$, is given by

$$(4) \quad P_{ij}^{MM} = P_{i.} \, P_{.j} \, .$$

The expected value of the observed relationship between the two classifiers is assumed to consist of two components. First, there is an estimate of the truematch proportion in the $(ij)^{th}$ cell of the earnings cross-tabulation, denoted P_{ij}^{TM}, times the fraction of the total sample that were truematches, denoted by $(1 - \alpha)$. The second term consists of the mismatch proportion P_{ij}^{MM} times the fraction of the total sample that were mismatches (i.e., "α"). Thus, we have that the observed cell proportions $\{\pi_{ij}\}$ can be expressed as

$$(5) \quad E\pi_{ij} = (1 - \alpha) \, P_{ij}^{TM} + \alpha \, P_{ij}^{MM}$$

From (4) this becomes

$$(6) \quad E\pi_{ij} = (1 - \alpha) \, P_{ij}^{TM} + \alpha \, P_{i.} \, P_{.j}$$

Since estimates of the mismatch rate α, the CPS marginal $\{P_{i.}\}$, and SSA marginal $\{P_{.j}\}$ were all readily available (tables 1 to 3), it was a simple matter to obtain estimates of the $\{P_{ij}^{TM}\}$ by substituting $\hat{\alpha}$, $\hat{P}_{i.}$, and $\hat{P}_{.j}$ in (6). The $\{\hat{P}_{ij}^{TM}\}$ so obtained were then used to produce the results in table 7. 5/

For the perfect agreement rule, the mismatching had only a small effect, but, for the other rules, changes in the percent with CPS and SSA earnings in the same interval were substantial. For the potentially usable rule, where the amount of mismatching was estimated to be greatest, that proportion increased by 1 percent, from 65.45 percent to 66.45 percent. Increases for the CPS-SER and surname rules were smaller but still sizable (0.3 and 0.4 percentage points, respectively). The range of the four estimates of the extent of agreement narrowed again as a result of this final adjustment (from 0.91 percent after raking to 0.59 percent). The "cost" of the mismatch adjustment was a very slight increase in the variance over that of the raked estimator.

Summary of Impact of Adjustments.--Overall, when we look at the combined affect of all three adjustments, we see that the range of earnings class agreement under the four rules has been reduced to less than one-third of what it was to begin with (i.e., from 2.0 percent to 0.6 percent). This narrowing of the range of agreement suggests that the techniques employed

Table 7. -- Percentage Distribution of Earnings Class Agreement Between CPS and SSA Reported Amounts Under Alternate Match Rules After All Adjustments, Including the Adjustment for Mismatching: Civilians 14 or Older with SSN's

Extent of Earnings Class Agreement	Perfect Agreement Rule	Surname Agreement Rule	CPS-SER Agreement Rule	Potentially Usable Rule
Total...............	100.00	100.00	100.00	100.00
SSA Earnings in Higher Interval than CPS..........	11.77	11.63	11.34	11.46
CPS and SSA Earnings Class Agree......................	66.03	66.25	66.62	66.45
CPS Earnings in Higher Interval than SSA..........	22.20	22.12	22.05	22.10

Note: Based on weighted sample counts for civilians, adjusted as explained in the text. Detail may not add to totals because of rounding.

may have been "moderately" successful in reducing the various biases which affect each rule (and may even have some merit in general). However, since the range in earnings-class agreement after adjustment is still about twice the standard deviation, it seems likely that residual uncorrected biases remain an important part of the total mean square error.

Except for the perfect agreement rule, the price that was paid for this bias reduction appears to be "small." The median increase in the standard errors was about 10 percent of the original standard errors. (However, since the sample sizes involved are so large, this amounted to only 0.025 percentage points.)

In the light of our computations, it might be of interest to comment on which match rule is "best." Because the final results are so close, this question has lost some of its force but is still worth pursuing. By and large, the results suggest that in this case, and for the statistics considered, the best choice of the four match rules examined is the potentially usable rule. 6/ It tends to have the smallest standard error after all adjustments; its initial and final estimates change the least; and, its initial and final estimates are the closest of any rule to the overall average for all rules after adjustment. Partly as a consequence of this finding, all subsequent public-use data tapes to be prepared from the 1973 Census-Social Security Study will be made available with all the potentially usable "matches" included. 7/ Also, since information on the extent of agreement on the confirmatory variables is available on these data tapes, another consequence of this decision is that users will have the option of choosing the match rule best suited for their purposes.

Conclusion.--Matched statistical samples have much in common with other surveys and, as we have seen, adjustment techniques normally encountered in standard practice (e.g., raking), can be applied successfully to linked data sets as well. The problems of choosing a suitable match rule and of dealing with mismatches are, however, unique to record linkage studies. Usually, in the literature on data linkage, match rules (and mismatching) have been dealt with in the context of the research design and how to choose "optimal" strategies for allocating resources. With few exceptions [17], there has been insufficient attention given to the analysis aspects of imperfectly matched samples. In the 1973 Census-Social Security Study, the administrative (and, to some extent, confidentiality) constraints imposed on the design and execution of the data linkage make these analysis issues particularly pointed.

Our approach to them has, of course, been quite applied. Obviously, theoretical examinations are warranted as an adjunct to the empirical work on matching commented on here. We invite participation in this endeavor.

FOOTNOTES

*The authors would like to thank Wendy Alvey and Gina Savinelli for their assistance, especially for helping to prepare the basic tabulations. Thanks also must be extended to Ben Bridges and Dean Leimer for their careful reading of an earlier draft.

1/ For details on the confidentiality precautions taken, see the invited paper session on the Reconciliation of Survey and Administrative Sources through Data Linkage shown elsewhere in these Proceedings.

2/ A paper is in preparation which provides more details on the procedures employed in estimating the number of mismatches with particular attention to other estimation methods.

3/ In the public-use file (with the CPS-SER match rule), the reweighting adjustment being made attempts to take account of most of these factors. See report nos. 5 and 6 in Studies from Interagency Data Linkages for details.

4/ The raking and mismatch adjustments were also carried out separately by CPS rotation group to make it possible to approximate their variance impact as well.

5/ The mismatch rates used were not those shown in table 1 but were calculated (by rotation group) in terms of the weighted data after having taken account of the adjustments for nonmatches.

6/ Readers should carefully note the qualifications on this "endorsement" of the potentially usable rule. While for the example chosen here the nonmatch and mismatch errors of this rule tended to cancel each other out, this would not always be the case. In fact, the potentially usable rule, if not adjusted for mismatches, in many situations might even be the worst rule one could choose.

7/ For reasons of confidentiality, social security information for CPS respondents who refused to provide their SSN's to the Census Bureau are not includable on the public-use files from this project, even though it was possible to find on account number for them. With the CPS-SER rule, 619 such cases were eliminated. With the potentially usable rule, 641 cases would have to be treated as nonmatches for this reason.

REFERENCES

[1] Vogel, L., and Coble, T., "Current Population Survey Reporting of Social Security Numbers," 1974 Amer. Stat. Assn. Proc. Soc. Stat. Sec., 1975, pp. 130-136.

[2] Kilss, B., and Tyler, B., "Searching for Missing Social Security Numbers," 1974 Amer. Stat. Assn. Proc. Soc. Stat. Sec., 1975, pp. 145-150.

[3] Cobleigh, C., and Alvey, W., "Validating Reported Social Security Numbers," 1974 Amer. Stat. Assn. Proc. Soc. Stat. Sec., 1975, pp. 137-144.

[4] Tepping, B. J., "A Model for Optimum Linkage of Records," J. Amer. Stat. Assn., vol. 63, 1968, pp. 1321-1332.

[5] Felleghi, I. P., and Sunter, A. B.,"A Theory for Record Linkage," J. Amer. Stat. Assn., vol. 64, 1969, pp. 1183-1210.

[6] DuBois, Jr., N. S. D., "A Solution to the Problem of Linking Multivariate Documents," J. Amer. Stat. Assn, vol. 64, 1969, pp. 163-174.

[7] Wells, B., Optimum Matching Rules, University of North Carolina, 1974.

[8] Nathan, G., "Outcome Probabilities for a Record Matching Process with Complete Invariant Information," J. Amer. Stat. Assn., vol. 62, 1967, pp. 454-469.

[9] Vaughan, D. R., and Ireland, C. T., "Adjusting for Coverage Errors in the March 1973 Current Population Survey," 1975 Amer. Stat. Assn. Proc. Soc. Stat. Sec.

[10] Mosteller, F., "Association and Estimation in Contingency Tables," J. Amer. Stat. Assn., vol. 63, 1968, pp. 1-28.

[11] Scheuren, F. J., and Oh, H. L., "A Data Analysis Approach to Square Tables," Comm. in Stat., July 1975.

[12] Alvey, W., and Cobleigh, C., "Exploration of Differences Between Linked Social Security and Current Population Survey Earnings Data for 1972," 1975 Amer. Stat. Assn. Proc. Soc. Stat. Sec.

[13] Johnston, M. P., "Evaluation of Current Population Survey Simulations of Payroll Tax Changes," 1975 Amer. Stat. Assn. Proc. Soc. Stat. Sec.

[14] Hansen, M., and Hurwitz, W., "The Problems of Non-Response in Sample Surveys," J. Amer. Stat. Assn., vol. 41, 1946, pp. 517-528.

[15] Deming, W. E., and Stephan, F. F., "On a Least Squares Adjustment of a Sampled Frequency Table When the Expected Marginal Tables are Known," Annals Math. Stat., vol. 11, pp.427-444, 1940.

[16] Feinberg, S. E., "An Iterative Procedure for Estimation in Contingency Tables," Annals Math. Stat., vol. 41, 1970, pp. 907-1017.

[17] Neter, J., Maynes, E. S., and Ramanathan, R., "The Effect of Mismatching on the Measurement of Response Errors," J. Amer. Stat. Assn., vol. 60, 1975, pp. 1005-1027.

AN APPLICATION OF A THEORY FOR RECORD LINKAGE

Richard W. Coulter, Department of Agriculture

I. INTRODUCTION

As part of the effort by the Statistical Reporting Service to build a master list sampling frame of farms in each State, a record linkage system is being developed for use in detecting duplication in a list. To build this master, lists from several sources are combined and duplication, both between and within the lists, is removed. In selecting a linkage technique, an important consideration was the paucity of identifying data on most records. The table below illustrates the information available for one fairly typical State.

As the table indicates, only given name, surname, and place name are guaranteed to be present. Address information for the rural population is scarce and most often is only a rural route number. The presence of identifier numbers is rare. It is estimated that in making comparisons, nearly 60 percent of the comparison pairs will have no information in addition to given name, surname, place name, and possibly route number. In an attempt to best use this limited information in linkage, a probability model is used which incorporates some of the concepts developed by Ivan Fellegi and Alan Sunter [1]. A number of modifications and extensions have been made to portions of the original theory. (See [3].) Some of these will be examined in the following. Prior to this some background information on the model is necessary.

Let L_A be the set of records, $\alpha(a)$, pertaining to the population A, with elements $a_i \varepsilon A$, under consideration.

Define $M = \{(a_i, a_j);\ a_i = a_j,\ i < j\}$

$U = \{(a_i, a_j);\ a_i \neq a_j,\ i < j\}$

as the matched and unmatched sets, respectively. Denote by $\gamma = (\gamma^k)$ the coded result of the comparison of the variables in the comparison pair $\left[\alpha(a_i),\ \alpha(a_j)\right]$ where the result of the comparison on the k^{th} component is denoted by γ^k.

The comparison space can be defined as the set of all realizations of γ generated as a result of the comparison of records associated with members of M or U. Two probabilities are estimated for each γ^k.

1. $m(\gamma^k) = P\{\gamma^k\left[\alpha(a_i), \alpha(a_j)\right];\ (a_i, a_j)\ \varepsilon\ M\}$

2. $u(\gamma^k) = P\{\gamma^k\left[\alpha(a_i), \alpha(a_j)\right];\ (a_i, a_j)\ \varepsilon\ U\}$

A component weight for each γ^k is defined by:

$$w(\gamma^k) = \log_{10}\left[m(\gamma^k)\ /\ u(\gamma^k)\right].$$

The component weights for those variables compared are then summed to yield a total weight, $w(\gamma)$, for each comparison pair.

Two threshold values are calculated to which the total weight is compared. If the total weight is less than the lower threshold, then the pair is classified as a nonlink. If the total weight is larger than the upper threshold, then the pair is classified as a link. Pairs with total weight between the two thresholds are classified as possible links.

As an illustration of this general technique, the specific calculations for surname - surname code will be examined. In addition, the manner in which several other variables are used will be briefly described. Since the same general technique is used for these, the specific

Table A.--Availability of Identifying Data

Variable	% Presence in File
Prefix	3 (82% of these are 'MR')
Given Name	100 (24% of these are an initial only)
Middle Name	52 (90% of these are an initial only)
Surname	100
Rural Route	76 (43% of these are 'RT 1')
Box Number	43
House Number	5
Street Name	8
Place Name	100
Social Security Number	0
Employer Identification Number	2
Telephone	4

computations (some of which are rather lengthy) will not be given at this time.

II. USE OF SURNAME - SURNAME CODE AS A MATCHING VARIABLE

Surname and surname code are used as a joint variable in the linkage model. (See [7].) When surnames agree, the appropriate weight is assigned and surname code is not considered. However, when surnames disagree, then surname codes are compared. Depending upon this outcome, the appropriate weight is assigned. Under the present blocking scheme, surname codes must agree and, thus, the weight assigned when surnames disagree will always be the weight for agreement on the particular surname code. The manner in which weights are calculated for this variable is described below.

A. Notation

Let, $X = \{x_j,\ j = 1,2,\ldots,n\}$ represent the set of all possible realizations of surnames in the file;

$Y = \{y_k,\ k = 1,2,\ldots,n'\}$ represent the set of all possible realizations of surname codes on the file;

$Y' = \{y_d,\ d = 1.2,\ldots,n''\}$ represent the subset of Y that consists of surname codes associated with more than one surname;

$f_{x_1}, f_{x_2}, \ldots, f_{x_n}$ denote the frequencies of the surname realizations;

$$\overset{n}{\Sigma} f_{x_j} = N$$

$f_{y_1}, f_{y_2}, \ldots, f_{y_n}$ denote the frequencies of the surname realizations;

$$\overset{n'}{\Sigma} f_{y_k} = N, \quad \overset{n''}{\Sigma} f_{y_d} = N'$$

e = P (surname in error in the file of records associated with the matched set);

e_T = P (error-free forms of the surnames in a pair associated with the matched set are different);

g_1 = P (a surname in error in a pair associated with the matched set receives the same code as the correct surname);

g_2 = P (a valid change in surname occurs in matched records and both receive the same surname code);

$m(\gamma_h) = P(\gamma_h \mid$ the pair represents records from M), h = 1,2,3; and

$u(\gamma_h) = P(\gamma_h \mid$ the pair represents records from U), h = 1,2,3;

where, γ_1 denotes agreement on surname,

γ_2 denotes agreement on surname code and disagreement on surname, and

γ_3 denotes disagreement on both surname and surname code.

B. Assumptions

1. The distribution of matching surnames (surname codes) in the matched set is the same as the distribution in the file.
2. The distribution of surnames (surname codes) in the unmatched set is the same as the distribution in the file.
3. The g_1 and g_2 probabilities are independent of surname code.

C. Calculations (for surname x_j and surname code y_d)

$$m\left[\gamma_1\ (x_j)\right] = (f_{x_j}/N)\ (1 - e)^2\ (1 - e_T)$$

$$u\left[\gamma_1\ (x_j)\right] = (f_{x_j}/N)^2$$

$$m\left[\gamma_2\ (y_d)\right] = (f_{y_d}/N')\ \Big[2g_1\ e(1 - e)(1 - e_T) + g_1^2\ e^2(1 - e_T) + g_2\ (1 - e)^2 \cdot e_T + 2g_1\ g_2\ e(1 - e)\ e_T + g_1^2\ g_2\ e^2\ e_T\Big]$$

$u\left[\gamma_2\ (y_d)\right]$ = u(agree on sn code) $\cdot$ u(disagree on sn | agree on sn code)
= u(agree on sn code) $\cdot$ [1 - u (agree on sn | agree on sn code)]

$$= (1/N^2)\left[f_{y_d}^{\ 2} - \sum_{j=1}^{n''_d} f_{x_j}^{\ 2}\right],$$

where n''_d = the number of surnames with surname code y_d

$$m(\gamma_3) = 2(1 - g_1)\ e(1 - e)(1 - e_T) + (1 - g_1^{\ 2})\ e^2(1 - e_T) + (1 - g_2)(1 - e)^2\ e_T + 2(1 - g_1g_2)e\ (1 - e)e_T + (1 - g_1^{\ 2}g_2)\ e^2e_T$$

$$u(\gamma_3) = 1 - \sum_{k=1}^{n'} (f_{y_k}/N)^2$$

$$\text{weight} = w(\gamma_h) = \log_{10}\left[m(\gamma_h)/u(\gamma_h)\right], \quad h = 1,2,3$$

Under the present blocking scheme, surname code is used as the first blocking factor and, thus, γ_3 does not occur; i.e., $m(\gamma_3)$ and $u(\gamma_3)$ are both zero. To fit the supplied probabilities to the actual situation, the probabilities for both m and u should be redistributed over γ_1 and γ_2.

For h = 1,2 the revised probability functions would be:

$$\begin{aligned} m(\gamma_h)' &= m(\gamma_h \mid \gamma_3 \text{ does not occur}) \\ &= m(\gamma_h) / \left[1 - m(\gamma_3)\right] \\ u(\gamma_h)' &= u(\gamma_h \mid \gamma_3 \text{ does not occur}) \\ &= u(\gamma_h) / \left[1 - u(\gamma_3)\right]. \end{aligned}$$

Since most of the probability for the unmatched set will be concentrated in γ_3, the net effect of this redistribution would be a significant reduction in the derived weights for exact matches on surname and surname code. For this reason, we have chosen to ignore this effect of blocking for weight calculation purposes. For example, in a test file of 150,000 records, a surname which occurs 1,000 times receives a weight for agreement of 2.16. The revised weight using the redistributed probabilities would be -.51.

The weight for γ_1 depends primarily on the frequency of the particular surname, with the more rare surnames receiving the larger weights. The weight for γ_2 depends on the frequency of the surname code, on the size of the error rates e and e_T and on the number of distinct surnames within that codes. Infrequent surname codes, large error rates and few different surnames all tend to make the weight for this condition large.

III. OTHER VARIABLES

Modifications have been made to other variables in an attempt to improve the linkage results. These will be outlined below.

A. Given Name - First Name

As part of the processing prior to linkage, each given name on the file is assigned a formal or first name. (See [8].) A dictionary of the most common given name is utilized for this purpose. For given names not in the dictionary, the given name will also serve as the first name. Common examples of given - first names are: Bill=William, Dick=Richard, Jack=John.

First name is used in the model in a manner similar to surname code. If given names agree, then first names are not compared. However, if given names disagree, then first names may either agree or disagree. Weight calculation routines have been developed for the three possible conditions using the same general technique as discussed for surname - surname code. An additional factor which has to be considered for this variable is that one name may be an initial, while the other may be a complete name. In this case, the initial is compared against the first letter of both the given and first names of the complete name. The probability of this occurring is estimated using frequencies of initials on the file and weights for the various outcomes are also calculated.

B. Place Name

A place name dictionary for each State is utilized to standardize all spellings and abbreviations of place names and to assign a latitude - longitude location to each. (See [11].) The standardization eliminates disagreement due to different spellings of place names. The location of each is, then, used to compute the distance between two places, in a comparison when the place names are different. This distance is classified into one of seven intervals, and a different weight is calculated for each interval. The intervals are:

1.	0 to 1 miles
2.	1 to 10 miles
3.	10 to 25 miles
4.	25 to 50 miles
5.	50 to 100 miles
6.	100 to 200 miles
7.	over 200 miles.

The m and u probabilities and subsequent weights for the agreement condition on place names are calculated in the same manner as is done for surname. The weight computation for place name disagreement is outlined below.

1. The m values are based on counts for each interval of matched pairs with place name disagreement taken from a sample. These are then fitted, using least squares estimates to a monotonically decreasing function of the form $y = ae^{bd}$. The fitted values form the distribution for m.

2. The u values are estimated from the file. Every pair of distinct place names is compared, their distance apart calculated, and the product of their relative frequencies summed in the appropriate interval. This yields the probability of getting place name disagreement in a particular interval by chance; i.e.,

   u(disagreement in Ith interval) = $2\,\Sigma(f_x/N)\,(f_y/N)$, where f_x, f_y are frequencies of place names whose distance apart is in interval I; and N = total number of records on file.

In practice, the further away two place names are, the larger their disagreement weight becomes.

C. Box Number and House Number

Disagreement weights for these variables are based on the amount of disagreement present. This is measured by comparing these on a character-by-character basis. (See [13].) Box and house number are up to five characters long and, thus, there are 15 different combinations of number of agreements - number of disagreements when the variable is present in both records and not identical. Different m and u probabilities and weights are calculated for each of these conditions. The key to the calculations is to estimate the appropriate probabilities for one character, given that data are present, and, then, to make the assumption that the probability of misreported data is independent of the particular character and is equal for each of them. In general, the more disagreement present, the larger the disagreement weight will be.

D. Social Security Number and Other Identifiers

Weights for identifier numbers, such as SSN, are also partitioned. Only one agreement weight is calculated for these. SSN, for example, is broken into four partitions which are assumed to be independent. (See [16].) The m and u values are calculated for one partition and independence assumptions allow these to be extrapolated to the entire number. For SSN, sixteen different weights are calculated for conditions ranging from complete agreement to complete disagreement.

See the following papers for additional information on identifier comparisons: [9] for derivation of the middle name comparison; [10] for a derivation of the negative weight to be used when one record has "Jr." and the other has no suffix; and [12] for a discussion of the additional negative weight when more than one address variable disagrees.

IV. ERROR RATES AND THRESHOLDS

Implicit in the use of the model is the assumption that the two error rates -- probability of a recording error and probability of a valid change for records associated with the matched set -- are known or can be estimated for each variable prior to processing the file through the linkage system. In the absence of prior knowledge, the current system is designed to process a sample of blocks through linkage in order to estimate these errors. (See [4] and [17].) Initial estimates are provided and the linkage decisions for the sample are manually reviewed and questionable decisions are resolved. Once this is completed, counts of error conditions are kept by variable for those pairs which are links. These are then used to estimate the necessary error rates.

To aid in this process, counts are maintained within the software for those pairs originally classified as definite links. As decisions are changed, based upon the review, these counts are updated. The importance of these estimates is demonstrated by the graph in Figure 1, which gives the frequency distribution of total comparison weights for three sets of error rates, where the rates were varied for four of the variables. As the graph indicates, the major effect of an increase in error rates (decrease in quality) is to shift the frequency curve to the right, particularly at the lower end of the scale, resulting in an increase in the number of pairs classified as possible links (weight between 5.0 and 7.5). That is, the model is unable to classify as many pairs as definite nonlinks. Pairs with small total weights are most affected, since it is in these pairs that there is the most disagreement in components, and the error rates affect most the weights assigned to the disagreement condition.

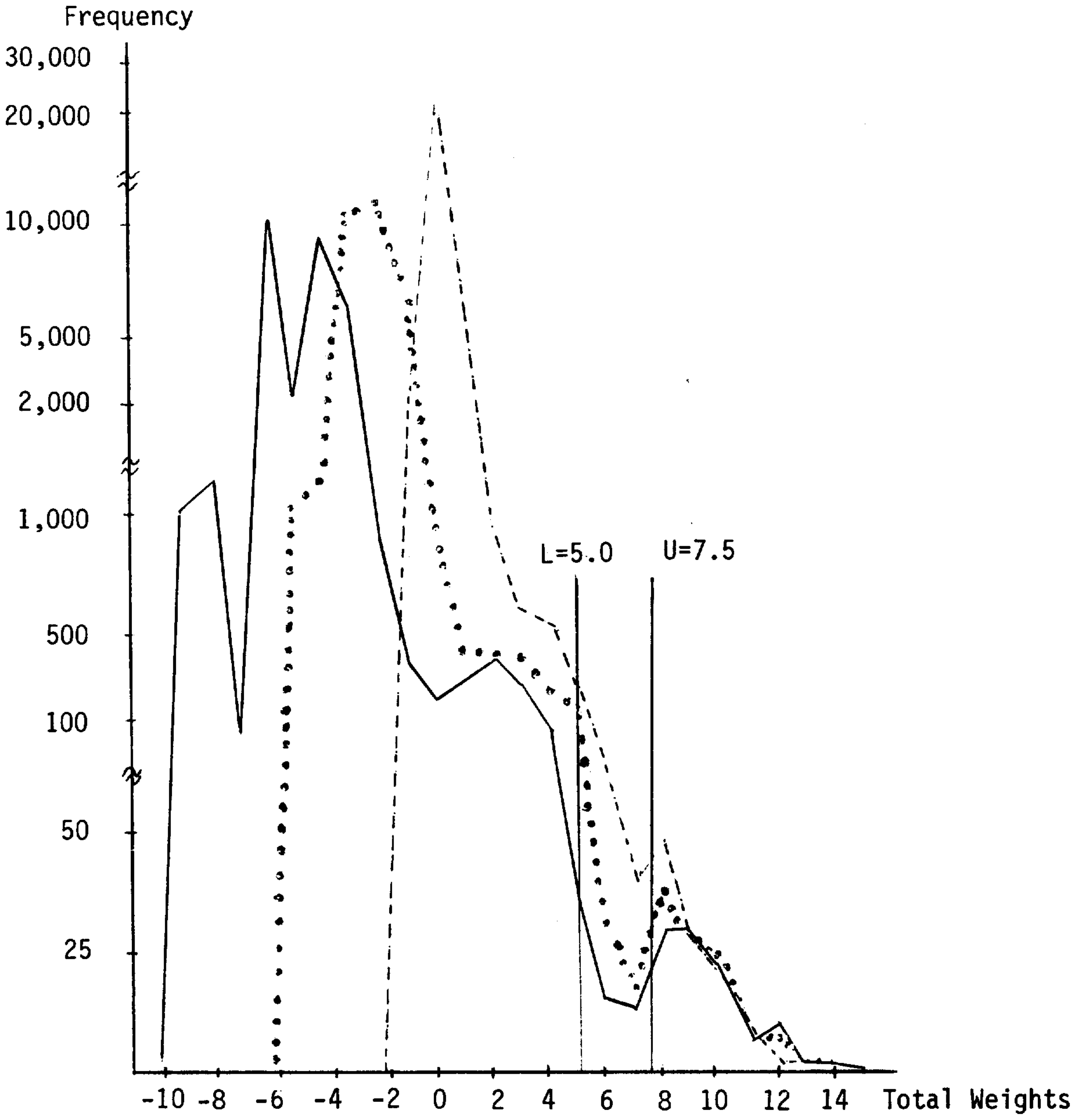
The final parameters to be supplied are the threshold values. It is these two values which ultimately determine the classification of each pair. Fellegi and Sunter suggest a technique of estimating these by sampling from the tails of the m and u probability distributions for the comparison pairs. In practice, a technique of initially estimating these -- based on a combination of weights for selected components-- and revising, as necessary -- as a result of the review of the sample used to estimate error rates -- has proven to be more satisfactory. The initial estimate of the lower threshold is made by summing the agreement weights for the most common given name, surname, and place name. This has proven to be an excellent "first guess." Another tool which can be useful in setting thresholds is the distribution of total weights. This distribution for one sample of 2,200 records is given in Figure 2. The thresholds could expect to be most efficiently set at points on either side of the lowest point on the u-shape portion of the curve (about a total weight of six in the example). The percentage of pairs classified as links after the manual resolution is also indicated for each interval in this example. Specifying the allowable rates of misclassification would, then, also determine where the thresholds will be set.

V. REMARKS

Research and analysis of results is continuing in order to further improve the procedure. For example, the possibility of using a coding procedure for given name is now being investigated. Also, questions concerning the stability of the error rates across States and, more generally the amount of preprocessing of a sample that is necessary are being investigated. The amount of manual review that is necessary after the automated procedure is also a concern. The limited amount of identifying data that is present on the lists necessitates using each item to the fullest extent possible, but it also implies that a manual review of, at least, some decisions will always be necessary.

Figure 1.--Total Weights by Frequency for Three Sets of Error Rates

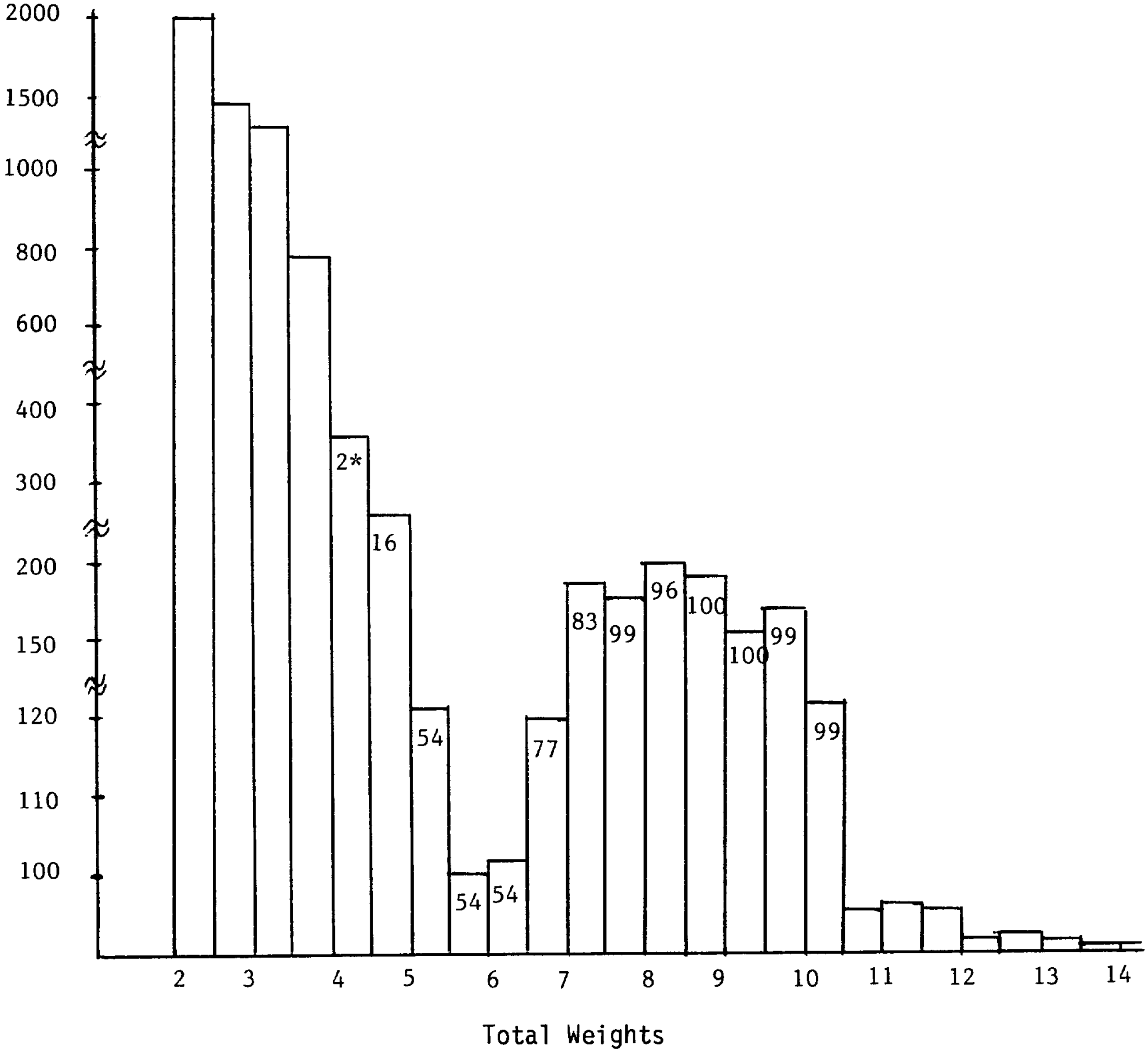
(Approximately 39,000 comparisons)



Key for Figure 1

Variable	Recording Error			Change Error		
	——	· · · ·	- - - -	——	· · · ·	- - - -
Given Name	.001	.01	.1	.001	.01	.1
Middle Name	.001	.01	.1	.001	.01	.1
Surname	.001	.01	.1	.001	.01	.1
Place Name	0	0	0	.001	.01	.1

Figure 2.--South Carolina Sample - Weight Distribution



*Numbers in each bar indicate the percentage of resolved pairs in that interval that were links.

The computed thresholds used prior to any resolution were 4.5 and 8.3.

NOTES AND REFERENCES

[1] Fellegi, Ivan P. and Sunter, Alan B. (1969) "A Theory for Record Linkage," Journal of the American Statistical Association, vol. 64, no. 328, pp. 1183-1210. (Also reprinted in this volume.)

Editors' Note:

This report is part of a series of Working Papers documenting the development of a record linkage system by the Statistical Reporting Service (SRS) of the U.S. Department of Agriculture (USDA). The collection represents various stages in the research and modification of matching theory to construct a master list sampling frame of farm operators by State. The work was begun under the direction of Max Arellano and later refined by Richard Coulter and others.

Thanks to the help of Nancy Kirkendall, we have added annotated references to this paper to tie it in with related reports prepared as part of the same series. With the exception of [6], none of the papers have been previously published, and they are only available in draft form from:

Henry Power
Statistical Reporting Service
U.S. Department of Agriculture
S. Agriculture Bldg., Room 5864
Washington, DC 20250.

It is the hope of the editors that interest generated by this Workshop will lead to the eventual publication of this valuable set of papers.

[2] Arellano, Max G. (1976) "Application of the Fellegi-Sunter Record Linkage Model to Agricultural List Files," SRS, USDA.

[3] Arellano, Max G. (1976) "The Development of a Linkage Rule for Unduplicating Agricultural List Files," SRS, USDA. This paper describes the differences between the USDA assumptions and the Fellegi-Sunter assumptions as applied to probabilistic matching. Major differences are in the definition of the error rates and the assumptions concerning errors in the files used to derive agreement weights. (6 pages)

[4] Arellano, Max G. (1976) "The Estimation of P(M)," SRS, USDA.

[5] Coulter, Richard W. and Mergerson, James W. (1977) "An Application of a Record Linkage Theory in Constructing a List Sampling Frame," SRS, USDA. From the Coulter paper reprinted here, one might think that the SRS record linkage system is strictly an application of the probabilistic matching procedures. In [5], Coulter and Mergerson describe the SRS system in more detail than is found in any of the other papers. This latter paper describes preprocessing and variable identification procedures; it, then, discusses the method used to classify records as being partnership, corporate or individual records. The partnership and corporate record linkages are handled manually. Only the individual records are processed through the probabilistic linkage. The overall system adjusts for some of the matches missed because of blocking on surname by identifying for manual review all of the record pairs which agree exactly on address. This paper gives a nice overview of the entire system. (29 pages)

[6] Lynch, Billy T. and Arends, William L. (1977) "Selection of a Surname Coding Procedure for the SRS Record Linkage System," SRS, USDA. This is the only paper in the series which was published by SRS. In it, Lynch and Arends describe the analysis of surname coding systems performed by USDA. These efforts led to the selection of a revised NYSIIS (New York State Identification and Intelligence System) coding system as the most appropriate system for SRS purposes. (31 pages)

[7] Arellano, Max G. and Coulter, Richard W. (1976) "Weight Calculation for the Surname Comparison," SRS, USDA. This paper provides the mathematical derivation for the weights used for the comparison of surname, including surname code. It details the assumptions and the error terms needed in the implementation. (6 pages)

[8] Arellano, Max G. and Coulter, Richard W. (1976) "Weight Calculation for the Given Name Comparison," SRS, USDA. This paper provides the mathematical derivation for the weights used for the comparison of given names. It recognizes nicknames and initials. As in [7], it details the assumptions. (9 pages)

[9] Arellano, Max G. and Coulter, Richard W. (1976) "Weight Calculation for the Middle Name Comparison," SRS, USDA. This paper provides the mathematical derivation for the weights used for the comparison of middle names. It also accounts for agreement on middle initial. As in [7], it details assumptions. (5 pages)

[10] Coulter, Richard W. (1976) "A Weight for 'Junior' vs. Missing," SRS, USDA. This paper derives the disagreement weight for the case when one record includes "Jr." and the other record does not. (4 pages)

[11] Arellano, Max G. (1976) "Weight Calculation for the Place Name Comparison," SRS,

USDA. This paper provides the mathematical detail for the comparison of place names. Disagreement weights for the place name comparison are based on how far apart the two different places are (as calculated by using the latitude and longitude for each place). This paper also details assumptions. (5 pages)

[12] Coulter, Richard W. (1976) "Processing of Comparison Pairs in Which Place Names Disagree," SRS, USDA. This paper compares addresses and their components -- street name, street number, etc. Since these variables are probably not independent, the paper derives an additional negative weight for use when there is a disagreement on more than one address variable. (4 pages)

[13] Arellano, Max G. (1976) "Calculation of Weights for Partitioned Variable Comparisons," SRS, USDA. This paper describes the calculation of agreement weights when variables are to be compared by splitting them into different partitions and comparing the pieces -- for example, if two 3-digit numbers were compared by examining one digit at a time. (This is how house number and box number are compared.) (10 pages)

[14] "Partitioned Variable Comparison/Algorithm for Identifying Configurations," SRS, USDA. This paper translates three outcome comparison configurations on n variables to integers in the interval from 0 to 2**(n+1)-2 for purposes of indexing. (1 page)

[15] Nelson, D.O. (1976) "On the Solution of a Polynomial Arising During the Computation of Weights for Record Linkage Purposes," SRS, USDA. The procedure described in [13] for determing weights for partitioned variables needs a root of a polynomial. This paper shows that a root in the appropriate range exists and that it can be evaluated numerically. (2 pages)

[16] Arellano, Max G. (1976) "Optimum Utilization of the Social Security Number for Matching Purposes," SRS, USDA. This paper presents the derivation of weights to be used in the comparison of social security numbers. The social security number is partitioned into four pieces (of length 2,2,2, and 3) for purposes of comparison. For more on this technique, see also [13]. (10 pages)

[17] Arellano, Max G. and Arends, William L. (1976) "The Estimation of Component Error Probabilities for Record Linkage Purposes," SRS, USDA. This paper describes the estimation of error rates used in calculating most of the agreement and disagreement weights for individual variable comparisons. There are three types of errors recognized in the USDA system: errors resulting from the erroneous reporting or recording of a value, errors which are a result of a valid change in the value of a variable, and missing values. (14 pages)

[18] Coulter, Richard W. (1975) "Sampling Size in Estimating Component Error Probabilities," SRS, USDA. This paper describes the determination of the sample size required to estimate the error rates described in [17]. It also refers to [4]. (12 pages)

A Generalized Iterative Record Linkage Computer System for Use in Medical Follow-up Studies*

G. R. HOWE

NCIC Epidemiology Unit, Faculty of Medicine, McMurrich Building, University of Toronto, Toronto, Ontario M5S 1A5, Canada

AND

J. LINDSAY

Vital Statistics and Disease Registries Section, Health Division, Statistics Canada, Ottawa, Ontario K1A OT6, Canada

The development of a generalized iterative record linkage system for use in follow-up of cohorts in epidemiologic studies is described. The availability of this system makes such large-scale studies feasible and economical. The methodology for linking records is described as well as the different modules of the computer system developed to apply the methodology. Two applications of record linkage using the generalized system are discussed together with some considerations regarding strategies for conducting linkages efficiently.

The primary focus of epidemiologic studies of chronic disease is the determination of factors which may be associated with increased risk of such diseases. Two classic approaches to identifying such factors are the case-control and cohort studies (*1*).

In a cohort or follow-up study one starts with a group of individuals some or all of whom may have been exposed to the factor under study, and ascertains their subsequent morbidity or mortality experience. In order to accumulate sufficient person-years of experience to provide a sufficiently powerful statistical test of any association between exposure and disease, it may be necessary to follow large groups of individuals for many years, and this is particularly true if the excess risk in question is a small one. However, even in the latter case it is possible that if exposure to the factor is widespread, the population attributable risk can be substantial and consequently the factor can be a significant health hazard. Conventional methods for following cohorts include personal contact, telephone, and mail inquiries (*1*) and when the cohort is large such methods can be prohibitively difficult, expensive, and time consuming.

*Reprinted with permission from Computers and Biomedical Research 14, Copyright © 1981 by Academic Press, Inc., pp. 327-340.

An alternative method for following cohorts is the use of computerized record linkage in which records of individual members of a cohort are compared with records from files of morbidity and mortality data (*2–4*). When a unique identification number (such, for example, as the Canadian Social Insurance Number or the U.S. Social Security Number) is present on both the exposure records and the morbidity or mortality records, such linkages simply involve sorting both files using the unique identifier as key and then directly matching records from the two files. However, such unique identifiers rarely exist, especially on data which have been assembled retrospectively. In this case, it is necessary to use identifying characteristics such as surname, given name, date of birth, etc. in order to link records from the two files, and this involves two practical problems. In the first place, such identifying items are not unique to a particular individual and even combinations of identifying items may not be unique; and in addition, identifying items may be misrecorded or missing on certain records. It is therefore necessary to devise algorithms for comparing the two records in order to produce some quantitative measure which is a function of the probability that those two records do indeed refer to the same individual. Secondly, given such algorithms, it is necessary to devise a computer system in order to efficiently carry out the data processing involved.

Considerable attention has been paid to the first of these two problems and the methods most widely used are those which have been developed by Newcombe and his associates (*5*) and Fellegi and Sunter (*6*). However, the implementation of these methods in terms of computer programs has generally been done on an ad hoc basis for each specific application. This paper describes some extensions of the Newcombe methodology, in particular to cope with the problem of partial agreement of identifying items, and also a generalized computer system which has been developed in order to carry out linkages between any two files of interest. The system may also be used to internally link records from a single file, where one individual may have more than one record, but again no unique identifier exists. The application of the system to two studies in cancer epidemiology is also described.

METHODOLOGY

A. Basic Principles

Conceptually carrying out a record linkage between two files A and B involves the following steps:

Step 1. Every record on file A is compared with every record on file B. The result of each comparison is a series of outcomes, one outcome resulting from each identifying item being used for linkage such as surname, first given name, year of birth, etc. An outcome may be defined as specifically as desired; for example, the two records agree on the first five characters of the surname and the value is SMITH, or the first given name agrees on first character irrespective of value, but remaining characters disagree.

Step 2. A statistic called the total weight (W^*) is calculated for the comparison of any two particular records. The weight is an estimate of the odds that the two records under consideration do in fact refer to the same individual, i.e., that they are linked (L) as opposed to referring to different individuals, i.e., they are not linked ($\bar{L}$).

Thus the weight is an estimate of:

$$\frac{P(L/{}_1O_2O_3O\ .\ .\ .)}{P(\bar{L}/{}_1O_2O_3O\ .\ .\ .)}, \qquad [1]$$

where $P(L/{}_1O_2O_3O\ .\ .\ .)$ is the probability that the two records are linked conditional that the outcome from comparing the first identifying item is ${}_1O$, etc. If one assumes that the values of the identifying items on the records are statistically independent then it follows that:

$$W^* = {}_1w + {}_2w + {}_3w\ .\ .\ . + \log_2 \frac{P(L)}{P(\bar{L})}, \qquad [2]$$

where ${}_1w$ is $\log_2$ of the estimate of the odds of obtaining outcome ${}_1O$ conditional upon the two records being linked. It is convenient as is customary in information theory to use $\log_2$ in Eq. [2] in order to make the equation additive.

In practice the final term in Eq. [2] is usually impossible to evaluate since it requires a priori knowledge of the number of links among the set of all comparisons and this is usually unknown. Thus a modified total weight may be defined as:

$$W = {}_1w + {}_2w + {}_3w\ .\ .\ .\ . \qquad [3]$$

If W can be estimated from Eq. [3] for all possible comparisons between the records on the two files and these comparisons are then ordered by the value of W, they represent potential links in decreasing order of believability, and, in particular, the difference $W1 - W2$ for any two particular comparisons is an estimate of $\log_2$ of the odds ratio. Thus, if two comparisons result in W's which differ by 1.0 the odds in favor of the first comparison being a true link are twice the odds for the second comparison being a true link. Details of weight calculations including the case of partial agreements are given below.

Step 3. Having ordered the comparisons by W, upper and lower threshold values are chosen. These are used to divide the set of all comparisons into three; namely, the "definite links"—those with a weight above the upper threshold, the "nonlinks"—those below the lower threshold, and the "possible links"—those between the thresholds. The possible links may be manually inspected and if possible resolved. If further identifying information is available which is not in machine-readable form, this may be used to supplement the data for the possible links in order to resolve them. If no such data are available, manual resolution is probably undesirable and one possible approach is to choose a single threshold value (*2*). Fellegi and Sunter (*6*) have developed a likelihood ratio test based upon the total weight statistic which leads to optimum values of the upper and lower thresholds. Alternatively, and frequently more conveniently, their values may be empirically assigned from inspection of the set of potential links.

B. Blocking

In order to compute W it is therefore only necessary to estimate ${}_1w$, ${}_2w$, ${}_3w$, etc. for each identifying item, for each possible outcome from comparing the possible values of that item. There is, however, a further practical consideration. When dealing with files of any appreciable size the total number of possible comparisons between records becomes extremely large and resulting computer costs are inordinate. It is therefore necessary to block the files using a combination of identifying items or derivatives of identifying items to define the blocks. Comparisons are then only carried out between records in corresponding blocks on the two files. The block identifier used in the applications described in the last section of this paper, for example, was the combination of sex and the NYSIIS code of surname (*7*). The NYSIIS code is an alphabetic code designed so that surnames of similar sound have the same code and frequently encountered errors of misreporting do not result in change in the NYSIIS code. Thus this blocking system will generally bring together records belonging to a single individual even when errors of recording have occurred. The effect of blocking on the calculation of weights is taken into account in the general formulation given below.

C. Derivation of Formulas for Weights

The w's of Eq. [3] may now be computed from simple probability theory. The general formulation proposed leads to slight modifications of the original formulas of Newcombe and Fellegi and Sunter as discussed subsequently.

It is convenient for illustrative purposes to consider a specific identifying item; the most useful one in the present context is surname since this involves a consideration of the blocking factor, namely, the NYSIIS code. Although the number and types of outcome in comparing the surnames from two records is arbitrary, we have found it most convenient to consider five possible types of outcome defined as follows. The subscript used to identify the particular identifying item is omitted from these formulas. (For outcomes 1 to 4 surname is assumed to be present on both records.)

(1) $O_{1=i}$: Surname agrees on first seven characters with value i.

(2) $O_{2=j}$: Surname agrees on first four characters with value j, but disagrees within next three characters.

(3) $O_{3=k}$: Surname agrees on NYSIIS code with value k, but disagrees within the first four characters.

(4) O_4: Surname disagrees on NYSIIS code.

(5) O_5: Surname is missing on one or both records.

The weight corresponding to O_5 is obviously zero unless the linked and unlinked set of records have different frequencies for the reporting or nonreporting of identifying items. If an estimate can be made of any differential reporting for the two sets, w_5 may be computed correctly from its definition. No further consideration need be given to missing data, as all probabilities and frequencies are assumed to be conditional upon a value for the identifying item in question being present.

In order to compute w_1 to w_4 it is necessary to specify the frequency with which surname is misreported. These frequencies, referred to as transmission rates, are defined as follows:

t_1: The probability that the surname on a particular record has the same first seven characters as the "true" value.

t_2: The probability that the surname has at least the first same four characters as its "true" value.

t_3: The probability that the surname has the same NYSIIS code as its "true" value.

By this definition there is a single set of transmission coefficients, t_1 to t_3, for each identifying item. It should be noted that the transmission coefficients correspond to the various possible outcomes listed above in the sense that if both records in a particular comparison are transmitted from the "true" value to the recorded value so that the first seven characters remain the same the outcome will be O_1 and the probability of such a transmission is t_1 for each record. It should also be noted that various components can contribute to the transmission coefficients, such as a genuine change in the "true" value of surname between the creation of the two records, errors of recording, etc. If such components can be identified and numerical values estimated, these values can be used to compute the transmission coefficients. The approach we have used is to compute the transmission coefficients in an iterative fashion from the records themselves as described subsequently.

In order to calculate the weights corresponding to each possible outcome the basic definition is used. For example, the probability of exact agreement on the first seven specific characters of a certain surname when the two records originate from the same individual is given by

$$t_1^2 f_i,$$

where f_i is the relative frequency of occurrence of the particular seven-character value among the individuals who give rise to the linked set. In order to estimate such frequencies it is usually necessary to use the frequencies as observed on the records in the files themselves. This involves a decision as to whether the frequencies on the linked set are most similar to the frequencies on file A or file B, and this obviously depends on the particular data sets under consideration and involves essentially an empirical decision. Given the particular file to be used for estimating the frequencies there are two possible models. In the first, it is assumed that errors in recording are such that the original "true" value is transmitted to some value that does not already exist within the linked set. This leads to the observed frequency value within the file being set equal to $t_1^2 f_i$, which is the formulation proposed by Fellegi and Sunter. Alternatively it may be assumed that when a recording error is made it results in some value which already exists within the linked set. If this process happens randomly the observed frequency within the file will be equal to f_i. We have used the second model since we feel it to be more realistic and since it leads to a formulation in which transmission and frequency components of the weights are separable and the weight for any particular outcome can be factorized into these two components.

The probability for any outcome with the unlinked set of comparisons is most simply determined from consideration of frequencies as they occur on the files. Thus the probability of agreement by chance on the first seven characters of surname in the unlinked set is given by:

$$ {}_A f_i \ {}_B f_i , $$

where ${}_A f_i$ and ${}_B f_i$ refer to the relative frequencies on files A and B, respectively. (The contribution to all possible comparisons from the linked set is negligibly small and is therefore ignored in this formulation.) Using this approach the weights for 1–4 above can be shown to be:

$$ w_{1=i} = \log_2 t_1^2 + \log_2 \frac{1}{{}_B f_i}, \qquad [4] $$

$$ w_{2=j} = \log_2(t_2^2 - t_1^2) + \log_2 \left[\frac{{}_A g_j}{{}_A g_j \ {}_B g_j - \sum_{i \epsilon j} {}_A f_i \ {}_B f_i} \right], \qquad [5] $$

$$ w_{3=k} = \log_2(t_3^2 - t_2^2) + \log_2 \left[\frac{{}_A h_k}{{}_A h_k \ {}_B h_k - \sum_{j \epsilon k} {}_A g_j \ {}_B g_j} \right], \qquad [6] $$

$$ w_4 = \log_2(O), \qquad [7] $$

where ${}_B f_i$ is as before; ${}_A g_j$ is the relative frequency of first four characters of surname equal to j, and ${}_A h_k$ is the relative frequency of NYSIIS code equal to k (for file A). Equation [7] is applicable only to the item used as a pocket identifier.

These formulas apply when the frequency distributions in the linked set are taken as being the same as those on file A.

In all the above expressions it will be seen that the transmission and frequency components of the weight are separable and their $\log_2$s are additive. It should be noted that the value for w_4 means that no two records from different blocks can link. In order to estimate the various values of t, we have used an iterative procedure as follows. The linkage is carried out using estimates for t, usually based on previous experience. Given an estimate of the upper threshold value, a sample of links may be drawn from the linked set and estimates made of the transmission coefficients from the number of times that full or partial agreements on surname occur within the linked set. These new values may then be used as the basis for another linkage and the process repeated iteratively until reasonably stable values for the transmission coefficients are obtained. Alternatively, as previously mentioned, the transmission coefficients may be estimated empirically.

System Design

The particular series of programs, which were written in order to apply the above methodological principles to specific data sets, relies heavily upon use of a data base system (Relational Access Processor for Integrated Data Bases (RAPID)) which is available within the facility where the programs were developed (Statistics Canada). The programs as such, therefore, are of no direct use in any other environment, but the principles of the system involved are readily generalizable to any other computer environment, and may be programmed within the particular limitations of the hardware/software available.

The system has been deliberately designed to be modular in nature. In particular, the most time-consuming element, namely, the comparison of all records within each block, was developed as a single module. Only one pass of the complete data is necessary, which will eliminate any comparisons which result in any obvious nonlinks and will produce a file of potential links with their corresponding outcomes. These potential links may then be subjected to a number of different weighting runs in order to refine the linkage results at a much lower cost than would be incurred by rerunning comparisons between the entire data files. This modular approach also facilitates the iterative process of calculating transmission weights. The modules involved in the system are shown in block diagram form in Fig. 1 and their specific functions are now described.

A. Preprocessing

This step involves editing and correcting of the original data files, including such functions as creating a unique sequence number for each record and the NYSIIS code of surname, left justifying fields such as given name, removing blanks within names, recoding variables, etc. Following the editing step the files are sorted by whichever identifying item is to be used as the pocket identifier, e.g., NYSIIS code.

B. Calculation of Frequency Component of Weights

Frequency counts are carried out on the preprocessed files for all levels of agreement and partial agreement for all identifying items. From these frequency distributions are computed the frequency components of the weights as given in Eqs. [4] to [7]. In practice it will often be found that for many items the frequency distribution is similar from one file to another and consequently a

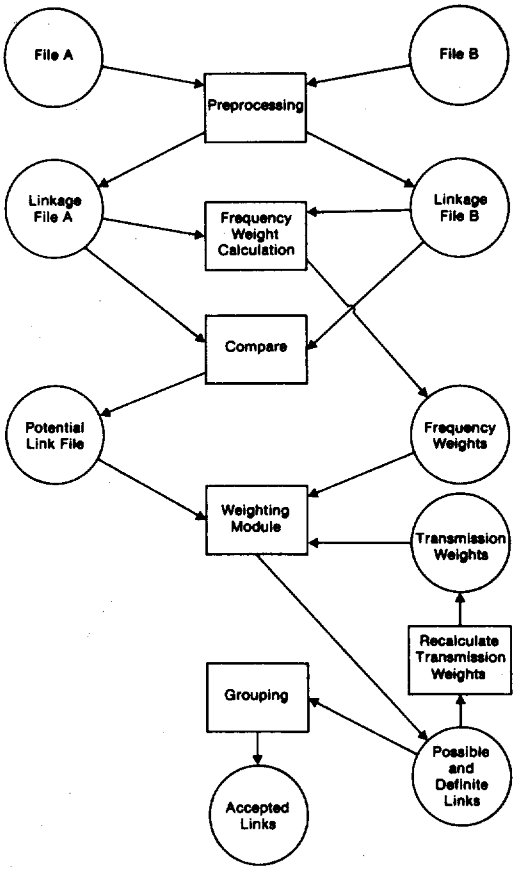


FIG. 1. Generalized iterative record linkage system.

single set of frequency weights will suffice. For other items, such as birth year, the distribution will vary considerably from file to file and may need recomputing each time.

C. Comparison Module

The function of the compare module as stated is to create a file of potential links and their corresponding outcomes and to eliminate all obvious nonlinks. In this module all records within a given pocket are compared with each other, each comparison giving rise to a series of outcomes such as, e.g., "seven character agreement on surname, and the value is Smith." Identifying items on the two records are compared in an order which is specified at execution time. This ordering is decided by two factors, the discriminating power of the identifying item and the CPU time necessary to make the comparison. An option is provided to carry a crude "running total of disagreement weights."

Each item is assigned an appropriate preliminary disagreement weight, and where a disagreement occurs, the running total is decremented by the disagreement weight for the item concerned. When the running total achieves a value below a preselected cutoff value, the comparison between the two records in question is abandoned and the module then proceeds to the next comparison. This procedure ensures that records which are in obvious disagreement are not considered as potential links. For any comparison which does not yield a value for the running weight below the critical, a "link record" is created consisting of the two record numbers and an outcome code and, where appropriate, a value for each identifying item in question. At the completion of this phase the link record file thus contains all potential links and further processing is concerned with this particular file.

D. Weighting Module

The function of this module is to add both frequency and transmission components of the weights to the link record file. Components may be added in separate passes as they are completely independent of each other as in the formulation of the previous section. The particular method used to add the weights will of course depend on the hardware configuration available. In general, the procedure will involve table lookups using the outcome code and value where appropriate as an index. Since the link records are ordered in the same sequence as the pocket identifier, the weights for the pocket identifier (e.g., NYSIIS of surname) may be added conveniently from a sequential file. For items with relatively limited numbers of values such as birth year the tables may be conveniently stored in core; for alphabetic data other than the pocket identifier, such as given name, random access disk files probably provide the most convenient means. As there are relatively few transmission coefficients these generally can be stored in core, and a weighting pass to change just the transmission coefficients can be carried out rapidly. Subsequent to applying the weights to the link record file, a sample of this can be printed out for manual inspection and this can be used to assign tested threshold values. Given these threshold values new estimates of transmission weights can be made using the set of links which are above the upper threshold. These new values can be applied to the links and the process repeated until some measure of consistency is achieved.

E. Grouping Module

The function of this module is to bring together all records which have linked with each other. The specific algorithm to be used is of course dependent upon the nature of the records concerned, and whether the linkage is two file or internal. For an internal linkage generally there is no limitation upon the number of records that can constitute a "group" corresponding to a single individual. Often in the case of two-file linkage only a one-to-one relationship is possible as for example in linking records for specific individuals to a file of death records. However, in the latter case, since some links will occur by chance, it is necessary to identify records which appear in more than one link.

For grouping records from an internal linkage we utilized the following method which involves starting with a single record, identifying all links to that record, then identifying all links to those links, and so on. We defined definite groups of records as those in which each member is linked to at least one other member of the group with a weight which is above the upper threshold (a definite link). Possible groups are then defined as being composed of a series of definite groups in which there is at least one possible link between members of the definite groups concerned. Any possible groups which are formed can then be printed out for visual inspection and a decision made as to whether the definite groups which constitute them should be amalgamated into a single group or whether the original definite groups should be maintained as individuals. The reservations concerning the utility of manual resolution when no further identifying data are available, expressed in the methodology section, should be taken into account when deciding whether to adopt such a grouping procedure.

In order to group links from a two-file linkage where only a one-to-one link is permissible, the links are sorted by weight, then proceeding from the link with the largest value downward, each link is checked to see whether either record concerned has appeared in a previous link. If either has, the link may be printed out as a conflict and the situation resolved by visual inspection. Alternatively, the link with the highest weight may be accepted.

Since processing up to this point has involved record numbers rather than the actual records themselves at this stage a number is assigned to each group or pair of records that has been linked. These group numbers may then be assigned sequentially using the record number of one of the original records, and sorting the records on this group number brings together those records which have been linked so they may thus then be processed further as desired. It should be noted that although the identifying items on any particular record which has entered into a possible link are essentially contained on the link record file, and are there available for inspection if needed, it is also desirable to provide a mechanism for accessing the original complete data records. In the system we have developed this is done by maintaining a parallel file containing those data records which have formed at least one link so that they may be accessed via the data base used.

Applications

The system described has been primarily developed for use in monitoring the morbidity and mortality experience of various groups of individuals with various exposures, by linking such exposure records to national morbidity and mortality files. Two such specific applications are now described in more detail.

Linkage of TB Patient File to Mortality File

Between 1930 and 1952 extensive use was made of collapse therapy in the treatment of tuberculosis. This involved considerable X-ray exposure from fluoroscopy machines which were extensively used for examination of the chest cavity. A major study of cancer mortality in relation to this radiation exposure is being conducted (*3*), by collecting data on individual patients from all existing hospital and sanitorium records in Canada.

The TB patient file was first internally linked using the generalized iterative linkage system described here to bring together treatment data from different institutions to form one complete treatment history per patient. The TB patient file containing 118,000 records was then linked to the national mortality file covering the years 1950 to 1977 containing 5,000,000 records. (1950 is the first year for which sufficiently well-identified mortality records are available in a format suitable for computerized record linkage.)

The identifying items used were the following: NYSIIS code and surname; first and second given names; day, month, and year of birth; place of birth; sex; NYSIIS of mother's maiden name; mother's first initial; mother's birthplace; father's first initial; and father's birthplace. Year of last contact on the TB records was compared with year of death on the mortality records in order to eliminate unnecessary comparisons. Use was made of the facility to incorporate partial agreements as follows: Surnames were considered to be in full agreement if they agreed on seven characters; the first level of partial agreement was on the first four characters and the second level of partial agreement, on NYSIIS only. Full agreement for given names was on the first four characters, and partial agreement, on initial only. Birth year was treated as being in full agreement if it was within plus or minus 1 year. The first level of partial agreement was within 5 years, and the second level, within 10.

The records were blocked by NYSIIS code of surname and sex. Alternate surname spellings and maiden names were also available. These were included as comparison items by creating duplicate records for alternate surnames at the preprocessing stage. Following the linkage, duplicate records were combined. The total file of TB patients was linked to 1 year of mortality records at a time. This provided the advantage of allowing the runs to be checked closely rather than risking costly errors over the entire linkage.

Initially, the number of potential links formed between the TB and mortality files was 787,800 for males and 554,800 for females, using a very conservative cutoff weight to ensure that no potential links were missed. The preliminary weights used were average values or approximations of the final weights. After the final weights were calculated and threshold values set, there were 82,828 possible and definite links generated by the male files and 67,490 by the female files. This was considered to be an application where only a one-to-one link was acceptable, i.e., one TB record could validly link with one death record. Following the application of the one-to-one rule, there remained 20,293 male links and 12,697 female links which were considered to be definite for the purpose of the subsequent statistical analysis.

The cost of this record linkage was just over $5000 (Canadian). This cost includes the comparison of the records, assignment of preliminary weights used to determine whether each link was a potential link, insertion of the final weights, setting of the thresholds and resulting classification of each link as definite, possible or rejected, the listing of a sample of links from each run, and resolution of duplicate links within each run. In addition, duplicate links involving records over different years of death were resolved. Over two-thirds of the cost was accounted for by the comparison of the records. As previously mentioned, this demonstrates the advantage of a modular system, where all other steps may be carried out iteratively at relatively minimal cost. The next most expensive step was the weighting which accounted for approximately 14%. The steps listed above took 179 min of CPU time for the males and 175 min for the females. It should be noted that testing was carried out first on a very small sample of the file consisting of a few blocks of records from the two files. At this point, the mortality records were selected from a single year of death. When preliminary testing was completed, an entire year of death records was linked with the TB records and further refinements made. For example, it was found that test runs where no cutoff weight was used were about 15% more expensive than those where a cutoff weight was used that was sufficiently low for no potential links to be missed. The cost of this linkage using the generalized system was substantially lower than the cost of linkages carried out previously using ad hoc programs.

Linkage of Occupational Cohort to Cancer Incidence

Between 1965 and 1971, data were collected by Statistics Canada for a 10% sample of the Canadian labor force (approximately 700,000 individuals). The data included identifying information together with the industry and occupation in which the individual was engaged in each particular year. In order to follow the mortality and cancer morbidity experience of this cohort with respect to their industrial and occupational exposure, these records were linked to the national mortality data base and the cancer incidence files. For the linkage to the cancer incidence files, Ontario occupational records were excluded, since identifiable cancer incidence records were not available for that province, leaving 476,174 occupational records.

The 287,786 male and 188,388 female occupational records were linked to 171,628 male and 215,651 female cancer incidence records covering the years 1969 to 1976. (Cancer incidence data were first collected nationally in 1969.) The identifying items available on both files were NYSIIS of surname; surname and alternate surname; first and second given names; day, month, and year of birth; and sex. As in the previous example, the records were blocked by NYSIIS of surname and sex. In this case only two separate runs were made since the files were split by sex, but not according to the year of diagnosis of cancer. The same levels of full and partial agreement were used as for the TB–mortality linkage.

The number of potential links generated was 96,100 from the male files and 82,482 from the female files. After the insertion of final weights and the setting of threshold values, and resolution of links of multiple occupation records to single cancer records, the number of possible and definite male links was 5315 and there were 2885 female links. In this case, multiple cancer incidence links to occupation records were considered acceptable since the cancer incidence file contains one record for each primary site of cancer. The number of occupation records involved in these links or the number of individuals linking to cancer records was 4953 men and 2747 women. The cost of this linkage was approximately $600 and the CPU time used was about 30 min for the males and 23 min for the females, including the same steps for which cost was calculated for the TB–mortality linkage. The proportion of time spent on the comparison of records and weighting was comparable to the TB–mortality linkage.

Strategy for Using Linkage System

There are three main factors which affected the cost of these linkage runs using the system described. The order in which comparisons are carried out is extremely important, as has been mentioned. Obviously it would be very costly to compare alphabetic fields first, knowing that at some point later in the comparison the records could be rejected as potential links. Efficiency can be maximized by first comparing numeric fields on the basis of which pairs of records can be immediately rejected. It may be decided, for example, that the quality of the two files concerned is sufficiently high that disagreement on birth year of more than 10 years means that the link would not possibly be believed. The second factor affecting cost is the extent to which records have missing identifying items of information. If one or both files contain many records with very little information present, these records will generate large numbers of potential links because there is little or no basis on which to reject these links, i.e., there will not be a sufficient number of disagreements to bring the disagreement weight below the cutoff weight. As a result, comparison of records takes longer since more records go through the comparison of all items and weighting will also be more expensive due to the volume of potential links. The third consideration is the setting of the cutoff weight. The apparent efficiency of a linkage may be increased by using a less strongly negative cutoff weight. However, depending on the purpose of the application, this may have subsequent adverse effects. If only the definite links are of interest, no problems may arise, but if the purpose of conducting the linkage is statistical analysis, it is then important to be able to identify the records or individuals whose status is unknown. This is the case with respect to the applications described here.

Conclusion

The system which was developed provides a very powerful tool for medical research in general, and the concepts can be implemented fairly readily on any medium-sized computer. Since the processing is sequential in general it can also be adapted to any small installation which has the facility for processing large volumes of sequential data.

ACKNOWLEDGMENTS

The authors wish to acknowledge the contributions of systems analysts Ted Hill and Steve Hobbs, and methodologists Simon Cheung, Mike Eagen, and Dave Binder.

REFERENCES

1. MACMAHON, B., AND PUCH, T. F. "Epidemiology Principles and Methods." Little, Brown, Boston, 1970.

2. HOWE, G. R., LINDSAY, J., COPPOCK, E., AND MILLER, A. B. Isoniazid exposure in relation to cancer incidence and mortality in a cohort of tuberculosis patients. *Int. J. Epidemiol.* **8**, 4, 305 (1979).

3. HOWE, G. R. Breast cancer mortality in relation to fluoroscopic X-ray exposure. Presented at the 4th International Symposium of the Detection and Prevention of Cancer, London, July 1980.

4. HOWE, G. R., LINDSAY, J., AND MILLER, A. B. A national system for monitoring occupationally related cancer morbidity and mortality. *Prev. Med.,* in press.

5. SMITH, M. E., AND NEWCOMBE, H. B. Methods for computer linkage of hospital admission-separation records into cumulative health histories. Methods of Information in Medicine **14**, 118 (1975).

6. FELLEGI, I. P., AND SUNTER, A. B. A theory for record linkage. *J. Amer. Stat. Assoc.* **64**, 1183 (1969).

7. LYNCH, B. T., AND ARENDS, W. L. "Selection of a Surname Coding Procedure for the SRS Record Linkage System." U.S. Department of Agriculture, Washington, D.C., 1977.

RELIABILITY OF COMPUTERIZED VERSUS MANUAL DEATH SEARCHES IN A STUDY OF THE HEALTH OF ELDORADO URANIUM WORKERS **

H. B. NEWCOMBE*, M. E. SMITH†, G. R. HOWE‡, J. MINGAY§, A. STRUGNELL§ and J. D. ABBATT§||

* P.O. Box 135, Deep River, Ontario, KOJ 1PO, Canada; † Vital Statistics and Disease Registries, Statistics Canada; ‡ National Cancer Institute of Canada Epidemiology Unit, University of Toronto; §Eldorado Nuclear Limited, Ottawa

Abstract—An epidemiological follow-up study of 16,000 uranium mine and refinery employees has made use of computerized techniques for searching a national death file. The accuracy of this computerized matching has been compared with that of corresponding manual searches based on one-eighth of the worker file. The national death file—Canadian Mortality Data Base—at Statistics Canada includes coded causes of death for all deaths back to 1950. The machine search was carried out using a generalized record linkage system based upon a probabilistic approach. The machine was more successful than the manual searchers and was also less likely to yield false linkages with death records not related to the study population. In both approaches accuracy was strongly dependent on the amount of personal identifying information available on the records being linked.

Uranium Radium Cancer Risks Follow-up Epidemiology
Industrial cancer Death searches Computer searches Automated follow-up

INTRODUCTION

Eldorado Nuclear Limited (E.N.L.) is conducting a retrospective epidemiological study of the health of its former employees. Eldorado operations involve the mining, milling and refining of uranium and these activities have been carried on continually from the early 1930s. Initially radium was extracted for medical and other purposes, and more recently uranium metal and nuclear fuel materials have become the main products.

The objectives of this study are:

(a) to identify former employees who may have a potential compensation claim, and to inform them or their survivors of these potential compensation claim rights, and

(b) to obtain dose-response data for evaluation of the risks to workers, especially with respect to atmospheres containing radon and radon-daughters.

The main study design and details regarding the assembly of the nominal roll have been described elsewhere [1]. The purpose of the present study, which serves both the short-term and the long-term aims of the broader investigation and of other similar studies, was to investigate the reliability of searches of all relevant death registration material using the study nominal roll and the Canadian Mortality Data Base (C.M.D.B.) operated by Statistics Canada. In an attempt to assess the reliability of machine record linkage for which the C.M.D.B. was designed [2, 3], the results of rapid computer searching and file linkage have been compared with manual searching and file linkage.

It has rarely if ever been possible to judge, much less quantify, how many false positive (incorrect) and false negative (missed) linkages result from conventional manual searches for death registrations where the dead or alive status of the members of the nominal roll is unknown. The present study is designed to provide quantitative information on both manual and machine file searching. The comparison has demonstrated the extent of the influence of an abundance or scarcity of personal identifiers on the efficiency of both types

**Reprinted with permission from Computers in Biology and Medicine, Vol. 13, No. 3, Copyright ©1983 by Pergamon Press Ltd., pp. 157-169.

Table 1. Manual matches of worker records with death records, by degree of assurance

Degree of assurance	Category	Number of worker records	
A	definite link	137	219
B+	very good possible	35	
B	good possible	47	
B−	unlikely possible	23	
C	poor possible	17	
D	not enough identification	10	
other	no link	1602	

From a sample of 1871 male worker records in which the surnames begin with the letters A or B.

of file matching. It has also demonstrated the greater efficiency of machine than manual matching.

The Eldorado study, although retrospective in nature, is being carried out with the intention of merging it into a prospective health monitoring instrument. It is the hope of many that similar prospective undertakings will come to be regarded in the future as desirable and feasible. Only thus can full use be made of available records to assess the adequacy of current standards of protection against delayed harm from the working experience.

MATERIALS AND METHODS

The Eldorado nominal roll used for the present study of linkage accuracy consists of a total of 16,658 names. These relate to past workers at the Port Radium mine (4526), Beaverlodge mine (9336), the Port Hope refinery (2514) and Research and Development (282), and involve employment as far back as 1932.

The Canadian Mortality Data Base file contains over five million death registrations with coded cause of death for the years 1950 to 1977.

For the computer linkage study, only E.N.L. records with a sex code equal to male or unknown (15,937) were used to initiate searches of the male half of the C.M.D.B. Searches for deaths relating to female workers (721) were not attempted because of the small numbers and the practical problems associated with changes of name at marriage. Such searches should be possible in the future, however, using the maiden surnames which occur on the death registrations of ever-married women, in the form of fathers' surnames.

For the manual linkage part of the operation, a sample of the E.N.L. file was used to initiate the searches representing all surnames of males beginning with the letters A and B (1871). A and B were chosen because they are known to provide a good sample of common and uncommon names (Andersons and Browns), and there is no evidence that they introduce a bias. The manual search used the C.M.D.B. microfiche listings.

The degree of assurance that a correct match has been achieved is assessed quantitatively by the computer. The decision is based upon prior information about the discriminating powers of various possible agreements and disagreements of the personal identifying information. The manual searchers assessed the degree of assurance subjectively and ranked the matches (links) they achieved on a scale that was qualitative (Table 1).

The principles are the same in both cases. Greater weight is attached to agreements of rare names, rare birthplaces, etc., than to agreements of their commoner counterparts. Similarly disagreements that occur only rarely, in a pair of records, argue more strongly against a correct match than will disagreements that are common. These fairly obvious inferences are taken into account by both the computer and the searcher. The chief difference is that the computer works from look-up tables that tell it by how much a given agreement, or disagreement, will shift the odds in favour of, or against, a correct match. The man relies on judgement with regard to the same matter, based on similar information and reasoning.

Table 2. Coincident identifiers in potentially matching worker records and death records (estimated)

Identifiers for searching and linkage	Percentage available in		
	Worker records alone	Death records alone	Both simultaneously (est.)
Surname plus at least one given name	100	100	100
plus a middle initial or name	50	47	23
Birth date in full	79	95	75
province or country	55	98	54
Parental initials, on? or more	23	87	20
birth province/country, one or both	8	87	7

The system used for searching the death records was developed by Statistics Canada and the Epidemiology Unit of the National Cancer Institute of Canada for use in medical studies at Statistics Canada [4] and is described as a Generalized Iterative Record Linkage System (GIRLS). It is an extension of the probabilistic approach to record linkage developed at Chalk River [5–8]. Record linkage has been described in detail in numerous other publications (see references [9–13] and for a complete bibliography [14]). The mathematical derivation of 'weighting factors', from the frequencies of the various identifier comparison outcomes (agreements, disagreements, etc.), in linked vs unlinked pairs of records, has been described in detail elsewhere [4–7]. The weighting factors serve to represent in numeric form the discriminating powers of different identifier comparisons and their outcomes.

The assurances calculated by the computer are conveniently expressed on a logarithmic scale using the base 2 as in information theory. On such a scale, zero represents odds of 1:1 that the linkage is a correct one, each added unit doubling the odds and each subtracted unit halving them. For example, +1 and +2 represent odds of 2:1 and 4:1 respectively, in favour of a correct match; whereas −1 and −2 represent odds of 1:2 and 1:4 and so argue against a correct match. With an abundance of personal identifying information common to a pair of records, the evidence for or against a correct match tends to become more decisive, and stronger positive or negative 'weights', as they are called, are likely to be associated with the comparisons. Thus, for genuinely linkable pairs of records, total weights of +10 to +20 may be common, representing favourable odds of 1000:1 to 1,000,000:1. For unlinkable pairs, the weights and the odds will tend to be similar in magnitude but opposite in direction.

The degrees of assurance of a correct match, in both approaches, may be expected to vary widely. In large part this is due to differences in the amount of personal identifying information common to a potentially linkable pair (Table 2). For example, without the full birth date, the name information alone will usually not carry enough discriminating power to enable the correct death record to be selected from among a million or so others. And in part it is due to differences in the rarity or commonness of the names, birthplaces and such. Assurance is similarly affected whether the search is carried out by computer or by man.

A major purpose in performing the analysis of the data yielded by the combined efforts of the computer and the human searchers is to determine to what degree the accuracy of the death searches depends upon the amount of personal identifying information which can be applied to the problem of distinguishing good matches from bad.

RESULTS AND DISCUSSION

Assurances associated with the computer and manual searches

As a result of the computer search, approximately 2000 of 15,937 Eldorado worker records were linked to matching death registrations with varying degrees of assurance (Table 3). As a result of the manual search, somewhat over 200 of the 1871 records from

Table 3. Computer matches of worker records with death records, by degree of assurance

Weight range	Category	Range of odds (inferred from weights)	Number of worker records	
+4 and over	positive link	(11:1 and over)	1490	2023
+1 to +3	probable link	(1.4:1 to 11:1)	362	
zero	possible	(1:1.4 to 1.4:1)	171	
−1 to −3	probable non-link	(1:11 to 1:1.4)	794	
−4 to −8	positive non-link	(1:256 to 1:11)	2339	
other	no link	—	10,781	

From a total of 15,937 records where sex is male or unknown.

the sample (relating to surnames beginning with A or B) were similarly linked (Table 1). In each case, the precise number of 'acceptable' links depends upon where one sets the 'threshold' for acceptability. If one places it where the implied odds in favour of a correct match are 50:50or better, either as calculated by the computer or as judged subjectively by the manual searchers, the precise number of 'acceptable' links would be 2023 and 219 respectively.

Because the setting of the threshold for acceptance is necessarily arbitrary in both cases, one must consider how best to estimate the numbers of accepted links that are in fact wrong, and the numbers of rejected matches that were correctly paired.

Estimating the false positive and false negative computer matches

There are two ways in which the accuracy of the computer linkages may be judged without reference to parallel manual searches. The first approach is based on the simple fact that where a worker's record links 'acceptably' to two different death records, only one of these links can be correct; the frequency of such instances tells us something about the potential for producing false positive outcomes. The second approach takes at face value the calculated odds, in favour of or against a correct match, and derives both an estimated number of false matches that lie above the threshold for acceptance, as well as another estimated number of potential correct matches that fall below the threshold for rejection.

Table 4. 'Runners up' as indicators of the potential for false positive linkages (computer searching)

Weight range	Range of odds (inferred from weights)	Number of worker records ('best' match for each)		Number of matches not the 'best' ('runners up')		'Runners up' (% of 'best')	
+10 and over	(724:1 and up)	1057	2023	10	325	1	16%
+4 to +9	(11:1 to 724:1)	433		64		15	
+1 to +3	(1.4:1 to 11:1)	362		150		41	
zero	(1:1.4 to 1.4:1)	171		101		59	
−1 to −3	(1:11 to 1:1.4)	794		680		86	
−4 to −8	(1:256 to 1:11)	2339		5053		216	

Note: (1) Weighting factors are rounded for simplicity, the precise dividing lines in the above table being +9.5, +3.5, +0.5, −0.5, and −3.5.

(2) In the '+10 and over' group, a substantial fraction carry weights in the region of +20 and even +30, representing odds of a million-to-one and a billion-to-one in favour of a correct linkage.

(3) Where such high weights occur among the 'runners up', which cannot be true links, they nevertheless correctly refer to similarities of identifying information which are exceedingly unlikely to have occurred by chance alone. Sometimes, such a pair of records will relate to two members of a family, one of whom was named after the other. Also, twins, who share the same birth date, are apt to turn up in such pairs of records, and so do members of small ethnic groups who share the same rare birth places and rare surnames. Manual searchers and the computer, both correctly tend to pay special attention to such non-random pairings of records, which signify correlations other than those due to the identity of the individual.

Table 5. Calculated 'weights' as indicators of probable false positives and false negatives (computer searching)

Weight range	Range of odds (inferred from weights)	Number of worker records ('best' matches)		Probable correct matches (est.)		Probable false matches (est.)	
+10 and over	(724:1 and up)	1057	2023	1057	1845	–	177
+4 to +9	(11:1 to 724:1)	433		424		9	
+1 to +3	(1.4:1 to 11:1)	362		279		83	
zero	(1:1.4 to 1.4:1)	171		85		85	
−1 to −3	(1:11 to 1:1.4)	794	3133	153	204	641	2929
−4 to −8	(1:256 to 1:11)	2339		51		2288	

Note: Whichever weight one chooses as representing a threshold for acceptance, those 'false matches' which fall above the threshold will become 'false positives', and those 'correct matches' which fall below the threshold will become 'false negatives'.

For the first approach, one may compare the numbers of 'best' matches with the numbers of 'runners up', broken down by the calculated 'weight' or odds in favour of a correct match (Table 4). The number of runners up increases with progressively lower weights. With the threshold for acceptance set just below zero, the 'runners up' (representing death records to which workers' records might have linked 'acceptably' if they hadn't found a better match) number sixteen per hundred 'best' matches. These are *potential* rather than actual false positives, but they indicate what might happen to the record of a worker who hadn't yet died and for whom there was therefore no correct matching death registration. This problem arises chiefly where the personal identifying information is limited.

For the second approach, the calculated weights (and their associated odds) were used to derive the probable numbers of links and non-links. For example, a weight of zero represents odds of 1:1 in favour of a correct linkage. Therefore half of the matches which have been assigned this weight, probably do relate to the same person and the other half do not. Taking the weighting factors at face value, the likely proportions of correct and false matches associated with each value of the total weights were calculated (Table 5). From this sort of calculation it was inferred that, for a threshold set just below zero weight, and with 2203 'accepted' links, 178 of these or just under 9% are likely to be false positives. In addition there are a probable 205 potential correct links that were not accepted, represent-

Table 6. Numbers of matches achieved by manual vs computer searching, by degree of assurance (based on worker records having surnames beginning with A or B)

Computer weight range	Degree of manual assurance						No man. match	Total
	A	B+	B	B−	C	D		
+10 and up	121	16	7	1	2	–	14	161
+4 to +9	13	8	9	1	1	–	21	53
+1 to +3	2	4	8	3	2	–	23	42
zero	–	1	3	1	–	–	11	16
−1 to −3	1	4	3	3	2	–	79	92
−4 to −8	–	1	9	10	5	9	266	300
no comp. match	–	1	6	5	7	–	1188	1207
Total	137	35	45	24	19	9	1602	1871

Note: (1) Where the thresholds for acceptance are set at zero and above for the computer, and at B and above for the manual searches, the following would be the result:

accepted by both = 192
accepted by computer only = 80
accepted by manual only = 25
rejected by both = 1574.

(2) The table includes cases in which the death record selected by the computer differs from that selected by the manual searcher (see next table).

Table 7. Computer – manual disagreements with respect to the death record selected
(Parentheses indicate which were judged correct on subsequent review.)

Computer weight range	Degree of manual assurance						Total
	A	B+	B	B−	C	D	
+10 and up	–	1(M)	1(C)	1(C)	1(C)	–	4
+4 to +9	–	1(?)	1(C)	1(?)	1(C)	–	4
+1 to +3	–	–	1(C), 1(X)	1(?)	2(?)	–	5
zero	–	–	–	–	–	–	–
– 1 to – 3	–	–	–	–	1(?)	–	1
– 4 to – 8	–	–	1(?), 1(X)	3(?), 2(X)	2(?), 1(X)	2(?)	12
Total	–	2	6	8	8	2	26

Note: These numbers are included in the previous table.
M = manual choice correct
C = computer choice correct
X = both manual + computer choices incorrect
? = uncertain

ing a false negative rate of about 10%. If the threshold were raised to get rid of the false positives the false negatives would increase, and lowering the threshold would have the opposite effect. With the threshold in the vicinity of zero the number of false positives and false negatives are expected to be about equal. The only way to simultaneously reduce the frequencies of false positives and false negatives is to obtain a greater amount of personal identifying information for each record.

The human searcher is faced with the same problem, except that in this case it is not quantified. For both the man and the computer there may be additional false negatives that arise because some of the worker records are grossly deficient in identifying information; e.g. an absent birth date may result in insufficient discriminating power to distinguish between multiple possibilities for linkage.

Comparisons of computer vs manual linkages

Further insights into the respective levels of accuracy may be gained from comparisons of the performance of the computer vs that of a human searcher. Specifically, where the two approaches fail to agree, (a) they may yield different deaths, (b) the human may appear to succeed and the computer not at all, and (c) the reverse may be the case.

It might be supposed that the ultimate test of the accuracy of the computer searching would be for a man to carry out the same searches as the machine to see where the computer had gone wrong. This assumes, without evidence, that the man is more accurate than the computer. Instead, however, the problem is actually quite symmetrical, because lack of specificity in the identifying information adversely affects the accuracy of both the computer and the human searcher, and it remains to be shown which is the more accurate in the present setting.

Direct comparisons serve to indicate where the two approaches have yielded the same

Table 8. Proportions of worker records linked with death records by the computer, when birth year is absent vs present

Birth year* (present/absent)	Linkages (weights zero and over)	Worker records	% linked
Absent	18	3323	0.5
Present	2004	12614	15.9
Total	2022	15937	12.7

* Note: Virtually all of the worker records that lack year of birth, also lack the rest of the birth date.

Table 12. Calculation of 'weighting factors' for place of death vs place of work

Place of death	Number in linked pairs	Expected for average Canadians	Ratio (inferred odds in favour of linkage)	Weighting factor ($\log_2$ of the ratio)
Port Radium and Beaverlodge workers (145 pairs)				
Que.–Atlantic	8	53	1:6.6	−2.7
Ont.	30	52	1:1.7	−0.8
Man.–Sask.	19	12	1.5:1	+0.6
Alta.–B.C.	51	27	1.9:1	+0.9
Y.T.–N.W.T.	8	0.4	20:1	+4.4
Edmonton	27	3.5	8:1	+3.0
Port Hope workers (59 pairs)				
Que.–Atlantic	–	22	1:43	−5.4
Ont.	44	21	2.1:1	+1.1
Man.–Sask.	3	5	1:1.7	−0.8
Alta.–B.C.	12	11	1.1:1	+0.1
Y.T.–N.W.T.	–	–	–	–
Port Hope	20	0.05	400:1	+8.7

Note: (1) Where no death occurred, the ratio is based on an assumed 0.5 deaths; the resulting 'weighting factor' will then tend to be conservative.

(2) The expected numbers 'for average Canadians' are based simply on the populations of the regions.

unlinkable pairs argue against linkage.) The conversion of this ratio into a logarithm to the base 2 is just a convenience to make the weights addable. The first of the two frequencies is obtained by direct observation of the linked pairs of records, and the second is normally calculated from the frequency of the particular value of an identifier in the files themselves.

Examples are given of the use of such data as derived from the present study after its completion. These have to do with (a) simple disagreement weights (Table 10), (b) weights for a spectrum of outcome values ranging from complete agreement through various degrees of partial agreement–disagreement to complete disagreement (Table 11), and (c) weights for the occurrence in matched pairs of records, of identifier combinations which are correlated but cannot be regarded as either agreeing or disagreeing (Table 12). The latter two tables represent relatively fine groupings of the full range of possible outcome values. Such breakdowns are designed to avoid unnecessary pooling of outcomes with high and with low discriminating power, which would degrade the usefulness of the identifiers (rather as the usefulness of panned gold dust is degraded by re-mixing it with the sand).

The setting of the 'zero point' on the weight scale has proved more complicated than originally expected. This is the point at which the total weight for a matched pair of records indicates 50:50 odds in favour of, or against, a correct linkage. The total weight as initially envisaged did not take into account either the increased likelihood of chance similarities where the file being searched is particularly large, or the degree to which age and sex may influence the likelihood that an individual will be represented in that file where it is a death file. The hope was that the zero point could be adequately pinpointed by manual examination of borderline linkages. However, the present extensive work of this sort leaves one less confident about use of the manual approach alone, for this purpose. Substantial biases are now suspected, from a human tendency to reject out-of-hand those troublesome pairs which lack sufficient identifiers on which to base a judgement but might non-the-less be correctly matched. For a total of the calculated weights to represent 'absolute odds', as distinct from just 'relative odds', components are required which will take into account (a) the size of the death file over a given period, (b) the likelihood of an individual dying in that period, and (c) the likelihood of his being alive at the start of the period so as to be 'available' to die within the period. This approach is now being developed as a result of the need indicated by the present manual studies. And ways of estimating, and perhaps correcting for, any biases in the total weights arising out of this approach are being considered.

outcomes, and where they have differed. But judgements concerning which is the correct outcome when the approaches disagree are necessarily subjective, except where an actual oversight/error of some kind can be detected, or where additional identifying information can be obtained and used. The comparisons between the outcomes of the computer vs the manual searches that will be considered relate to the sample of 1871 Eldorado worker records in which the surnames began with A or B.

The degree of assurance of a correct linkage with a death record, or of a non-linkage, was variable both for the computer and for the manual searches. To a large extent, where the computer was 'very sure' that a correct decision had been made, so was the manual searcher, but the correlation is a fairly loose one when all degrees of assurance are considered (Table 6).

The conclusions one may draw from this comparison are best described in terms of a possible arbitrary threshold for 'acceptance' as a linkage, or 'rejection' as a non-linkage. Suppose, for example, that this threshold is set so that computer weights of zero and above, and manual assurances of B and above, are taken to indicate acceptable linkages. Then for 94% of worker records the outcomes from the two types of search both indicate either an appropriate linkage (192 cases or 10.3% of the records) or a non-linkage (1574 cases or 84.1% of the records).

For about 6% of the worker records the computer and the manual searcher were in disagreement as to whether an appropriate matching death record had been found (Table 6). If the results of the human searching are believed the computer approach resulted in 80 false positives and 25 false negatives (i.e. 4.3% and 1.3%, respectively, of the 1871 worker records, or, when based on the 219 manual linkages, 37% and 11% of the potentially linkable records). If the results of the computer searching are believed, the manual approach is similarly inaccurate and results in 25 false positives and 80 false negatives (out of 1871 worker records, or, when based on the 272 computer linkages, 9% and 29% of the potentially linkable pairs). This comparison serves chiefly to suggest that both approaches may involve considerable inaccuracy where the personal identification lacks discriminating power. And, of course, such comparisons cannot indicate how many relevant death records were missed by both kinds of searching.

There is evidence, however, that the computer searching results in fewer false negatives than does the manual searching. Thus, in Table 6 there are only seven cases of 'acceptable' manual matches of which the computer was apparently unaware, as against 69 cases of 'acceptable' computer matches of which the manual searchers were seemingly unaware.

Evidence that the computer is likewise less prone to the production of false positive linkages, may be obtained from those instances in which both approaches appeared to be successful but each identified a different death record as the appropriate one. For all 26 examples of disagreement of this kind, the source documents (E.N.L. work records and death certificates) were re-examined for additional information with which to resolve alternative choice 'matches' (Table 7). The resulting 'final' judgements are not infallible, but they do show that the computer is more reliable than the manual searchers where the two find different death records. The computer 'accepted' thirteen matches for the 26 ENL records, later judged to consist of six 'right', two 'wrong', and five 'doubtful'. The manual searchers 'accepted' just eight matches, later judged to consist of one 'right', five 'wrong', and two 'doubtful'.

From the above evidence, the computer searches appear to result in substantially fewer false positive and false negative outcomes than do the manual searches. Appropriate empirical tests and procedural adjustments will further improve the quality of machine linkage. Some of the proposed procedural changes will be described in what follows.

DISCRIMINATING POWER AS A LIMITING FACTOR

Since record linkage in the absence of unique identifier numbers depends upon multiple identifiers, it follows that discrimination decreases rapidly as personal identifying inform-

Table 9. Effects of differences in the availability of identifying particulars on the estimated proportions of false positives and false negatives (matched pairs with computer weights of zero and above being 'accepted' as 'linked')

Available identifiers	Number of matched pairs	Calculated false positives		Calculated false negatives	
		No.	% of accepted	No.	% of accepted
Year of birth, but not month and day					
Accepted	291	47.8	16.4	–	–
Rejected	805	–	–	54.2	18.6
Full birth date					
Accepted	1684	122.9	7.3	–	–
Rejected	2092	–	–	136.6	8.1
Birth date and place, plus two given names					
Accepted	166	4.8	2.9	–	–
Rejected	89	–	–	5.2	3.1

Note: (1) Columns headed 'No.' contain estimated numbers. They will therefore not be integers. For the method of estimation, see Section on 'Estimating the false positive and false negative computer matches'.

(2) For the purpose of this table an identifier is said to be 'available' as a basis for linkage when it is present on both a worker record and the death record to which it is matched, regardless of whether it agrees or disagrees.

(3) Where not specifically mentioned, an identifier may be either available or unavailable.

ation diminishes in abundance. In other words, the number of false negatives increases disproportionately as identifying information decreases.

Some indication of the quantitative importance of different amounts of identifying information may be gained from a few comparisons. For example, where information on birth year was present on the ENL records, some 16% were successful in finding a matching death record. But when it was absent, the success rate was only 0.5% (Table 8).

A better comparison involves three different levels of discriminating power in records that have the birth year (Table 9). 'Full identifying information' results in an estimated 3% of false positives and 3% of false negatives. Records reduced to birth date without place, etc., double both error rates to 7 and 8% each. Records with year of birth only again double the error rates to 16 and 19%. The comparisons are not precise, because different data sets are involved. But, in the absence of more elaborate and expensive tests, it would be unwise to disregard the practical guidance from such internally consistent evidence, of the need for multiple identifiers.

A redundancy of identifiers may be needed for a rather different reason. Strictly speak-

Table 10. Frequency of discrepancies in personal identifying information, and the 'weighting factors' derived from these frquencies (based on 269 matched pairs of worker and death records, with weights of zero and up)

Kind of identifier	Discrepant	Total linked pairs	Frequency in linked pairs	Weight for discrepancy ($\log_2$ freq.)
Surname spelling	12	269	1/22	−4.5
First initial	27	269	1/10	−3.3
First given name	74	268	1/3.6	−1.8
Second initial	19	119	1/6	−2.6
Second given name	18	65	1/3.6	−1.8
Birth province or country	7	114	1/16	−4.0
Parental initials	18	73	1/4	−2.0
Parental birth province/ country	11	25	1/2.3	−1.2

Note: For simplicity, the frequency of the discrepancy in unlinked pairs is taken to be virtually unity. Thus, $\log_2$ of the frequency in linked pairs approximates closely, $\log_2$ of the ratio of the frequencies in linked/ unlinked pairs.

Table 11. Calculation of 'weighting factors' for birthdate discrepancies

Degree of discrepancy	Number in linked pairs	Expected in unlinked pairs	Ratio (inferred odds in favour of linkage)	Weighting factor ($\log_2$ of the ratio)
Year of birth (268 pairs)				
0	170	2	85:1	+6.4
1	45	4	11:1	+3.5
2–3	38	8	5:1	+2.3
4–9	8	24	1:3	−1.6
10+	7	230	1:33	−5.0
Month of birth (243 pairs)				
0	219	20	11:1	+3.5
1	10	37	1:3.7	−1.9
2–3	8	64	1:8	−3.0
4–9	5	112		−4.3
10–11	1	10	1:20	
Day of birth (241 pairs)				
0	189	8	24:1	+4.6
1	11	16	1:1.5	−0.6
2–3	10	29	1:2.9	−1.6
4–9	17	76	1:4.5	−2.2
10+	14	112	1:8	−3.0

Note: The numbers expected in unlinked pairs are calculated as follows:

For exact agreements the expectation is taken to be n/n^2 times the number of matched pairs, where n is the number of different values of the identifier.

For discrepancies of degree d, the expectation is taken to be $2(n-d)/n^2$ times the number of matched pairs.

These equations represent approximations based on the assumption that the different values are equal in frequency. Where they are not equal, a more detailed calculation is required and this has been carried out in the case of year of birth.

ing, total weights reflect only the likelihood or unlikelihood that the observed similarity of identifying information on pairs of records has arisen other than by chance. But the ruling out of chance does not necessarily establish that the same person is involved:

Family members may be named after each other, and twins may be confused because of a common birthplace, birth date, and perhaps because of similar given names.

There are fashions in given names with small communities, and surnames repeat in localized ethnic groups and communities.

In short, similar or identical identifiers occasionally refer to attributes associated with particular groups of people, but not uniquely with any individual person.

The above kinds of problems can be minimized by abundant information, and to some extent by manual resolution using additional identifiers.

IMPROVING THE WEIGHTING PROCEDURES

The present manual/machine matching study has revealed needs for improvements in the weighting procedures used by the machine, and has provided some of the data required for the purpose. Such improvements would have to do in particular with (a) putting to use more of the potential discriminating power that could otherwise remain latent in the available identifiers, and (b) finding a better way of setting the 'zero-point' on the weighting scale.

The data used for calculating the weighting factors consists of the frequencies of various identifier comparison outcomes (agreements, disagreements, etc.) in pairs of records judged to be correctly linked, together with the corresponding frequencies for unlinkable pairs. Quite simply, the ratio between these two frequencies indicates the degree of assurance associated with a particular comparison outcome. (Outcomes that are more fashionable in linked pairs argue for linkage, and those that are more fashionable in

Table 13. Discrepancies of given names, by kind of discrepancies (based on 92 discrepancies of the first and second names combined, among 333 given names compared in record pairs with weights of zero and above)

Kind of discrepancy	Examples	
All discrepancies (92 cases)		
Position only, same spelling	(John – William John)	24
Different initial and name	(John – Fred)	16
Different spelling, same initial	(Louie – Louis)	52
Spelling discrepancies (52 cases)		
Vowel change only	(Ralph – Rolph)	15
Shortened only	(Fred – Frederick)	11
Nicknames, not just shortened	(John – Jack)	5
Phonetic similarities	(Ouide – Ovide)	4
Anglicizations	(Kenneth – Kazimie)	3
Double consonants	(Riser – Risser)	2
Other	(Bjom – Bjorvi)	12

Note: Of 46 disagreements of first or second initials, 11 were associated with simple reversals of the sequence on one of a matched pair of records as compared with the other (inversions), and 22 were due to one of the initials being transposed from first to second place (frame shifts).

Various other possible improvements in the weighting system, which will not be described here, are under development as a result of the present manual comparisons. Some of these have to do with (a) the handling of given name similarities where precise agreement is lacking (see examples in Table 13), (b) comparisons involving inverted sequences (e.g. of initials, and of birth month and day), and (c) practical means for making better use of the discriminating powers of very rare surnames, without recourse to excessively long look-up tables of weights.

IMPLICATIONS FOR ALL RETROSPECTIVE AND PROSPECTIVE STUDIES

Safety standards

(1) It is in everyone's interests to know where problems of safety are greatest and where they are least.

(2) Neither workers, management nor society in general benefit where undue emphasis is directed to non-problems, while real problems are neglected because they remain undetected.

(3) The limited public funds available earmarked for administration and enforcement of safety standards ought to be used so that attention to low-risk situations never results in the neglect of higher risks.

Fears about possible loss of privacy have tended recently to further reduce the specificity of personal identification on personnel records, notably on application forms for employment. At the same time, the public has increasingly demanded investigations of the delayed risks in various work situations, and has emphasized the right of the worker to know the risks.

To detect and measure delayed personal harm of almost any sort, and resulting from almost any kind of 'exposure', individual people require to be identified in a reasonably unambiguous fashion. This is true whether one follows exposed individuals forward to look for harm, or sick individuals backward in time to look for exposures. With both approaches, the most serious stumbling block is often a lack of sufficient specificity and redundancy in the personal identifiers (names, birth dates and such) by which people are known and represented on their various records, including their work records.

SUMMARY

Computerized searching of a national death file has been tested and compared for accuracy with the corresponding manual searches. The test formed a part of an

epidemiological follow-up study of some 16,000 former Eldorado employees, in which employment records are being used to initiate the searches for related death registrations contained in the Canadian Mortality Data Base at Statistics Canada. This facility includes the coded cause for all deaths back to 1950. The computer searching was guided by a generalized record linkage program, based on a probabilistic approach; the program was developed by Statistics Canada and the Epidemiology Unit of the National Cancer Institute of Canada. The corresponding manual searches used microfiche printouts from the Mortality Data Base tapes.

The results from the test showed the machine to be more accurate than the manual searchers. Not only was it more successful in extracting the relevant deaths, but it was also much less likely to yield false linkages with death records not relating to members of the study population. For both approaches, however, accuracy was strongly dependent on the amount of personal identifying information available on the records being linked.

Acknowledgments—The authors wish to thank Mr. J. Silins, Mrs. C. Poliquin and Mrs. M. Warner for their invaluable assistance. The advice and criticism of many other colleagues, particularly Mr. S. E. Frost and Mr. R. G. Dakers, is gratefully acknowledged.

REFERENCES

1. J. D. Abbatt, The Eldorado Epidemiology Project; Health Follow-up of Eldorado Uranium Workers. Eldorado Nuclear Limited, Ottawa, Ont. (1980). (Available on request by writing to Eldorado Nuclear Limited, 255 Albert Street, Suite 400, Ottawa, Ont. K1P 6A9.)
2. M. E. Smith and H. B. Newcombe, Automated follow-up facilities in Canada for monitoring delayed health effects, *Am. J. pub. Hlth* **70**, 1261–1268 (1980).
3. G. W. Beebe, Record linkage systems–Canada vs the United States, *Am. J. pub. Hlth* **70**, 1246–1247 (1980).
4. G. R. Howe and J. Lindsay, A generalized iterative record linkage computer system for use in medical follow-up studies, *Comput. biomed. Res.* **14**, 327–340 (1981).
5. H. B. Newcombe, J. M. Kennedy, S. J. Axford and A. P. James, Automatic linkage of vital records, *Science* **130**, 954–959 (1959).
6. H. B. Newcombe and J. M. Kennedy, Record linkage: making maximum use of the discriminating power of identifying information, *Communs Ass. Comput. Mach.* **5**, 363–566 (1962).
7. H. B. Newcombe, Record linking: the design of efficient systems for linking records into individual and family histories, *Am. J. hum. Genet.* **19**, 335–359 (1967).
8. M. E. Smith and H. B. Newcombe, Methods for computer linkage of hospital admission–separation records into cumulative health histories, *Meth. Inf. Med.* **14**, 118–125 (1975).
9. E. D. Acheson, Record Linkage in Medicine, E. and S. Livingstone, Edinburgh. (1968).
10. J. A. Baldwin, Linked medical information systems, *Proc. R. Soc.* **184**, 403–420 (1973).
11. P. Beauchamp, H. Charbonneau and B. Desjardins, La reconstitution automatique des familles; un fait acquis, dans la mesure des phénomènes démographiques, Homage à Louis Henry, Popul 1977, numéro spécial (mars 1977).
12. M. E. Smith, Record linkage of hospital admission–separation records, Chalk River Nuclear Laboratories, Chalk River, Ont. Publication No. AECL-4507 (Sept. 1973).
13. M. E. Smith and H. B. Newcombe, Accuracies of computer versus manual linkages of routine health records, *Meth. Inf. Med.* **18**, 89–97 (1979).
14. G. Wagner and H. B. Newcombe, Record linkage: Its methodology and application in data processing (a bibliography), *Meth. Inf. Med.* **9**, 121–138 (1970).

About the Author—Howard B. Newcombe, B.Sc. (Acadia University 1935), Ph.D., D.Sc., F.R.S.C. Born 1914. Dr. Newcombe was a Research Scholar at the John Innes Horticultural Institute in 1939 and after wartime service as a Lieutenant, R.N.V.R., 1941–46, from 1947–79 was Head of the Biology Branch and later Population Research Branch, Atomic Energy of Canada Limited, Chalk River, Ontario. He was Visiting Professor of Genetics to the University of Indiana in 1963, Member of the International Commission on Radiological Protection and is the author of numerous scientific papers (mutations in microorganisms; effects of ionizing radiations; methods of study of human population genetics).

Dr. Newcombe is a Past President of the American Society of Human Genetics and the Genetics Society of Canada. At the present time he is Consultant to Eldorado Nuclear Limited and Statistics Canada.

About the Author—MARTHA SMITH received her B.Sc. from the University of Manitoba and her M.Sc. in Computing and Information Science from Queen's University in 1973. She was employed for several years in the Biology and Health Physics Division at Atomic Energy of Canada Limited, working with Dr. H. B. Newcombe on the British Columbia Record Linkage Study. This work involved developing new computer record linkage techniques for studying the effects of radiation on human populations. In 1978 she joined Statistics Canada and is currently Head of the Occupational and Environmental Health Research Unit. She is involved in planning and setting up some of the national files and facilities required to do long-term medical follow-up studies.

About the Author—GEOFFREY R. HOWE, B.Sc. (University College, London 1965), Ph.D. 1969. Born 1942. Dr. Howe was initially a Research Chemist with I.C.I. in England. He has subsequently been Research Fellow at Brock University and is now Senior Biostatistician to the N.C.I.C. Epidemiology Unit, University of Toronto. In addition, Dr. Howe is Professor in the Department of Preventive Medicine and Biostatistics at the University of Toronto and a Faculty Member of the School of Graduate Studies, University of Toronto. He is the author of numerous scientific papers, mostly on epidemiology and computerized record linkage.

He is a Fellow of the Chemical Society of London, Consultant to Eldorado Nuclear Limited and Atomic Energy of Canada Limited, Member of the American Statistical Association, Biometric Society and the Scoiety for Epidemiologic Research.

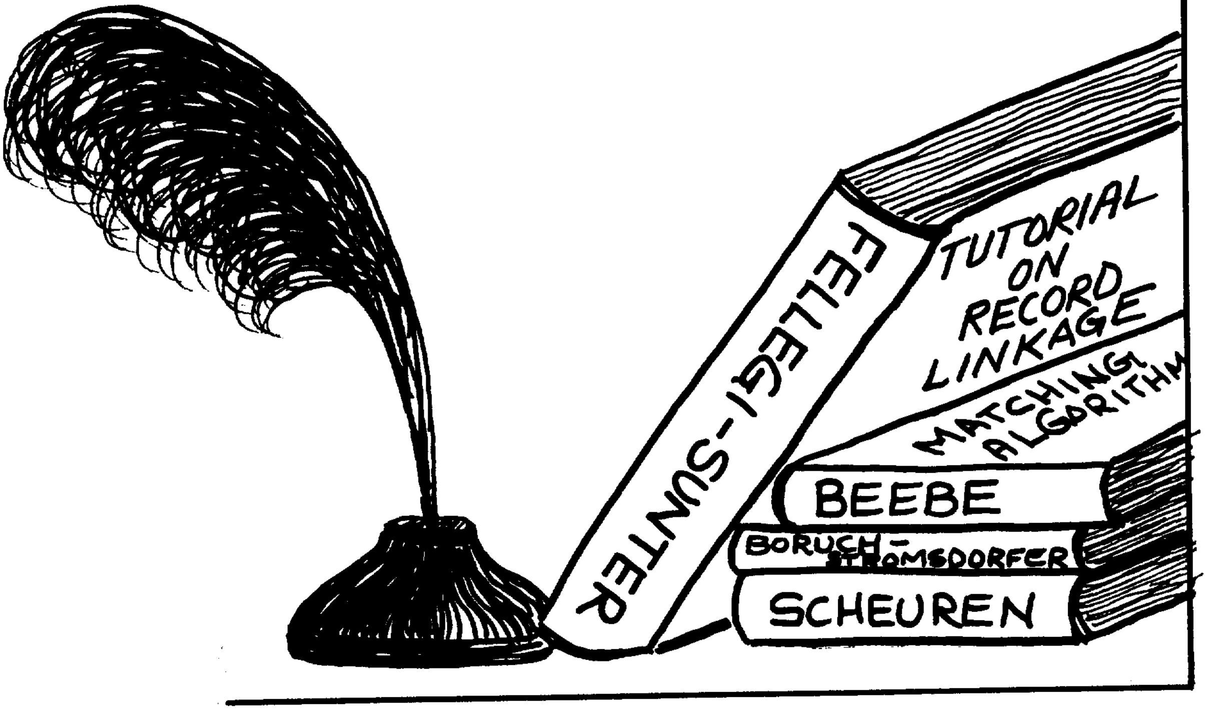
About the Author—JANE MINGAY received her Bachelor of Journalism degree in 1977 from Carleton University, having received practical experience in journalism. She subsequently worked on contract on a number of data collection and editing projects including medically oriented studies. After one of these projects organizing an historical research project for Eldorado Nuclear Limited, she became an Occupational Health Researcher on the E.N.L. Epidemiology Project. She is now on the staff of Maclean Hunter.

About the Author—ARLENE STRUGNELL graduated from business college in Montreal, Quebec in 1965 and worked as Secretary in the Department of Meteorology, McGill University until 1971. After working for a number of years in Toronto and Belleville, Ontario, she subsequently moved to Ottawa and is presently Research Assistant with the Epidemiology Project at Eldorado Nuclear Limited.

About the Author—JOHN D. ABBATT, B.Sc., M.B., Ch.B. (University of Edinburgh 1945), D.M.R., C.C.B.O.M. Born 1923. After wartime service with R.A.F.V.R. and hospital appointments in Edinburgh was Member of the U.K. M.R.C. External Scientific Staff at Hammersmith Hospital and Consultant Radiotherapist. After subsequent service as a Canadian Federal Civil Servant, he retired as Director General of Laboratory Centre for Disease Control, D.N.H.&W. and is now Medical Adviser to Eldorado Nuclear Limited, Ottawa.

Author of numerous scientific papers on the early applications of nuclear medicine and the therapeutic effects of radiation in man and animals, followed by epidemiological studies on human radiation effects.

Section II: Overview of Applications and Introduction to Theory



TUTORIAL ON THE FELLEGI-SUNTER MODEL FOR RECORD LINKAGE

Ivan P. Fellegi, Statistics Canada

EDITORS' NOTE

The following exhibits, numbered 1 to 22, were used at the Workshop on Exact Matching Methodologies (in the form of transparencies) as the basis for a presentation of the essential features and some of the consequences of the Fellegi-Sunter model and theory for record linkage. Many Workshop participants commented favorably on the exhibits and requested copies. The exhibits are presented here, without additional commentary, for the benefit of those who would like to have a convenient summary of the main points. The following chart shows the relationship between groups of exhibits and specific sections of the article, "A Theory for Record Linkage," which can be found on pages 51-78 of this volume.

Figure 1.--Exhibits for Fellegi-Sunter Article

Exhibit Numbers	Topic	Section of Article	Pages
1 to 6, 7a	Basic model and theory	2	52-57
7b, 8 to 10	Method of constructing an optimum linkage rule; consequences	2.1	54-57
11 to 14	Assumptions used in estimating weights	3.2	57-59
15 to 17	Calculation of weights, Method I	3.3.1	60-62
18	Calculation of weights, Method II	3.3.2	62-63
19, 20	Blocking	3.4	64-65
21	Choice of comparison space	3.6	66-67
22	Calculation of threshold values	3.7	67-68

Exhibit 1

Two sets of units: $A = \{a\}, \quad B = \{b\}$

Vector of characteristics $\alpha(a)$, $\beta(b)$ associated with units.

$L_A = \{\alpha(a);\ a \varepsilon A\}, \qquad L_B = \{\beta(b);\ b \varepsilon A\}$ (lists)

$L_A \times L_B = M + U$

where $M = \{[\alpha(a), \beta(b)];\ a = b,\ a \varepsilon A,\ b \varepsilon B\}$

$U = \{[\alpha(a), \beta(b)];\ a \neq b,\ a \varepsilon A,\ b \varepsilon B\}$

$L_A \times L_B$ unmanageable.

Exhibit 2

Code results of comparing $\alpha(a)$, $\beta(b)$: $\gamma(a, b)$

$\gamma[\alpha(a), \beta(b)] = \gamma(a, b) = (\gamma^1, \gamma^2, \ldots, \gamma^k)(a, b)$

Examples: $\gamma_i = 0$ if sex is same

1 if sex is different

Exhibit 3

γ_j = 0 if name is same and is Brown

1 if name is same and is Smith

2 if name is same and is Jones

3 if name is same and not Brown, Smith, Jones

4 if name is different

5 if name is missing on either record

$\Gamma = \left\{ \gamma(a, b) \right\}$: comparison space.

Exhibit 4

Linkage rule: decision regarding match status of (a, b) based on $\gamma(a, b)$

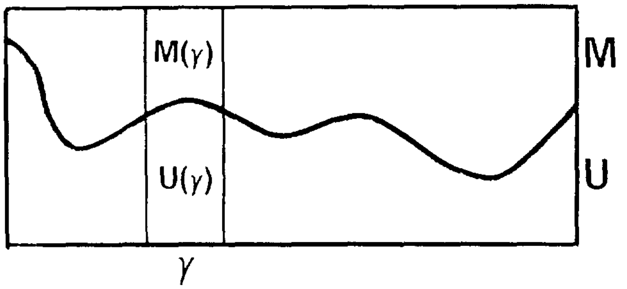
$d(\gamma) = A_1$: link (inference is “match”)

$d(\gamma) = A_2$: possible link (“don’t know”)

$d(\gamma) = A_3$: non-link (inference is “unmatched”)

Exhibit 5

$\gamma(a, b) = \gamma_0$ is a subset of $L_A \times L_B$

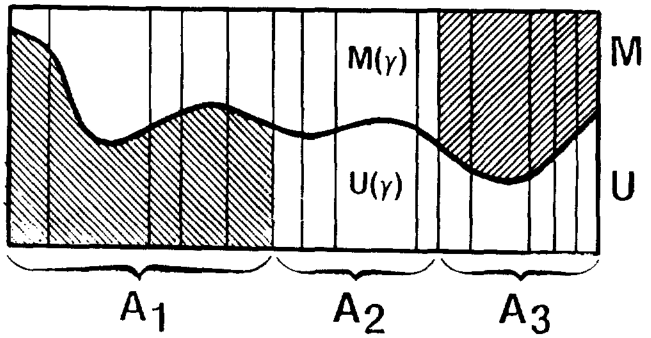


$$m(\gamma) = P\{\gamma(a, b) \mid (a, b) \varepsilon M\} = \frac{\| M(\gamma) \|}{\| M \|}$$

$$u(\gamma) = P\{\gamma(a, b) \mid (a, b) \varepsilon U\} = \frac{\| U(\gamma) \|}{\| U \|}$$

Exhibit 6

A linkage rule partitions $L_A \times L_B$:



For any $\gamma \varepsilon A_1$ all record pairs in $U(\gamma)$ are linked in error.

$\mu = P(A_1 \mid U) = \sum_{\gamma \varepsilon A_1} u(\gamma)$ proportion of linked record pairs in U

$\lambda = P(A_3 \mid M) = \sum_{\gamma \varepsilon A_3} m(\gamma)$ proportion of unlinked record pairs in M

Exhibit 7

a) Definition: Consider all linkage rules R on Γ with error levels μ_0, λ_0. Then R^1 is optimal if $P(A_2 | R^1) \leqq P(A_2 | R)$ for all R.

b) Heuristic: arrange $L_A \times L_B$ so that $m(\gamma)$ monotone decreases and $u(\gamma)$ increases. Choose A_1, A_3 to correspond to desired μ, λ. Then this linkage rule is optimal.

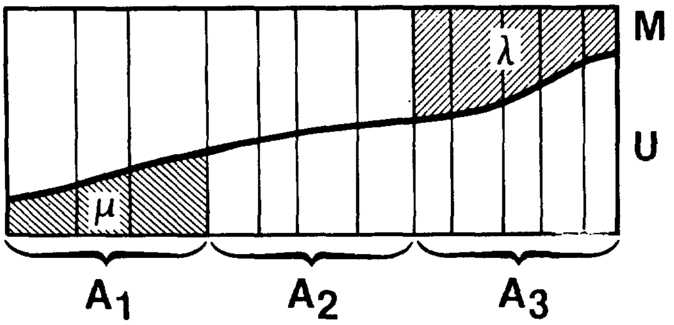


Exhibit 8

Optimal rule: order γ by decreasing values of $m(\gamma)/u(\gamma)$.

A_1 if $T_\mu \leqq m(\gamma)/u(\gamma)$

A_2 if $T_\lambda < m(\gamma)/u(\gamma) < T_\mu$

A_3 if $m(\gamma)/u(\gamma) \leqq T_\lambda$

T_μ chosen so that $\mu = \mu_0$, T_λ so that $\lambda = \lambda_0$

Likelihood ratio tests: A_1 at level μ, A_3 at level λ.

Uniformly most powerful.

Tepping's test (JASA, 1968) functionally equivalent.

Exhibit 9

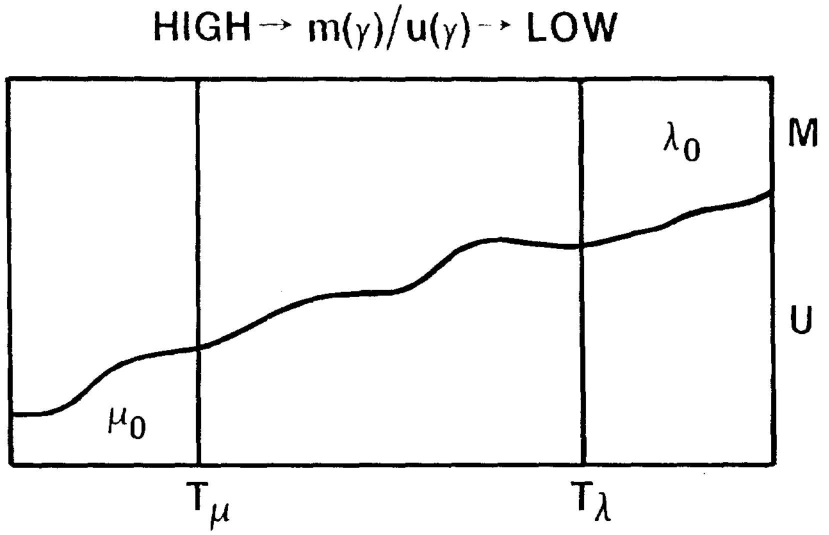


Exhibit 10

1. Trade-off between decreasing μ_0, λ_0 <u>or</u> A_2

2. A_2 can be eliminated if $T_\mu = T_\lambda$

3. Typically $\mu_0 << \lambda_0$ should hold. If N is the number of matched record pairs, $(N_A N_B - N)$ the number of unmatched record pairs, then condition for number of linked record pairs to be N is

$$N(1 - \lambda_0) + (N_A N_B - N)\mu_0 = N.$$

$$\text{True if } \mu_0 = \frac{N}{N_A N_B - N} \lambda_0$$

4. Randomized decision may be needed to achieve $\mu = \mu_0$, $\lambda = \lambda_0$ <u>exactly</u>.

Exhibit 11

Estimating m/u

If $\gamma = (\gamma^1, \gamma^2, \ldots, \gamma^K)$

γ^k has n_k values

then γ has $n_1 \cdot n_2 \ldots n_K$ values.

Simplifying assumption:

$$m(\gamma) = m(\gamma^1).\ m(\gamma^2) \ldots m(\gamma^K)$$

$$u(\gamma) = u(\gamma^1).\ u(\gamma^2) \ldots u(\gamma^K)$$

Components of γ are conditionally independent w.r. to m and u.

Exhibit 12

Matched records: Without errors, all γ^k should show "agreement". Hence independence → errors in different ident. variables of a and b are independent.

Unmatched records: accidental agreement on one variable (e.g. name) is independent of accidental agreement on another (e.g. address).

Estimands: $m(\gamma^1), m(\gamma^2), \ldots m(\gamma^K)$ -- $n_1 + n_2 + \ldots + n_K$

(also for u).

Exhibit 13

Need care in defining γ :

$$\gamma^1 = \begin{cases} \text{agreement on female given name} \\ \text{agreement on male given name} \\ \text{disagreement on given name} \\ \text{given name missing on either record} \end{cases}$$

$$\gamma^2 = \begin{cases} \text{agreement on sex} \\ \text{disagreement on sex} \\ \text{sex missing on either record} \end{cases}$$

Accidental agreement on $\gamma^1 \rightarrow$ agreement on γ^2. Independence might hold if first two codes of γ^1 combined.

Exhibit 14

Prefer to use log (m/u) - monotone incr. function of (m/u).

$\log (m/u) = w^1 + w^2 + \ldots + w^k$ where

$$w^k = \log \left[m(\gamma^k) / u(\gamma^k) \right]$$

We have

$$w^k \gtrless 0 \quad \text{if} \quad m(\gamma^k) \gtrless u(\gamma^k)$$

(intuitively appealing).

Similar to Newcombe-Kennedy (Communications of ACM, 1962).

Exhibit 15

METHOD 1 FOR WEIGHT CALCULATION (ILLUSTRATION)

Weights for "name" component.

Let proportions of different names in A, B and $A \cap B$ be

$p_A(i)$, $p_B(i)$, $p(i)$ ($\Sigma\, p - 1$). For simplicity:

$$p_A(i) - p_A(i) - p(i)$$

e_A, e_B: prob. of misreporting name in A, B respectively

p observable, e separately to be estimated.

Exhibit 16

w (agreement on jth name) $\approx$ log $(1/p_j)$

- Positive
- The smaller p(j), the larger w
- I.e. large positive weight for agreement on rare characteristic

w(agreement) $\approx$ log (1/p) where $p - \sum_j p_j^2$

- Large for uniformly well discriminating variable
- p decreases fast if common outcomes are separated.

Exhibit 17

$$w\ (\text{disagreement}) = \log \frac{e_A + e_B}{1-p}$$

- Typically negative
- The smaller the error, the larger the negative weight
- I.e. disagreement on well reported variable $\longrightarrow$ large negative weight
- E.g.: sex. Don't restrict linkage variables to high discrimination.

$$w\ (\text{name missing on either file}) = 0$$

- neutral contribution.

Exhibit 18. SECOND METHOD (ILLUSTRATION)

Assume only three components; each coded to two states: "agreement", "disagreement".

Conditional probabilities of "agreement" are m_h, u_h.

$$N_A N_B U_h = N\, m_h + (N_A N_B - N)\, u_h \qquad h = 1,2,3$$

where U_h: proportion of record pairs with "agreement" in h-th component.

U_h, N_A, N_B observable; N, m_h, u_h unknown.

Above 3 equations can be supplemented by other 4; all involve observable quantities + 7 unknown variables.

Solvable; generalizable; heavy dependence on independence.

Exhibit 19

Blocking

Objective: reduce number of comparisons.

Implicit assumption: comparisons not made are non-linked (A_3).

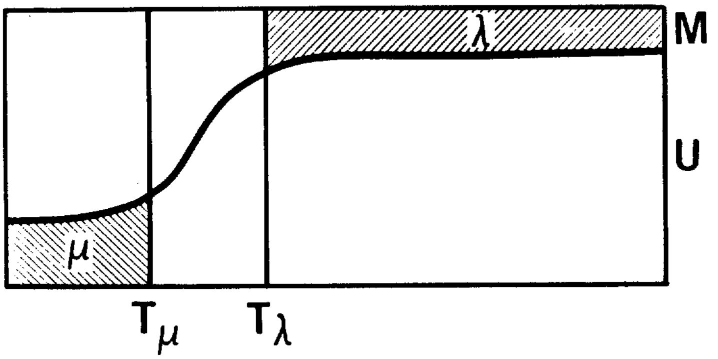


Exhibit 20. IDEAL BLOCKING VARIABLE

1. If a variable is such that disagreement results in very large negative weight -- corresponding e_A, e_B very small. Does not increase λ.

2. High discrimation results in maximum file blocking (comparisons restricted to records which agree on the blocking variable).

Frequent compromise: coded name where code is designed to reduce impact of misspellings.

Additional use of any well reported variable, even of low discrimination (e.g. sex), is net bonus.

Exhibit 21. CHOICE OF COMPARISON SPACE

1. How many separate values to recognize for agreement?

   Trade-off between complexity and reduction in Σp_j^2

2. How many of the variables common to both files should we use?

   Generally: the more the better.

3. w is positive for agreement, negative for disagreement almost certainly.

4. If $e_A + e_B < \frac{1}{2} < 1\text{-}p$, then each additional variable increases total weight for matched records, decreases total weight for unmatched records -- both with probability $> \frac{1}{2}$.

Exhibit 22. ESTIMATING THRESHOLDS

1. Select at random one value of each γ^k. Higher probabilities for high $|w|$;

2. Combine into γ; compute corresponding weight (w);

3. Repeat n times;

4. Arrange γ by decreasing w;

5. Set T_μ , T_λ as in Γ, but counting each γ with inverse of probability of selection.

WHY ARE EPIDEMIOLOGISTS INTERESTED IN MATCHING ALGORITHMS?

Gilbert W. Beebe, National Cancer Institute

INTRODUCTION

Both public and scientific concerns about hazards to health determine the agenda of epidemiology. The more we learn about health hazards the more there is to be learned, it seems, and the more the public comes to recognize health hazards the more it demands risk identification, risk estimates, and control measures. In recent decades new chemicals have been entering the environment at a very rapid pace. Under the Toxic Substances Control Act [1], passed in 1976, the Environmental Protection Agency (EPA) has been receiving over 1,000 pre-manufacture notices annually. There is now a list of about 30 chemicals and industrial processes recognized by the International Agency for Research on Cancer (IARC) as carcinogens for man, and another 61 thought to be probable carcinogens [2]. Another 103 are known to be carcinogenic for experimental animals, but IARC has reviewed only somewhat more than 600 chemicals and industrial processes on which there is adequate published information. I think we must assume that the carcinogens for man are far from identified and that the pace of industrial change exceeds our capacity for refined etiologic studies. We need inexpensive surveillance systems that will tell us where to look for significant hazards to health, and we need alert medical practitioners and industrial physicians to spot the unusual and unexpected [3].

The public is increasingly concerned with risks of a size that would have passed unnoticed in earlier years, risks associated with ionizing radiation, foods, drugs, toxic wastes, non-ionizing radiation, and the quality of our air and water. The MMR vaccine against measles, mumps, and rubella may cause brain damage in only one in a million vaccinees, but this risk is now sufficient to discourage manufacture of the vaccine because of the burden of litigation [4]. To identify small risks requires large samples, which in some instances may not be possible.

Ours has been aptly called an information society. Our capacity for recording, storing, transmitting, and manipulating information has been growing by leaps and bounds under the impetus of the computer revolution. I commend to you the recent (26 April 1985) computer issue of Science. The epidemiologist contributes to our understanding by bringing together for examination facts about individuals derived from different contexts. Increasingly these facts, or leads to them, are to be found in computer files. And since his unit of study is generally the individual, the epidemiologist wants to link files, which means matching, and to transfer data from files other than his own. And when he matches files he wants to be sure he is identifying the same person in each file.

In the U.S. we are experiencing a budgetary crunch. Funds for research are being reduced and staffs are being cut. The use of administrative records in research through record linkage, which means computer matching, is often the most economical way of obtaining information. For reasons of economy alone we should be looking more to record linkage as an adjunct to the more expensive procedures that we may have been following.

THE SPECTRUM OF EPIDEMIOLOGIC INTERESTS

The following illustrations are drawn from the field of chronic disease epidemiology with which I am more familiar, but record-matching routines are also of interest to epidemiologists working in the infectious diseases.

Etiology. -- (1) The cause of multiple sclerosis remains an enigma but epidemiologists are developing a great deal of information on differentials in risk; and (2) we may be getting closer to an understanding of the role of viruses in human cancer. There are animal cancers of known viral etiology and several human cancers are now being linked to viruses.

Risk Estimation. -- (1) There is a widespread desire to know the carcinogenic risk of exposure to low doses of ionizing radiation; and (2) we are interested in the hazards of certain prescription drugs such as oral contraceptives.

Value of Early Diagnosis. -- A prime example is breast cancer. At issue is the value of a screening regimen that includes mammography.

Prevention of Disease. -- (1) Epidemiologists are involved in intervention trials to prevent coronary heart disease, as illustrated by the Multiple Risk Factor Intervention Trial (MRFIT) program of the National Heart, Lung and Blood Institute; and (2) numerous intervention trials are also being conducted against cancer; for example, the National Cancer Institute (NCI) has trials in righ-risk areas of China where micronutrients, principally vitamins, beta-carotene, and minerals, are being prescribed on a controlled basis.

Treatment. -- Breast cancer is a recent example. At issue are the extent of the surgery and the value of adjuvant drugs and radiation.

Natural History. -- Acquired Immune Deficiency Syndrome, or AIDS, is a current example.

RECORD LINKAGE

Whether epidemiologists are working retrospectively or prospectively, in case-control or cohort mode, or are testing hypotheses or generating new ones, they are typically trying to link together, within the lives of individuals, events that are displaced in time and independently recorded. This underlies our dependence on record linkage; i.e., on matching and data-transfer. Matching requires rules of agreement, an algorithm, whether it be done manually or electronically.

Epidemiologists create their files from their own observations and from such records as are available to them. Often they must reach out to administrative record files of large organizations such as medical care providers, insurers, state government agencies, and even the Federal agencies, for some of the facts they need to complete the history of the individual subject. It may even be necessary, for example, to go to the Internal Revenue Service (IRS) to obtain addresses needed to locate subjects for examination or interview.

Agencies with large files tailor their matching algorithms to the identifying information they characteristically deal with and understand. One cannot, for example, go to IRS for an address or to the Social Security Administration (SSA) for a mortality check, without a social security account number. The Health Care Finance Administration (HCFA), on the other hand, can search its files for addresses on the basis of a name and date of birth, after first passing the incoming file through a nominal index file that provides the SSNs essential for the address search of its Medicare file. The Veterans Administration (VA) has a very flexible approach to matching with algorithms that will work on almost any variable or combination of variables the requestor may provide. Epidemiologists often do not have any number other than the date of birth, and lack of a SSN will often keep Federal agency files beyond their reach.

Matching algorithms must depend on the identifiers available but they also reflect the scientific imagination and experience of those responsible for the programming. Newcombe has stressed the importance of experience in the manual matching of representative records as preparation for designing programs for matching by computer. He also emphasizes the value of redundancy in identifying variables when matching is involved. It was his 1959 paper, more than any other single contribution, I believe, that paved the way for technically adequate machine matching in the absence of a central ID number like the SSN [5]. With a number like the SSN it is possible to insist on an exact match. Even though the SSN is not precisely a unique number and lacks a check digit, it is nevertheless a very good number in most situations requiring linkage. If you transpose digits of your SSN in your tax return you will soon receive a query from the IRS. Names may be abbreviated to 4-6 letters of the surname if main reliance is placed on the SSN, but in other contexts the surname may be coded phonetically in New York State Identification and Intelligence System (NYSIIS) or Soundex fashion.

The investigator wants the benefit of a matching algorithm that minimizes both false positive and false negative matches but he may have no idea of the false negative rate in the absence of formal tests such as are being made on the National Death Index of the National Center for Health Statistics (NCHS) [6]. If the false positives are frequent, and in some applications NCHS algorithms have returned two false positives for each true positive match, the consumer may be hard put to evaluate the output without a weighting scheme such as Newcombe has devised.

Record linkage is now often being required on such large files that matching must be performed electronically or not at all. One cannot think of the IRS file of individual taxpayers being searched for addresses in any fashion except electronically. I am told the file contains 155 million records and takes three weeks to run. And if you want to locate a large roster of subjects under age 65 and 20-40 years after some occupational exposure, alternative sources of addresses would probably be expensive and inefficient.

THE BACKGROUND OF MY OWN INTEREST

From the medical experience of World War II came the suggestion, by Dr. Michael E. DeBakey, the heart surgeon, that a medical research program be established to follow up the injuries and diseases of the war [7]. We both served as staff for a committee of the National Research Council (NRC) that looked into his idea and I wound up in charge of the statistical work of the group known today as the Medical Follow-up Agency of the NRC. Knowing that work with records would be a large part of the effort, one of the first persons I hired was Nona-Murray Lucke. She had been working with Dr. Halbert Dunn, then director of the Vital Statistics Division of the Bureau of the Census and originator of the term "record linkage," on his scheme for matching birth and death records at the state level [8]. Although there were Army punchcard indices to the entire medical experience of the war, the cards contained Army serial numbers but not names. A manual look-up was required to obtain the corresponding names that we could then match to the nominal VA Master Index in order to find VA claim numbers and to locate the offices having custody of the hard-copy VA files. All the linkage was manual, but usually there was enough detail beyond name and Army serial number to rule out misidentification. Identification was a problem in only about 2-4 per cent of the cases and records were unavailable in less than one percent. Starting in 1972 we benefitted from automation of the VA Master Index, now the Beneficiary Identification and Records Locator Subsystem (BIRLS) file, as well as from the automated record systems for hospital discharges and for compensation and pension status. Tape-to-tape matching has long been the rule. But the detailed medical records, not only those of World War II but also those generated today as well, are available only in hard copy.

One of the matching efforts I personally directed was a test of the completeness of VA information on the mortality of war veterans, matching known deaths obtained from NCHS against the military files in St. Louis to determine veteran status, and then submitting the resulting file intermingled with living veterans to the VA for a blind search [9]. We learned that the VA had about 95 percent of the mortality information on WW II veterans.

At the Atomic Bomb Casualty Commission (ABCC) in Japan, where I directed the epidemiologic and statistical work for some years, we followed two main samples of 55,000 and 110,000 for mortality, using the Japanese family registration system devised in 1871 [10]. Each Japanese citizen has a place of family residence (his honseki), and the city office for that place keeps a running family record, the koseki, that shows vital events for all the family members, no matter where in Japan these events take place or where the individuals live. The koseki tells where any death certificate is retained and for the cause of death one must go there. To enter the system both the name and the honseki must be known. There is very little slippage in this system, but it is manually operated. At ABCC mortality was checked every three years on a rotational scheme that levelled out the workload.

An interesting matching problem arose in the late 1950's when I first went to Japan. The U.S.-Japan Joint Commission had created a file of about 14,000 records of its medical investigations in 1945 that were stored at the Armed Forces Institute of Pathology (AFIP) in Washington. To recapture the 1945 observations for the ABCC files we obtained blow-ups of microfilm copies retained at AFIP. For the Hiroshima portion of the sample, names were written in the Romanized fashion, not in the Japanese ideographs, or kanji. Location at the time of the bomb was given in terms of a numbered radial zone and the direction from the hypocenter, not in terms of a postal address, and age was usually given in the Japanese style which is equivalent to the western style plus one year. That is, in Japan, children are one year old at birth. Under Seymour Jablon's supervision this file was later matched to the ABCC records so that the 1945 data could be added to the ABCC files that represented largely individuals alive in 1950. About 42 percent could be matched, largely because of the considerable ancillary detail on both record sources. The false negatives could not be assessed but tests showed that the false positives probably numbered no more than 5 percent. The matching rate in Nagasaki, for which the records did contain the name in kanji and the postal address, was higher, 60 percent.

At the National Institutes of Health I have also been very much concerned with record linkage, trying to make it easier to link some of the large files of Federal agencies in the furtherance of medical research [11]. We need to restore access to the IRS address file for a broader class of investigators than just National Institute for Occupational Safety and Health (NIOSH) investigators who are concerned with occupational health, and Federal investigators studying the occupational hazards of military service, these being the privileged classes under current law. We also need to restore the kind of freedom we had before the Tax Reform Act of 1976, when SSA was willing to define industrial employment cohorts and determine their mortality. With Dr. Scheuren's help I have been trying to learn how to strengthen the Continuous Work History Sample of SSA so that it might provide some national mortality data by both industry and occupation. In addition, I'm engaged in a research project that has involved extensive matching to the files of the VA, IRS, and HCFA.

POSSIBLE LIMITATIONS OF COMPUTER-LINKED DATA

If the only observations available to the epidemiologist derive from the linkage of administrative files, his study may be useful for screening a large experience or for developing working hypotheses, but it will probably not illuminate the meaningful aspects of exposure or define end-points precisely. If we link files as part of a larger process, e.g., to obtain addresses so that we can examine or interview subjects, or to learn that deaths have occurred and where we can find the death certificates, such limitations do not apply. Even as an index to hard-copy records, however, a large computer file may prove disappointing: recently I found that a VA diagnostic index I must depend on contains so much coding error for the cancer I am investigating that I will have to review the underlying hard-copy records for validity of diagnosis.

LANDMARK STUDIES BASED ON MATCHING RECORDS

Any list of landmark studies is bound to be very selective and the following is further limited by my own reading and knowledge of the field:

- Framingham Heart Study [12];
- Follow-up Studies of War Injuries and Diseases, and Registry of Veteran Twin Pairs, NRC Follow-up Agency [7];
- Mancuso's Studies of Occupational Risks Based on Industrial Employment Rosters of the SSA [13];
- Studies of A-bomb Survivors in Japan [10];
- Court-Brown and Doll's Study of Ankylosing Spondylitis Patients Treated by X Ray [14];
- Dorn's Study of the Health Effects of Smoking, WW I Veterans [15];
- Oxford Record Linkage Project [16];
- Selikoff's Study of Asbestos Workers [17];
- The Mayo Clinic Studies of Olmstead County, Minnesota [18];
- The Canadian Studies of Newcombe, Statistics Canada, and the National Cancer Institute of Canada [19]; and
- The British Office of Population Surveys and Statistics Longitudinal Study [20].

SOME OF THE LARGER COMPUTER FILES OF INTEREST TO THE EPIDEMIOLOGIST

It would be fruitless to enumerate all the files used by epidemiologists but generated independently of their own efforts. They cover a wide range of classes: employment, medical care, vital records, finance, life insurance, disability, city directories, licensing, etc. But some examples follow in Table 1.

Table 1. Some Large Files Used by Epidemiologists

Name of File	Millions of Records
IRS, File of Individual Taxpayers	155
SSA, Master Beneficiary Record (MBR File)	35-40
HCFA, Medicare Beneficiaries	30
VA, BIRLS	35
National Archives Records Agency, "Registry" File of Military Records in National Personnel Records Center, St. Louis	30
NCHS, National Death Index	10
SSA, File of Deceased	30
California Automated Mortality Linkage System (CAMLIS)	3.6
Army WW II Hospital Diagnosis Index	12

SOME CURRENT EPIDEMIOLOGIC STUDIES TAPPING LARGE COMPUTER FILES

Apart from current studies that are already represented on our program today, some that I am particularly familiar with include:

The Johns Hopkins Study of Nuclear Shipyard Workers. -- The investigators are sampling the 700,000 nuclear shipyard worker population, stratifying on radiation dose, and seeking to relate cause of death to radiation dose, demographic characteristics, occupation, and other specific risk factors. External linkage has been established with the VA BIRLS file, the SSA MBR file, state death files, the NDI file of NCHS, and OPM files. In addition there is considerable internal file linkage to unduplicate the eight yards and to update study files with radiation dose, job classification, and the like. About 90,000 deaths have been ascertained.

Study of X-Ray Technologists. -- The NCI Radiation Epidemiology Branch has initiated a study, together with NIOSH investigators and epidemiologists of the University of Minnesota, of about 160,000 x-ray technologists in the U.S. whose exposure has long been monitored by radiation badges. Investigative interest centers not only on the carcinogenic effect of low doses of radiation, but also on the highly fractionated character of their exposure. Linkage will involve the SSA MBR file, the NDI file of the NCHS, the HCFA Medicare file, the IRS address file, and possibly other files.

Hepatitis B Virus and Primary Liver Cancer. -- In the NCI Clinical Epidemiology Branch I am doing a study with 6 VA hospitals and the Medical Follow-up Agency of the National Research Council to learn whether the contaminated yellow fever vaccine that led to 50,000 cases of acute hepatitis in the Army in 1942 has also produced excess liver cancer among the vaccinees. Record linkage has involved the Army World War II diagnostic index, the National Archives "Registry" file in St. Louis, the VA BIRLS file, the IRS address file, and the HCFA Medicare file. About 60,000 men are under study.

Study of Atomic Veterans. -- The NRC Medical Follow-up Agency is completing a study of 50,000 "atomic veterans" exposed in weapons tests in the Pacific and at the Nevada Test Site. Rosters of exposed individuals assembled by the Department of Defense were linked with the VA BIRLS file, the VA Master Index (a microfilm file), the NDI file of NCHS, and various military service files. This is another low-dose study, stimulated by the earlier finding of some excess leukemia among men exposed to the Smoky shot.

Study of Cancer from Fallout from the Weapons Tests. -- Epidemiologists at the University of Utah, under a contract with the NCI, are studying leukemia and thyroid cancer among Utah residents downwind from the Nevada Test Site, trying to establish whether fallout from the atmospheric tests of the 1950's caused excess cancer. Linkage involves two files of the Church of Jesus Christ of Latter-Day Saints (Mormons), one of about two million members registered in church censuses, the other of 400,000 deceased members. Matching also extends to the state mortality files and to the population-based cancer registry in the state of Utah.

Health Effects of Agent Orange and Service in Vietnam. -- The Centers for Disease Control have under way a complex investigation of the effect of the exposure of servicemen to Agent Orange in the Vietnam War. A sample of about 30,000 men is under study and record linkage procedures involve the IRS address file, the SSA MBR file, the VA BIRLS file, and the NCHS NDI file.

OUTLOOK FOR THE FUTURE

I think we can expect the computer to play an ever larger role in future epidemiologic studies through record linkage. There will be no let-up in the demand of society to know its risks and to learn how to control them, and no let-up in the forward march of computer science. We can expect to find more and more data in computer files, with less dependence on them as mere indexes to hard-copy records. And matching algorithms will provide the key to the record linkage. But there are obstacles and there will be missed opportunities. Files that might have been useful for epidemiologic research may not be so because insufficient identifying information will have been collected. For the epidemiologist a critical item is often the social security number but SSA policy seems to be against its widespread use as concern for privacy and confidentiality has led to restraints on access to data that have been placed without regard for the special needs for epidemiologic information on health risks. These restraints are made doubly difficult to deal with by the fractionation of Federal statistical programs and responsibilities, each agency collecting its own statistics in support of its own narrow mission and having laws to limit access to its data. We might wish for a Statistics USA akin to Statistics Canada, but I doubt that day will ever come.

The concern for privacy stems in part from a public fear of "data banks" on the ground that they could too easily be misused. But record linkage need not imply the necessity for huge data banks. It requires only that communication be permitted between files on an ad hoc basis under restrictions that reflect the public interest in both privacy and adequacy of information.

REFERENCES

[1] PL 94-469, Oct. 11, 1976.

[2] Tomatis, L., "Exposure Associated with Cancer in Humans," J. Cancer Res. Clin. Oncol. 108:6-10, 1984.

[3] Miller, R.W., "The Alert Practitioner As a Cancer Etiologist," Cancer Bull. 29:183-185, 1977.

[4] Medical News, "AMA Offers Recommendations for Vaccine Injury Compensation," J. Am. Med. Assn. 252:2937-2946, 1984.

[5] Newcombe, H.B., Kennedy, J.M., Axford, S.J. and James, A.P., "Automatic Linkage of Vital Records," Science 130:954-959, 1959.

[6] Wentworth, D.N., Neaton, J.D. and Rasmussen, W.L., "An Evaluation of the Social Security Administration Master Beneficiary Record File and the National Death Index in the Ascertainment of Vital Status," Am. J. Public Health 73:1270-1274, 1983.

[7] DeBakey, M.E. and Beebe, G.W., "Medical Follow-up Studies on Veterans," J. Am. Med. Assn. 182:1103-1109, 1962.

[8] Dunn, H.L., "Record Linkage," Am. J. Public Health 36:1412-1416, 1946.

[9] Beebe, G.W. and Simon, A.H., "Ascertainment of Mortality in the U.S. Veteran Population," Am. J. Epidemiol. 89:636-643, 1969.

[10] Beebe, G.W., "Reflections on the Work of the Atomic Bomb Casualty Commission in Japan," Epidemiol. Rev. 1:184-210, 1979.

[11] Beebe, G.W., "Record Linkage and Needed Improvements in Existing Data Resources," Banbury Report 9, Cold Spring Harbor, New York, Cold Spring Harbor Laboratory, 1981, pp. 661-673.

[12] Dawber, T.R., Kannel, W.B. and Lyell, L.P., "An Approach to Longitudinal Studies in a Community: The Framingham Study," Ann. N.Y. Acad. Sci, 107:539-556, 1963.

[13] Mancuso, T.F. and Coulter, E.J., "Methods of Studying the Relation of Employment and Long-term Illness--Cohort Analysis," Am. J. Public Health 49:1525-1536, 1959.

[14] Court-Brown, W.M. and Doll, R., "Mortality from Cancer and Other Causes After Radiotherapy for Ankylosing Spondylitis," Brit. Med. J. 2: 1327-1332, 1965.

[15] Dorn, H.F., "The Mortality of Smokers and Nonsmokers," Proc. Soc. Statist. Sec. Am. Statist. Assoc., 1958, pp. 34-71.

[16] Acheson, E.D., "Medical Record Linkage," London, Oxford Univ. Press, 1967.

[17] Selikoff, I.J., "Cancer Risk of Asbestos Exposure," In Origins of Human Cancer (Hiatt, H.H., Watson, J.D. and Winsten, J.A., eds.), Cold Spring Harbor, New York, Cold Spring Harbor Laboratory, 1977, pp.1765-1784.

[18] Kurland, L.T. and Molgaard, C.A., "The Patient Record in Epidemiology," Sci. Am. 245:54-63, 1981.

[19] Howe, G.R. and Lindsay, J., "A Generalized Iterative Record Linkage Computer System for Use in Medical Follow-Up Studies," Comput. Biomed. Res. 14:327-340, 1981.

[20] Office of Population Censuses and Surveys, "Cohort Studies: New Developments," Studies in Medical and Population Subjects No. 25, London, Her Majesty's Stationery Office, 1973.

EXACT MATCHING OF MICRO DATA SETS IN SOCIAL RESEARCH: BENEFITS AND PROBLEMS

Robert Boruch, Northwestern University
Ernst Stromsdorfer, Washington State University

1. INTRODUCTION

The first objective here is to review some applied social research projects that have benefited from exact matching. The examples are merely illustrative but stem from a variety of disciplines.

The second objective is to discuss the negative aspects of matching. In particular, our argument is that, by espousing the opportunity to match too ardently, we may constrain or misdirect our ability to respond to other research issues and problems. An issue of special interest here is obtaining unbiased estimates of the effects of manpower projects.

The idea of matching records in the interest of science has a long pedigree. For instance, R.A. Fisher lectured at a Zurich public health congress in 1929, arguing the usefulness of public records supplemented by (and presumably linked with) family data, in human genetics research (Box, 1978, p. 237). Earlier, Alexander Graham Bell exploited geneological records, administrative records on marriages, census results and others, apparently linking some sources, to sustain his familial studies of deafness (Bruce, 1973; Bell, 1906).

2. HOW AND WHY HAS MATCHING BEEN HELPFUL

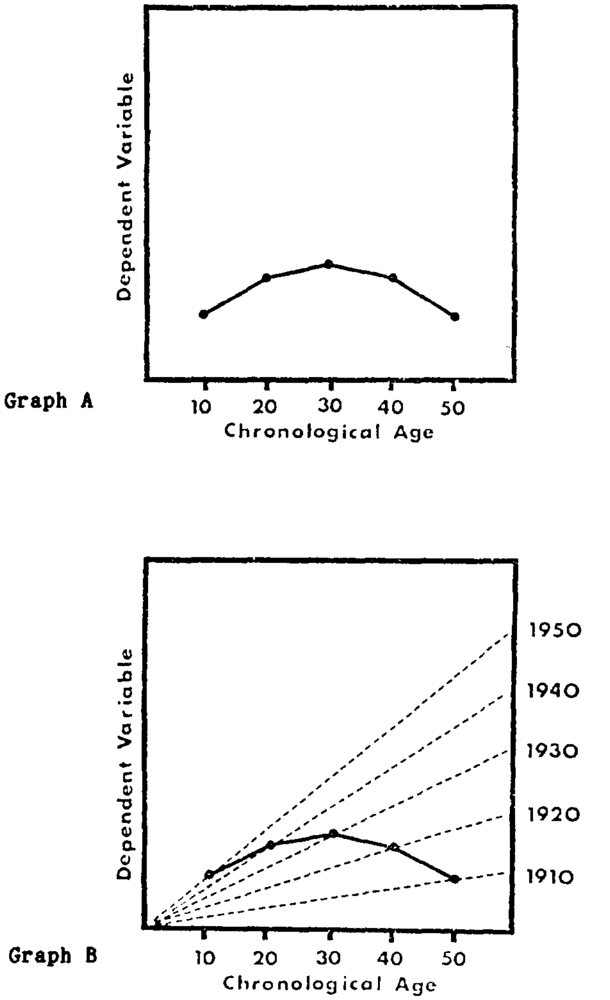
The fundamental reasons that matching has been useful do not differ appreciably from those implied by the above examples. Nor do the reasons differ much across the social and behavioral sciences. The following illustrations are taken from Boruch and Cecil (1979); unless otherwise noted, specific references are given there.

2.1 Matching to Understand Phenomena and Avoid Egregious Error

In psychology, for example, graphs of the sort used in Figure 1A were commonly used during the 1940's and 50's to describe the gradual increase in IQ with age, an IQ plateau and gradual decrease in IQ with age. The data are based on cross-sectional surveys.

The ability to match, as in linking individuals' records obtained at one point in time to those collected at another to generate longitudinal files, yielded an entirely different picture of behavior. This, given in Figure 1B, tells us that earlier declines in IQ are an artifact of cross-sectional studies and that cohort differences are important and account for the misleading interpretations of the earlier data.

Figure 1. Confounding of Age and Cohort Differences in Cross-sectional Research.



From: Boruch, R.F., and Cecil, J.S. <u>Assuring the Confidentiality of Social Research Data</u>. Philadelphia: University of Pennsylvania Press, 1979.

Lest you think the example confined to a quantitatively naive discipline, consider an economic example. Table 1, based on simple cross-sectional surveys, suggests that a graph similar to Type A is appropriate for earnings data as well as IQ data. Such earnings data were commonly used during the 60's to describe increases, plateau, and gradual decline in income. Table 2 gives cohort earnings obtained in longitudinal surveys, matching on individuals. It shows a different picture, one that is less dramatic and more similar to the Type B figure.

Studies that try to separate genetic and environmental influences in schizophrenia are bound to be more controversial. But they are important and worth pursuing... So, for example,

Table 1.--Estimates of Mean Annual Income in Dollars for Men Aged 25-64
(Data is based on independent samples taken in 1947, 1948, and 1949.)

Year	Age			
	25-34	35-44	45-54	55-64
1947	2,704	3,344	3,329	2,795
1948	2,898	3,508	3,378	2,946
1949	2,842	3,281	3,331	2,777

From: Boruch, R.F., and Cecil, J.S. Assuring the Confidentiality of Social Research Data Philadelphia: University of Pennsylvania Press, 1979.

Table 2.--Estimates of Mean Annual Income in Dollars Over Ten-Year Intervals for Six Cohorts

Year	Ages		
	25-34	35-44	45-54
1. 1947	2,704 (1947)	5,300 (1957)	8,342 (1967)
2. 1948	2,898 (1948)	5,433 (1958)	8,967 (1968)
3. 1949	2,842 (1949)	5,926 (1959)	9,873 (1969)
Year	Ages		
	35-44	45-54	55-64
4. 1947	3,344 (1947)	5,227 (1957)	7,004 (1967)
5. 1948	3,508 (1948)	5,345 (1958)	7,828 (1968)
6. 1949	3,281 (1949)	5,587 (1959)	8,405 (1969)

Note: Each cohort was surveyed every ten years. The first cohort, for example, contains individuals who were 25-34 years of age in 1947 and had an average income of $2704; in 1967, when they were 45-54 years of age, their mean income was $8342.

From: Boruch, R.F., and Cecil, J.S. Assuring the Confidentiality of Social Research Data. Philadelphia: University of Pennsylvania Press, 1979.

Danish-U.S. collaboration supported by the National Institute of Mental Health (NIMH) has involved intensive record matching to determine how children born of schizophrenic parents fare when they are adopted and reared by non-schizophrenic, foster parents. Matching among records of hospitals, surveys, and psychiatric systems was required to execute the research. The work appears to confirm a genetic component in that incidence of schizophrenia among such children is higher than its incidence among adopted children born of nonschizophrenic parents, including children adopted by schizophrenic parents.

That use of matched records can improve scientific analysis seems clear from studies of the economic impact of education. Paul Samuelson, for example, has argued that returns on higher education are substantial. Christopher Jencks has analyzed various survey data sets to argue that the returns are marginal. Fagerlind used Swedish data that were better than data available to either Samuelson or Jencks: matching individual records from military screening; birth registries, tax registries on earnings of the respondent, census records on occupational mobility. These analyses favor Samuelson's theory.

Neither the schizophrenic study nor the Samuelson-Jencks-Fagerlind work is unambiguous, of course. There has been considerable debate about the models exploited in each. The main point is that improvements in data, notably through linkage of records from a variety of sources, can enhance the analyst's ability to explore ideas and test hypotheses. The "sources" may be additional survey panels in a longitudinal design. Or they may be administrative records that are at least as good as survey data.

2.2 Matching to Avoid Aggregation Error and Ecological Fallacy

We often compute correlations between X and Y based on aggregate data, being cautious, of course, in generalizing to the individual level. The opportunity to match individual records often gives us the opportunity to entirely avoid the problems and caution engendered by aggregation.

One of the oldest illustrations is still the most dramatic. At a particular point in time, the correlation between literacy rate and color (black vs. white) computed on the basis of nine census regions in the United States was .95. When the data are aggregated by State instead of region, the correlation becomes .77. Finally, access to individual records led to a correlation of .20.

2.3 Matching Records in Randomized Tests of Social and Education Programs

In Middlestart education programs at Oberlin College, for instance, a series of experiments was undertaken to understand whether precollege programs worked for promising but poor adolescents. The evaluators relied on randomization to assure statistically unbiased estimates of long-run program effect. They relied on records matched among surveys, high school records, and standardized precollege records to avoid the problem of low validity in student reports of grades, and to enhance the statistical power of the tests.

Randomized field experiments, designed to understand how one can increase compliance with food stamp registration rules, have been mounted by the Office of Analysis and Evaluation of the U.S. Department of Agriculture's Food and Nutrition Service (1984). These tests depend on matches of records among participant reports and records of State Employment Security agencies and the Food Stamp Agency. Results show remarkable decreases in food stamp costs and employment benefits for certain innovative approaches to compliance assurance.

Police research is relevant, too, of course. In the Minneapolis Domestic Violence Experiments, the object was to understand how to handle domestic violence effectively, for example, immediate arrest versus referral to social services, within limits. Undertaken by the Police Foundation, the experiment involved matching among police patrolman records, precinct arrest records, and the experimenters' records. Arrest, incidentally, seems to work in the sense of reducing subsequent incidence of domestic violence (Sherman and Berk, 1984).

Motor vehicle research is pertinent to matching, too. Work done some years ago by the Insurance Institute for Highway Safety, for example, involved linking an experimenter's observations on vehicle registration, the drivers' seat belt use, and advertisements on the topic, to motor vehicle records that contained data on the drivers' residence area. The residence area match with the other information made it possible to determine how effective alternative TV commercials, directed to different areas, were in encouraging seat belt use.

Program Implementation and Validity of Reporting

The New Jersey Negative Income Tax Experiments attended to the potential problem of overpaying welfare recipients. This set a standard for validity studies in later experiments. Overpayment of benefits in such experiments was critical insofar as (a) other sources of assistance were available to participants in the experiment, and (b) they might receive such assistance illegitimately through error (welfare rules are complicated) or deceit (crime is still a bastion of the free enterprise system). All participants reported their income based on recall. Matching these reports with administrative records helped to assure reasonable implementation of the program and to assess quality of reporting.

For example, welfare audits were created to reduce or prevent the problems: these depended heavily on the experimenters' ability to match research records with records of welfare boards. Internal Revenue Service (IRS) W-2 forms were required of families and permitted comparisons between IRS-reported income and income reported to the experiment. (Underreports of income to the experiment relative to IRS appear to have been less than 15 per cent). The Social Security Administration (SSA) cooperated by taking the experimental data, matching to its own records on individuals, and providing aggregate earnings data (not individual records) to permit estimates of underreporting of earnings in the experiment (Kershaw and Fair, 1979). (The SSA comparison suggests that about 80% of families underreport to researchers by 15% or less even when they have incentives to misreport.)

In the Seattle and Denver Income Maintenance Experiments (SIME/DIME), research records were matched to public agency records on food stamp purchase, rent subsidy, and wages. The experiment produced some small surprises through evidence that public records on rent support and food stamps were less accurate than respondents' reports in the experiments, evidence that was later strengthened by independent investigation. Underreporting of wages appeared in the expected direction based on matches with IRS records (Halsey, 1980).

In the New Jersey Negative Income Tax Experiments, Mercer County Welfare Board records were used in a pilot test to determine composition, work history, and residential mobility of families that attrited from the experiment and could not be interviewed without great difficulty. More generally, the attrited families in five cities were traced through post office change-of-address cards, motor vehicle registration agencies, welfare boards, prisons, and community groups. Apparently, face-to-face interviews with former neighbors were most productive (Kershaw and Fair, 1979).

The use of administrative records to trace attriters and assess misreporting in all the income maintenance experiments is an important but underexamined topic. The experiments themselves were well run, relative to any pragmatic standard. They cover a sufficient number of sites to tantalize any scholar with an interest in regional differences in record accuracy, misreporting models and so on. Sample sizes for validity studies were small, however. This may account partly for the disinterest of scholars. Still, it is a bit distressing to some that otherwise thoughtful commentators such as Hausman and Wise (1985) fail to recognize the policy import of misreporting and the methodological contributions of randomized tests of economic programs to this area.

2.4 Matching and Testing New Ways to Elicit Information

Innovative ways to elicit information, such as randomized response, need to be tested despite their cleverness. We are unaware of any individual match studies in this arena. But studies that compare marginals or point estimates for individuals on whom both responses and archival records are available are done.

So, for example, Bradburn, Locander and Sudman found that a randomized response method worked at times to reduce response distortion on sensitive topics such as drunk driving arrests. The basis for comparison was administrative records on the same individuals, e.g., arrest records. Individual records were not matched; comparisons are based on marginal counts or averages. But matching in this and related research is possible in principle. And it may be useful insofar as it helps us to understand how response distortion varies with sensitivity of the traits that are being examined and characteristics of individual respondents.

A fascinating example of a near match study on reporting energy use to the Census Bureau was given by Tippett (1984) in recent 1984 Proceedings of the ASA. Her experiment involved encouraging utility companies to send a randomly assigned group of individuals a statement of the year's utility bills. A randomly assigned comparison group was not sent the statement. The statements were sent prior to the 1980 census to understand whether providing such records could enhance quality of respondents' reports of utility costs to Census. Both groups overstated costs; the "primed" group overstated costs appreciably less than the control group. Again, matching could be helpful in understanding how degree of reporting error varies with the true state of the individual.

2.5 Matching Records to Understand Validity of Response and Inferential Errors

We know that error in measurement of a response variable degrades statistical power. More important, it can lead to invidious biases in covariance analyses based on fallibly measured covariates. That is, the analyses can make programs look useless when their effects are in fact slightly positive, and can make programs look harmful when indeed they are merely useless (Riecken et al., 1974). The recent work by Andersen, Kasper, Frankel and their colleagues (1979) on total survey error clarifies the effect of imperfections in observational studies generally.

The point is that understanding validity of the measures is important in applied social research, especially policy research, as well as in basic work. Matching studies undertaken in education and supported by the National Institute of Education and the National Center for Education Statistics, for instance, show that females are appreciably more accurate than males in responding to questions about their own grades and coursework, and more accurate in reporting on income and education levels of parents. There are race differences as well as gender differences in respondents' ability and willingness to furnish information. Failure to recognize these differential validities can lead to errors in understanding which programs work and for whom. Matching helps us to avoid those errors merely by showing which subgroup differences in reporting quality may account for differences in performance.

Imperfect measures of employment and occupation can produce similar biases in explanatory models of income gain and other response variables. Matching studies of the sort undertaken by Mathiowetz and Duncan (1984) in which private employer records are linked to survey records of the Panel Study on Income Dynamics are not common. But they have potential for revising ideas about error structure. Errors in retrospective reporting on employment and occupation seem to depend less on time or recency than on salience of events in a particular month (e.g., a raise) and task difficulty (e.g., a single unemployment spell vs. multiple spells). Gender and race differences in reporting error are reduced when these variables are taken into account.

3. WHEN BENEFITS OF MATCHING ARE NEGATIVE OR AT LEAST NOT SO CLEAR

Having the option to capitalize on existing records and to match so as to obtain a better file is important because the idea and the relevant technology have been so useful. For instance, the 1984 Proceedings of the ASA, Section on Survey Research Methods contains over 30 articles that concern exact matching methods or analysis or depend heavily on matching for conclusions (validation studies, capture-recapture, others). Unlike the 1984 Proceedings, the 1978 Proceedings of the same section contained no sessions on using administrative records in conjunction with surveys or on quality control of statistical systems (partly through linkage).

The Interagency Linkage Study participants --Internal Revenue Service, Census, and Social Security Administration--deserve special credit for advances in this arena. Other agencies have worked at least as vigorously and as often, however, e.g., the National Center for Education Statistics and the National Center for Health Statistics. And a good many research projects undertaken with support of the U.S. Department of Labor's Employment and Training Administration, the National Institute of Justice, the National Center for Health Services Research (and the Department of Health and Human Services more generally) have made use of matching where it has been useful and legally possible to match.

Matching is a seductive option, however. That is, we may capitalize on matching existing records to obtain estimators that are efficient and cheaply produced, but wrong. They are wrong at times partly on account of the administrative system in which matching must take place. They are wrong partly because the matched data (observational data more generally) are inappropriate despite their accessibility and ostensible relevance.

Consider a recent case, one in which the role of matching is important.

3.1 The Case at Hand

Estimating the effect of manpower employment and training programs in this country is a significant policy issue. Since 1965 or so, most estimates have been based on observational data, i.e., sample surveys. Two kinds of observational data are most relevant here--the Continuous Longitudinal Manpower Survey (CLMS) and the Current Population Survey (CPS). Both are based on large, well-designed samples. Both have been augmented by matching respondent records with social security (SSA) earnings records.

The CLMS-SSA match works as follows. The Bureau of the Census, under agreement with the Department of Labor, designs the CLMS probability sample and collects the data. The record on each individual includes identifying information and social security number. A list of respondent SSA numbers is given to the SSA which then searches

SSA files for records on the relevant individuals. The SSA records include the social security number, earnings, birth year, six letters of surname, and other bits of information. These SSA records are then given to Census for matching to the CLMS survey records under an interagency agreement that assures confidentiality of both sets of files. Census matches the records, deletes identifying information and geographic area related characteristics. The geographic data are deleted to prevent deductive disclosure.

Recently, the U.S. Department of Labor contracted for two kinds of analyses bearing on the impact of manpower programs and based on these files. In the first kind, different, well regarded contractors were asked to use such data to estimate the effects of training programs (Westat, 1984; Dickinson, et al., 1984; Bassi, et al., 1984). In the second kind of study, estimates based on observational survey data, similarly constructed, were compared to estimates yielded by randomized field experiments. In particular, the models used on CLMS and CPS data were used to construct quasi-experimental comparison groups. The performance of these comparison groups was compared to randomized control groups generated in the National Supported Work Demonstration (Fraker & Maynard, 1985).

The results of three independent analysts generating models and using them to estimate program effects based on CLMS and CPS data yielded the following results:

(a) Effects of training on earnings are positive and significant, especially for females and all post Comprehensive Employment and Training Act follow-up years (Westat, 1984, p. 61).

(b) Effects on earnings for men are not generally significant; effects on women's earnings are significant (Bassi, et al., 1984, p. xv).

(c) Effects on earnings for men tend to be significant and negative, but effects on women are positive and significant but small (Dickinsen, et al., 1984, p. xiii).

We have oversimplified here, of course. "Significance" is emphasized too much and the statements are misleadingly blunt. But the conclusions are as they appear in the final reports.

Comparing estimates of control group performance similarly constructed to estimates of control group behavior based on randomized experiments had the following results: depending on the particular model and matching strategy used, estimated effects on earnings range from minus 2000% of "true" earnings to plus 50% of "true" earnings, "true" being estimated from the randomized trial.

These results should be a bit disconcerting. They are indeed puzzling and potentially embarrassing. The Labor Department deserves praise for scholarship in disclosing the puzzle and for its political fortitude in willingness to tolerate potential embarrassment.

More to the point, what are the reasons for the discrepancies? Sampling variations may account for some of the differences. But it is not likely to account for all. In the next section, the reasons engendered by another line of argument are discussed, in the interest of understanding the strength and weakness of the argument.

3.2 Line of Argument

The critic can propose that part of the reason for discrepant results lies in relying---

(a) solely on observational data, matched or otherwise, and
(b) on models whose validity is untestable with the data at hand.

Critics who are more blunt may further suggest that the CPS, SSA, and CLMS are used because they are available and seemingly appropriate and not because they are sufficient.

Finally, the administrative system in which matching occurs demands that one give up some opportunities that should not be given up if the object is to produce good estimates of program effects.

To illuminate the contentions, consider SSA earnings matches with observational data from surveys. Problems similar to ones discussed here occur in other contexts. The material that follows is based on thoughtful reports by Bassi, et al. (1984), Dickinson, et al. (1984), and Westat (1984), that is, the producers of the estimates of manpower program effects.

State Identifiers and Areas as Missing Data

Welfare laws differ appreciably among states. These laws determine who gets welfare and how much they get. It makes sense to incorporate such data into any analysis of the way a federal employment program is used by the poor and what the impact of the program is. Local labor market information is also crucial to thoughtful analyses of why people do or do not get jobs as a consequence of programs.

Yet such information is absent from public use microdata files that are released after matching records. The result is that the economist must be content with data that are bound to generate estimates of program effect that are likely to be biased. That is, important major variables are left out of the left hand side of explanatory equations because they are deleted from public use files or remain unmeasurable variables. The incompleteness of the model is responsible for biased estimates of effect.

Why are they left out of such files? Because their inclusion will permit deductive disclosure. That is, it becomes possible to deduce the identity of anonymous respondents if information about geographic area is supplied. The Census, for example, cannot countenance the possibility of deductive disclosure of information that it has collected, and invokes Title 13 to justify its position. Census perspective on this matter is important not only for this case: The Bureau "performs a major portion of its survey work on a reimbursable basis for other Federal agencies" (Cox, et al., p. 1, 1985). It is important as a survey agency and as a model of virtue in this respect.

Exclusion of relevant data seems to us to be the most serious consequence of our use of Census-SSA in data collection and matching. From such a matching system, we cannot produce credible estimates without the appropriate variables.

Earnings not Covered by SSA

Many public sector jobs are not covered by SSA reporting. Insofar as the employment and training program leads to jobs that are public sector and not covered, two problems occur. When earnings are a dependent variable, estimates of impact will be understated when the comparison groups jobs are more likely to be SSA covered. When earnings are used as a covariate, e.g., "prior base year," estimates of program impact will be biased because the covariate is fallible.

One way to assess the problem is by looking at interview-based earnings reports and SSA earnings, of course. Dickinson, et al. (1984) did so. They found substantial error in CLMS interview reports, e.g., 33% of CLMS respondents who said they did not work in 1977 had positive SSA earnings reported. The rate for CPS is about 10%. We still have a dilemma: SSA is clearly better than self-reports of earnings, although they are imperfect.

SSA earnings data are also truncated at both ends. For example, the maximum earnings subject to SSA tax is the maximum recorded earnings level. Dickinson, et al. (1984) examined interview earnings and SSA cap earnings to find no appreciable difference between analyses using each. i.e., estimates of program effect are about the same (p. 98).

Updatedness: A Possibly Tractable Problem

As of 1983-84, the period of DOL analyses of interest here, 1979 SSA records merged with CPS and CLMS data are incomplete. That is, not all 1979 SSA earnings for members of these samples were available. A "zero" entry for the missing data means we cannot tell how much missing data there is. Bias cannot be estimated. Still, this problem seems tractable.

Program Participation not Measured: A Possibly Tractable Problem

The CPS does not now measure participation in employment programs. Consequently, a public use file will not permit construction of a comparison group that is "uncontaminated." Among youth in the CPS comparison group, for example, it has been estimated that between 1975-78 30% entered CETA. So the contamination issue seems important. It, too, seems tractable but not without substantial effort.

Alignment Problems

According to Dickinson, et al. (1984), in Westat's analysis of the FY76 cohort, SSA earnings in calendar year 1975 were used to match individuals, despite the fact that calendar year 1975 earnings included up to six months of post-enrollment earnings for some CLMS members, (p. 35). Dickinson, et al., used calendar year cohorts rather than fiscal year cohorts. The disadvantage is in potentially missing the preprogram drop in earnings.

4. RESTATEMENT OF THE PROBLEMS AND POSSIBLE SOLUTIONS

4.1 Core Problems

There are two kinds of problems implicit in the case just presented. The first concerns reliance solely on surveys coupled to administrative records to understand relative effects of programs. Problems engendered by relying on such data affects not only efforts to estimate impact of manpower training programs, of course. They also appear in health services research, psychiatric and mental health services evaluations, assessments of court procedures, tax compliance, and police procedures (Riecken, et al., 1974). We attribute the problems partly to the seductiveness of matching and partly to the more dangerous problem of untestable models.

The second kind of problem stems from our inability to use all the data in ways that permit confidence that the analysis is statistically unbiased. Denial of access to micro-records on account of deductive disclosure affects research by Bureau of Labor Statistics (Plewes, 1985) as well as the DOL Employment and Training Administration, by the National Institute of Justice (e.g., in victimization studies), and others. The issue is also likely to affect newer statistical programs, e.g., the Survey of Income and Program Participation (David, 1984). We attribute this problem to the administrative environment in which matching technology must be exploited.

4.2 Resolving the First Kind of Problem and Exacerbating the Second

A scientifically reasonable solution to the first kind of problem is to actively experiment. That is, we need to run randomized trials of projects, project components, or project variations. The research policy option that seems worth exploring is routinely adjoining randomized experiments to the longitudinal studies and/or record files that are matched. See for instance, the Hollister, et al. (1985) report on evaluating the effectiveness of youth employment programs.

Exercising the option of randomized experiments can exacerbate the second problem, i.e., of deductive disclosure. That is, experiments generally involve a smaller number of individuals than national probability samples and more detailed information on each individual. This makes deductive disclosure easier. It also makes it difficult to adopt sampling rates as a partial index of likelihood of deductive disclosure (Cox, et al., 1985). If an agency with restrictive rules is involved in data collection then no public use tapes with sufficient detail will be released and no sensible competing analyses will be done.

Apart from the information demands of randomized experiments, the demand for microdata is increasing. Cox, et al. (1985) recognize that this increase has strong implications for Census policy on disclosure and they provide a thoughtful analysis.

4.3 Resolving the Second Problem

The possible resolutions to the disclosure problems are of at least three kinds: procedural, statutory, and empirical. The following options illustrate each.

Avoiding Restrictive Agencies

One may stay away from agencies that have data worth matching but that also have restrictive disclosure policies. Indeed, it is not hard to argue that private agencies are as capable of producing good data with equal privacy protection for the respondent and fewer constraints on the research than a government agency. The case is especially arguable for controversial topics of research such as AIDS, but it is also relevant here (Boruch, 1984).

Still, doing without micro-records from agencies such as the Census Bureau, Social Security Administration, or others, and doing without their capacity to serve as a broker for linking records from independent sources, is not an attractive prospect. We may gratuitously abandon opportunities to do socially useful and reliable research by foregoing collaboration with such agencies. So it is sensible to consider other options in addition to this one.

Proactive Change in Law and Policy

Alteration of law and more feasibly the interpretation of law is possible and seems desirable. The battles for statistical enclaves suggest, however, that this war will not be won easily, if at all. Still, sensible work has been done and some progress in clarifying issues has been made (Alexander, 1983). Assaults on Census's stewardship of Title 13 seem not to have been productive, for example (Plewes, 1985). Still, working toward legitimate reinterpretation of law seems an effort worth making, especially if more empirical research can be brought to bear on the issue of perceived risks of disclosure to populations. This brings us to the next option.

Empirical Research

Research on the role that privacy and consent have in record matching contexts seems sensible. How much the assurance of confidentiality means to respondents and how it influences the cooperation rate has received some attention from empiricists. For example, randomized field tests have been run under the auspices of the NAS Committee on National Statistics to understand whether people attend to assurances about privacy (Panel on Privacy and Confidentiality, as Factors in Survey Response, 1979). We agree with Thomas Plewes (1985) of the Bureau of Labor Statistics (BLS) in urging that more related work needs to be done.

In particular, obtaining respondent consent to disclose and link records for research purposes is an avenue for resolving deductive disclosure/confidentiality problems at Census, SSA, and elsewhere. We are aware of no good field experiments to determine effective strategies to elicit consent or their consequences. The BLS has been successful, according to Plewes, in eliciting consent for disclosure of its data to the Department of Agriculture, for instance, so that better sampling frames for forms could be developed. But this evidence is anecdotal and few hard data from controlled trials are available.

Both Cox, et al. (1985) at Census and Plewes (1985) at BLS recognize that public perceptions of government agencies are important in this context. That is, public confidence in government affects cooperation in surveys and resultant public data.

This chain of reasoning is plausible. But our agreement is a matter of intuition, not hard evidence. Moreover, the politicians' view of the idea and its implications for a bureaucracy and votes seem important. Neither the Census Bureau nor BLS (nor other agencies) can work on this tangle of issues with impunity, at least not always. Academic researchers have some responsibility to do so if they expect to have access to good data. We know of very few who are involved in such work, e.g., Flaherty, Hanis, and Mitchell (1979) in Canada, Mochmann and Muller (1979) and Damman and Simitis (1977) in Germany.

Research: Analytic

The Department of Labor's support of competing analyses, and of comparisons of the results of randomized tests to the results of nonrandomized assessments, is admirable. Research in the same spirit on matching and disclosure is warranted.

The thoughtful observer ought to admire the work by Nancy Spruill and Joe Gastwirth (1982) on microaggregation and masked data and work by George Duncan and Diane Lambert (1985) on disclosure limited dissemination. Their analysis helps to actualize a balance between privacy needs and the need to assure quality of released data. The thoughtful observer will also recognize, however, that not much work has been done on the costs, traps, flaws, and benefits of using the suggestions of these analysts. We ought to know more about these issues. And so we ought to invest some resources routinely in the design of side studies to illuminate the limits on the utility of their work.

The importance of this matter stems partly from the fact that the effects of social programs in tax compliance, police, training, and employment effects are usually small. Expecting small effects, we should then be better able to anticipate the effects of micro-aggregation, random perturbation (contamination), random rounding, collapsing, and other strategies used to transform data so as to make it suitable for public use. All such tactics are used by the Census and other agencies to protect individual (and at times institutional) privacy (Cox, et al., 1985). But very little has been published about their implications for the validity of inferences based on analyses of such public use data.

Administrative Procedures

Suppose that we create a matching system under which public use tapes that are first expurgated or "adjusted" to reduce deductive disclosure problems are used for crude analyses. These analyses are eventually verified using the unexpurgated records by the agency that maintains the more detailed micro-records. The procedure achieves a balance between privacy concerns and scientific demands for quality in analysis.

But it demands substantial resources, i.e., a sequential system of crude analyses, based on public use tapes, followed closely by confirmatory analyses, based on within-agency analysis of micro-records. Still, the option seems worth considering especially because the procedure seems generalizable, e.g., to matching economic variables in the Survey of Income and Program Participation (David, 1984).

For example, 1976 Annual Housing Survey data on energy use were matched on geographic area to local utility company data. Census created the file. To protect against deductive disclosure, the Census adjusted the accuracy of energy use data "prior to release to guard against the possibility that the utility companies could uniquely identify individuals on the released file from their reported cost data" (Cox et al., 1985, p. 22). The adjustment involved random perturbation (that can be accommodated up to a point in analyses, given the perturbation parameters) and rounding. We are unaware of any formal benefit-cost analysis of this case. We believe that some sort of evaluation of such cases should be undertaken and published.

5. REPRISE AND CONCLUSION

There is no doubt that matching can be and has been useful in a variety of social research projects. Moreover, the analytic work on the topic by Felligi and Sunter (1969) and others is remarkable for its thoughtfulness. The technology for matching, considered apart from the matching system (organization and data), has stimulated fascinating research by academic and bureaucratic scholars. But solutions to the problem of getting the benefit of matching without reducing interpretability of data are not yet clear.

The ingeniousness of a matching algorithm is one thing. The system in which the algorithm is applied is quite another. It is clear that the administrative environment of the matching system can lead to invidious problems in analysis at the policy level. The problems lie not so much in matching technology as in other elements of the matching system: the data and rules under which it was collected, the institutional vehicle for matching and the rules governing it, and the procedures one uses to understand the errors we make based on analyses of matched data. The problems are severe enough to warrant the serious concern of applied statisticians and social scientists. Unless attention is dedicated to the matter we will do far less than we should for science, society, and the profession.

ACKNOWLEDGMENTS

The background research for this paper was supported by the National Science Foundation during 1984-85 to the Social Science Research Council (Measurement and Methodology Division, NSF) and 1985-86 to Northwestern University (RII-8418179 Ethics and Values in Science and Technology, NSF). We are indebted to Robert Pearson of the Social Science Research Council and Richard Beatty of Westat for background information. Readers should be grateful, as I am, to Theodore Clemence of the Census Bureau for graciously providing references on Alexander Bell.

REFERENCES

Alexander, L. Proposed legislation to improve statistical and research access to federal records. In R.F. Boruch and J.S. Cecil (Eds.), Solutions to ethical and legal problems in social research. New York: Academic Press, 1983, 273-292.

Andersen R., Kasper, J., Frankel, M.R. and Associates. Total survey error. San Francisco: Jossey Bass, 1979.

Bassi, L.J., Simms, M.C., Burbridge, L.C., and Betsey, C.L. Measuring the effect of CETA on youth and the economically disadvantaged. Washington, D.C.: The Urban Institute, April 1984.

Bell, A.G. The deaf. In: U.S. Department of Commerce and Labor, Bureau of the Census. Special Reports: The blind and the deaf, 1900. Washington, D.C.: U.S. Government Printing Office, 1906.

Boruch, R.F., and Cecil, J.S. Assuring the confidentiality of social research data. Philadelphia: University of Pennsylvania Press, 1979.

Boruch, R.F. Should private agencies maintain federal research data? IRB, 1984, 6(6), 8-9.

Box, J.F. R. A. Fisher: The life of a scientist. New York: Wiley, 1978.

Bruce, R.V. Alexander Graham Bell and the conquest of solitude. Boston: Little, Brown, and Company, 1973.

Cox, L.G., Johnson, B., McDonald, S.K., Nelson, D., and Vazquez, V. Confidentiality issues at the Census Bureau. Presented at the first Annual Census Bureau Research Conference. Reston, Virginia, March 20-23, 1985.

Damman, U., and Simitis, S. Bundesdatenschutzgesetz. Baden-Baden: Nomos Verlagsgesellschaft, 1977.

David, M. Discussion. Proceedings of the American Statistical Association: Social Statistics Section. Washington, D.C.: ASA, 1984, pp. 534-536.

Dickinson, K.P., Johnson, T.R., and West, R.W. An analysis of the impact of CETA programs on components of earnings and other outcomes. Menlo Park, CA: SRI International, November 1984.

Duncan, G.T., and Lambert, D. Disclosure limited data dissemination. Journal of the American Statistical Association. 1985, in press.

Fellegi, I.P., and Sunter, A.B. A theory for record linkage. Journal of the American Statistical Association, 1969, 64, 1183-1210.

Flaherty, D.G., Hanis, E.H., and Mitchell, S.P. Privacy and access to government data for research. London: Mansell, 1979.

Fraker, T., and Maynard, R. The use of comparison group designs in evaluations of employment related programs. Princeton, N.J.: Mathematica Policy Research, 1985.

Halsey, H.I. Data validation. Chapter 2 of P.K. Robins, R.G. Spiegelman, S. Weiner, and J.G. Bell (Eds.) A guaranteed annual income: evidence from a social experiment. New York: Academic, 1980, pp. 33-55.

Hausman, J.A. and Wise, D.A. (Eds.) Social experimentation. University of Chicago Press, 1985.

Hollister, R. and others (Eds.) Report of the Committee on Youth Employment Programs. Washington, D.C.: National Academy of Sciences, 1985.

Kershaw, D. and Fair, J. The New Jersey income maintenance experiment: Operations surveys, and administration, Volume I, New York: Academic Press, 1979.

Locander, W., Sudman, S. and Bradburn, N.M. An investigation of interview method, threat, and response distortion. Journal of the American Statistical Association, 1976, 71, 269-275.

Mathiowetz, N.A., and Duncan, G.J. Temporal patterns of response errors in retrospective reports of unemployment and occupation. Proceedings of the American Statistical Association: Section on Survey Research Methods. Washington, D.C.: 1984, 652-657.

Mochmann, E., and Muller, P.J. (Eds.), Data protection and social science research. Frankfurt/New York: Campus Verlag. 1979.

Newcombe, H.B., Smith, M.E., Howe, G.R., Mingay, J.Strugnell, A. and Abbatt, J.D. Reliability of computerized versus manual death searches in a study of the health of Eldorado uranium workers. Computers in Biology and Medicine, 1983, 13(3), 157-169.

Panel on Privacy and Confidentiality as Factors in Survey Response. Committee on National Statistics. Privacy and confidentiality as factors in survey response. Washington, D.C.: National Academy of Sciences, 1979.

Plewes, T.J. Confidentiality principles and practice. Paper presented at the first Annual Census Bureau Research Conference. Reston, Virginia, March 20-23, 1985 (available from the author, Bureau of Labor Statistics).

Riecken, H.W., and others. Social experimentation: A method for planning and evaluating social programs. New York: Academic, 1974.

Sherman, L.W., and Berk, R.A. The specific deterrent effects of arrest for domestic assault. American Sociological Review, 1984, 49, 261-272.

Spruill, N.L., and Gastwirth, J. On the estimation of the correlation coefficient with grouped data. Journal of the American Statistical Association, 1982, 77, 614-620.

Tippett, J.A. An experimental project to improve reporting of selected costs in the 1980 census. Proceedings of the American Statistical Association: Survey Research Methods Section. Washington, D.C.: ASA, 1984, 323-328.

U.S. Department of Agriculture, Food and Nutrition, Office of Analysis and Evaluation. Food stamp work registration and job search demonstration. (Contract No. 533198085). Alexandria, VA: DOA, 1984.

Westat, Inc. Continuous Longitudinal Manpower Survey: Net impact results. Rockville, MD: Westat, April 1984.

METHODOLOGIC ISSUES IN LINKAGE OF MULTIPLE DATA BASES

Fritz Scheuren *

Data linkage offers several obvious benefits in studying the dynamics of aging. Retrospective and prospective approaches are possible. Many ad hoc epidemiological studies could serve as examples here (e.g., Beebe, 1985). Perhaps of even more importance are broad-based statistical samples composed of linked administrative records, either used alone or in conjunction with survey data (e.g., Kilss and Scheuren, 1980; Scheuren, 1983).

In general, linked administrative records, when structured longitudinally (e.g., Buckler and Smith, 1980), can be very effective in tracing changes with age in income and family relationships--including the onset of some forms of morbidity (e.g., Klein and Kasprzyk, 1983); and, with the advent of the National Death Index, mortality as well (e.g., Patterson and Bilgrad, 1985).

Survey data can be used, among other things, to explore the underlying causal mechanisms for these administratively recorded outcomes. The design challenge, of course, is how to build a data collection process which exploits the comparative advantages of both administrative and survey information.

The present paper examines settings where linkages of U.S. federal government records for individuals are feasible and of interest in the study of the dynamics of aging. Both administrative and survey records will be considered. Our focus will be on the barriers to and benefits from data linkages, with examples drawn from studies conducted using records from the Social Security Administration (SSA), the Health Care Financing Administration (HCFA), the National Center for Health Statistics (NCHS), the Bureau of the Census and, of course, the Internal Revenue Service (IRS).

Organizationally, the paper has been divided into three main sections. Structural questions (e.g., legal and procedural) in the development of a data linkage system are taken up first (Section 1). Technical issues in the matching process itself are discussed next (Section 2). The paper concludes (in Section 3) with some recommendations on areas for future study. An extensive set of references is also provided, along with some additional bibliographical citations (See Appendix A).

1. STRUCTURAL DESIGN CONSIDERATIONS

During the last several decades numerous data systems have been built by linkage techniques in an attempt, among other objectives, to study various aspects of the aged population. Some of these, like the Continuous Work History Sample, remain enormously valuable (e.g., Kestenbaum, 1985) but are no longer fully exploited because of access problems and severe resource constraints (e.g., Cartwright, 1978). Others, notably the Retirement History Survey (Irelan and Finegar, 1978), have not been continued. Many studies had an ad hoc character to begin with. While successful, they have not been repeated (e.g., The 1973 Exact Match Study, Kilss and Scheuren, 1978; the Survey of Low Income Aged and Disabled, Barron, 1978). Still other studies originally envisioned as standalone survey systems have not exploited available data linkage opportunities to extend their useful life beyond the point at which interviewing has stopped (e.g., the National Longitudinal Survey, Parnes, et al., 1979). What can we learn from these experiences and others that are similar--

- First, agency support for the activity has to be very strong and continuing. Social Security, which supported most of the projects listed above, has moved away from such general research efforts and shifted towards examining improvements in program operations (Storey, 1985). A sustained long-run commitment to basic research simply may not be possible in what is inherently a policy-oriented environment (President's Reorganization Project for the Federal Statistical System, 1981).

- Second, strong user support is essential. The products must have high, perceived public value, be delivered in a timely manner and with sufficient regularity to sustain continued interest. Start-up problems with the Retirement History Survey caused it some major difficulties from which it may never have been able to fully recover (Maddox, Fillenbaum, and George, 1978). The Continuous Work History Sample has, especially in recent years, been unable to sustain user interest outside of Social Security because of access issues raised by the 1976 Tax Reform Act. Also, the emphasis on employee-employer relationships, long a main feature of the Continuous Work History Sample, may not have been seen to be as important as the resource commitment required to maintain it.

- Third, start-up costs may be high for data linkage systems, especially if based in part on survey data. Linkage systems tend to be easily maintained at low cost unless

*Prepared for the Panel on Statistics for an Aging Population and presented September 13, 1985. Reprinted with permission from the National Academy of Sciences, Committee on National Statistics (to appear in their forthcoming report).

continued surveying is done; however, certain data problems, due to insufficient attention in obtaining good matching information, can cause continuing expense and difficulty at the analysis stage. Obviously also, as turned out to be the case with the Continuous Work History Sample, data quality limitations in the administrative records may necessitate considerable additional expense.

- Fourth, data linkage systems employ methods that may not be seen as entirely ethical (e.g., Gastwirth, 1986) or that have confidentiality constraints that make the systems hard to maintain as with the Retirement History Survey or hard to use as with the Continuous Work History Sample (e.g., Alexander, 1983). These controversial elements in data linkage techniques, it may be speculated, could be one of the reasons linkages to the National Longitudinal Survey (NLS) have never been attempted (despite the collection of social security numbers in the NLS).

It is only with the last of these points that we touch on risks that data linkage systems encounter, which are not also encountered to some degree in more conventional data-capture approaches. The force of these concerns will be discussed below.

Confidentiality and Disclosure Concerns

Data linkage operations bring us face-to-face with a "dense thicket" of laws, regulations and various ad hoc practices justified on heuristic grounds. There are statutory considerations which apply either to the particular statistical agencies involved or to the federal government, as a whole. These include the Privacy Act; the Freedom of Information Act; special legislative protections afforded to statistical data, for example, at the Census Bureau and the National Center for Health Statistics; and, of course, legislative protections afforded to administrative data, notably the 1976 Tax Reform Act. The paper by Wilson and Smith (1983) gives a good summary of the legal protections afforded tax data. For a more general treatment of legal issues and one which advocates change, see Clark and Coffey (1983); also see Alexander and Jabine (1978).

The regulations and practices of each federal statistical agency differ too, not only because of the different legislative statutes under which they operate, but also because of the varying approaches that they have taken in the accomplishment of their missions. Indeed, interagency data sharing arrangements almost defy description; they vary, among other reasons, depending on which agencies are sharing whose data and for what purpose. One excellent, albeit incomplete, taxonomy of current practice is found in the work of Crane and Kleweno (1985).

Despite the complexity of this topic, several general trends emerge that are worth noting:

- First, the American People are at best ambivalent about letting their government conduct linkages across data systems, specifically between different agencies and for purposes not obviously central to the missions of both agencies. For example, in a recent survey, questions were asked about the sharing of tax records with the Census Bureau, something which is a longstanding practice specifically permitted by law. Three-fourths of those surveyed did not support this use of administrative records even though an attempt was made to put the matter in a very favorable light, arguing for it on efficiency grounds. (Gonzalez and Scheuren, 1985; see also Appendix B for exact question wording).

- Second, bureaucratic practices which do not respect this general unease about linkage may need to be reexamined (e.g., Gastwirth, 1986). It is the duty, after all, of government statisticians to uphold both the letter and the spirit of the law. The whole tenor of the post-Watergate, Privacy Act and Tax Reform Act era has been to limit administrative initiatives (both big and little "a") and only to permit the expansion of access after the enactment of positive law. The failed initiative regarding Statistical Enclaves illustrates this point quite nicely. The Enclave proposal (Clark and Coffey, 1983) sought what many regarded as a degree of reasonable discretion on data linkage and data access; however, the authority requested was too broad for the current political climate. The arguments put forward in the proposed legislation's defense, for example, that it would increase efficiency and bring order to a patchwork of disparate practices, simply did not carry the day. In summary, we do not seem to be even close to a general solution on access to data for statistical purposes.

- Third, absent new legislation, many statistical agencies have begun to reexamine their traditional access arrangements and tighten still further their practices (e.g., Cox et al., 1985). For example, the use of special Census agents to facilitate linkages or to improve their subsequent analysis has been drastically curtailed resulting in a clear short-run loss in the utility to outsiders of linkage methods at the Census Bureau. On the other hand, new linkage practices have emerged from such reviews which may be superior to what otherwise might have been done. The linkage between the Current Population Survey and the National Death Index is an excellent example (Rogot, et al.,1983). Neither the Census Bureau nor the National Center for Health Statistics felt it could give up access of its data to the other agency; however, a compromise was worked out where joint access was maintained during the linkage operation and this has proved satisfactory. In fact, similar arrangements have been made successfully between the Center and the Internal Revenue Service as part of a study of occupational mortality (Smith and Scheuren, 1985b).

- Fourth, the extent to which public use files can be made available from linked data sets has been greatly curtailed because of new concerns about what is called the "reidentification" problem (Jabine and Scheuren, 1985). Simply put, this means that if enough linked data are provided in an otherwise unidentifiable (public-use) form, then each contributing agency could reidentify at least some of the linked units, almost no matter what efforts at disguise are attempted (Smith and Scheuren, 1985b). The only major exception occurs when the data made public from the contributing agencies are extremely limited (Oh and Scheuren, 1984; Paass, 1985); but then, usually, the incentives for cooperation on the part of the contributing agencies are limited as well. In practice, of course, there is almost no incentive for the contributing agencies to reidentify; thus, legally binding contractual obligations might be entered into that could stipulate that there was no such interest. Contractual guarantees, however, may not satisfy all parties to the linkage, because of the public perception issues mentioned earlier. It is conceivable, moreover, that no degree of legal or contractual reassurance would be adequate at the present time to permit the release of certain public use linked data sets--for example, those involving Census surveys linked to Internal Revenue Service information. Historically it was only the impossibility of reidentification which made the release of matched CPS-IRS-SSA public use files possible (Kilss and Scheuren, 1978).

It goes almost without saying that confidentiality and disclosure concerns pose the greatest barriers to the development of data linkage systems for studying aging. We will, however, defer to Section 3 a discussion of what might be done to deal with such issues and go on to explore the technical side of matching.

2. MATCHING DESIGN CONSIDERATIONS

This section is intended to provide a brief discussion of matching design questions that must be looked at in developing data linkage systems. We begin with some historical background and then focus specifically on "person" matches, where the social security number is a possible linking variable. Linkage systems based in part on survey information are emphasized. Analysis problems also are covered, particularly ways of estimating and adjusting for errors arising from erroneous links or nonlinks.

Historical Observations

The main theoretical underpinnings for computer-oriented matching methods were firmly established by the late nineteen sixties with the papers of Tepping (1968) and especially Fellegi and Sunter (1969). Sound practice dates back even earlier, at least to the nineteen fifties and the work of Newcombe and his collaborators (e.g., Newcombe, et al., 1959).

The Fellegi-Sunter approach is basically a direct extension of the classical theory of hypothesis testing to the problem of record linkage. A mathematical model is developed for recognizing records in two files which represent identical units (said to be matched). As part of the process there is a comparison between all possible pairs of records (one from each file) and a decision made as to whether or not the members of the comparison-pair represent the same unit, or whether there is insufficient evidence to justify either of these decisions. These three decisions can be referred to as a "link," "non- link" or "potential link."

In point of fact, Fellegi and Sunter contributed the underlying theory to the methods already being used by Newcombe and showed how to develop and optimally employ probability weights to the results of the comparisons made. They also dealt with the implications of restricting the comparison pairs to be looked at, that is of "blocking" the files, something that generally has to be done when linking files that are at all large.

Despite the early seminal work of Newcombe, Fellegi and others, ad hoc heuristic methods abound. There are many reasons for this state of affairs:

- First, until recently (and maybe even now) there have been only a handful of people whose main professional interest is data linkage. This means, among other things, that most of the applied work done in this field has been carried out by individuals who may be solving matching problems for the first time. Because the basic principles of matching are deceptively simple, ad hoc solutions have been encouraged that could be far from optimal.

- Second, statisticians typically get involved very late in the matching step, often after the files to be matched have already been created. Even when this is not the case, little emphasis may be placed on the data structures needed for linkage because of other higher priorities. Design opportunities have, therefore, been generally limited to what steps to take given files which were produced largely for other purposes.

- Third, until the late nineteen seventies good, portable, general-purpose matching software had not been widely available (e.g., Howe and Lindsay, 1981), despite some important early attempts (e.g., Jaro, 1972). Even in the presence of general-purpose software, the uniqueness of each matching environment may lead practitioners to write complex customized programs, thereby absorbing resources that might have been better spent elsewhere.

- Fourth, especially for matches to administrative records, barriers to the introduction of improved methods have existed because cruder methods were thought to be more than adequate for administrative purposes.

- Fifth, the analysis of linked data sets, with due consideration to matching errors, is still in its infancy (Smith and Scheuren, 1985a). Qualitative statements about such limitations typically have been all that practitioners have attempted.

More will be said below concerning these issues in the context of computerized person matching.

Person Matching

Typically in a computerized matching process there are a number of distinct decision points:

- First, design decisions have to be made about the linking variables that are to be used, including the extent to which resources are expended to make their reporting both accurate and complete. (This step may be the most important but it is likely also to be the one over which statisticians have the least control, especially when matching to administrative records.)

- Second, decisions have to be made about what preprocessing will be conducted prior to linkage. Some of the things done might include correcting common spelling errors, calculating SOUNDEX or NYSIIS Codes, etc. (Winkler, 1985). Decisions about how to sort and block the files also fall here (Kelley, 1985).

- Third, decisions about the match rule itself come next. If a probabilistic approach is taken, as advocated by Fellegi and Sunter (1969), then we have to estimate a set of weights that represent the extent to which agreement on any particular variable provides evidence that the records correspond to the same person (and conversely, the extent to which disagreements are evidence to the contrary).

- Fourth, invariably there are cases where status is indeterminate regardless of the approach taken and a decision has to be made about excluding them from the analysis, going back for more information, etc.

To give some realism and specificity to our discussion, let us consider potential linkage settings in which we could bring together two files based on common identifying information: name, social security number, sex, date of birth, and address. As appropriate we will contrast the linkage as taking place either entirely in an administrative context or between survey and administrative data.

Linking Variables--The social security number (SSN) is the most important linking variable that we in the United States have for person matching purposes. SSNs were first issued so that the earnings of persons in employment covered by the social security program could be reported for eventual use in determining benefits. SSNs were also used as identifiers in state-operated unemployment insurance programs but no other major uses developed until 1961 when the Internal Revenue Service decided to use the SSN as the taxpayer identification number for individuals. Other uses by federal and state governments followed rapidly and now the social security number is a nearly universal identifier. The Privacy Act of 1974 placed restrictions on the use of SSNs but exempted those formally established prior to 1975. So far these restrictions have had only a minor impact on the widespread use of the social security number by governments and private organizations (Jabine, 1985).

The social security number is nearly a unique identifier all by itself and extremely well reported, even in survey settings, as well as on records such as death certificates (e.g., Cobleigh and Alvey, 1974; Alvey and Aziz, 1979). In survey contexts, error rates may run to 2 or 3 percent; but this depends greatly on the extent to which respondents are required to make use of records in order to provide the requested information. Typically, driver's licenses, pay stubs, and the like are excellent sources (in addition to the use of the social security card itself).

Both administrative and survey reporting of social security numbers are subject to possible mistakes in processing, but these can be guarded against by using part of the individual's surname as a confirmatory variable. For example, IRS and SSA use this method as one way of spotting keying errors.

A difficulty with current administrative approaches is that name changes (especially for females) may lead to considerable extra effort in confirming (usually through correspondence) that the social security number was indeed correct to begin with. (It is a requirement of the social security system that notification is to be made when name changes occur, but many people fail to do this until the omission is called to their attention.)

One disadvantage of the social security number is the absence of an internal check digit allowing one to spot errors by a simple examination of the number itself. At the time the social security system started in the mid-thirties, the widespread use of the SSN as an identifier was not envisioned. Indeed, there is not a one-to-one correspondence between individuals and the social security numbers they use. In some instances more than one person uses the same social security number. Historically, the most important cases of this type arose because SSN's were used by advertisers in promotional schemes. Perhaps the best known such instance is the number 078-05-1120 (Scheuren and Herriot, 1975). It first appeared on a sample social security number card contained in wallets sold nationwide in 1938. Many people who purchased the wallets assumed the number to be their own. The number was subsequently reported thousands of times by different individuals; 1943 was the high year, with 6,000 or more wage earners reporting the number as their own.

While there have been over 20 different "pocketbook" numbers, like 078-05-1120, they are probably no longer the main cause of multiple use of the same number. Confusion can arise (and go largely undetected) when one member of a family uses the number of another. Also, there are incentives for certain individuals, like illegal aliens, to simply "adopt" the social security number of another person as their own. The extent to which these problems exist is unknown, but they are believed, at least by some authorities, to be less prevalent than the opposite problem--issuances of multiple numbers to the same person (HEW Secretary's Advisory Committee, 1973).

Until 1972, applicants for SSNs were not asked if they had already been issued numbers, nor was proof of identity sought. This led to perhaps as many as 6 million or more individuals having two or more social security numbers (Scheuren and Herriot, 1975). A substantial fraction of the multiple issuances have been cross-referenced so that multiple reports for the same individual can be brought together if desired. Based on work done as part of the 1973 Exact Match Study, it appears that, despite the frequency of the problem, multiple issuances can largely be ignored unless one is looking at longitudinal information stretching back to the early days of the social security program. (In other words, people tend consistently to use only one of the numbers they have been issued.)

While the social security number is nearly ideal as a linking variable it is not always available. For example, in the Current Population Survey for adults the number is missing between 20 and 30 percent of the time (Scheuren, 1983). Evidence exists, however, from work done in connection with the Survey of Income and Program Participation, suggesting that with a modest effort the SSN missed rate can be lowered significantly, to less than 10% in Census surveys (Kasprzyk, 1983). Recent experience with death certificates shows a missed rate of about 6% for adults (Patterson and Bilgrad, 1985).

What, then, do we do when the SSN is missing or proves unusable? We are obviously forced either to seek more information or to try to make a match using the other linking variables. Now, as a rule, none of these other linking variables is unique alone and all of them, of course, are subject in varying degrees to reporting problems of their own. Some examples of the problems typically encountered are--

- Surname--As already mentioned, name changes due to marriage or divorce are, perhaps, the main difficulty. For some ethnic groups, there can be many last names and the order of their use may vary.

- Given Name--The chief problem here is the widespread use of nicknames. Some are readily identifiable ("Fritz" for "Frederick") but others are not (like "Stony" for "Paul").

- Middle Initial--People may have many middle names (including their maiden name) and the middle name they employ may vary from occasion to occasion. Often, too, this variable may be missing (Patterson and Bilgrad, 1985).

- Sex--This is generally well reported and, except for processing errors, can be relied upon. The main difficulty with this variable is that it is not always available in administrative records. (IRS does not have this variable except through the recoding of first names which simply cannot be done with complete accuracy.)

- Date of Birth--Day and month are generally well reported even by proxy respondents. Year can be used with a tolerance to good effect as a matching variable. Again, as with "sex," this item is not available on all the administrative files we are considering.

- Address--This is an excellent variable for confirming otherwise questionable links. Disagreements are hard to interpret, however, because of address changes; address variations (e.g., 21st and Pennsylvania Avenue for 2122 Pennsylvania Avenue); and, of course, differences between mailing addresses (usually all that is available in administrative files) and physical addresses (generally all that is obtained in a household survey). Recent research on this variable has been done by Childers and Hogan (1984).

Still other linkage variables could have been discussed, for example, race and telephone number. Race is a variable that is similar to sex except not nearly as well reported (unless it is recoded as black, nonblack (e.g., U.S. Bureau of the Census, 1973). Telephone numbers have problems similar to addresses and, while potentially of enormous value eventually, are not now widely available in administrative files.

Preprocessing Steps--In general, any method of standardization of identifier labels, such as names and addresses, will improve the chances of linking two records that should be linked during the actual matching process; however, it will also, to an unknown degree, result in some distortion and loss of information in the identifying data and may even increase the likelihood of designating some pairs of records as a positive link when, in fact, the pair is not a match.

Typically, for person matches to SSA or IRS information, two preprocessing steps have been undertaken: (1) to validate reported social security numbers; and (2), if missing or unusable, to search for SSNs using surname and other secondary linking variables. Both of these steps have had to be conducted largely within the existing administrative arrangements. The cost of mounting a wholly separate effort has been judged to be prohibitive. (The data sets involved are simply enormous: Social Security has roughly 300 million SSNs now issued. In recent years IRS has been processing about 100 million individual income tax returns annually, containing well over 150 million taxpayer social security account numbers.)

The "Validation Step" itself consists of two parts: first, a simple match on SSN alone is attempted; and, if an SSN is found, then secondary information from Social Security or Internal Revenue records is made available on the output computer file. Further processing then takes place so that the confirmatory matching information (names, etc.) can be examined and coded as to the extent of agreement. It is possible that this part of the current administrative procedure can be readily modified to accord with modern matching ideas. What is needed is to institute probability-based weights for the agreements (disagreements) found. At present administrators and statisticians alike simply employ a series of ad hoc rules to separate what will be considered a link from cases that have questionable SSNs (e.g., Scheuren and Oh, 1975; Jabine, 1985).

The "Search Step" is an elaborate and fairly sophisticated computerized procedure (which differs in detail at SSA and IRS). The files used are in sort; and, for the most part, the only possible links that can be looked at are cases that agree on surname. Since other blocking variables are used as well, the current administrative methods tend to be very sensitive to small reporting errors. This is believed to be true despite the fact that the computer linkage procedures go to great lengths to protect against more common reporting errors (such as those mentioned above). At Social Security they do this by systematically varying the linking information on the record for which an SSN is being searched. An extensive set of manual procedures also exists for cases where computer methods prove unsuccessful.

Unlike the "Validation Step," it may not be possible to bring the "Search Step" into full accord with modern practice. First of all, we would need to reexamine the decisions about what blocking variables to use (Kelley, 1985). Ideally we want variables that are without error themselves, or nearly so, in both sources (Fellegi, 1985) and that divide the files into blocks or "packets" of reasonably small size, within which we can look at all possible linkage combinations (e.g., Smith, 1982). Research is now underway in both agencies to find ways of improving the blocking variables, but it is unlikely that the current deterministic methods will ever be replaced by probability-based ones and for good reason. Linkage techniques for administrative purposes must be employed with high frequency in a great variety of situations and hence be extremely efficient in the use of computer time since the basic files involved are so large.

A compromise that naturally arises within the world of large computer files is to employ some form of multiple, albeit still deterministic, scheme. This is the approach taken with the National Death Index. The NDI currently employs over a dozen different combinations of matching variables. Some give a primary role to the social security number, some to the surname; still others place primary emphasis on the given name or on date of birth (Patterson and Bilgrad, 1985). Adopting the NDI approach at SSA or IRS, if feasible, might be one way to make a real advance.

Match Rules--Usually the computerized matching phase in a data linkage system consists of three steps: (1) comparisons between the linkage variables on the files being matched; (2) generation of codes which indicate the extent to which agreements exist or disagreements are present; and (3) decisions regarding the status of each comparison pair. This structure is the same, whether probability-based methods are being implemented (e.g., Howe and Lindsay, 1981) or heuristic approaches are taken (e.g., Scheuren and Oh, 1975).

- Comparison Step--In a sense, we have already discussed this step earlier. It depends heavily on what linkage variables are present; the reformatting, etc., done of those variables to facilitate comparisons; and the degree to which blocking is required because of resource or other considerations. What is desired here conceptually is to compare every record on each file with every record on the other. Blocking, of course, limits (sometimes severely) the extent to which such comparisons can be carried out. Any recoding of the linkage variables (say SOUNDEX for surname) may possibly, as we have noted, reduce the utility of this step. Generally, if resources permit, all the linking variables should be used in the computer comparisons. When this is not possible, they can still be employed later in manually settling cases where the outcome might otherwise be indeterminate. However, it almost goes without saying that manual intervention needs to be carefully limited and closely controlled. Manual matching is extremely costly and, while individual manual decisions can sometimes be better than with computer matching, usually humans lack consistency of judgment and can be distracted by extraneous information, such that they act more decisively than the facts would warrant.

- Coding Step--As a result of the comparison step, a series of codes can be generated indicating the degree of agreement which has been achieved. These agreement outcomes may be defined quite specifically, e.g., "Agrees on Surname and the value is GILFORD." They might be defined more generally: agree, disagree or unknown (the last arising because of missing information, perhaps).

  It becomes very difficult to talk about the coding step without looking ahead to the decision step and the specific approach that will be taken there. Nonetheless, some general observations can be made. Obviously, when we have, in fact, brought together records for the same person, we would like the agreement coding structure not to obscure this point. For example, to protect against trivial spelling errors, we might use the same agreement code even though there are transposition or single-character differences in the name. (The preprocessing of the files should have taken care of some of this but it may, again, be a consideration in the agreement coding itself.)

In most applications of the Fellegi-Sunter approach the assumption is made that agreement (or disagreement) on one linking variable is independent from that on any other, conditional only on whether or not the records brought together are, in fact, for the same person. To aid in making this assumption plausible, special care needs to be taken in structuring agreement codes for such variables as sex and first name, which are inherently related (Fellegi, 1985).

- Decision Step--An assessment can now be made as to the extent to which an agreement on any particular linking variable, or set of variables, constitutes evidence that the records brought together represent the same person. Conversely, an assessment can be made as to the extent to which disagreements are due to processing or reporting errors or are evidence that the records do not represent information for the same person. Typically, the records are divided into those (1) where a positive link is deemed to have been "definitely" established, (2) where a "possible" link may exist but the evidence is inconclusive, and (3) where it can "definitely" be said that no link exists.

In probability-based methods a statistical weight function is calculated to order the comparison pairs. The weights are developed by examining the probability ratio--

$$\frac{\text{Prob (result of comparison, given match)}}{\text{Prob (result of comparison, given nonmatch)}}$$

The numerator represents the probability that comparison of two records for the same person would produce the observed result. The denominator represents the probability that comparison of records for two different persons, selected at random, would produce the observed result. In general, the larger the ratio, the greater our confidence that the two records match, i.e., are for the same person.

Let us consider a particular example in which we are matching on both sex and race; where sex is always represented as either male or female and where race has been recoded black or nonblack. Further suppose the proportion of males and females is each 50% and that blacks constitute 10% of the population and nonblacks 90%. Also suppose that the chances of a reporting error on race are 1/100 and for sex 1/1000. Finally, we will assume that sex and race are independently distributed in the population and that reporting errors are independent as well.

With these stipulations and assumptions, we have the following table of possible probability or "odds" ratios, say for blacks. Usually, given the independence assumption, the probability ratio is broken up into a series of ratios, one for each agreement or disagreement, and logs are taken (to the base 2). One is now working with simple sums, such that the larger (more positive) the total, the more likely that the pair is a match; conversely, the more negative the sum, the greater the likelihood that the two records are not for the same person.

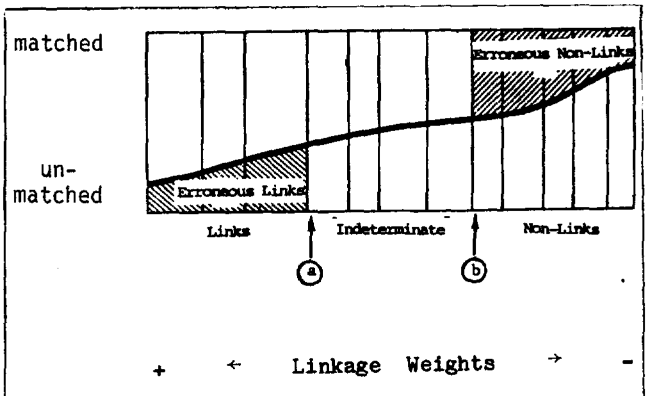
Outcome	Probability Ratio	Base 2 Log of Ratio
Race and sex agree:		
Race is black...........	197.8020	7.6279
Race is nonblack........	2.4420	1.2881
Race agrees, sex does not:		
Race is black...........	0.1980	-2.3364
Race is nonblack........	0.0024	-8.7027
Sex agrees, race does not.	0.1110	-3.1714
Neither agree.............	0.0001	-13.2877

See Computational Note at end of paper.

In our particular example it is only when both sex and race agree that the sum of the logs is positive. If the race is black, the log is between +7 and +8, moderately strong evidence in favor of a match. If the race is nonblack, however, the log is only slightly more than +1. As one would expect, the strongest evidence in favor of a nonmatch occurs when both race and sex disagree; for this outcome the log of the probability is about -13. (Parenthetically, it might be noted that this example illustrates nicely the fact that outcomes that are frequent in the population do not add very much to one's ability to decide if the pair should be treated as a link; but if there are disagreements on such variables and reporting is reasonably accurate, then the variable may have a great deal of power in identifying comparison pairs that represent nonlinks.)

Now it can be shown in general, as by Fellegi and Sunter (1969) or by Kirkendall (1985), that we can divide the weight distribution into three parts, as seen in figure A. The points "a" and "b" optimally divide the distribution of weights so that we can simultaneously minimize the error of accepting as a positive link cases that we should not have matched, plus minimize the error of rejecting as nonlinks cases that we should have kept. Assumptions, like independence, must be made, as a rule, and formidable computational problems exist. Nonetheless, the approach is entirely workable, especially since the development of the Generalized Iterative Record

Figure A.--Hypothetical Distribution of Linkage Weights



(adapted from Fellegi, 1985; comparison pairs above the line are matched , those below nonmatched)

Linkage System (GIRLS), which provides a state-of-the-art solution to the major computational problems (Howe and Lindsay, 1981). Other notable approaches in advanced linkage software include the work of Jaro and his collaborators (Jaro, 1985).

Indeterminate Outcomes--Virtually all computerized record linkage schemes may leave at least some cases where the status is indeterminate. Three kinds of indeterminacy might be distinguished:

- Nonlinks--Cases that were "definitely" determined by the method to have no suitable match, given the approach taken, but which might have been matched if another technique had been used (e.g., if we had employed a different set of blocking variables). The difficulty here is that, while all the potential links that get looked at may have proved inadequate, not all possible links are examined and we cannot tell the difference necessarily between a case that should have been a link and one that should not. The only way this issue can be skirted directly is in the implausible situation when the probability of a match between blocks is zero. (An indirect "solution" to this problem can be developed using contingency table ideas as will be discussed below.)

- Multiple Links--These can occur in the Fellegi-Sunter formulation; that is, there may be more than one comparison pair for a unit whose match weight or score exceeded the threshold for acceptance. In some cases, these many-to-one links might be appropriate but, usually, a further step has to be taken to select "the best" one. This problem also can occur with some frequency in administrative contexts and with the National Death Index. Manual resolution is usually the approach taken, especially if further information is going to be sought or is available to help make the selection. Jaro (1985) offers a computerized transportation algorithm to solve multiple linkage problems. His approach is most effective when all the linking information has already been computerized and when there are contention problems in the linkages, that is, "n" records on one file are matching "m" records on another. Smith and Scheuren (1985a) suggest ways of carrying through the statistical analysis using all the links.

- Potential Links--This type may be the largest form of indeterminacy. These are the cases that fall in the middle area in figure A. The usual advice, resources permitting, is to collect more information to resolve the match status. If statistical estimates are to be made, and the resources needed to seek further information are not available, the potential links may be treated as nonlinks and a survey-type nonresponse adjustment may be made (Scheuren, 1980). It is possible, also, to consider keeping some of the potential links and then conducting the analysis, with an adjustment being made for mismatching (Scheuren and Oh, 1975).

Often, the difficulty with indeterminate cases can be traced back to a design flaw in the data linkage system. For example, not enough linking information may have been obtained on one or both files to assure uniqueness. Maybe the degree of redundancy in the identifiers was insufficient to compensate completely for the reporting errors. In an administrative context, the linkage process may be so constrained for operational reasons that, even if there are sufficient linkage items, they cannot be brought fully to bear.

Analysis Issues

Statements about the nature of the matching errors are typically provided in data linkage studies; generally, however, there is no real attempt to quantify the implications of matching errors for the specific inferences being drawn. Data linkage systems, like other survey-based or sample-based techniques, need to be "measurable" and to be structured to be as robust as possible in the face of departures from underlying assumptions. What can be done to achieve this is a separate and sizable subject (Smith and Scheuren, 1985a). For our present purposes it may be enough to sketch some of the issues and indicate general lines of attack.

- Linkage Documentation--Documentation should routinely be provided which tabulates the results of the match effort along dimensions that turned out to be important in the analysis. A distribution of the weights would be one example, perhaps shown for major subgroups. If a public-use file is being created, then the match weight might be placed in the file along with summary agreement codes, so that secondary analysts can "second-guess" some of the decisions made. Providing potential links, at least near the cut-off point, is another example of good practice. Most of the above, by the way, were part of the documentation and computer files made available from the 1973 Exact Match Study (Aziz, et al., 1978).

- Adjusting for Nonlinks--It is generally worthwhile to consider reweighting the linked record pairs actually obtained to adjust for failures to completely link all the proper records to each other (Scheuren, 1980). Conventional nonresponse procedures can be followed (Oh and Scheuren, 1983). Imputation strategies are also possible, but may be less desirable because they tend to disturb the estimated relationships across the two files being brought together (Oh and Scheuren, 1980; Rodgers, 1984). An important problem in this adjustment process, however conducted, is in being able to estimate whether a link should have occurred. Sometimes, by the nature of the problem, we know all the records should have been linked. In other cases (Rogot et al., 1983), one of the key things we are interested in is, in fact, the linkage rate. Elsewhere (Scheuren, 1983; Smith and Scheuren, 1985a), we have advocated a capture-recapture approach to this estimation problem. Such an approach, in the presence of blocking, will actually allow us to improve the links obtained, as well as make it possible to measure the extent to which our best efforts still lead to erroneous nonlinks. Capture-recapture ideas are well described in the literature (e.g., Bishop et al., 1975; Marks et al., 1974). Here we will only indicate the application.

If we employ more than one set of blocks and keep track for each blocking procedure whether we would have found (and linked) the case in every other blocking scheme, then for any subpopulation of linked records we can construct the usual 2^n table, where we look at the link/nonlink status for each blocking (with "n" being the number of separate blocking schemes). To estimate the number of records not caught by any scheme, three or more sets of blocks are recommended; otherwise, the assumptions made may be unrealistically strong. (The National Death Index, or NDI, already employs many more than this, as we have noted earlier.) For best results the blocks need to be as independent functionally and statistically as is possible, given the linkage information. (Improvements in the current NDI would be recommended here, but these seem to be coming in any case.) Application of these ideas in an IRS or SSA context seems worthy of study (Scheuren, 1983), although the expense of developing such an approach, say at SSA, may never be incurred unless there were a compelling administrative need.

- Adjusting for Mismatches--In most linkage systems practitioners have operated in what they considered to be a conservative manner with regard to the links they would accept. Sometimes this may have meant heavy additional expense in obtaining more information or the risk of seriously biasing results by leaving out a large number of the potential links. In any event, further research is needed on how to apply more complex analytic techniques that take explicit account of the mismatch rate, possibly by use of errors-in-variable approaches where the mismatch rate is estimated, e.g., as in Scheuren and Oh (1975), so that a correction factor can be derived. We must also attempt to find ways of estimating the mismatch rate that make weaker assumptions than those made in most Fellegi-Sunter applications. (Some further ideas on this are found in Smith and Scheuren, 1985a).

In summary, the main issues in the analysis of linked data sets are that, at a minimum, we need to examine the sensitivity of the results to the assumptions made in the linkage process. Where possible, we need to quantify uncertainties in the results; specifically, indeterminacies in the linkages should translate into wider confidence intervals in the estimates. To achieve these goals we need to bring in techniques from other areas of statistics and apply them creatively to linked data sets. Examples here include information theory, error-in-variable approaches and contingency table (capture-recapture) ideas.

3. SOME CONCLUSIONS AND AREAS FOR FUTURE STUDY

In this paper we have dealt with the topic of data linkage in abroad conceptual framework, using examples from recent practice. It is appropriate now to draw out the implications of the point of view expressed for studies of aging and to use that summary as a basis for recommending further research.

Overall Perspective

We have argued elsewhere that the potential for the statistical use of data linkage systems is truly enormous (e.g., Kilss and Scheuren, 1980; Jabine and Scheuren, 1985). The suggestion has even been made that data linkages among administrative records (with some supplementation) might eventually replace conventional censuses in the United States (Alvey and Scheuren, 1982). Such ideas are not new, certainly not to Europeans, where many developed nations have been rapidly moving in this direction (e.g., Pedfern, 1983). Indeed some countries, like Denmark (Jensen, 1983), may have "already arrived."

In the United States there has been some reluctance and resistance to accepting the inevitability of such a future. Grave concerns have been expressed (Butz, 1985) about moving too fast or in the wrong way. After all, while Denmark has succeeded in its efforts, other countries (notably West Germany) have encountered major problems which did grave damage to their statistical programs.

In view of what has happened elsewhere and, especially, given the current state of public opinion, we would caution that any planned use of data linkage systems be grounded firmly in existing practice and not be based on new legislation designed to expand on what it is currently possible to do. On the other hand, it is important to conceptually integrate what is now possible with what might be possible ten or twenty years from now. Some further observations are--

- First, if a data linkage approach is going to be taken, it should be a necessary means, not just a sufficient one, for achieving some required specific purpose. It is simply not enough to argue the need for data linkage on efficiency grounds.

- Second, the linkage should be seen as important by all the cooperating agencies and part of their mission. It is simply not enough that the law can be interpreted to permit such linkages. Positive law, and indeed social custom, must exist which encourages the research, at least in broad outline (Cox and Boruch, 1985).

- Third, strong continuing user support is essential if a long-term basic research effort is to be successful. Program agencies cannot be relied on for really long-run undertakings without this support. Opportunity costs are simply too high. If the linkage system is to be placed in a statistical agency, user involvement is, again, essential (from the outset, if possible). Without strong user involvement, statistical agencies will tend to emphasize continuity of measurement over relevance (while program agencies tend to the reverse).
- Fourth, cost considerations suggest that most data linkage systems be based on, or augment, an existing survey or administrative system. Further, maintenance costs should be low so that in the long run most of the resources can be focussed on exploiting the analytic potential of the system.
- Fifth, access to the results of the linkage system must be basically open not only to the primary user(s), but to secondary users as well. Ways to solve the "reidentification" problem must be built into the undertaking from the beginning and firmly rooted in the best statistical practice.

Still other considerations come to mind, such as adequate physical security during the linkage operation and minimizing the risks by removing identifiers from working files as soon as possible (Kilss and Scheuren, 1978; Steinberg and Pritzker, 1967; Cox and Boruch, 1985; and Flaherty, 1978).

Many _ad hoc_ efforts have succeeded without strictly adhering to one or more of the above; nonetheless, if one is working towards a future which encompasses still more data linkages, it is essential that the strategy taken be absolutely sound and above reasonable reproach.

Potential Data Systems Deserving Further Study

Within the framework just given, there seems to be a clear need to intensively examine the potential of particular data linkage systems to answer certain questions. We will illustrate this point by looking at one of the most pressing areas in the United States where better data are needed -- this is on our rapidly growing aged population. Even if we confine ourselves to this single area, many subsidiary issues must be addressed. For example, where are the greatest gaps: in data on health, general demographic information, financial data, or the extent to which federal programs provide support? In what follows, there has been no attempt to answer this question. To do so, we would go well beyond the scope of the present paper. Instead, there is a discussion of four data linkage environments that, depending on the answer to the question, may warrant further study. Special emphasis has been placed on the limitations of working in each of these settings and of the role that a strong outside user might play in overcoming those limitations.

Social Security and Health Care Financing Administrations -- The Social Security (SSA) and Health Care Financing Administrations (HCFA) are unlikely to take the lead in building and maintaining _general purpose_ statistical data inkage systems, in part because of a reduced emphasis on basic and applied research. Nevertheless, the program-oriented statistical activities of these agencies will continue to give them an important role in data linkage efforts which are consistent with agency missions. The potential at SSA and HCFA for providing improved sources of statistics on the aging population depends on the extent to which they are able to: (1) maintain major in-house data linkage efforts, like the Continuous Work History Sample (e.g., Buckler and Smith, 1980) and the Medicare Statistical System (U.S. Health Care Financing Administration, 1983); (2) continue to sponsor or co-sponsor periodic or _ad hoc_ surveys; and (3) cooperate in linkage studies sponsored elsewhere (for example, in the Survey of Income and Program Participation or in the Health Interview Survey) if they are in support of the agencies' missions.

However, these efforts would need to be coupled with strong outside user support. At SSA and HCFA, there may be a particularly pressing need for outside users to aid in the resumption of some form of public release of subsets, at least, of the administrative samples now being employed almost solely for in-house purposes.

Internal Revenue Service -- It seems pointless to speculate upon the degree to which interagency data linkages can or should take place involving Internal Revenue Service (IRS) data. Formidable statutory barriers narrowly limit access to tax records and, even when the legal requirements can be met, many other agencies, notably the Census Bureau, feel they would be unable to engage in a cooperative study because of concerns about public perception. American social customs, particularly concerns about "Big Brother," stand as nearly insurmountable obstacles in the short run.

It is possible, though, to use IRS records essentially all by themselves as a basis for studying the aged population. This may seem surprising because the statistical program of the Internal Revenue Service is not looked at typically as a source of such information. Certainly the _Statistics of Income_ publication series has focused very little on the aged, and then mainly through the use of the age exemption to identify taxpayers 65 years or older (e.g., Holik and Kozielec, 1984). Broader-based research has been possible through occasional linkages between the IRS's Individual Income Tax Model File and Social Security information. In a few cases, these linkages have resulted in public-use files (DelBene, 1979). What has not been done is to look at the aging population longitudinally, although this is fairly straightforward, at least back to 1972. Furthermore, with the recent addition of complete SSA year-of-birth information to IRS files, it will be possible to routinely study age cohorts by means other than the age exemption. It is also noteworthy of mention that linkages between IRS files and the recently instituted National Death Index have just been successfully instituted (Bentz, 1985).

Tax returns probably represent the single best source of financial information and could, therefore, prove of value in studying the aging process. There are, however, three main limitations to their use:

- First, the income data, while of exceedingly high quality (relative to surveys), are incomplete since certain nontaxable incomes have been omitted (e.g., tax-exempt bond interest and welfare payments). Until recently, social security benefits were unavailable but they are now potentially taxable (beginning with 1984).

- Second, the population coverage of income tax returns is incomplete. In fact, only about half the population ages 65 years or older show up as taxpayers on income tax returns. Again, recent changes have a bearing here since information documents, notably Forms 1099 from Social Security, are filed with the Internal Revenue Service for all social security beneficiaries. This change permits an expanded population concept that could be essentially complete for the aged population.

- Third, the tax return is exceedingly awkward as a unit of analysis for some purposes since it does not always conform to conventional family and household concepts (Irwin and Herriot, 1982). It is possible though, using information documents like Forms W-2 (for wages), Forms W-2P (for private pensions), and Forms 1099 (for social security payments, dividend, interest, etc.), to develop approximate financial profiles of virtually all individuals aged 65 or older. (Major gaps would exist, of course, for supplemental security income recipients and recipients of veterans disability benefits.) There does not appear to be much hope in inferring changes in lifestyles directly from the current IRS information, although the proposed addition of dependent social security numbers could lead to real progress (Alvey and Scheuren, 1982).

Depending on its extent, the cost of maintaining an IRS data linkage system to study aging could be quite modest. Public-use files are possible; but, as with the Social Security and Health Care Financing Administrations, strong outside support would be needed.

National Center for Health Statistics -- Recent changes (Sirken and Greenberg, 1983) at the National Center for Health Statistics suggest that the Center may be assuming a leading role in sponsoring data linkage systems. Naturally and appropriately, the focus of these systems will be quite narrow, looking almost solely at health concerns. The National Health Interview Survey (HIS), involving about 40,000 households annually, appears to be the Center's main survey vehicle for the approach it is planning to take. Continued periodic matching to Medicare records seems planned (Cox and Folsom, 1984) and, of course, the National Death Index can be expected to be fully exploited (Patterson and Bilgrad, 1985). Still other linkage efforts are underway (e.g., Johnston, et al., 1984) which, taken together, suggest that the Center is pursuing a coherent, fully integrated approach, both among its surveys and towards needed vital record systems.

When the social security number question was added to the HIS a few years ago, it was largely for matching to the National Death Index. Great care initially was given to securing informed consent from respondents before obtaining the information. This approach proved tedious and expensive. Now the social security number question is simply asked without much explanation; and, only if requested, are reasons given for why the information needs to be obtained (see Appendix C). Response rates are quite high, about 90%, and it appears that the HIS may constitute a major vehicle for a successful data linkage approach to studying aging. Concerns exist about the reidentification problem, but exactly how the Center will deal with this factor is unclear.

Bureau of the Census -- Historically, the Census Bureau has played a major role in federal data linkage systems involving surveys, sometimes as the sole sponsor (e.g., Childers and Hogan, 1984), but often as a partner in conducting a particular study (e.g., as with Social Security, Bixby, 1970). Much of this work has focussed on the Current Population Survey (Kilss and Scheuren, 1978). Of more promise in future studies of aging has been the development of the Survey of Income and Program Participation (SIPP), which has as one of its design elements the notion that data linkages would be attempted, at least to Social Security information (Kasprzyk, 1983). SIPP, which may settle down to a sample size of about 30,000 households annually, is certainly of sufficient size and scope to look at many general demographic, financial and program related questions concerning aging. The SSN reporting rate is on the order of 90%; hence, the needed resources to "perfect" the linkage (and the analysis problems resulting from faulty or incomplete linkage) should be entirely manageable. Oversampling is possible for particular subgroups (e.g., those aged 65 or older); however, unfortunately, SIPP, like the HIS, is confined to the noninstitutional population and for studies of the very old it may not be suitable alone.

Two difficulties exist with SIPP that further research may resolve. First is the extent to which informed consent is being obtained when the social security number is being secured (SIPP's approach is similar to that in the HIS--see Appendix D). Related to this concern, of course, is the extent to which such consent is felt to be needed. The second issue, and one that seems exceedingly troublesome to the Census Bureau, is the "reidentification" problem. (Briefly stated, the reidentification problem is particularly acute where linkage is concerned, because the cooperating agencies might have enough data on the linked file to reidentify virtually all of the individuals linked.)

The Census Bureau appears to be searching for a solution that involves either simply not releasing public-use files of linked data or releasing public-use files where only very limited linked data have been provided and some kind of masking technique has been employed to prevent reidentification. Given these restrictions, it must be said, there seem to be real difficulties in concluding that there are sufficient benefits to outside users of a SIPP-based data linkage system. Some further comments on this dilemma and ways a general research program could address it are given below.

General Issues Deserving Further Study

Further research is needed on a wide range of data linkage issues, both structural and technical. Four, in particular, stand out from the rest and deserve special attention: ethical and legal concerns, public perception questions, finding solutions to the reidentification problem, and finally, analysis issues in the presence of matching errors.

Ethical concerns such as those raised by Gastwirth (1986) seem to need a more specific answer than they have been given so far (e.g., as by Dalenius, 1983). What might be done is to obtain some data directly bearing on how respondents actually think about data linkage. We could approach this in a way similar to the earlier study by the Committee on National Statistics concerning confidentiality guarantees (Committee on National Statistics, 1979). Within the context of current survey efforts in HIS and SIPP it might be extremely valuable to know how often respondents ask for clarification before providing social security numbers and to code the cases accordingly so we can look at differential refusal rates, for example. Again, exactly what is said (by respondents and interviewers) typically when respondents do ask? Legal and procedural issues abound here, too. For example, how long, even assuming informed consent, can the consent be treated as binding? Social Security practices with outside researchers (when they obtain consent to gain access to individual records) is to treat the consent as binding potentially only once; thus, requests for information on the same subjects may require a renewal of the consent. Signed consent agreements are also required of outside researchers. Such a requirement has never been imposed, say, in Census Bureau surveys, but should it be? If it were, what would be the costs of such a practice in interview time, reduced response, and cooperation generally?

Public perception concerns deserve to be examined in depth. To what extent are we already violating the public's sense of the social customs within which statisticians are supposed to work? The public opinion polling results reported in Gonzalez and Scheuren (1985) need to be followed up. It does not seem defensible simply to speculate about whether this or that approach to data linkage would be acceptable to the public. While we can never use opinion polling to answer all the many specific issues that exist here, much can be done. Of particular interest may be the extent to which the public knows or assumes such linkages take place now and for what purposes; the perceived legitimacy of actual and perceived purposes; whether statutory or contractual prohibitions against efforts at reidentification would be seen to be adequate; and so on.

We do not believe that an entirely satisfactory technical solution to the reidentification problem is possible; but a great deal more can be done to allow for at least limited release of linked information. The work of Paass (1985) and Smith and Scheuren (1985a) is suggestive here. The line of attack that appears most promising is what might be termed a three-step process. First, "slice" the data up into small enough bits so that each of the "bits" can be adequately masked. (The data, for example, might be divided up into disjoint subsets and for each subset of observations, say, only 2 to 4 different items of administrative data would be provided.) Second, if the slices are chosen appropriately, then one can "splice" back together the complete data set using statistical matching; but in a setting where the conventional--and usually false conditional--independence assumption (e.g., Rodgers, 1984) does not have to be made. Finally, the masking step can add "noise" to the data set in such a way that certain analytic results are either invariant under the noise transformation or correction factors can be calculated and readily applied.

There are some serious losses in this approach. For example, the effective sample size of the linked data items may have shrunk considerably. In any case more research on this problem is definitely warranted, (maybe even if contractual and legal solutions turn out to be eventually possible). Either way, public access to the linked data sets must be seen as a key objective when such studies are undertaken and, to the extent possible, release practices should be as open as with any other data set (Committee on National Statistics, 1985).

Finally, a number of analysis issues have been mentioned which deserve further research, especially in measuring matching errors and adjusting the matched results accordingly. In particular, we need to find a way to escape the historical dilemma that the dissemination and growth of sound theory and practice have been retarded by the perceived uniqueness of many linkage problems (and the customized solutions this perception has led to). The profound nature of the common sense principles upon which good practice is based are not widely enough appreciated. Insufficient attention has been paid to the analysis issues in data linkage systems, perhaps because so much creative energy and financial resources typically go into the linkage steps (Smith and Scheuren, 1985a). It may be too optimistic to suppose that things are now changing, but there is some evidence to this effect in the success of the 1985 Washington Statistical Society Workshop on Exact Matching Methodologies (Kilss and Alvey, 1985). In any case, it is time to stop treating matching as a *necessary but dirty business, isolated from* other parts of statistical theory and practice.

ACKNOWLEDGMENTS AND AFTERWORDS

The ideas in this paper owe much to my associations with other professionals in the field of matching. Particular thanks are due to Dan Kasprzyk, for his useful remarks, and, especially, Tom Jabine, whose insightful comments were much appreciated, even though I was unable to incorporate them all in the present version. Tom also acted as a discussant when this paper was originally given and, among other things, corrected a computational error in the calculation of the probability ratios shown in the example. All the remaining errors are, of course, my responsibility.

COMPUTATIONAL NOTE

The Probability Ratios shown in the table above were calculated as follows:

Race and Sex Agree (Race is Black)

$$\frac{99}{100}\cdot\frac{999}{1000}\Big/\left(\frac{1}{10}\cdot\frac{1}{10}\right)\left(\frac{1}{2}\cdot\frac{1}{2}+\frac{1}{2}\cdot\frac{1}{2}\right)= 197.8020$$

Race and Sex Agree (Race is Nonblack)

$$\frac{99}{100}\cdot\frac{999}{1000}\Big/\left(\frac{9}{10}\cdot\frac{9}{10}\right)\left(\frac{1}{2}\cdot\frac{1}{2}+\frac{1}{2}\cdot\frac{1}{2}\right)= 2.4420$$

Race Agrees, Sex Does Not (Race is Black)

$$\frac{99}{100}\cdot\frac{1}{1000}\Big/\left(\frac{1}{10}\cdot\frac{1}{10}\right)\left(\frac{1}{2}\cdot\frac{1}{2}+\frac{1}{2}\cdot\frac{1}{2}\right)= 0.1980$$

Race Agrees, Sex Does Not (Race is Nonblack)

$$\frac{99}{100}\cdot\frac{1}{1000}\Big/\left(\frac{9}{10}\cdot\frac{9}{10}\right)\left(\frac{1}{2}\cdot\frac{1}{2}+\frac{1}{2}\cdot\frac{1}{2}\right)= 0.0024$$

Sex Agrees, Race Does Not

$$\frac{1}{100}\cdot\frac{999}{1000}\Big/\left(\frac{9}{10}\cdot\frac{1}{10}+\frac{1}{10}\cdot\frac{9}{10}\right)\left(\frac{1}{2}\cdot\frac{1}{2}+\frac{1}{2}\cdot\frac{1}{2}\right)= 0.1110$$

Neither Agree

$$\frac{1}{100}\cdot\frac{1}{1000}\Big/\left(\frac{9}{10}\cdot\frac{1}{10}+\frac{1}{10}\cdot\frac{9}{10}\right)\left(\frac{1}{2}\cdot\frac{1}{2}+\frac{1}{2}\cdot\frac{1}{2}\right)= 0.0001$$

REFERENCES

Alexander, L. and Jabine, T.
1978 Access to Social Security Microdata Files for Research and Statistical Purposes: An Overview, Social Security Bulletin, U.S. Social Security Administration.

Alexander, L.
1983 There Ought to be a Law..., Proceedings, Section on Survey Research Methods, American Statistical Association.

Alvey, W. and Aziz, F.
1979 Mortality Reporting in SSA Linked Data: Preliminary Results, Social Security Bulletin, U.S. Social Security Administration.

Alvey, W. and Scheuren, F.
1982 Background for an Administrative Record Census, Proceedings, Social Statistics Section, American Statistical Association.

Aziz, F., et al.
1978 Studies from Interagency Data Linkages (Report No. 8), U.S. Social Security Administration.

Barron, E.
1978 The Survey of Low-Income Aged and Disabled: Survey Design and Data System, Policy Analysis with Social Security Research Files, U.S. Social Security Administration.

Beebe, G.
1985 Why Are Epidemiologists Interested in Matching Algorithms? Record Linkage Techniques--1985, U.S. Internal Revenue Service.

Bentz, M.
1985 The Intergenerational Wealth Study: Prospects for Data Analysis and Methodological Research, presented at the Canadian Conference in Tax Modelling, September 1985.

Bishop, Y., et al.
1975 Discrete Multivariate Analysis: Theory and Practice, MIT Press: Cambridge.

Bixby, L.
1970 Income of People Aged 65 or Older: Overview from the 1968 Survey of the Aged, Social Security Bulletin, U.S. Social Security Administration.

Buckler W. and Smith, C.
1980 The Continuous Work History Sample (CWHS): Description and Contents, Economic and Demographic Statistics, U.S. Social Security Administration.

Butz, W.
1985 The Future of Administrative Records in the Census Bureau's Demographic Activities, Journal of Business and Economic Statistics, American Statistical Association.

Cartwright, D.
1978 Major Limitations of CWHS Files and Prospects for Improvement, Policy Analysis with Social Security Research Files, U.S. Social Security Administration.

Childers, D. and Hogan, H.
1984 Matching IRS Records to Census Records: Some Problems and Results, Proceedings, Section on Survey Research Methods, American Statistical Association.

Clark C. and Coffey, J.
1983 How Many People Can Keep a Secret? Statistical Data Exchange Within a Decentralized System, Proceedings, Section on Survey Research Methods, American Statistical Association.

Cobleigh, C. and Alvey, W.
1975 Validating the Social Security Number, Studies from Interagency Data Linkages (Report No. 4), U.S. Social Security Administration.

Committee on National Statistics
1985 Sharing Research Data, National Academy of Sciences.

Committee on National Statistics
1979 Privacy and Confidentiality as Factors in Survey Response, National Academy of Sciences.

Cox, B. and Folsom, R.
1984 Evaluation of Alternate Designs for a Future NMCUES, Proceedings, Section on Survey Research Methods, American Statistical Association.

Cox, L., et al.
1985 Confidentiality Issues at the Census Bureau, Proceedings of the First Annual Census Bureau Research Conference, U.S. Bureau of the Census.

Cox, L. and Boruch, R.
1985 Emerging Policy Issues in Record Linkage and Privacy, presented at the 45th Session of the International Statistical Institute.

Crane, J. and Kleweno, D.
1985 Project LINK-LINK: An Interactive Database of Administrative Record Linkage Studies, Record Linkage Techniques--1985, U.S. Internal Revenue Service.

Dalenius, T.
1983 Informed Consent or R.S.V.P., Incomplete Data in Sample Surveys (Volume I), Academic Press.

DelBene, L.
1979 1972 Augmented Individual Income Tax Model Exact Match File, Studies from Interagency Data Linkages (Report No. 9), U.S. Social Security Administration.

Fellegi, I.
1985 Tutorial on the Fellegi-Sunter Model for Record Linkage, Record Linkage Techniques--1985, U.S. Internal Revenue Service.

Fellegi, I. and Sunter, A.
1969 A Theory of Record Linkage, Journal of the American Statistical Association, vol. 64, pp. 1183-1210.

Flaherty, D.
1978 The Bellagio Conference on Privacy, Confidentiality and the Use of Government Microdata, New Directions in Program Evaluation, vol. 4, pp. 19-30.

Gastwirth, J.
1986 Discussion comments to paper by George Duncan and Diane Lambert, A Model for Statistical Disclosure Control Based on Predictive Distributions and Uncertainty Functions, Journal of the American Statistical Association, American Statistical Association.

Gonzalez, M. and Scheuren, F.
1985 Future Work by the Conference of European Statisticians on Population and Housing Censuses, presented before the Thirty-Third Plenary Session of the U.N. Conference of European Statisticians.

Holik, D. and Kozielec, J.
1984 Taxpayers Age 65 or Older, 1977-81, Statistics of Income Bulletin, U.S. Department of the Treasury, Internal Revenue Service.

HEW Secretary's Advisory Committee
1973 Records, Computers and the Rights of Citizens, U.S. Department of Health, Education and Welfare.

Howe, G. and Lindsay, J.
1981 A Generalized Iterative Record Linkage Computer System for Use in Medical Follow-up Studies, Computer and Biomedical Research, vol. 14, pp. 327-340.

Irelan, L. and Finegar, W.
1978 Surveys Relating to Retirement and Survivorship, Policy Analysis with Social Security Research Files, U.S. Social Security Administration.

Irwin, R. and Herriot, R.
1982 An Initial Look at Preparing Local Estimates of Household Size from Income Tax Returns, Proceedings, Section on Survey Research Methods, American Statistical Association.

Jabine, T.
1985 Properties of the Social Security Number Relevant to Its Use in Record Linkages, Record Linkage Techniques--1985, U.S. Internal Revenue Service.

Jabine, T. and Scheuren, F.
1985 Goals for Statistical Uses of Administrative Records: The Next Ten Years, Journal of Business and Economic Statistics, American Statistical Association.

Jaro, M.
1985 Current Record Linkage Research, Record Linkage Techniques--1985, U.S. Internal Revenue Service.

Jaro, M.
1972 UNIMATCH--A Computer System for Generalized Record Linkage Under Conditions of Uncertainty, AFIPS-Conference Proceedings.

Jensen, P.
1983 Towards a Register-Based Statistical System--Some Danish Experience, Statistical Journal of the United Nations, vol. 1, pp. 341-365.

Johnston, D. et al.
1984 1980 AHA Hospital and National Natality/Fetal Mortality Survey Linkage Methodology, Proceedings, Section on Survey Research Methods, American Statistical Association.

Kasprzyk, D.
1983 Social Security Number Reporting, the Use of Administrative Records and the Multiple Frame Design in the Income Survey Development Program, Technical, Conceptual and Administrative Lessons of the Income Survey Development Program, Social Science Research Council: New York.

Kelley, R.
1985 Advances in Record Linkage Methodology: A Method for Determining the Best Blocking Strategy, Record Linkage Techniques--1985, U.S. Internal Revenue Service.

Kestenbaum, B.
1985 The Measurement of Early Retirement, Journal of the American Statistical Association, vol. 80, pp. 38-45.

Kilss, B. and Alvey, W.
1985 (Ed.) Record Linkage Techniques -- 1985, U.S. Department of the Treasury, Internal Revenue Service.

Kilss, B. and Scheuren, F.
1980 Goals and Plans for a Linked Administrative Statistical Sample, Proceedings, Section on Survey Research Methods, American Statistical Association.

Kilss, B. and Scheuren, F.
1978 The 1973 CPS-IRS-SSA Exact Match Study, Social Security Bulletin, U.S. Social Security Administration.

Klein, B. and Kasprzyk, D.
1983 Designing an Integrated Disability Data System from Social Security Administrative Records, Proceedings, Section on Survey Research Methods, American Statistical Association.

Kirkendall, N.
1985 Weights in Computer Matching: Applications and an Information Theoretic Point of View, Record Linkage Techniques--1985, U.S. Internal Revenue Service.

Maddox, G.: Fillenbaum, G. and George, L.
1978 Extending the Uses of the LRHS' Data Set, Policy Analysis with Social Security Research Files, U.S. Social Security Administration.

Marks, E., et al.
1974 Population Growth Estimation: A Handbook of Vital Statistics Measurement, The Population Council: New York.

Newcombe, H., et al.
1959 Automatic Linkage of Vital Records, Science, vol. 130, pp. 954-959.

Oh, H. L. and Scheuren, F.
1984 Statistical Disclosure Avoidance, presented before a May 1984 meeting of the Washington Statistical Society.

Oh, H. L. and Scheuren, F.
1983 Weighting Adjustments for Unit Nonresponse, Incomplete Data in Sample Surveys (Volume 2), Panel on Incomplete Data, National Academy of Sciences.

Oh, H.L. and Scheuren, F.
1980 Differential Bias Impacts of Alternative Census Bureau Hot Deck Procedures for Imputing Missing CPS Income Data, Proceedings, Section on Survey Research Methods, American Statistical Association.

Paass, G.
1985 Disclosure Risk and Disclosure Avoidance for Microdata, presented at the May 1985, meetings of the International Association for Social Service Information and Technology (IASSIST).

Parnes, H., et al.
1979 From the Middle to Later Years: Longitudinal Studies of the Preretirement and Postretirement Experiences of Men, Ohio State University.

Patterson, J. and Bilgrad, R.
1985 The National Death Index Experience: 1981-1985, Record Linkage Techniques -- 1985, U.S. Department of the Treasury, Internal Revenue Service.

President's Reoganization Project for the Federal Statistical System
1981 Improving the Federal Statistical System: Issues and Options, Statistical Reporter.

Redfern, P.
1983 A Study of the Future of the Census of Population: Alternative Approaches, commissioned by the Statistical Office of the European Communities.

Rodgers, W.
1984 An Evaluation of Statistical Matching, Journal of Business and Economic Statistics, American Statistical Association, vol. 2, pp. 91-102.

Rogot, E., et al.
1983 The Use of Probabilistic Methods in Matching Census Samples to the National Death Index, Proceedings, Section on Survey Research Methods, American Statistical Association.

Scheuren, F.
1983 Design and Estimation for Large Federal Surveys Using Administrative Records, Proceedings, Section on Survey Research Methods, American Statistical Association.

Scheuren, F.
1980 Methods of Estimation for the 1973 Exact Match Study, Studies from Interagency Data Linkages (Report No. 10), U.S. Social Security Administration.

Scheuren, F. and Herriot, R.
1975 The Role of the Social Security Number in Matching Administrative and Survey Records, Studies from Interagency Data Linkages (Report No. 4), U.S. Social Security Administration.

Scheuren, F. and Oh, H. L.
1975 Fiddling Around with Nonmatches and Mismatches, Proceedings, Social Statistics Section, American Statistical Association.

Sirken, M. and Greenberg, M.
1983 Redesign and Integration of a Population-Based Health Survey Program, presented at 44th Session of the International Statistical Institute.

Smith M.
1982 Development of a National Record Linkage Program in Canada, Proceedings, Section on Survey Research Methods, American Statistical Association.

Smith, W. and Scheuren, F.
1985a Multiple Linkage and Measures of Inexactness: Methodology Issues, presented at the Workshop on Exact Matching Methodologies, Arlington, Virginia, May 9-10, 1985.

Smith, W. and Scheuren, F.
1985b Some New Methods in Statistical Disclosure Avoidance, presented at the 1985 Annual Meetings of the American Statistical Association, in a session sponsored by the Section on Survey Research Methods.

Steinberg, J. and Pritzker, L.
1967 Some Experiences with and Reflections on Data Linkage in the United States, Bulletin of the International Statistical Institute, vol. 42, pp. 786-805.

Storey, J.
1985 Recent Changes in the Availability of Federal Data on the Aged, report prepared for the Gerontological Society of America.

Tepping, B.
1968 A Model for Optimum Linkage of Records, Journal of the American Statistical Association, vol. 63, pp. 1321-1332.

U.S. Bureau of the Census
1973 The Medicare Record Check: An Evaluation of the Coverage of Persons 65 Years of Age and Over in the 1970 Census, PHC(E)-7.

U.S. Health Care Financing Administration
1983 Medicare Statistical Files Manual.

Wilson, O. and Smith, W.
1983 Access to Tax Records for Statistical Purposes, Proceedings, Section on Survey Methods, American Statistical Association.

Winkler, W.
1985 Preprocessing of Lists and String Comparison, Record Linkage Techniques--1985, U.S. Internal Revenue Service.

Appendix A

SUPPLEMENTAL BIBLIOGRAPHIC SOURCES

In this paper we have cited some of the literature on exact and statistical matching when the discussion warranted. Further bibliographic material can be found in the following publications:

- Record Linkage Techniques--1985 (1985), U.S. Internal Revenue Service. (Edited by Beth Kilss and Wendy Alvey.) Many of the citations in the present paper come from this volume, which contains the proceedings of the Workshop on Exact Matching Methodologies, held May 9-10, 1985, in Arlington, Virginia.

- Statistical Working Paper Series (1977-1985), Federal Committee on Statistical Methodology. (Produced under the general editorial guidance of Maria Elena Gonzalez.) See especially, No. 5, on "Exact and Statistical Matching," and No. 6, on the "Statistical Uses of Administrative Records." Some of the publications in the Series were prepared by the U.S. Department of Commerce; more recently the publications have been issued by the U.S Office of Management and Budget.

- Statistics of Income and Related Administrative Record Research (1981-1984), U.S. Internal Revenue Service. (Edited by Beth Kilss and Wendy Alvey.) This annual publication series contains numerous papers on record linkage topics and is a successor to the Social Security publications: Statistical Uses of Administrative Records With Emphasis on Mortality and Disability Research (1979) and Economic and Demographic Statistics (1980), which also may be useful.

- Statistical Uses of Administrative Records: Recent Research and Present Prospects (1984), U.S. Internal Revenue Service. (Edited by Thomas Jabine, Beth Kilss and Wendy Alvey.) This handbook of recent work includes many papers on data linkage, most of which are also found in the series listed above.

- Studies From Interagency Data Linkages (1973-80), U.S. Social Security Administration. (Produced under the general editorial supervision of Fritz Scheuren.) Of special interest may be the bibliography by Scheuren, F. and Alvey, W. (1975), "Selected Bibliography on the Matching of Person Records from Different Sources," which will be found in Report No. 4 in the Series, pages 127-136.

- Policy Analysis with Social Security Research Files (1978), U.S. Social Security Administration. (Edited by Wendy Alvey and Fritz Scheuren.) Most of the research files described are based on data linkage methodologies.

- Accessing Individual Records from Personal Data Using Non-Unique Identifiers, National Bureau of Standards, NBS Special Publication 500-2.

Additional citations to the recent literature on disclosure which may be of value are given below. Some of these are of interest as general background; others focus specifically on disclosure barriers to data linkage.

Crank, S. (1985)
Evaluation of Privacy and Disclosure Policy in the Social Security Administration, Social Security Bulletin, U.S. Social Security Administration.

Dalenius, T. (1985)
Privacy and Confidentiality in Censuses and Surveys, Proceedings, Section on Survey Research Methods, American Statistical Association.

Hansen, M. (1971)
The Role and Feasibility of a National Data Bank, Based on Matched Records and Alternatives, Federal Statistics, Report of the President's Commission (vol. II).

Spruill, N. (1984)
Protecting Confidentiality of Business Microdata by Masking, The Public Research Institute: Alexandria, VA.

Spruill, N. (1983)
The Confidentiality and Analytic Usefulness of Masked Business Microdata, Proceedings, Section on Survey Research Methods, American Statistical Association.

Young, P. (1984)
Legal and Administrative Impediments to the Conduct of Epidemiologic Research, Task Force on Environmental Cancer and Heart and Lung Disease: Washington, DC.

Appendix B

TAXPAYER OPINION QUESTION ON SHARING IRS DATA

Yankelovich, Skelly and White, Inc. (1984)
1984 General Purpose Taxpayer Opinion Survey

60a. As you may know, the IRS has been required by law to keep all of their records confidential. However, some people feel the IRS should share this information with other government departments in order to save money and reduce bureaucratic waste since those departments also need this information to do their work. Others feel that the taxpayer's right to privacy is more important. For which, if any, of these departments or purposes do you think it would be all right for the IRS to provide information?

a. The Census Bureau	24%
b. Major criminal investigations (such as drugs and organized crime)	43%
c. Investigations of illegal aliens	34%
d. Welfare fraud investigations	48%
e. Draft Boards or Selective Service	17%
f. Other U.S. Federal departments	12%
g. State governments	13%
h. Child support investigations	38%
i. Fraud and embezzlement investigations	43%
j. Other	1%
k. None (should keep records private)	31%
l. Don't know/no answer	4%

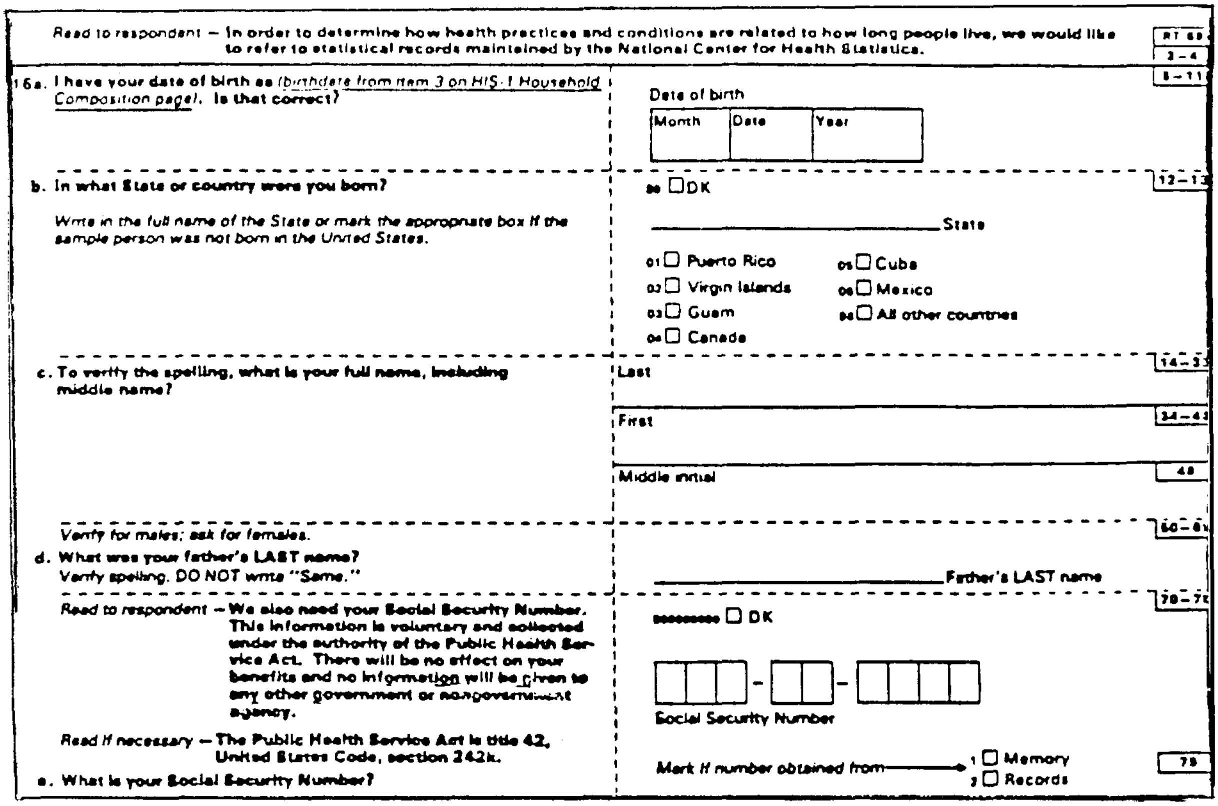
Author's Note:

Tom Jabine, Dan Kasprzyk and others have commented on the many problems this question may have had when it was asked. In my opinion the responses are far from definitive, but they do make the main point I wished to make--that we need more and better research on this issue.

Appendix C

RECORD MATCHING INFORMATION FOR HIS

(Question 16)



Read to respondent – In order to determine how health practices and conditions are related to how long people live, we would like to refer to statistical records maintained by the National Center for Health Statistics.

RT 69
3-4

16a. I have your date of birth as *(birthdate from item 3 on HIS-1 Household Composition page)*. Is that correct?

Date of birth

Month	Date	Year

5-11

b. In what State or country were you born?

Write in the full name of the State or mark the appropriate box if the sample person was not born in the United States.

99 □ DK

____________________ State

01 □ Puerto Rico
02 □ Virgin Islands
03 □ Guam
04 □ Canada
05 □ Cuba
06 □ Mexico
98 □ All other countries

12-13

c. To verify the spelling, what is your full name, including middle name?

Last — 14-33

First — 34-48

Middle initial — 49

Verify for males; ask for females.

d. What was your father's LAST name?
Verify spelling. DO NOT write "Same."

____________________ Father's LAST name

50-69

Read to respondent – We also need your Social Security Number. This information is voluntary and collected under the authority of the Public Health Service Act. There will be no effect on your benefits and no information will be given to any other government or nongovernment agency.

Read if necessary – The Public Health Service Act is title 42, United States Code, section 242k.

e. What is your Social Security Number?

□ DK

□□□-□□-□□□□
Social Security Number

70-78

Mark if number obtained from → 1 □ Memory
2 □ Records

79

Instructions

1. Read the introductory statement above item 16 to explain the purpose of obtaining the information.

*2. When asking 16a, insert the birthdate from the HIS-1, Household Composition Page. If the birthdate recorded in the HIS-1 is in error, make no changes to the HIS-1 entry, but enter the correct birthdate in the answer space in 16a and note "Date verified." If you determine that the person is actually under 55 years of age, footnote the situation and continue the interview. Do not make any changes to the HIS-1(D16-2) or to the supplement. Mark Check Item S2 in Section S based on the original HIS-1 age.

3. Enter the full state name on the line in 16b; do not use abbreviations. If the sample person was not born in one of the 50 states or the District of Columbia, mark the appropriate box in 16b, leaving the state line blank.

4a. If questions arise in 16c, we want the name the sample person is legally known by. If the person has more than one middle name, enter the initial of the first one given. Some women use their maiden name as a middle name: accept the response as given. Be sure to verify the spelling and record the last name first in this item.

*4b. It is acceptable to record an initial as the first name in 16c if this is how the person is legally known. Even if such a person uses their full middle name, only the middle initial is necessary. For example, G. Watson Levi would be recorded as Levi, G., W. in 16c. Do not record name suffixes such as "Sr.," "Jr.," "III," etc.

5a. When verifying 16d for males, ask "Was your father's last name ________?" Always ask the question for females, regardless of their marital status. Be sure to verify the spelling.

5b. Enter the last name of the sample person's father in the answer space, whether it is the same as the person's name or not. Always verify the spelling, even if the names sound alike. If it is volunteered that the person was legally adopted, record the name of the adoptive father.

NOTE: Take special care to make the entries in 16b-d legible. Printing is preferred.

6. Read the introduction to 16e to all respondents. If you are asked for the legal authority for collecting social security numbers, cite the title and section of the United States Code, as printed below the introduction. If you are given more than one number, record the first 9-digit number the respondent mentions, not the first one issued. If the number has more than 9 digits, record the first 9-digits. Do not record alphabetic prefixes or suffixes.

7. After recording the social security number, mark the appropriate box indicating whether the number was obtained from memory or records.

* Revised February 1984

SENSITIVE QUESTIONS

There are no questions considered to be sensitive on either the core series of items or the supplement. However, certain information may be considered sensitive and the following explanation of the need for the data is provided regarding social security number and the subject of incontinence.

- Social Security Number and National Death Index Match

So that in the future the National Center for Health Statistics (NCHS) may investigate the relationship between the results of the "Supplement on Aging" data and causes of death, the supplement collects the appropriate information (items 11a-11e of questionnaire Section 3, Occupation/Retirement), particularly the social security number, that will enable monitoring the National Death Index records for sample persons.

The cost-effectiveness of this supplement is enhanced by the availability of the National Death Index (NDI). Data on the future mortality of the survey population will be available with minimum expenditures by means of a computer search of the NDI. Information on age at death, cause of death, residence at time of death and place of death can be easily ascertained from a copy of the death certificate obtained from the appropriate vital records office. This additional information can be integrated with data from the original survey to greatly enrich the scope of the analysis. Extensive information on the health status of the elderly is being collected on the original survey. Information obtained from death certificates will allow investigators to relate these health status measures to longevity and cause of death. It will also be possible to determine whether selected behavioral and socioeconomic factors collected at the time of the original survey, such as living arrangements, affect the relationship between health characteristics and mortality.

Several years after the data collection and preparation is completed, a list of all survey respondents will be submitted to the NDI and a search made to determine which respondents had died during the interim period. Additional searches of the NDI will be carried out on a periodic basis. In order to optimize the successfulness and reduce the cost associated with these searches, the following information must be collected as part of the original survey: social security number, full (legal) name, Date of birth, State of birth, race, sex, and marital status. Ascertainment of social security number is most essential. A search of the NDI which uses social security number should produce only one match if the subject is deceased. The other information is then used to verify the match. The result of such a match identifies a death certificate which can be obtained from the State with reasonable certainty that it is in fact for the subject. If a social security number is not available, multiple matches within the age range established will occur, especially for common names. This would necessitate obtaining death certificates from several States and attempting to determine whether any of them is for the subject. These false positives would add both acquisition costs and staff costs to the death search process, as well as introducing error.

Interviewers will verify the person's name and birth date (which may have been provided by the household respondent on the core questionnaire), and obtain the last name of the person's father. The social security number will also be requested and if the person is unable to recall the number, he or she will be asked to check their card. This information is not thought to be sensitive; however, respondents will be reminded of the voluntary and confidential nature of the survey, the purpose of the data collection, the legislative authority under which the information is being collected, and the absence of any penalty for refusal. Nonresponse to any of these items will not affect most of the analyses planned for the supplement; however, provision of social security numbers allows for future epidemiologic research for this population without the necessity of conducting a separate longitudinal or followback survey.

- Incontinence

NCHS's and NIA's interests in general physical problems of older people, which relate directly to their quality of life, include questions on urination and bowel control (Pretest Questionnaire Section V, Items 6a-6e, 7a-7e). One issue is the relationship of incontinence to the aging process. In this case, incontinence can be viewed as a health problem, independent of other illnesses. In order to examine this issue, it will be necessary to collect data from all persons in the 55-and-over age group (so that their effects can be examined) and from people both with and without other illnesses.

In addition, a substantial part of the interest in the problem of incontinence results from the relationship between incontinence and institutionalization. It is the view of some experts consulted that incontinence is one of the main reasons for the decision to institutionalize an older person.

Considerable effort went into wording these questions both to minimize sensitivity and to assure comparability with similar items proposed for the 1984 National Nursing Home Survey. Attachment VIII presents planned analysis of comparable data for both the institutionalized and noninstitutionalized populations from the two surveys.

Appendix D

RECORD MATCHING INFORMATION FOR SIPP
(Question 33)

CARD B - Continued
COMMON QUESTIONS AND SUGGESTED ANSWERS

I thought that the Bureau of the Census operated only every 10 years, when they counted people. What is the Bureau of the Census doing now?

In addition to the decennial census, which is conducted every 10 years, the Bureau collects many different kinds of statistics. Other censuses required by law are conducted on a regular basis including the Census of Agriculture, the Censuses of Business and Manufactures, and the Census of State and Local Governments. In addition, we collect data on a monthly basis to provide current information on such topics as labor force participation, retail and wholesale trade, various manufacturing activities, trade statistics, as well as yearly surveys of business, manufacturing, governments, family income, and education.

Why does the Census Bureau want to know my Social Security Number?

We need to know your Social Security Number so we can add information from administrative records to the survey data. This will help us avoid asking questions for which information is already available and help to ensure the completeness of the survey results. The information we obtain from the Social Security Administration and other government agencies will be protected from unauthorized use just as the survey responses are protected.

O.M.B. No. 0607 0425. Approval Expires June 30, 1984

FORM SIPP-4001
U.S. DEPARTMENT OF COMMERCE
BUREAU OF THE CENSUS

CONTROL CARD

SURVEY OF INCOME AND PROGRAM PARTICIPATION

NOTICE – Your report to the Census Bureau is confidential by law (title 13, U.S. Code). It may be seen only by sworn Census employees and may be used only for statistical purposes.

PGM 2

1 REGIONAL OFFICE CODE

2 CONTROL NUMBER — PSU | Segment | Segment suffix | Serial | Sample 8 | 8 | 4

3 ADDRESS I.D.

4 SEGMENT TYPE
1 □ Address
2 □ Unit
3 □ Permit
4 □ Area
5 □ Special place

5a Interviewer code

5b Letter sent

WAVE	1	2	3	4	5	6	7	8	9
5a Interviewer code									
5b Letter sent	□ Yes □ No	□ Yes □ No	□ Yes □ No	□ Yes □ No	□ Yes □ No	□ Yes □ No	□ Yes □ No	□ Yes □ No	□ Yes □ No

6a EXTRA UNIT Original unit serial number

6b Sheet ___ Line ___

6c OFFICE USE ONLY

7 Wave for which Control Card first prepared

HOUSEHOLD RECORD (Card ___ of ___)

FILL ITEMS 17 - 20 FOR ALL PERSONS LIVING OR STAYING HERE

FILL (OR UPDATE AS APPROPRIATE) ITEMS 23 - 33 FOR HOUSEHOLD MEMBERS ONLY - Ask each item for entire household before asking next item

17 ENTRY ADDRESS I.D.

18 PERSON NUMBER — Assign 101, 102, etc. for wave 1; 201, 202, etc. for wave 2

19a HOUSEHOLD ROSTER — FIRST INTERVIEW AT ORIGINAL ADDRESS – What are the names of all persons living or staying here? Start with the name of the person or one of the persons who (owns, rents) this home. Please include middle and maiden names. Last, first, middle, maiden. FIRST INTERVIEW AT MOVER'S NEW ADDRESS for mover's new address – Fill items 17 – 18a, 24 – 33c before the interview.

● 19b RELATIONSHIP TO REFERENCE PERSON (RP) — Ask if not apparent: What is ...'s relationship to (Read name of RP)? Example: Reference person, wife or husband, son, daughter, son-in-law, sister, foster son, partner, lodger, lodger's wife. See codes in Information Booklet

20 HOUSEHOLD MEMBER — Does ... usually live here? 1-Yes 2-No – Probe for URE

● 23 DATE ENTERED OR LEFT — Complete if instructed by item 21, enter code and date. See codes below

● 24 BIRTH DATE/AGE — a. What is ...'s date of birth? Example: 01-20-1963, 12-01-1924. Verify age using flashcard. b. ... is now (Read age) years old. Is that correct?

● 25 PERSON NUMBER OF PARENT — Refer to 18 and 19b. If none enter "999"

ASK IF 15 OR OLDER

● 26a MARITAL STATUS — Is ... now – Married 1-Sp. Pres. 2-Sp. Abs. 3-Widowed 4-Divorced 5-Separated OR 6-Never married?

● 26b PERSON NUMBER OF SPOUSE — Refer to 18, 19b, and 26a. If none enter "999"

● 27 DESIGNATED PARENT OR GUARDIAN — Fill for "Never married" persons under 18. If no parent, ask: Who in this household is responsible for ...? Enter person number

28 SEX — Ask if not apparent: Is ... 1-Male or 2-Female?

29 RACE — SHOW FLASH-CARD. What is the race of each person in this household? See codes below

30 ORIGIN — SHOW FLASH-CARD. What is the origin or descent of each person in this household? See codes below

ASK IF 15 OR OLDER — EDUCATION / ARMED FORCES

● 31a What is the highest grade or year of regular school ... has ever attended? See codes below

● 31b Did ... complete that grade (year)? 1-Yes 2-No

32a Did ... ever serve on active duty in the U.S. Armed Forces? 1 Yes 2 No

If "Yes" in 32a ask:

32b When did ... serve? See codes below

32c Is ... now in the Armed Forces? 1-Yes 2-No

GO TO ITEM 33a

PGM 3

17	18	19a	19b	20 Code	20 Update	20 Yes	20 No	23 First occurrence Code	Mo.	Day	23 Update Code	Mo.	Day	24 Mo.	Day	Year	Age	Age update	25	26a Code	26a Update	26b	27	28 M	28 F	29	30	31a Code	31a Update	31b Yes	31b No	32a Yes	32a No	32b	32c Yes	32c No
						1□	2□																	1□	2□					1□	2□	1□	2□		1□	2□
						1□	2□																	1□	2□					1□	2□	1□	2□		1□	2□
						1□	2□																	1□	2□					1□	2□	1□	2□		1□	2□
						1□	2□																	1□	2□					1□	2□	1□	2□		1□	2□
						1□	2□																	1□	2□					1□	2□	1□	2□		1□	2□
						1□	2□																	1□	2□					1□	2□	1□	2□		1□	2□
						1□	2□																	1□	2□					1□	2□	1□	2□		1□	2□
						1□	2□																	1□	2□					1□	2□	1□	2□		1□	2□
						1□	2□																	1□	2□					1□	2□	1□	2□		1□	2□
						1□	2□																	1□	2□					1□	2□	1□	2□		1□	2□
						1□	2□																	1□	2□					1□	2□	1□	2□		1□	2□
						1□	2□																	1□	2□					1□	2□	1□	2□		1□	2□

21 HOUSEHOLD ROSTER COVERAGE

● FIRST INTERVIEW

		Yes	No
21a I have listed (Read names from item 18a), have I missed – If "Yes," to any of the following, ask name and complete items 17 – 20 above. PGM 4			
– Any babies or small children?	0010	1□	2□
– Any lodgers, boarders, or persons you employ who live here?	0012	1□	2□
– Anyone who usually lives here but is away now – traveling, at school, or in a hospital?	0014	1□	2□
– Anyone else staying here?	0016	1□	2□

● FIRST INTERVIEW AT MOVER'S NEW ADDRESS

		Yes	No
21b I have listed ... (Read names from item 18a), as now living at this address. Is this correct? If NO, ask – Which person did not move here?	0018	1□	2□
21c When did ... move in here? Enter appropriate code and date in item 23			
21d Is there anyone else living or staying here now? If YES, ask name and complete items 17 – 19a, then ask: When did ... begin living here? If moved in before sample person(s), enter code "4" in item 23 and print date. If moved in at the same time or after sample person(s), enter appropriate code and date in item 23.	0020	1□	2□
21e Which of these persons (owns/rents) this home? Assign as reference person in item 19b			

22 HOUSING UNIT COVERAGE

● FIRST INTERVIEW

22a INTERVIEWER CHECK ITEM
Unit is: □ In a special place – SKIP to 24 above
□ Not in a special place – ASK 22b

Ask if not apparent:

22b Do all persons in this household live OR eat together?
□ Yes – ASK 22c
□ No, neither live nor eat together – Fill Table X for the person or group of persons not living or eating with RP

Ask if not apparent:

22c Does any other household on the property live OR eat with this household?
□ Yes – Redefine the unit to include space occupied by all persons who live or eat together. Apply merged unit procedures if appropriate.
□ No
GO TO ITEM 24 ABOVE

● SUBSEQUENT INTERVIEWS

	W2	W3	W4	W5	W6	W7	W8	W9
21f I have listed ... (Read names from item 19a). Are all of these persons still living or staying here? If NO, ask: Who no longer lives here? When did ... leave? Enter appropriate code and date left in item 23. Probe if necessary, for reason left. If entire household moved complete item 23, determine new address (item 34), and end interview.	0022 1□ Y 2□ N	0026 1□ Y 2□ N	0030 1□ Y 2□ N	0034 1□ Y 2□ N	0038 1□ Y 2□ N	0042 1□ Y 2□ N	0046 1□ Y 2□ N	0050 1□ Y 2□ N
21g Is anyone else living or staying here, including new born babies? If YES, list name and complete items 17 – 20 above then ask When did ... begin living here? Enter appropriate code and date in item 23. Anyone else?	0024 1□ Y 2□ N	0028 1□ Y 2□ N	0032 1□ Y 2□ N	0036 1□ Y 2□ N	0040 1□ Y 2□ N	0044 1□ Y 2□ N	0048 1□ Y 2□ N	0052 1□ Y 2□ N

CODES FOR 23

Entered
1 – Birth
2 – Marriage
3 – Other
4 – (Use only with item 21d)

Left
5 – Deceased
6 – Institutionalized
7 – Living in Armed Forces barracks
8 – Moved outside of country
9 – Separation or divorce
10 – Person number 201+ no longer living with sample person
11 – Other

CODES FOR 30

01 – German	10 – Norwegian	19 – Central or South American (Spanish speaking)
02 – English	11 – Russian	20 – Other Spanish
03 – Irish	12 – Ukrainian	21 – Afro-Amer. (Black or Negro)
04 – French	13 – Welsh	30 – Another group not listed
05 – Italian	14 – Mexican-American	39 – Don't know
06 – Scottish	15 – Chicano	
07 – Polish	16 – Mexican	
08 – Dutch	17 – Puerto Rican	
09 – Swedish	18 – Cuban	

CODES FOR 31a

00 – Never attended or kindergarten
01-08 – Elementary
09-12 – High school
21-26 – College (Academic)

CODES FOR 29

1 – White
2 – Black
3 – American Indian, Eskimo or Aleut
4 – Asian or Pacific Islander
5 – Other – Specify below

Person No.	Specify race

CODES FOR 32b

Enter code in descending order of priority. Thus if person served in Vietnam and in Korea, enter "1".

1 – Vietnam Era (Aug. '64 – April '75)
2 – Korean Conflict (June '50 – Jan. '55)
3 – World War II (Sept. '40 – July '47)
4 – World War I (Apr. '17 – Nov. '18)
5 – May 1975 or later
6 – Other Service (All other periods)

SOCIAL SECURITY

33a Person number — Enter person No. from item 18

★ 33b What is ...'s Social Security (or Railroad Retirement) number? If don't know ask: If I call back, would ... be able to provide this information? If "Yes," mark reminder card.

33c Nonresponse — Mark the appropriate box: 1-Don't know 2-Refused 3-None

33a	33b	33c
		1□ 2□ 3□
		1□ 2□ 3□
		1□ 2□ 3□
		1□ 2□ 3□
		1□ 2□ 3□
		1□ 2□ 3□
		1□ 2□ 3□
		1□ 2□ 3□
		1□ 2□ 3□
		1□ 2□ 3□
		1□ 2□ 3□
		1□ 2□ 3□

INTERVIEWER: Go to first questionnaire interview.

Section III: Current Theory and Practice

PREPROCESSING OF LISTS AND STRING COMPARISON

William E. Winkler, Energy Information Administration

1. INTRODUCTION

By combining data on entities from different sources, researchers are often able to perform analyses that would not be possible if they were to use data from individual sources separately.

When a unique common identifier (such as a verified Social Security Number) is available on individual sources of data, matching files merely involves using the unique identifier as the sort key and then directly matching records from the two files.

When a unique common identifier is not available, it is necessary to use other identifying information. Characteristic identifying information might consist of surname, street address, or ZIP code in matching files that contain name and address information. Use of such information involves several practical problems.

First, if the precise locations of identifiers (such as first name and surname) are not consistent from record to record, computer matching using the identifiers cannot be performed. Second, some identifiers may be miscoded or missing on some records. Third, such identifiers, or even combinations of them, are not unique for individuals or businesses.

This paper presents examples of some of the solutions for problems arising in preparing name and address information for use in matching files.

Most of the work described has taken place at the U.S. Bureau of the Census, the Statistical Reporting Service in the U.S. Department of Agriculture, the Energy Information Administration, and Statistics Canada. The problems, examples, and resultant methodologies should be representative of problems that arise in general.

2. BACKGROUND

2.1. Why Preprocessing is Needed

Match/merge strategies generally perform better (i.e., have lower rates of erroneous matches and nonmatches) when address lists have been preprocessed to produce more consistent formats and spellings and to delineate records representing different types of entities (such as records associated with individuals/ sole proprietorships, partnerships, and businesses).

2.2. Definitions

As the terminology of matching is not always consistent from reference to reference, we present definitions.

A match is a pair of records that represent the same unit and a nonmatch is a pair of records that do not. Blocking is a procedure for subdividing files into a set of mutually exclusive subsets under the assumption that no matches occur across blocks. Each mutually exclusive subset consists of records agreeing on the blocking characteristics.

A positive link is a pair of records that is designated as a match. A positive nonlink is a pair of records that is designated as a nonmatch. A possible link is a pair of records that is not designated as a positive link or nonlink. Additional steps, such as manual review or collection of additional information, are needed to designate it as a positive link or nonlink.

A Type I Error is the designation of a pair of records as a positive nonlink when it is a match. Type I Errors have been referred to as erroneous or false nonmatches (U.S. Department of Commerce, 1980). A Type II Error is the designation of a pair of records as a positive link when it is a nonmatch. Type II Errors have been referred to as erroneous or false matches.

2.3. Nature of the Problem

The specific types of match/merge procedures adopted depend on the identifiability and consistency of corresponding information in the address lists to be merged. For instance, if an address list were in free format, then merging would have to be done manually because computer software could not use corresponding information such as NAME or ZIP for blocking pairs of records.

Even if fields such as NAME, ADDRESS, CITY, STATE, and ZIP are identified (possibly using manual techniques), it may not be possible to block records accurately if words in corresponding fields do not contain consistent spellings. For instance, the STATE field and words such as 'COMPANY,' 'CORPORATION,' 'P O BOX,' and 'STREET' should be spelled or abbreviated in a consistent manner.

If subfields such as FIRST NAME, MIDDLE INITIAL(S), SURNAME, STREET NUMBER, STREET NAME, PO BOX NUMBER, ROUTE NUMBER, and SUITE NUMBER are identified and placed in fixed locations, then they can be used for delineating true and false matches. If FIRST NAME and SURNAME subfields are in inconsistent order within the NAME fields of two lists, then it will not be possible to block records accurately using the NAME field.

2.4. Match/Merge Stages

As the need for specific types of preprocessing is closely connected to different match/ merge strategies, these strategies and their relationship to specific data needs will be summarized.

Matching records within or across lists consists of two stages. In the blocking stage, pairs of records are blocked into sets of pairs using a few common characteristics with substantial discriminating power. Some such characteristics are the SOUNDEX abbreviation of SURNAME (see e.g. Bourne and Ford (1961)) or ZIP code. Records for which such common characteristics do not agree are assumed to represent different entities.

In the discrimination stage, blocked pairs are categorized as positive links, positive nonlinks, or potential links using all available discriminating characteristics within blocked pairs of records.

At both stages preprocessing can play an important role. For instance, if records of individuals are blocked using the SOUNDEX abbreviation of the surname, the location of surname needs to be identified and the spelling of surnames needs to be moderately accurate. If records of establishments or businesses are blocked using ZIP code, then ZIP codes need to be accurate.

If the first name, first four characters of the street address, and state abbreviation are used for designating links and nonlinks within a set of blocked pairs, then those fields and subfields need to be located and accurate.

2.5. Topics Addressed in Paper

The remainder of this paper presents examples of the kinds of name and address lists that are encountered and the types of preprocessing that are performed. The third section presents examples illustrating problems with names and addresses in lists that are normally available for updating. The fourth section presents a summary of the various types of preprocessing software and procedures to identify different types of entities, clean up fields and subfields, and identify subfields of the NAME and STREET ADDRESS fields.

The fifth section describes methods for comparing strings that are used to overcome some spelling variations and to create sort keys. The final section poses some problems for further research.

3. EXAMPLES OF PROBLEMS IN NAME AND ADDRESS LISTS

In addition to the problem of locating sources of lists for use in updating, there are problems associated with lists that can make them difficult to use. Problems can include transferral of hardcopy lists to computer files, identification of fields and subfields, and different name and/or address representation of similar entities or similar representation of different entities.

This section provides examples of the problems that affect a list's suitability for use as an update source.

3.1. Keypunch Error in Consistently Formatted Subfields

Addresses in a source list might contain a significant number of typographical errors -- which do not seriously affect manual processing -- while the computerized mailing list does not. The following two pairs of names and addresses representing two entities, from source lists and mailing lists being updated, respectively, illustrate the problem.

(a)	J K Smoth	114 E Main Stret
	J K Smith	114 Main St
(b)	Southside Feul	898 Northwst Hghwy
	Soth Side Fuel	8895 Northwest Hwy

3.2. Unidentified Fields

Address records in which the five fields NAME, STREET, CITY, STATE, and ZIP occur in free format generally cannot be placed in consistent formats using straightforward computer code. They must be reformatted manually. Free format records often exist as address labels in which the five fields occur in no fixed format.

The following examples illustrate the problem of free formats:

(a) A A Fuel Oil
c/o Marvel Distribution Co
PO Box 519
Laramie, Wyoming 66519

(b) Smith Distributing
5632 Westheimer
Suite 43
Houston TX 77514

(c) ABC Oil, PO Box 54
Grand Rapids
Michigan 49506

In example (a) the name occurs on the second line whereas in examples (b) and (c) it occurs on the first. The STREET/PO BOX field appears on the third, second, and first lines of examples (a), (b), and (c), respectively. The CITY field appears in the second to last line in example (c) but on the last line in examples (a) and (b).

3.3. Inconsistently Formatted Subfields

If formatting conventions within subfields of the name and address field vary substantially, merging procedures may not perform as well as in the situation in which corresponding subfields can be readily identified using computer software. For instance, one or more lists might contain records with names and addresses in the following forms:

(a)	J K Smith Co	113 Main
	Smith J K Co	113 E Main St
	Smith Jonathon K Co	PO Box 16
(b)	A A Fuel Co	PO Box 105
	AA Fuel Distribution Inc	Drawer 105
(c)	R Smith Fuel Co	1171 Northwest Highway
	Robert Smith	Highway 65 West
	Smith Co	Route 1

In the first two lines of example (a), both SURNAME and STREET NAME are not obvious matches using a straightforward computer comparison and the billing address in the third entry makes it difficult to determine if the three entries represent the same company.

In example (b), the COMPANY NAME subfields cannot be easily identified and the ADDRESS fields may be difficult to compare. In the example (c), SURNAMES may not be identified and the equating of street addresses of the first two entries requires specific geographic information. Without additional information, it is difficult to determine whether the third entry represents the same company as that given by the first two entries.

3.4. Name and Address Representation

3.4.1. Same Entity, Different Name and Address

Entities in some potential update sources are represented in substantially different forms than the entities are represented in the main mailing list. When this happens, it is difficult to determine those records representing entities that are out-of-scope or duplicates to records in the main mailing list.

For instance, a list of individuals licensed by a state to sell petroleum products might be considered as an update source for a list of businesses selling petroleum products in the state. The reason that the list of owners might be considered is that sending a form to either the owner of a small fuel oil dealership or the appropriate corporate billing address (which might exist in the main mailing list) could yield correct sales information.

Combining such a list of owners with a list of businesses can yield difficulties. Without a suitable additional data source, it may be impossible to identify records representing the same entity that take the following form:

```
J K Smith          116 Main St
  Anytown      66591
A A Fuel           PO Box 68
  Othertown    66442
```

3.4.2. Same or Different Entity, Similar Name, Different Address

If the purpose of a mailing list is to provide one address record for each corporate entity, then additional difficulties can arise. Businesses often maintain substantially different mailing addresses, sometimes even requiring survey forms to be sent to locations in different states. For instance, addresses could take the following form:

```
ABC Fuel Co            116 Main St
  Anytown     CA 96591
ABC Fuel Oil           PO Box 534
  Othertown   NY 10091
J K Smith ABC Co       PO Box 68
  Sometown    KS 66442
```

The first two records could represent the same corporate entity, independent but affiliated companies, or unaffiliated companies. The third address could represent a subsidiary of one of the companies represented by the first two records, a subsidiary of an unidentified company, or an affiliated but independent distributor of products for some ABC Co.

3.4.3. Different Entity, Identical Address and/or Phone

With some lists, different entities may be represented as follows:

```
(a) Pargas of Illinois    PO BOX 661
      NY 10015  202/664-2139
    Pargas of Ohio        PO BOX 661
      NY 10015  202/664-2139
(b) ABC Distributing      1345 Westheimer
      TX 71053  703/789-5439
    Lone Star Oil         1345 Westheimer
      TX 71053  703/789-5439
```

Example (a) illustrates a situation in which a parent company reports separately for two subsidiaries. Example (b) could represent a situation in which an accountant reports for two different companies. The address and phone number could be the accountant's.

Example (b) could also represent different companies which are both located in the same office building or two different companies, one of which has gone out of business. If companies are matched using TELEPHONE, manual followup may be required to determine whether one has gone out of business or is an affiliate of the other.

4. PREPROCESSING METHODS

Methods of preprocessing, using manual procedures or software, have been developed to (1) delineate corresponding classes of records such as those associated with corporations, partnerships, or individuals within a list of businesses; (2) identify corresponding subfields such as HOUSE NUMBER, STREET NAME, and PO BOX; (3) make consistent the spelling of words such as 'STREET,' 'CORPORATION,' and 'ROUTE;' and (4) clean up ZIP codes.

4.1. Identification of Individuals, Partnerships, and Corporations

As records associated with individuals/sole proprietorships, partnerships, and corporations within a list of businesses have different characteristics, they are sometimes distinguished and processed separately. The U.S. Department of Agriculture/Statistical Reporting Service (USDA/SRS, 1979) and the U.S. Department of Commerce (1981) have developed software and/or procedures for identifying individuals, partnerships, and corporations in lists of farms.

It appears that partnerships are identified as those records having '&' in the NAME field. Corporations are those records having words such as 'CORP,' 'CO,' 'INC,' 'FARMS,' and 'DAIRY' in the NAME field. Individuals are those records not classified as partnerships or corporations.

Records associated with partnerships are more difficult to process (may require more manual followup) because partnerships can be erroneously matched more times than records associated with individuals and because partnership records can take the following inconsistent forms:

```
Smith John A & Mary B
Smith John & Jones Lee
Smith John A, Smith Mary B, & Lee Jones
Smith Mary B & Jones Lee
Smith Mary B & Smith John A
```

The first entry contains only one SURNAME entry while others contain one SURNAME for each partner. The third entry represents a partnership of three individuals while the others represent only two. Due to ordering differences in entries two through four, it is difficult to determine if Jones or Lee is the individual's surname.

4.2. Formatting and CLeanup of the Name Field Subfields

Cleanup of the name field consists of replacing common words such as 'COMPANY,' 'INCORPORATED,' 'LIMITED,' 'FARMS,' 'BROTHERS,' 'SALES,' and 'DISTRIBUTOR' with standard spellings or abbreviations and replacing common variations of first names such as 'ROBERT,' 'BOB,' 'ROB,' 'ROBT' with standard spellings or abbreviations.

The standardization is typically done using lookup tables that contain previously identified spelling variations. Such lookup tables are easily updated when new spelling variations are encountered. Lookup tables are in use at USDA/SRS (1979), the U.S. Department of Commerce (1978b, 1981), the Energy Information Administration (EIA) (Winkler, 1984), and Statistics Canada (1982).

Formatting of name fields associated with individuals involves manually identifying the subfields FIRST NAME, MIDDLE INITIAL(S), and SURNAME and either placing them in fixed locations (USDA/SRS, 1979) or in fixed order (U.S. Dept. of Commerce, 1981). If NAME subfields are in fixed order, then software can be used to identify individual subfields.

4.3. Formatting and Cleanup of the Street/Mailing Address Field

Cleanup of the street/mailing address involves replacing such commonly occurring words as 'STREET,' 'PO BOX,' 'RURAL ROUTE,' 'DRAWER,' 'AVENUE,' and 'HIGHWAY' with standard spellings or abbreviations. Such standardization typically involves lookup tables that are easily updated as new spelling variations are encountered.

Various spellings of large cities in the CITY field can also be standardized using lookup tables. Such standardization may only be partially effective because of the large differences in spelling and abbreviations used for core cities and suburbs in large metropolitan areas.

Formatting can also involve placing subfields such as STREET NAME, STREET NUMBER, PO BOX NUMBER, RURAL ROUTE in fixed locations (USDA/SRS, 1979; U.S. Dept. of Commerce, 1978b; Statistics Canada, 1982).

ZIPSTAN software (U.S. Dept. of Commerce, 1978b) has been developed to identify pertinent subfields of the STREET field in files of individuals. The following examples show representative EIA records before and after ZIPSTAN processing:

Figure 1. -- Before ZIPSTAN

1. EXCH ST
2. HWY 17 S
3. 1435 BANK OF THE
4. 2837 ROE BLVD
5. MAIN & ELM STS
6. CORNER OF MAIN & ELM
7. 100 N COURT SQ
8. 100 COURT SQ SUITE 167
9. 2589 WILLIAMS DR APT 6
10. 15 RAILROAD AVE
11. 2ND AVE HWY 10 W
12. MAIN ST
13. 184 N DU PONT PKWY
14. 1230 16TH ST
15. BOX 480

Figure 2. — After ZIPSTAN

No.	House No.	Prefixes		Street Name	Suffixes		Unit
		1	2		1	2	
1.				EXCH	ST		
2.		HW		17TH	S		
3.	1435			BANK OF THE			
4.	2837			ROE	BL		
5.				MAIN ELM STS			
6.				CORNER OF MAIN ELM			
7.	100	N		COURT	SQ		
8.	100	CT	SQ	*** NO NAME ***			RM 167
9.	2589			WILLIAMS	DR		AP 6
10.	15			RAILROAD	AV		
11.				2ND	AV	HW	10
12.				MAIN	ST		
13.	184	N		DU PONT	PW		
14.	1230			16TH	ST		
15.	480			*PO BOX*			

ZIPSTAN is able to identify accurately subfields in 13 of 15 cases. The two exceptions are cases 2 and 8. In case 2, 'HWY' is moved to a prefix position and '17' is placed in the STREET NAME position. In case 8, 'COURT,' the STREET NAME, is placed in a prefix location.

Although ZIPSTAN accurately identifies the subfields associated with intersections (cases 5, 6, and 11), such identification may not allow accurate delineation of duplicates in comparisons of various lists. Some lists may contain STREET ADDRESS in the following forms, none of which is readily comparable with the forms in examples 5, 6, and 11.

5. 34 Main St
5. Elm and Main Streets
11. Hwy 10 W
11. 7456 Richmond Hwy

5. METHODS OF STRING COMPARISON

If comparable strings have been identified (see sections 3.4, 4.2, and 4.3), then it is useful to compute a distance between them in blocked pairs of records. If properly devised, string comparators can overcome minor spelling errors.

5.1. Abbreviation Methods

Abbreviation methods (see e.g., Bourne and Ford, 1961) are intended to maintain some information needed for identifying a record while alleviating problems due to spelling variations. As an example, the SOUNDEX abbreviation method will be described and illustrated.

The SOUNDEX abbreviation of an alphabetic word consists of four characters. The first SOUNDEX character agrees with the first character in the word. All nonleading vowels and the letters H, W and Y are deleted. Similar sounding consonants are mapped into integer codes as follows:

B, F, P, V -> 1,
C, G, J, K, Q, S, X, Z -> 2,
D, T -> 3,
L -> 4,
M, N -> 5, and
R -> 6.

Repeating integer codes are deleted and SOUNDEX abbreviations of less than four characters are zero filled on the right.

Comparison of SOUNDEX abbreviations of words induces a metric in which agreeing SOUNDEX abbreviations are assigned distance 0 and disagreeing 1.

5.2. General String Comparators

As common abbreviation methods (section 5.1) are not able to deal with typical coding errors, more exotic methods for string comparison have been introduced.

An early comparator is the Damerau-Levenstein (D-L) metric (see e.g., Hall and Dowling, 1980, pp. 388-390). The basic idea of the metric is as follows. Any string can be transformed into another string through a sequence of changes via substitutions, deletions, insertions, and possibly reversals. The smallest number of such operations required to change one string into another is the measure of the difference between them.

The minimum value that the D-L metric can assume is 0 (character-by-character agreement) and the maximum is the maximum number of letters in the two words being compared. For instance, the D-L distance between 'ABCDEFG' and 'WXYZ' is 7.

Using the Damerau-Levenstein metric or various straightforward extensions of it (see e.g., Hall and Dowling, 1980) is difficult because: (1) the dynamic programming necessary for computing the metric is cumbersome and (2) neighborhoods of given strings contain too many unrelated strings (i.e., the metric does not have good distinguishing power, see section 5.3).

5.3. Jaro's String Comparator

Jaro (see e.g., U.S. Dept. of Commerce, 1978a, pp. 83-108) introduced a string comparator that is more straightforward to implement than the Damerau-Levenstein metric and more closely relates to the type of decisions a human being would make in comparing strings.

The string comparator is a weighting function for pairs of strings denoted as reference file strings and data file strings. It is defined as follows (U.S. Dept. of Commerce, 1978a, p. 108):

```
    W = wgt_cd*c/d + wgt_rd*c/r +
        wgt_tr*(c-tr)/c
 where
  wgt_cd = weight associated with characters in
           the data file string but not in the
           reference file string;
  wgt_rd = weight associated with characters in
           the reference file string but not in
           the data file string;
  wgt_tr = weight associated with
           transpositions;
    d    = length of the data file string;
    r    = length of the reference file string;
    tr   = number of transpositions of
           characters; and
    c    = number of characters in common in
           the two strings.
```

Two characters are considered in common only if they are no further apart than (m/2 - 1) where m = max(d,r). Characters in common from two strings are said to be assigned. Other characters from the two strings are unassigned. Each string has the same number of assigned characters because each assigned character represents a match.

The number of transpositions are computed as follows: The first assigned character on one string is compared to the first assigned character on the other string. If the characters are not the same, half of a transposition has occurred. Then the second assigned character on one string is compared to the second assigned character on the other string, etc. The number of mismatched characters is divided by two to yield the number of transpositions.

If two strings agree on a character-by-character basis, then the Jaro weight, W, is set equal to wgt_cd+wgt_rd+wgt_tr, which is the maximum value that W can assume. The minimum value that the Jaro weight, W, can assume is 0, which occurs when the two strings being compared have no characters in common (subject to the above definition of common).

5.4. Manual Comparison

The purpose of different string comparators is to assign a value to the quality of comparison in a manner that mimics how a human being might make a decision. Because of this, it is useful to describe how manual review decisions can be quantified. In section 5.5, the manual review decisions will be compared to results obtained using the string comparators of sections 5.1-5.3.

Quantification of manual review decisions can be performed as follows:

1. have a number of individuals compare pairs of corresponding substrings such as SURNAMEs;
2. score comparisons using the scale: 1-no match, 2-likely false match, 3-possible true match, 4-likely true match, and 5-true match; and
3. average results of the comparisons over individuals and compute the corresponding coefficients of variation.

5.5. Comparison of String Comparators

Table 1 provides a comparison of the measures of agreement using the SOUNDEX abbreviation, the Damerau-Levenstein metric, Jaro's string comparator, and a weight based on manual review. To make the values in the table easier to compare, all measures were transformed to a scale from 0 to 1. A value of 0 represents nonmatch and a value of 1 represents match.

The transformations are performed as follows:

```
1. SOUNDEX=1-SOUNDEX;
2. D_L    =(5-D_L)/5;
3. JARO   =JARO/900; and
4. MAN    =(MAN-1)/4.
```

In equations 1-4 the measures on the right-hand side (as defined in sections 5.1-5.4) are replaced by the scaled measures. As the basic Damerau-Levenstein metric D-L (section 5.2) on the right-hand side of equation 2 varies from 0 (total agreement) to 5 (substantial disagreement) for the examples in Table 1, the scaled

D-L metric is transformed into a weight in which 0 and 1 represent nonmatch and match, respectively.

In computing the Jaro weight, JARO, the weights wgt_cd, wgt_rd, and wgt_tr (section 5.3) are each given the values 300 which are the same as the default values given in the Census software (U.S. Dept. of Commerce, 1978a, p. 88). As the basic JARO weight on the right hand side of equation 3 varies between 0 and 900, dividing by 900 changes the scale from 0 to 1.

In Table 1, with the exception of example (h) (completely different words), all examples represent similar character strings that disagree because of minor transcription/keypunch errors. Each pair of surnames is taken from EIA files. With the exception of example (h), the surnames represent the same entity.

Overall, we can see that the SOUNDEX weight is high for only 5 of 9 matching surname pairs; D-L weights are generally moderately high to high for 8 of 9; Jaro weights are consistently high; and the manually estimated weights vary significantly with no apparent consistency. It is important to note that, with the exception of example (h), all weights should be consistently high.

In comparing the D-L metric and the Jaro weight, we see that the Jaro weight gives additional weight to longer, but similar, strings. For instance, with short strings in which one character disagrees (examples (f) and (i)), the D-L and Jaro weights are about the same. With longer strings in which one character disagrees (examples (d) and (e)), the Jaro weight is higher than the D-L weight.

For example (g), it is interesting to note that the manually estimated weight of 0.88 is lower than the weight of 1.0 provided by each of the other string comparators. Human beings are able to make use of the auxiliary information that "Smith" is a commonly-occurring word and downweight their judgements accordingly. Such downweighting is inherent in the application of the Fellegi-Sunter model which utilizes frequency of occurrence of character strings (see e.g., Rogot, Schwartz, O'Conor, and Olsen, 1983, p. 324).

6. NEEDED FUTURE WORK

Although it is intuitive that preprocessing can both identify information that should correspond and make such information more consistent, few, if any, studies have been set up to determine its effectiveness. We do not know how much different types of preprocessing reduce matching error rates, nor do we know the extent to which they lower amounts of manual processing.

Effective evaluation may require the creation of data bases with all matches identified and suitably connected to entities used for mailing purposes. Fellegi and Sunter (1969) indicate that error rates obtained using samples are subject to substantial variability unless the samples are very large. Winkler (1984) provides examples of rates of erroneous nonmatches based on samples of size 1,800 for which the estimated sampling error exceeds the estimated error rate.

A key issue that needs to be addressed is whether the results obtained by empirical evaluation of methodologies on one data set are likely to be relevant to a different data set. Specific research problems follow.

6.1. Effects of Spelling Standardization

How much does standardization of the spelling of words such as 'COMPANY,' 'CORPORATION,' 'PO BOX,' 'STREET,' and 'EAST' reduce the error rates associated with a given matching strategy? What errors can certain types of standardization induce?

Some matching strategies consist of blocking files of individuals using the SOUNDEX or New York State Intelligence and Identification (for NYSIIS, see Lynch and Arends, 1977) abbreviations of surnames. When compared with blocking using surname, how much does blocking using abbreviated surnames reduce the rate of erroneous nonmatches and can such abbreviations provide information useful for delineating matches and nonmatches within the set of blocked pairs?

Some matching strategies consist of blocking files of businesses using the ZIP code and first few characters of the NAME field. How much effort is involved in cleaning up ZIP codes and how much do the cleaner ZIP codes reduce rates of erroneous nonmatches? Should the ZIP codes in a given metropolitan area all be mapped into one sort key used for blocking records?

How much can the delineation of true and false matches be improved if the spelling and formatting of the CITY field are made more consistent? What are the best strategies for correcting inconsistencies in the CITY field?

6.2. Effect of Formatting of Subfields

How much does the identification of SURNAME, FIRST NAME, HOUSE NUMBER, STREET NAME, and PO BOX help reduce error rates? What subfields provide the greatest reduction? Are the subfields providing the greatest reduction different in files of businesses than in files of individuals?

6.3. Abbreviation Methods Used in Blocking

What are the best methods for blocking files of individuals? Blocking on surnames abbreviated using methods such as SOUNDEX and NYSIIS will usually designate as nonmatches those matches containing errors due to miskeying, insertions, deletions, and transpositions.

In comparing methods of abbreviation and blocking, we need to consider rates of erroneous nonmatches, total number of pairs in all blocks, and computing requirements if some blocks are large. Given these evaluation criteria, are there methods of abbreviation and blocking that would perform better than SOUNDEX or NYSIIS?

6.4. Effect of String Comparison

How much does the string comparator of Jaro (section 5.3) that is used for computing agreement weights for corresponding subfields such as SURNAME, FIRST NAME, and STREET NUMBER (U.S. Dept. of Commerce, 1978a) help reduce rates of erroneous matches? Are there better algorithms for string comparison? What measures should be used in comparing the effectiveness of two string comparators?

ACKNOWLEDGMENTS

The author wishes to thank Yvonne M. Bishop, Thomas Jabine, and Nancy J. Kirkendall for a number of helpful comments.

REFERENCES

Bourne, C. P., and Ford, D. J. (1961), "A Study of Methods for Systematically Abbreviating English Words and Names," J. ACM 8, 538-552.

Damerau, F. J. (1964), "A Technique for Computer Detection and Correction of Spelling Errors," Communications of the ACM. 7, 171-176.

Fellegi, I. P., and Sunter, A. B. (1969), "A Theory for Record Linkage," JASA 40, 1183-1210.

Hall, P. A. V. and Dowling, G. R. (1980), "Approximate String Matching," Computing Surveys 12, 381-402.

Lynch, B. T. and Arends, W. L. (1977), "Selection of a Surname Coding Procedure for the SRS Record Linkage System," U.S. Department of Agriculture, Statistical Reporting Service.

Morgan, H. L. (1970), "Spelling Correction in Systems Programs," Communications of the ACM, 13, 90-94.

Newcombe, H.B. and Kennedy, J.M. (1962), "Record Linkage," Communications of the ACM, 5, 563-566.

Rogot, E., Schwartz, S., O'Conor, K., and Olsen, C. (1983), "The Use of Probabilistic Methods in Matching Census Samples to the National Death Index." ASA 1983 Proceedings of the Section on Survey Research Methods, 319-324.

Statistics Canada/ Systems Development Division (1982), "Record Linkage Software."

U. S. Department of Agriculture/ Statistical Reporting Service (1979), "List Frame Development: Procedures and Software."

U. S. Department of Commerce, Bureau of the Census/Agriculture Division (1981), "Record Linkage for Development of the 1978 Census of Agriculture Mailing List."

U. S. Department of Commerce, Bureau of the Census/Survey Research Division (1978a), "UNIMATCH: A Record Linkage System."

U. S. Department of Commerce, Bureau of the Census/Survey Research Division (1978b), "ZIPSTAN: Generalized Address Standardizer."

U. S. Department of Commerce, Office of Federal Statistical Policy and Standards (1980), "Statistical Policy Working Paper 5: Report on Exact and Statistical Matching Techniques."

Winkler, W. E. (1984), "Issues in Developing Frame Matching Procedures: Exact Matching Using Elementary Techniques." Presented to the ASA Committee on Energy Statistics in April 1984. A summary appeared in Statistics of Income and Related Administrative Record Research: 1984 U.S. Department of the Treasury, Internal Revenue Service, Statistics of Income Division, 171-176. The summary also appeared in the ASA 1984 Proceedings of the Section on Survey Research Methods, 327-332.

Table 1: Comparison of String Comparator Metrics Using Surnames that are Generally Similar

	Surnames	Maximum string length	SOUNDEX	D-L	Jaro	Manual	CV 1/
(a)	Tranisano Traivsano	9	0.00	0.60	0.93	0.35	40.3
(b)	Alexander Aleander	9	0.00	0.80	0.96	0.63	15.1
(c)	Nuzinsky Newzinski	9	1.00	0.40	0.81	0.42	39.2
(d)	Smthfield Smithfeld	9	1.00	0.60	0.93	0.63	20.2
(e)	Bachman Bahcman	8	1.00	0.80	0.96	0.63	30.9
(f)	Dixon Nixon	5	0.00	0.80	0.87	0.13	35.1
(g)	Smith Smith	5	1.00	1.00	1.00	0.88	24.0
(h)	Smith Jones	5	0.00	0.00	0.00	0.00	0.0
(i)	Ouid Ovid	4	0.00	0.80	0.83	0.55	13.2
(j	Boc Boco	4	1.00	0.80	0.92	0.32	29.3
Number of values above 0.5		NA	5	8	9	5	NA

1/ Coefficient of variation associated with estimate based on manual review by nine individuals.

WEIGHTS IN COMPUTER MATCHING: APPLICATIONS AND AN INFORMATION THEORETIC POINT OF VIEW

Nancy J. Kirkendall, Energy Information Administration

This paper summarizes the historical development of computerized match/merge procedures and describes the test statistic used to classify record pairs as a match or nonmatch in terms of its information theoretic interpretation. Current match/merge software procedures are compared and contrasted based on their differing approaches to estimation.

INTRODUCTION

The match/merge procedures discussed in this paper are those which are intended to perform exact matching. Exact matching has been defined (U.S. Department of Commerce, 1980) as the linkage of records from two or more files containing units from the same population. The intention of exact matching is to link data for the same unit (e.g., person) from different files. If units which do not represent the same individual are linked, the result is a false match or type 2 error. If units which do represent the same unit are not linked, the result is a missed match, or type 1 error.

There are many different purposes in exact matching. Examples range from obtaining more data elements for an individual by merging information from different surveys, to creating a more comprehensive name and address list by merging the names and addresses from many sources. In the first case, it is important to make sure that matching is done accurately so that the merged data constitute a multivariate observation from a single individual (see Kelley, 1983). In the second case, the merging is intended to ensure as complete a list as possible while eliminating duplication.

The most significant paper on the theory and practice of matching is by Fellegi and Sunter (1969). Their paper documents the derivation of a test statistic and a critical region for deciding whether or not a pair of records is a match. In addition, it discusses some of the assumptions necessary for practical application and describes approaches for estimating the probabilities which are used to calculate the test statistic. Most of the probabilistic match/merge procedures in use today are based on an application of the techniques described in the Fellegi-Sunter paper.

Although the Fellegi-Sunter paper was the first publication of the theoretical background for match/merge procedures, many of the ideas and techniques embodied in the methodology had been used since the late 1950's by Howard Newcombe et al. Newcombe's papers from that time period describe the use of the test statistic for which the derivation was later presented by Fellegi and Sunter. (See Newcombe et al., 1959 and Newcombe and Kennedy, 1962.)

THEORETICAL BACKGROUND

Assume that two files, A and B, are to be merged. Each file contains at least one record for each unit (person or establishment) in the file. Each record contains a set of attributes for that unit. These attributes may include: numerical identifiers with very good identifying characteristics such as the social security number; standard identifiers such as name and address; characteristic information such as sex or date of birth; or any other data which might be available on survey files or administrative record files.

In the matching process, each record in file A can be compared to each record in file B. The comparison of any such pair of records can be viewed as a set of outcomes, each of which is the result of comparing a specific attribute from the record in file A with the same attribute in the record from file B. Outcomes may be defined as specifically as desired. For example, one might define an outcome of a comparison to be simply that the attributes agree or that they disagree. Or, one might define the agreement outcome more specifically, based on the possible values that attribute can take. For example, one outcome might be that the surnames agree and equal "Smith," while another might be that the surnames agree and equal "Zebra," etc.

"Comparison of attributes" is usually interpreted to mean that the same attribute is recorded on each record and that they can be compared directly. However, it is possible to "compare" different attributes which are known to be correlated or to use information from only one record in conjunction with general information from the other file. An example is given in Smith, Newcombe, and Dewar (1983). In their application, records from a file of patients diagnosed as having cancer are linked with records in a death file. The variable "cause of death" in the death file is used in conjunction with general statistics concerning the cause of death among cancer patients and the cause of death among the general population to provide a different sort of "comparison of attributes."

In the above, it was implied that every record from file A is compared to every record from file B. In practice, with large files this would require an extremely large number of comparisons, the vast majority of which would not be matches. To make the size of the problem more manageable, files are generally "blocked" using one or more of the available attributes, and record pairs are assumed to be a possible match and subject to the detailed attribute comparison only if they agree on the blocking attribute. In using a blocking procedure, there is necessarily a higher rate of unmatched duplicates (type 2 error) because records which do represent the same unit, but disagree on the blocking attribute, are automatically rejected as possible matches. However, the gains in the form of reduced processing are significant. See Kelley (1985) for a probabilistic approach to selecting blocking strategies.

THE PROBLEM

Probabilistic test procedures are based on evaluating record pairs one at a time and subjecting each pair to a decision as to its match status. The procedure does not consider the expected number of matches or nonmatches in a merging of two files, and does not make use of the result of the classification of any previous record pairs.

In this section the test statistic and the critical region are described based on an information theoretic argument. Details of the derivation are presented in the Appendix. The resulting test statistic and critical region are exactly the same as those derived by Fellegi and Sunter. One advantage of the information theoretic approach is that the inclusion of the log of the prior odds of a match, as described by Howe and Lindsay (1981) and by Newcombe and Abbatt (1983) can be directly related to the methodology. Calculation of this test statistic yields a value which is commonly referred to as the "weight" for or against a match.

Given any pair of records, we want to make a decision as to whether they match (H_o -- the null hypothesis) or do not match (H_1 -- the alternative hypothesis). This decision will be based on the observed comparison of the attribute items on the two records. The set of all outcomes resulting from this comparison is the random variable, x_i, which takes values according to the outcomes which were specified for all of the attributes.

The discrete random variable, x_i, can take any of n different values. The number n can be very large, either because a large number of attributes are compared, or because a large number of outcomes are possible for any one attribute comparison. The probabilities with which x_i takes any of the n values under both H_o and H_1 are assumed to be known. The question of estimating these probabilities is addressed later. The decision process is formalized by considering the following two hypotheses:

H_o: The event that two records represent the same unit (i.e., a match). Under H_o, the frequency function of the random variable, x_i, is denoted $P(x_i/H_o) = p_{oi}$ for i=1, ... n.

H_1: The event that the two records represent different units (i.e., a nonmatch.) Under H_1, the frequency function of the random variable, x_i, is denoted $P(x_i/H_1) = p_{1i}$ for i=1, ... n.

AN EXAMPLE OF A COMPARISON VARIABLE

Assume that two records are being compared and that a decision will be made as to their match status based on a comparison of three attributes: surname, first name, and sex. For each attribute there will be two possible outcomes: either they agree or they do not agree. Thus, the comparison set can take any of 2**3 = 8 (n=8) possible values. For simplicity we also assume that the probabilities of agreement or disagreement of the attributes are independent under both H_o and H_1. Thus, given the following table of probabilities, the frequency function of the comparison vector can be calculated under both hypotheses.

TABLE I
PROBABILITIES OF AGREEMENT

Attribute	Under H_o	Under H_1
Surname	.90	.05
First name	.85	.10
Sex	.95	.45

In the following let $x=(a_1,a_2,a_3)$, where $a_i = 0$ if item i disagrees, and $a_i=1$ if item i agrees. The comparison of surname is represented by a_1, the comparison of first name by a_2, and the comparison of sex by a_3. Thus, the random variable, x_i, has the frequency functions given by p_{oi} (under H_o) and p_{1i} (under H_1) in the following table.

TABLE II
PROBABILITIES FOR COMPARISON VARIABLE

i	x_i	p_{oi}	p_{1i}
1	(0,0,0)	.0008	.4703
2	(1,0,0)	.0068	.0248
3	(0,1,0)	.0043	.0523
4	(0,0,1)	.0143	.3848
5	(1,1,0)	.0383	.0028
6	(1,0,1)	.1283	.0203
7	(0,1,1)	.0808	.0428
8	(1,1,1)	.7268	.0023

THE TEST STATISTIC

As shown in the Appendix, the test statistic

$$T(x_i) = \log(p_{oi}/p_{1i}) = I(o:1;x_i). \qquad (1)$$

is a sufficient statistic for discriminating between H_o and H_1. The number log (p_{oi}/p_{1i}) is an information number. It provides a measure of the information for discriminating for H_o and against H_1 which was gained by observing the random variable, x_i.

$T(x_i)$ is the log of the ratio of the probability of the outcomes, denoted by x_i, under H_o to the probability of the *same* set of outcomes under H_1 (the log of the likelihood ratio.) Note that if these probabilities are the same then $T(x_i)=0$, and this set of outcomes has no discriminating power for identifying whether records represent the same unit. If p_{oi} is larger than p_{1i}, then $T(x_i)$ will be positive for that category. The larger $T(x_i)$, the stronger is the possibility that observation of this set of outcomes indicates that the records represent the same unit. If p_{oi} is smaller than p_{1i}, then $T(x_i)$ is negative. The smaller $T(x_i)$, the stronger is the possibility that this set of outcomes indicates that the records do not represent the same unit.

DETERMINING THE CRITICAL REGION

The final part of the matching problem is to determine cut-off values, c_1 and c_2, so that H_1 is rejected if $T(x_i)$ is greater than c_2 and H_o is rejected if $T(x_i)$ is less than c_1. If $T(x_i)$ falls between these two values, the test is inconclusive and the record pair may be subject to manual follow up.

In standard applications of testing simple hypotheses, there are only two outcomes: accept the null hypothesis or reject it. Here, the three region test comes from the union of two tests. First, consider a test of H_o vs. H_1. For a test with significance level alpha, this leads to the critical region defined by c_1. Next, consider the test of H_1 vs. H_o with significance level beta. This leads to a critical region defined by c_2. Individually, according to the Neyman-Pearson Lemma, these tests are the best tests at their respective significance levels. The first test rejects H_o if $T(x_i)$ is less than c_1. The second test rejects H_1 if $T(x_i)$ is greater than c_2. Since c_1 is generally less than c_2, the union of these two tests yields the three region test described above.

This is illustrated below with our previous example. In Table III the column labeled $T(x_j)$ is the log of the ratio of p_{oj} and p_{1j} from Table II, but here the table is arranged so that the $T(x_j)$ are in ascending order. The next to last column presents the cumulative probability under H_o of observing $T(x_i)$ less than or equal to the given $T(x_j)$. It is used to specify c_1. In this example, if alpha is equal to .05, then c_1 is equal to -1.9. The last column is the cumulative probability under H_1 of observing $T(x_i)$ greater than or equal to the given $T(x_j)$. It is used to specify c_2. In this example, if beta is equal to .05 then c_2 is equal to 2.7.

TABLE III

THE DISTRIBUTION OF THE TEST STATISTIC

j	x_j	$T(x_j)$	p_{oj}	p_{1j}	$\sum_{k=1}^{j} p_{ok}$	$\sum_{k=j}^{n} p_{1k}$
1	(0,0,0)	-9.2	.0008	.4703	.0008	1.0004
2	(0,0,1)	-4.8	.0143	.3848	.0151	.5301
3	(0,1,0)	-3.6	.0043	.0523	.0194	.1453
4	(1,0,0)	-1.9	.0068	.0248	.0262	.0930
5	(0,1,1)	.9	.0808	.0428	.1070	.0682
6	(1,0,1)	2.7	.1283	.0203	.2353	.0254
7	(1,1,0)	3.8	.0383	.0028	.2736	.0051
8	(1,1,1)	8.3	.7268	.0023	1.0004	.0023

Thus, if alpha and beta both equal .05, we would classify a pair as a match if we observe vectors (1,0,1), (1,1,0), or (1,1,1). We would classify pairs as a nonmatch if we observe (0,0,0), (0,0,1), (0,1,0), or (1,0,0). If we observed (0,1,1): agreement on sex and first name, but disagreement on surname, we would be unable to classify the pair as either a match or a nonmatch.

The test statistic and critical region defined in this way are the same as those developed by Fellegi and Sunter (1969), although that paper also included a discussion of randomization to achieve the type 1 and type 2 error levels exactly. They develop the decision rule for accepting H_o or H_1 based on minimizing the probability of not making a decision. That is: minimizing the probability that $T(x_i)$ falls between c_1 and c_2 for a given alpha and beta.

THE POSTERIOR ODDS RATIO

The development presented here and in Fellegi-Sunter (1969) use the test statistic defined in equation (1). However, equation (A2) can be rewritten as

$$\log P(H_o/x_i)/P(H_1/x_i) = \log p_{oi}/p_{1i} + \log P(H_o)/P(H_1). \quad (2)$$

Here the log of the posterior odds ratio is written as the sum of the information number and the log of the prior odds ratio. Howe and Lindsay (1981) call equation (2) the "total weight" for a match, but acknowledge that the prior odds ratio is difficult to evaluate. The most recent papers by Newcombe and Smith include procedures for estimating the prior odds ratio in some unique situations (see Newcombe and Abbatt, 1983 and Smith, Newcombe, and Dewar, 1983). Note that the prior odds ratio reflects any information available regarding the match status of a given record pair before the attribute comparison. If the prior odds of a match were the same for each record pair then the test statistic and critical region for the comparison of attributes would both be shifted by the same value. In such a case the inclusion of the prior odds ratio would not change the outcome of the statistical test. However, the posterior odds ratio has the advantage that it can be interpreted directly as the odds that the record pair matches.

In the Smith, Newcombe, and Dewar paper, the prior odds ratio is calculated based on a life table analysis of the severity of cancer diagnosed, an attribute available in the search file, and the year of the death file being searched. In their example, the prior probability of a match is different for each individual in the search file and instead of applying specifically to a record pair, it applies to the individual record initiating the search and to an entire one year death file.

INDEPENDENCE OF ATTRIBUTES -- A SIMPLIFYING ASSUMPTION

In the original pages of this discussion, x_i was defined to be a discrete random variable which was the intersection of m attribute comparisons. If the result of each attribute comparison is denoted as t_j for j=1, ..., m, then x_i can be written as the intersection of the t_j:

$$x_i = t_1 \cap t_2 \cap \dots \cap t_m.$$

If $t_1, \dots, t_m$ are statistically independent, then equation (1) can be written as:

$$I(o:1;x_i) = \sum_{j=1}^{m} I(o:1;t_j).$$

Thus, if the set of attribute variables, t_j, are statistically independent, the weights (i.e., the information) for each t_j can be calculated separately, and the overall weight (the information contained in the intersection of the t_j) is just the sum of the weights for each t_j.

In the previous example, the three attributes were assumed to be independent. Hence, the weight for any observed vector can be calculated as the sum of the information associated with agreement or disagreement on each attribute. For example, for $x_i=(0,1,1)$ the weight can be calculated as the sum of the information associated with disagreement on surname,

$$T(a_1=0) = \log(.1/.95) = -3.25;$$

the information associated with agreement on first name,

$$T(a_2=1) = \log(.85/.1) = 3.09;$$

and the information associated with agreement on sex,

$$T(a_3=1) = \log(.95/.45) = 1.08.$$

The sum of these weights is .92, as shown in Table III for the weight (the value of $T(x_j)$) associated with the observation (0,1,1). Thus, if it is reasonable to assume that the outcomes of attribute comparisons for different attributes are statistically independent, then the calculation of the test statistic is simplified because the weights can be calculated separately and summed.

In this example, it is reasonable to assume that agreement on surname is independent of agreement on either first name or sex. However, if there is agreement on first name, it is likely that there will be agreement on sex. Hence, in this example, the assumption of independence does not really hold. To incorporate this dependence, one would need to consider the probabilities associated with the bivariate random variable.

AN EXAMPLE OF A MULTIPLE OUTCOME COMPARISON

The following is a vastly simplified example of defining the specific outcomes of attribute comparison by making use of the values they can assume. This type of "frequency" argument results in lower weights for agreement on common items and higher weights for agreement on rare items. It is a simplified version of the treatment of frequencies and error structures presented in the Fellegi-Sunter paper, pages 1192 and 1193 (pp. 60 and 61 in this volume).

Here, assume that surnames are being compared in a pair of records. Assume that there are only two frequently occurring names in the file, "Smith" and "Jones"; the other names (m of them) all occurring with roughly the same low frequency. Thus, we define the following set of outcomes of the comparison of surname:

$$x = \begin{cases} \text{"Smith"} & \text{if the two variables agree and both equal "Smith,"} \\ \text{"Jones"} & \text{if the two variables agree and both equal "Jones,"} \\ \text{"other"} & \text{if both variables agree but do not equal either "Smith" or "Jones,"} \\ \text{"disagree"} & \text{if the items disagree.} \end{cases}$$

(Note that the set of outcomes defined for item comparison must specify a partition of the set of all possible results into mutually exclusive and exhaustive subsets.)

Further assume that: 1) surnames in the two files under consideration are both random samples from the same population, and that in this population, "Smith" occurs with probability p_a, "Jones" occurs with probability p_b, and each of the other m error-free names in the file occurs with probability p_o; and 2) the only errors in the name fields are keypunch errors, which occur at the same rate, 1%, in both files, independent of the particular name.

Under H : A pair of records is a match. Names agree unless there is a keypunch error. Thus, the probability of agreement on Smith is $p_{o1} = p_a*(.99)**2$ (the probability of observing "Smith" times the probability that the value was keypunched correctly on both files). Similarly, the probability of agreement on Jones $p_{o2} = p_b*(.99)**2$, and the probability of agreement on one of the other names is $p_{o3}=p_o*(.99)**2$. The probability of disagreement on name when the record pairs represent the same individual is $p_{o4}= 1-p_{o1}-p_{o2}-m*p_{o3}$ $= (1-(.99)**2)*(p_a+p_b+m*p_o)$ $= 1-(.99)**2=.02$.

Under H_1: The records do not represent the same individual and any agreement on name occurs at random. The probability of agreement with name "Smith" is $(.99*p_a)**2$; the probability of agreement with name "Jones" is $(.99*p_b)**2$; the probability of agreement with some other name is $(.99*p_o)**2$; and the probability of disagreement on name is $1-.99**2*(p_a**2+p_b**2+m*p_o**2)$. (We have assumed that the probability that a keypunch error results in some valid name is negligible.)

Thus, from equation (1) the weight for the various outcomes is:

If x*=Smith,
$T(x*)=\log(.99**2*p_a/.99**2*p_a**2)=\log(1/p_a)$.

x*=Jones,
$T(x*)=\log(.99**2*p_b/.99**2*p_b**2)=\log(1/p_b)$.

x*=other,
$T(x*)=\log(.99**2*p_o/.99**2*p_o**2)=\log(1/p_o)$.

x*=disagree,
$T(x*)=\log$
$(.02/(1-*.99**2*(p_a**2+p_b**2+m*p_o**2)))$.

Newcombe, Kennedy, Axford, and James (1959) noted that in frequency based matching, if an item, a, is found in a master file with probability p_a, and if the two files being matched can be viewed as a sample from that master file, then, when a record pair is a match, the probability that the items agree and equal "a" is proportional to p_a. When the record pair is a nonmatch the probability is proportional to p_a**2 with the same constant of proportionality. Thus, the weight for a match when item a is observed is $\log(p_a/p_a**2) = \log(1/p_a)$. This is illustrated in the example above. Most of the Smith and Newcombe papers describe calculation of the weights for agreement on a particular item as the log of the inverse of the frequency of occurrence of that item.

The Fellegi-Sunter paper presents a derivation of the frequency based weights for specific agreement in the presence of several types of errors. Their procedure still leads to weights for agreement of $\log(1/p_a)$ because, as in the above example, the error terms impact the probability of agreement under H_o and the probability of agreement under H_1 in the same way.

VARIATIONS IN PRACTICE

Probabilistic matching techniques (based on the Fellegi-Sunter paper) have been implemented in many software systems, including the Generalized Iterative Record Linkage System (GIRLS) from Statistics Canada (see Smith and Silins, 1984) which is now called the Canadian Linkage System (CANLINK); UNIMATCH from the U.S. Bureau of the Census (see Jaro, 1972); the Statistical Reporting Service's (SRS) Record Linkage System from the U. S. Department of Agriculture (USDA); and the California Automated Mortality Linkage System (CAMLIS) from the University of California at San Francisco. Work by Rogot et al. (1983) at the National Center for Health Statistics has also used probabilistic matching techniques.

The two major references for this section are a paper by Howe and Lindsay (1981), which describes a version of the GIRLS system, and a number of unpublished papers by Richard Coulter, Max Arellano, William Arends, Billy Lynch, and James Mergerson dated 1976 and 1977, which describe the SRS Record Linkage System. These two systems were included in this review because they are applications of a modified Fellegi-Sunter approach and because the available documentation was thorough.

The GIRLS system was developed to support epidemiological research. Thus, it is primarily intended to link records for a cohort group to morbidity or mortality data. Attributes available for comparison usually include first name, surname, middle initial, sex, date of birth, place of birth, parents' names and places of birth. Some of the application-specific items, such as blocking attribute and definition of outcomes for attribute comparison, are not fixed in the system. They can be specified by the user. In the following, the specific applications by Howe and Lindsay are described.

The SRS record linkage system is intended to support development and maintenance of state-level sampling frames for agricultural surveys. Here, the primary intent of the linkage system is to unduplicate a list created by merging multiple lists. The most commonly available attributes are surname, first name, and address. In addition to the probabilistic matching procedure, record pairs which have identical address fields are reviewed manually to identify matches. This system is not a general-purpose matching system. It was developed and is used solely to maintain the USDA frames.

Blocking

In these applications, both systems block first on surname code -- a variation of the New York State Identification and Intelligence System (NYSIIS) code. A surname code is an alphabetic code designed so that the most similar names and the names with the most frequently encountered errors of misreporting will have the same code. See Lynch and Arends (1977) for a description of surname codes and the rationale used by SRS to select the NYSIIS code for their system. If the resultant block size is too big, SRS uses secondary blocking on first initial and tertiary blocking on location code. The Howe and Lindsay application blocks first on NYSIIS code, then on sex. In neither case are the weights changed to reflect the impact of blocking.

Weights for Agreement

Both systems make extensive use of frequency-based weights, and both systems use the files being matched to calculate the frequencies. Both systems also assume that these frequencies include keypunch errors, recording errors, and legitimate name changes. This is different from the Fellegi-Sunter approach, which assumed that the frequencies were based on an error-free name file.

The SRS approach handles partial agreements by calculating a weight for agreement on specific surname and a weight for agreement on specific NYSIIS code with disagreement on surname. The Howe-Lindsay paper extends the accounting for partial agreement by specifying agreement on specific first seven characters of surname; agreement on specific first four characters with disagreement on the next three characters; and agreement on specific NYSIIS code with disagreement on the first four characters of surname. In both systems, pairs with disagreement on NYSIIS code will never be considered because of the blocking.

Estimation of Error Rates

Both systems use an iteration scheme to provide final estimates for the required error rates. First, initial estimates are provided, a sample of records is processed through the matching algorithm, and a preliminary set of matched record pairs is identified. These pairs are assumed to be true matches and are used to estimate the error rates, as discussed below. These revised estimates for the error rates are input to the system; the sample is processed again and the newly matched pairs are used to reestimate the error rates. The iteration is continued until the estimates for the error rates converge.

The errors are handled in the Howe-Lindsay paper as transmission rates:

t_1 = the probability that the first seven characters of surname are equal to the "true" value;

t_2 = the probability that the first four characters are equal to the "true" value but the next three characters are different; and

t_3 = the probability that the surname code is equal to the "true" surname code, but that the surnames disagree in the first four characters.

These transmission rates can be estimated from a sufficiently large set of pairs which represent true matches by using the following counts: the number of pairs which agree on the first seven characters; the number of pairs which agree on the first four characters not on the next three, and the number which do not agree on the first four characters. The assumption is made that this set of matched pairs is representative of all possible matched pairs. Note that t_3 will be underestimated because of the blocking.

In the SRS system, the error rates used are:

e = the probability that a name is misreported or misrecorded

e_T = the probability that in a record pair which does represent the same unit, the names are correct but different.

These definitions of the error rates are the same as those used in the Fellegi-Sunter paper. The overall weights for specific agreement are different because the frequencies themselves are derived under different assumptions, as mentioned above. In the SRS system, the error rates are estimated from the set of pairs which represent true matches by using: the number of pairs which have the same name; the number which have different names; and the number which have similar names (where "similar" was not defined). Here, e_T will necessarily be underestimated because the blocking procedure assures that records will be compared only if they agree on NYSIIS code.

The Critical Region

Both systems use an empirical procedure to determine the critical region. That is, a frequency distribution of the weights for a sample of record pairs is plotted, and the critical values are selected based on the shape of the curve. As an alternative, the SRS system also calculates an initial lower critical region as the sum of the weights for agreement of the most common surname, first name, and location. The initial upper critical region is estimated as the initial lower critical region plus the weights for agreement on the most common middle name, route and box number. These calculated upper and lower regions are used during the iteration to estimate error rates. They are conservative since both are positive.

System Considerations

In the Howe-Lindsay approach, an initial blocking and comparison are done before the frequency based agreement weights are calculated. At this stage, only weights for disagreement are summed and as the accumulated weight becomes too negative, the record pair can be rejected as a possible match before all attributes have been compared. With this approach the order of adding in attributes is important, with those having the greatest negative weight for disagreement entering first. If the total disagreement weight is above the threshold, the record pair is a possible match. A separate file is created containing those possibly matched pairs. For each such pair, this file contains one record with the identification numbers of the two records, the results of the comparison of attributes, and the values taken (if needed for the weight calculation). This potential linked file is then sent to a separate subroutine for calculation of the weights.

Grouping

Both systems create groups consisting of all records which have been linked with each other. (Here linked means that the calculated test statistic is above the upper critical value.) As described in the Howe and Lindsay paper, the group is formed by first taking a single record and adding to the group any records which have been linked to it, then adding all records which were linked to those records, and so on. Additional subgroupings are considered when two records from different groups have a weight between the two critical values.

Interpretation of the groups depends on the application. In the SRS application, members of a group could all be duplicates to each other. In the SRS system, subgroups are analyzed manually. In some of the applications described by Howe and Lindsay, neither input file has any duplication, and there is at most one matched record for a given record in the search file. In this case the groups are analyzed to pick the pair which represents the most likely match, usually the pair with the highest weight.

SUMMARY

This paper has described the probabilistic matching procedures discussed by Fellegi and Sunter (1969) from an information theoretic point of view. This approach gives additional insight into the calculation of the posterior odds ratio as mentioned by Howe and Lindsay, and as implemented in the recent work of Newcombe and Smith. Additionally, it has described some of the differences between two of the major systems which have been implemented based on the Fellegi-Sunter paper. Major differences between systems are in accounting for partial matches, the definition of the error rates, and in the handling of groups of record pairs which are all linked to each other. The major differences between these systems and the Fellegi-Sunter approach are 1) that these systems base their frequency counts on files which are acknowledged to contain errors, and 2) that they use an empirical procedure to determine the critical region for the statistical test.

REFERENCES

Arellano, Max and Arends, William, "The Estimation of Component Error Probabilities for Record Linkage Purposes," Statistical Reporting Service, U.S. Department of Agriculture, Unpublished, May 1976.

Arellano, Max, "Optimum Utilization of the Social Security Number for Matching Purposes," "Weight Calculation for the Place Name Comparison," "Calculation of Weights for Partitioned Variable Comparisons (Trailing Blanks Case)," "Estimation of Component Error Probabilities for Record Linkage Purposes," "Development of A Linkage Rule for Unduplicating Agricultural Lists," Statistical Reporting Service, U.S. Department of Agriculture, Unpublished papers, 1976 and 1977.

Arellano, Max and Coulter, Richard, "Weight Calculation for the Given Name Comparison," "Weight Calculation for the Middle Name Comparison," "Weight Calculation for the Surname Comparison," Statistical Reporting Service, U.S. Department of Agriculture, Unpublished papers, 1976 and 1977.

Coulter, Richard, "An Application of a Theory for Record Linkage," Statistical Reporting Service, U.S. Department of Agriculture, Unpublished, March 1977.

Coulter, Richard and Mergerson, James, "An Application of a Record Linkage Theory in Construction of a List Sampling Frame," Statistical Reporting Service, U.S. Department of Agriculture, Washington, D.C., April 1977.

Fellegi, Ivan and Sunter, Alan, "A Theory for Record Linkage," Journal of the American Statistical Association, 1969, pp. 1183-1210.

Howe, G. R. and Lindsay, J., "A Generalized Iterative Record Linkage Computer System for Use in Medical Follow-up Studies," Computers and Biomedical Research 14, Academic Press, 1981, pp. 327-340.

Jaro, Matthew, "UNIMATCH--A Computer System for Generalized Record Linkage Under Conditions of Uncertainty," AFIPS Conference Proceedings, Vol. 40 for Spring Joint Computer Conference, May 1972, pp. 523-530.

Jaro, Matthew, "UNIMATCH--Generalized Record Linkage Applied to Urban Data Files," Proceedings of the American Statistical Association.

Kelley, Patrick, "A Preliminary Study of the Error Structure of Statistical Matching," Proceedings of the American Statistical Association, Social Statistics Section, 1983.

Kelley, Patrick, "Advances in Record Linkage Methodology: A Method for Determining the Best Blocking Strategy," Record Linkage Techniques--1985, Internal Revenue Service.

Kullback, Soloman, Information Theory and Statistics, Dover Publications, Inc., New York, New York, copyright 1968.

Lynch, Billy and Arends, Williams, "Selection of a Surname Coding Procedure for the SRS Record Linkage System," Sample Survey Research Branch, Research Division, Statistical Reporting Service, U. S. Department of Agriculture, Feb 1977.

Newcombe, Howard, "Record Linking: The Design of Efficient Systems for Linking Records into Individual and Family Histories," American Journal of Human Genetics, Vol 19, No. 3, Part I, (May) 1967.

Newcombe, Howard, and Kennedy, James, "Record Linkage: Making Maximum Use of Discriminating Power of Identifying Information," Communications of the Association for Computing Machinery 5, 1962, pp. 563-566.

Newcombe, H., Kennedy, J., Axford, S., and James, A.,"Automatic Linkage of Vital Records," Science, 130, 1959, pp. 954-959.

Newcombe, H., and Abbatt, J., "Probabilistic Record Linkage in Epidemiology," Report Prepared for Eldorado Resources, Ltd., Oct. 1983.

Rogot, Eugene, Schwartz, Sidney, O'Connor, Karen and Olsen, Christina, "The Use of Probabilistic Methods in Matching Census Samples to the National Death Index," Proceedings of the American Statistical Association, Section on Business and Economic Statistics, 1983.

Smith, Martha, Newcombe, Howard, and Dewar, Ron, "Proposed Procedure for the Alberta Cancer Registry Death Clearance," Health Division, Statistics Canada, (OEHRU-No. 1), March 1983.

Smith, Martha, Newcombe, Howard, and Dewar, Ron, "The Use of Diagnosis in Cancer Registry Death Clearance," Health Division, Statistics Canada, (OEHRU-No. 2), April 1983.

Smith, Martha and Silins, John, "Generalized Iterative Record Linkage System," (An excerpt), Statistical Uses of Administrative Records: Recent Research and Present Prospects, Department of the Treasury, Internal Revenue Service, July 1984.

U. S. Department of Commerce, Office of Federal Statistical Policy and Standards, Report on Exact and Statistical Matching Techniques, Statistical Policy Working Paper 5, 1980.

APPENDIX

This appendix presents a derivation of the test statistic for determining whether a record pair is a match or a nonmatch using an information theoretic approach (see Kullback, 1968).

WHAT IS AN INFORMATION NUMBER?

Given the prior probabilities associated with a match and a nonmatch, $P(H_o)$ and $P(H_1)$, we use Bayes theorem to calculate the posterior probabilities of H_o and H_1 based on the observed attribute comparison, x_i:

$$P(H_o/x_i) = P(H_o)*p_{oi}/(P(H_o)*p_{oi} + P(H_1)*p_{1i})$$

$$P(H_1/x_i) = P(H_1)*p_{1i}/(P(H_o)*p_{oi} + P(H_1)*p_{1i}).$$

Dividing these gives the posterior odds ratio:

$$P(H_o/x_i)/P(H_1/x_i) = P(H_o)*p_{oi}/(P(H_1)*p_{1i}),$$

and taking the logarithm (to any base) gives:

$$\log P(H_o/x_i)/P(H_1/x_i) = \log p_{oi}/p_{1i} + \log P(H_o)/P(H_1). \tag{A1}$$

This is the log of the posterior odds ratio or equivalently, the log of the posterior likelihood ratio. It can be rearranged to get:

$$\log p_{oi}/p_{1i} = \log P(H_o/x_i)/P(H_1/x_i) - \log P(H_o)/P(H_1). \tag{A2}$$

This number is the difference between the log of the posterior odds ratio and the log of the prior odds ratio. Thus, it provides a measure of the information for discriminating in favor of H_o against H_1 which was gained by observing the random variable x_i.

For this reason, the information gained by the set of outcomes of the attribute comparison, x_i, is defined to be:

$$I(o:1;x_i) = \log p_{oi}/p_{1i}. \tag{A3}$$

THE MEAN INFORMATION

The mean information for discriminating in favor of H_o against H_1 is the expected value of $I(o:1;x_i)$ under H_o, or

$$\begin{aligned} I(0:1) &= E_o(\log p_{oi}/p_{1i}) \\ &= \sum_{i=1}^{n} p_{oi} * \log p_{oi}/p_{1i}. \end{aligned} \tag{A4}$$

Here E_o represents the expectation under H_o. Note that the mean information is simply the expected value of the log of the likelihood ratio under H_o.

One useful mathematical fact is that I(o:1) is always greater than or equal to zero, with equality only when $p_{oi} = p_{1i}$ for all i = 1, ..., n. This gives an approach to selecting between the two hypotheses. Given any sample, it is possible to evaluate the sampling distribution under both hypotheses, and to calculate the mean information between the sampling distribution and the hypothesized distribution. The hypothesized distribution which was closer to the sampling distribution, as measured by the mean information, would be preferred.

THE TEST STATISTIC

When we compare the attributes associated with any two records, the result is one of the n possible values taken by x_i. We denote this observed random variable as x^*. The probability of observing $x^*=x_i$ is p_{oi} under H_o and p_{1i} under H_1. Thus, the sampling distribution of x^* is simply;

$$p_i=1 \text{ if } x^* = x_i,\ p_i=0 \text{ if } x^* \text{ ne } x_i.$$

We can write the mean information between the sampling distribution and H_o as

$$I(x^*:H_o) = \log(1/p_{oi}) \text{ for } x^*=x_i,$$

and the mean *information* between the sampling distribution and H_1 as

$$I(x^*:H_1) = \log(1/p_{1i}) \text{ for } x^*=x_i.$$

The decision rule, as described in Kullback (1968, chapter 5), is to pick the hypothesis which has the smallest mean information relative to the sampling distribution. That is, we accept the hypothesized distribution which is closest to the sampling distribution.

Thus, the procedure would be to accept H_o if $I(x^*:H_1)-I(x^*:H_o)$ is positive (or "sufficiently large.") and accept H_1 if it is negative (or "sufficiently small.")

This yields the test statistic, $T(x^*)$, where

$$T(x^*) = I(x^*:H_1)-I(x^*:H_o)$$

$$= \log(p_{oi}/p_{1i}) \text{ for } x^*=x_i. \qquad \text{(A5)}$$

$T(x^*)$ is the log of the *ratio* of the probability of the set of outcomes, x^*, under H_o to the probability of x^* under H_1. Note that if these probabilities are the same then $T(x^*)=0$, and this set of outcomes has no discriminating power for identifying whether records represent the same unit. If p_{oi} is larger than p_{1i}, then $T(x^*)$ will be positive for that category. The larger $T(x^*)$, the stronger is the possibility that observation of this set of outcomes indicates that the records represent the same unit. If p_{oi} is smaller than p_{1i}, then $T(x^*)$ is negative. The smaller $T(x^*)$, the stronger is the possibility that this set of outcomes indicates that the records do not represent the same unit.

Since $T(x^*) = \log(p_{oi}/p_{1i})$ with probability p_{oi} under H_o, and with probability p_{1i} under H_1, the ratio of the probability that $x^*=x_i$ and the probability that $T(x^*) = T(x_i)$ is equal to 1.

Since the ratio of the probability function of x_i and the probability function of $T(x_i)$ does not depend on the p_{oi} or p_{1i}, $T(x_i)$ is a sufficient statistic for discriminating between H_o and H_1.

ADVANCES IN RECORD LINKAGE METHODOLOGY: A METHOD FOR DETERMINING THE BEST BLOCKING STRATEGY

R. Patrick Kelley, Bureau of the Census

I. INTRODUCTION

The term record linkage, as it will be used in this paper, is a generic term for any process by which the set of reporting units common to two or more files of data is determined.

Historically, government agencies have been the primary users of record linkage techniques. The reasons such agencies carry out record linkage projects are as varied as the purpose and scope of the agencies themselves. Consider the following examples:

a) The United States Department of Agriculture uses record linkage to update mailing lists (see Coulter and Mergerson, 1977).
b) Statistics Canada uses record linkage as a tool in epidemological research(see Smith, 1982).
c) The United States Census Bureau uses record linkage as a tool in coverage and content evaluation (see Bailar, 1983).

For a more detailed discussion of the history and and use of record linkage by United States government agencies see U.S. Department of Commerce (1980).

As an area of study, Record Linkage, with its associated statistical problems, is a special case of a larger area of concern. This area makes use of various mathematical and statistical techniques to study the problems involved in the classification of observed phenomena.

Discriminant analysis, discrete discriminant analysis, pattern recognition, cluster analysis and mathematical taxonomy are some of the specific fields which study various aspects of the classification problem. While record linkage contains its own specific set of problems it also has a great deal in common with these other fields.

The basic unit of study in the linking of two files F1 and F2 is F1XF2, the set of ordered pairs from F1 and F2. Given F1XF2, our job is to classify each pair as either matched or unmatched. This decision will be based on measurements taken on the record pairs. For example, if we are linking person records, a possible measurement would be to compare surnames on the two records, and assign the value 1 for those pairs where there is agreement and 0 for those pairs where there is disagreement. These measurements will yield a vector, Γ, of observations on each record pair.

The key fact which will allow us to link the two files is that Γ behaves differently for matched and unmatched pairs. Statistically we model this by assuming that Γ is a random vector generated by $P(\cdot \mid M)$ on matched pairs and $P(\cdot \mid U)$ on unmatched pairs. Thus, the Γ value for a single randomly selected record pair is generated by $pP(\cdot \mid M)+(1-p)\,P(\cdot \mid U)$ where p is the proportion of matched records.

This model for the record linkage problem is the same as the one used in discriminant analysis. In particular, as Γ is almost always discrete, the literature on discrete discriminant analysis is extremely useful (see for example Goldstein and Dillon, 1978). There are, however, several areas of concern that seem to be a great deal more important for record linkage than for the other classification techniques.

Our topic of discussion in this paper, blocking, arises from consideration of one of these problem areas. That area concerns the extreme size of the data sets involved for even a relatively small record linkage project. The size problem precludes our being able to study all possible record pairs. So, we must determine some rule which will automatically remove a large portion of record pairs from consideration. Such a rule is referred to as a blocking scheme since the resulting subset of record pairs often forms rectangular blocks in F1XF2.

The literature on the blocking problem is not extensive. Brounstein (1969), Coulter and Mergerson (1977) and U.S. Department of Commerce (1977) contain discussions of the practical aspects of choosing a blocking scheme; however, they provide no general framework within which to make such a selection. Jaro (1972) provides a framework for the selection of a blocking scheme but doesn't discuss the errors induced by blocking. Many other papers, particularly those on clerical matching, contain implicit information on blocking. But so far there has been no systematic study of this area.

To provide such a study we begin with the following three questions:

1) What are the benefits and costs involved in blocking and how do we measure them?
2) How do we select between competing blocking schemes? Is there a best scheme?
3) How do the various computing restrictions effect our blocking scheme selection?

These three questions will serve as a guideline for our investigation of the blocking problem. But, before we begin this investigation, we need to consider some background material on record linkage.

II. BACKGROUND

Again, our job in linking the two files F1 and F2 is to classify each record pair as either matched or unmatched. In practice, however, we usually include a clerical review decision for tricky cases. So, our set of possible decisions is

A1: the pair is a match
A2: no determination made - clerical review
A3: the pair is not a match.

Now, consider the class of decision functions $D(\cdot)$ which transform our space of comparison vector values, elements of which we will denote by γ, to the set of decisions $\{A1,A2,A3\}$. Given two or more decision functions in this class, what criterion will we use to choose between them?

In Fellegi and Sunter (1969) the argument is put forward that, as decision A2 will require costly and error prone clerical review, we should pick a decision procedure which will minimize the expected number of A2 decisions while keeping a bound on the expected number of pairs which are classified in error. Since the unconditional distribution of the comparison vector is the same for any randomly chosen pair, this reduces to picking that decision procedure which will minimize P(A2) subject to $P(A1|U) <= \mu$ and $P(A3|M) <= \lambda$.

Given that you know $P(\cdot\ |M)$ and $P(\cdot\ |U)$, Fellegi and Sunter prove that the decision procedure which solves this problem is of the form

$$(1)\quad D(\gamma) = \begin{cases} A3 \text{ if } \ell(\gamma) <= t1 \\ A2 \text{ if } t1 < \ell(\gamma) < t2 \\ A1 \text{ if } \ell(\gamma) >= t2 \end{cases}$$

where $\ell(\gamma) = P(\gamma\ |M)/P(\gamma\ |U)$, t1 is the largest value in the range of $\ell(\cdot)$ for which $P(A3|M) <= \lambda$ and t2 is the smallest value in the range of $\ell(\cdot)$ for which $P(A1/U) <= \mu$.

It is this decision procedure that forms the basis for our study of the blocking problem.

III. MEASUREMENT OF THE COST AND BENEFIT OF BLOCKING

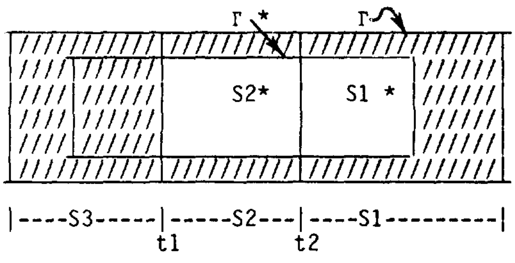
In the past sections we have outlined the more general aspects of record linkage and defined the blocking problem. In this section we will discuss blocking in the context of the decision procedure given in section II.

We base our general blocking strategy on the fact that the proportion of matched pairs in F1XF2 is small. So we will concentrate on blocking rules in which the pairs removed by the rule will be assigned the status of unmatched.

Fellegi-Sunter (1969) provides a formal model for blocking. This model defines a blocking scheme to be a subspace, say Γ^*, of the comparison space. Kelley (1984) provides a preliminary study of selected methods of measuring cost and benefit. The method found to have the most intuitive appeal is one that is based on the following amended decision procedure:

$$(2)\quad D'(\gamma) = \begin{cases} A3 \text{ if } \ell(\gamma) <= t1 \text{ or } \gamma \in \Gamma^{*C} \\ A2 \text{ if } t1 < \ell(\gamma) < t2 \text{ and } \gamma \in \Gamma^* \\ A1 \text{ if } \ell(\gamma) >= t2 \text{ and } \gamma \in \Gamma^* \end{cases}$$

A Venn diagram of this situation is given by



where S3* is represented by the shaded region.

In this design Si and Si* are the regions of Γ values for which we make decision Ai under decision functions given by (1) and (2), respectively.

The error levels for this amended decision rule are given by

$$P(S3^* \mid M) = P(S3 \mid M) + P(S3^* - S3 \mid M)$$

$$= \lambda + P(S3^* - S3 \mid M).$$

and

$$P(S1^* \mid U) = P(S1 \mid U) - P(S1 \cap S3^* \mid U)$$

$$= \mu - P(S1 \cap S3^* \mid U).$$

These equations give us a means to compute a cost incurred by blocking on the subspace Γ^*, namely, $P(S3^* - S3 \mid M)$, the increase in probability of a false nonmatch. The benefit gained from blocking on Γ^* takes the form of a decrease in the number of pairs which will have to be processed. We will measure this benefit by the unconditional probability that a randomly chosen record pair yields a Γ vector in the block.

Now, given two blocking schemes which both have cost less than or equal to a fixed amount, the preferred scheme is the one with greatest benefit. Thus, we define the best blocking scheme to be that scheme which minimizes $P(\Gamma^*)$ subject to $P(S3^*-S3|M) <= w$, where w is an independently determined upper bound on blocking costs.

IV. COMPUTING THE BEST BLOCKING SCHEME - THE ADMISSIBILITY CONCEPT

Since the comparison vector is discrete, the computation of the best blocking scheme will require a comparison of all competing schemes. So, it's in our best interest to reduce the number of competing schemes. To make this reduction we note that if $\Gamma1^*$ and $\Gamma2^*$ are two competing schemes such that $\Gamma1^*$ is a subset of $\Gamma2^*$ then $\Gamma1^*$ is uniformly better than $\Gamma2^*$. So, we can remove $\Gamma2^*$ from the set of competing blocking schemes. The following definition formalizes this example:

> Γ^* will be said to be an admissible blocking scheme at w = w0 if
> a) $P(S3^* - S3 \mid M) <= w0$ and
> b) for every Γ^{**} that is a subset of Γ^*
> $P(S3^{**} - S3 \mid M) > w0$.

The concept of an admissible blocking scheme given by this definition is analogous to the concept of an admissible decision procedure. It serves to reduce, hopefully to a reasonable size, the number of blocking schemes competing for best. But, unfortunately, when actually applied to the task of computing the set of admissible blocking schemes, this definition is very cumbersome. The following lemma gives necessary and sufficient conditions for admissibility which are more favorable to algorithm development:

Lemma 1:

Γ^* is admissible at w = w0 if and only if
$\Gamma^* \cap S3 = \emptyset$ and
$P(\gamma|M) > w0 - P(S3^*-S3|M) \geq 0$ for all γ in Γ^*.

Proof:

If Γ^* is admissible then $P(S3^*-S3|M) <= w0$. Further, for $\Gamma^{**}= \Gamma^* - \{\gamma\}$ we have $P(S3^{**} - S3|M) > w0$. But $S3^{**} - S3 = (S3^* - S3) \cup (\{\gamma\}-S3)$. So, $P(\{\gamma\}-S3|M) + P(S3^* - S3|M) > w0$.

From this relationship we see that if γ is in S3 then $P(S3^*-S3|M) > w0$; thus, $\Gamma^* \cap S3 = \emptyset$. So we have $P(\gamma|M) > w0 - P(S3^*- S3|M)$ for all γ in Γ^*.

Conversely, we first note that $P (S3^*-S3|M) <= w0$. Next, let Γ' be a proper subset of Γ^* then Γ' is a subset of $\Gamma^*- \{\gamma\}$ for some γ. So, $P(S3'-S3|M)> = P(S3^*-S3|M) + P(\{\gamma\}-S3|M)$. Thus, we have $P(S3'-S3|M) >= P(S3^*-S3|M) + P(\gamma|M) > w0$. Hence, Γ^* is admissible.

Now, in theory, we can use the result of lemma 1 to compute all admissible schemes. However, since the minimum number of dimensional Γ vector values is 2**n, we would have to generate and classify on the order of 2**(2**n) subsets.

For n=5 this yields 4,294,967,300 subsets, which is clearly too large for practical consideration. So, while the admissibility concept is helpful in reducing the number of competing schemes, it hasn't served to provide us with a practical algorithm for the computation of the best blocking scheme. In the next section, we will give more attention to the development of such an algorithm.

V. IMPLEMENTATION CONSIDERATIONS

The previous section provides a general framework for studying blocking; however, it doesn't give us much insight into the practical side of determining a block of records for possible linkage. If we keep in mind that I/O and computing the comparison vector are the biggest consumers of time in the linkage operation we see that admissible blocking schemes that require the computation of a Γ vector value for each record pair are not practical. Thus, though a scheme might be theoretically admissible it might not be feasible.

One solution for this problem is to block by using certain fields on the record (such as soundex code of surname or address range) as sort keys. The blocks would be determined by those record pairs with equal keys. Thus, the match status of unmatched pairs would be implicitly assigned to all record pairs with unequal keys.

Restricting our study to blocking schemes which are determined by sort keys implies that the comparison vector we want to use will consist of dichotomous components measuring agreement on the record identifier fields. We will further assume that the components of the comparison vector are stochastically independent for both matched and unmatched record pairs.

Now, letting $mi = P(\Gamma i=1|M)$, $ui=P(\Gamma i=1|U)$ and Γ^* be the blocking scheme determined by sorting on components $i1,\ldots,ik$ we have the following result:

Lemma 2:

Suppose that $mi>1/2$ and $ui<mi$ for all i then Γ^* is admissible at w0 if and only if

a) $w0 - P(S3^*-S3|M) >= 0$
b) $P(\gamma^*|M) > \mathrm{Max}\ \{t1P(\gamma^*|U),$
$w0 - P(S3^*-S3|M)\}$,

where γ^* is such that $\gamma i1^* = 1,\ldots, \gamma ik^* = 1$, $\gamma ik+1^* = 0, \ldots, \gamma ip^* = 0$.

Proof:

First suppose that Γ^* is admissible at w0 then conditions a) and b) follow directly from lemma 1 and the fact that $P(\gamma|M) > t1\ P(\gamma|U)$ for all γ in $S3^c$.

Now, to establish the converse we first note that, since $mi > 1/2$ for all i, $P(\gamma^*|M) = \min_{\gamma\varepsilon\Gamma^*} P(\gamma|M)$. So $P(\gamma|M) > w0 - P(S3^*-S3|M) >= 0$ for all γ in Γ^*. Next we need to prove that $\Gamma^* \cap S3 = \emptyset$. To prove this we note that $ui < mi$ implies that $mi/ui > (1-mi)/(1-ui)$. So, $P(\gamma|M)/P(\gamma|U) > P(\gamma^*|M)/P(\gamma^*|U)$ for all γ in Γ^*. Thus, $\Gamma^* \cap S3 = \emptyset$. The converse follows from lemma 1.

In comparing lemma 2 with lemma 1, we see that lemma 2 has a definite computational advantage above and beyond the reduction in competing schemes gained by restricting attention to those schemes based on sorting. That advantage lies in the requirement to check for admissibility at only one point in the blocking scheme, namely γ^*. This results in tremendous savings in computing time and simplifies algorithm construction and coding considerably. In the next section we apply lemma 2 to a simple numeric example.

VI. AN EXAMPLE

As an example, let's consider matching two files of records based on the identifiers surname, first name, and sex.

Suppose we have determined beforehand that,

for surname	$m1 = .90$ and	$u1 = .05$,
for first name	$m2 = .85$ and	$u2 = .10$,
and for sex	$m3 = .95$ and	$u3 = .45$.

Retaining the assumption of the previous section our discriminant function is given by

$$L(\gamma)= \ln2(l(\gamma)) = \sum_{i=1}^{3} [\gamma i\ \ln2\ (mi/ui) +(1-\gamma i)\ \ln2\ ((1-mi)/(1-ui))].$$

To compute the Fellegi-Sunter decision procedure we first compute L for each agreement pattern and then we order the patterns on increasing L. The following table gives the results of this operation:

Pattern	Sum of $P(\cdot|M)$	One minus sum of $P(\cdot|U)$	L
(0,0,0)	.00075	.52975	-9.29
(0,0,1)	.01500	.14500	-4.76
(0,1,0)	.01925	.09275	-3.62
(1,0,0)	.02600	.06800	-1.87
(0,1,1)	.10675	.02525	.92
(1,0,1)	.23500	.00500	2.67
(1,1,0)	.27325	.00225	3.79
(1,1,1)	1.00000	0.00000	8.34

Using this table it is clear how one would compute t1 and t2 for given λ and μ .

For example, if we let λ = .05 and μ = .05 then t1 = -1.87 and t2 = 2.67. The actual values of λ and u are .026 and .02525, respectively. We will use this decision procedure to discuss the blocking problem.

Consider our space of admissible blocking schemes based on sorting. We note that since no single component blocking scheme is admissible, we have a total of four schemes to test. Now, for convenience let B1 denote blocking on surname and first name, B2 denote blocking on surname and sex, B3 denote blocking on first name and sex, and B4 denote blocking on all components.

The following table gives the information necessary to determine the admissibility of Bi:

Bi	P(S3*-S3|M)	P(γ*|M)	values of w0 for which Bi is admissible
B1	.209	.03825	.209 ≤ w0 < .24725
B2	.119	.12825	.119 ≤ w0 < .24725
B3	.1665	.08075	.1665 ≤ w0 < .24725
B4	.24725	.72675	.24725 ≤ w0 < .974

Before we go on it is interesting to note that the minimum w0 value for which any of the B_i is admissible is .119. Thus, the minimum loss we can incur by blocking is an increase in false non-match probability of .119.

Looking at the admissible blocking schemes as a function of w0, we have the following:

1. For .119 ≤ w0 < .1665 B2 is admissible.
2. For .1665 ≤ w0 < .209 B2 and B3 are admissible.
3. For .209 ≤ w0 < .24725 B1, B2, B3 are admissible.
4. For .24725 ≤ w0 < .974 B4 is admissible.

Now, to compute the best admissible blocking scheme we must determine which of the competing schemes has the smallest probability of occurrence. The probability of occurrence of schemes Bi, say P(Bi), is given by pP(Bi|M)+(1-p)P(Bi|u), where p is the proportion of matched record pairs. Thus, in general, the best admissible scheme will be a function of p.

To compute the best blocking scheme for cases 2 and 3 consider the following table:

	P(Bi|M)	P(Bi|U)
B1	.765	.005
B2	.855	.0225
B3	.8075	.045

So, for case 2, B2 is the best blocking scheme for values of p <= .3214 and B3 is the best blocking scheme for p > .3214. For case 3, B1 is uniformly the best blocking scheme.

At this point, we have demonstrated how to select the best blocking scheme for a fixed value of w0. But it still is unclear how one would use this information to actually make a decision about which scheme to use. To study this question let's consider the nature of such a decision. To select a blocking scheme we need to balance the cost with the overall benefit. Let's redo our example this time for several different values of w0 and compare the benefits for the resulting schemes.

The following is the first part of the list of the best blocking schemes for all values of w0. This list is presented in increasing order of w0. The expected benefit, in terms of the percent of F1XF2 that would be examined, is given for each scheme. To compute this benefit the approximate sizes of F1 and F2 are required. We used F1 size = 200,000 and F2 size = 100,000 in this example.

1. Admissible blocking schemes at w0=0.0492501 are as follows:
   The scheme determined by sorting on sex.
   The expected percent of the cross product of this blocking scheme would examine is bounded above by 45.00005%.
2. Admissible blocking schemes at w0=0.0992500 are as follows:
   The scheme determined by sorting on surname.
   The expected percent of the cross product this blocking scheme would examine is bounded above by 5.00009%.
3. Admissible blocking schemes at w0=0.1442501 are as follows:
   The scheme determined by sorting on surname and sex.
   The expected percent of the cross product this blocking scheme would examine is bounded above by 2.25008%.
4. Admissible blocking schemes at w0=0.149250 are as follows:
   The scheme determined by sorting on first name.
   The scheme determined by sorting on surname and sex.
   Of these, the best blocking strategy, as a function of the proportion of matched pairs, is as follows:

   For p=0.000000000 to p=0.939394700 sort on components surname and sex.
   For p=0.939394700 to p=1.000000000 sort on components first name.
   The expected percent of the cross product this blocking scheme would examine is bounded above by 2.25008%.

To use this list for decision-making purposes one would have to have some idea about how much data they can afford to look at and how large a false non-match rate they could tolerate. For example, in looking at the scheme determined by sorting on sex, we have a small (though maybe not small enough) w0 value but the number of record pairs we would have to look at would be around 9x10**10, which is clearly not feasible. Sorting on surname has a slightly higher w0 value, but reduces the number of records to 10**10. If we are willing to accept an even higher w0, then we can sort on surname and sex, which further reduces the number of record pairs to 4.5x10**9.

Another important piece of information that we shouldn't overlook is the number of record pairs we can hold in memory at any one time. We don't want to select a blocking scheme for which the individual block sizes are too large. So not only is the total number of pairs in the block important but so is the number of states of the sorting variable and the distribution of that variable over those states.

VII. SUMMARY

The blocking problem is intrinsic to record linkage. As such, before a link between files is attempted a decision must be made concerning the appropriate blocking method.

In this paper we study this decision, along with its costs and benefits, through the record linkage methodology developed in Fellegi and Sunter (1969). This methodology applies classic decision theory techniques to the record linkage problem, constructing the optimum classifer under a loss function analogous to that of hypothesis testing.

The result of our study is a method which can be used to balance the cost and benefit of blocking. This method involves maximizing benefit subject to an upper bound on cost. The measurement of cost and benefit is based on the Fellegi-Sunter method and, as such, makes use of a similar loss function.

NOTES AND REFERENCES

Bailar, Barbara A. (1983), Counting or Estimation in a Census -- A Difficult Decision, Proceedings of the American Statistical Association, Social Statistics Section, pp. 42-49.

Brounstein, S. H. (1969), Data Record Linkage Under Conditions of Uncertainty, delivered at the Seventh Annual Conference of the Urban and Regional Information Systems Association.

Coulter, Richard W. and Mergerson, James, W. (1977), An Application of a Record Linkage Theory in Constructing a List Sampling Frame. List Sampling Frame Section, Sample Survey Research Branch, Statistical Reporting Service, U.S. Department of Agriculture.

Fellegi, Ivan and Sunter, Alan (1969), A Theory for Record Linkage, Journal of the American Statistical Association, vol. 64, pp. 1183-1210.

Goldstein, Matthew and Dillon, William (1978), Discrete Discriminant Analysis, Wiley.

Jaro, M. A. (1972), UNIMATCH - A Computer System for Generalized Record Linkage Under Conditions of Uncertainty, AFIPS - Conference Proceedings, vol. 40, pp. 523-530.

Kelley, Robert Patrick (1984), Blocking Consideration for Record Linkage Under Conditions of Uncertainty, Statistics of Income and Related Administrative Record Research: 1984, Department of the Treasury, Internal Revenue Service, pp. 163-165.

Smith, Martha E. (1982), Value of Record Linkage Studies in Identifying Population at Genetic Risk and Relating Risk to Exposures. Progress in Mutation Research, vol. 3, pp. 85-98.

U. S. Department of Commerce, National Bureau of Standards (1977), Accessing Individual Records from Personal Data Files Using Non-Unique Identifiers.

U.S. Department of Commerce, Office of Federal Statistical Policy and Standards (1980), Statistical Policy Working Paper 5 - Report on Exact and Statistical Matching Techniques.

DISCUSSION

Eli S. Marks, Consultant

WINKLER

This paper discusses Bill Winkler's presentation on "Preprocessing of Lists and String Comparison."

Key factors in "Preprocessinq of Lists" are:

1. The objectives of the system and the costs of various levels and types of matchinq error.
2. Costs of attaininq a given matchinq accuracy level by preprocessing vs. other alternatives (e.g., suitably tailored "tolerances").
3. The nature of the matching system--manual, computerized, "mixed," etc.
4. How preprocessinq is performed.

1. Objectives

The objectives of the system and the costs of matching error are intimately related. For example, if the objective is to estimate undercoverage of the U.S. census in each state, city, county, township, place, etc. for purposes of allocation of representation in Congress and state legislatures, city/county councils, etc. and for allocating federal and state funds to state and local jurisdictions, a uniform level of matching error everywhere is more important than the absolute level of matching error. Thus, preprocessing may have little value if its effect is to reduce the different types of matching errors by the same percentages in all jurisdictions. On the other hand, if preprocessing reduces urban matching error more than rural, it may be desirable or undesirable, depending upon whether the level of urban matching error without preprocessing is greater or less than the level of rural matchinq error without preprocessinq.

2. Alternative Techniques

The objective of preprocessing (i.e., reduction of matching errors) can be attained by other means (e.g., the prescription of matching "tolerances"); and these techniques may cost less than preprocessing. For example, soundex coding is a form of "matching tolerance." That is, all disagreements of vowels and some disagreements of consonants are ignored in determining whether a pair of records match on the soundexed "identifier." One can, in fact, combine some preprocessinq with tolerances (and, perhaps, other error-reducing techniques) to get a more efficient matching system than either can give alone. For example, one can prescribe standard abbreviations for the address suffixes "Avenue," "Street," "Road," "Drive," "Place," "Boulevard," etc., but also provide that an address match where the suffixes differ will be accepted unless there is another address match where the suffixes agree. For example, "Sutton Drive" would match "Sutton Road" unless either file contains both "Sutton Road" and "Sutton Drive."

Standard spelling of name and address may be achieved more accurately and more cheaply by controlling data collection, recording and "keying" (to put the data in machine readable form) than by preprocessing. This would, for example, avoid most of the errors of preprocessing by ZIPSTAN exhibited by the examples shown in the paper. Preprocessinq errors can also be reduced or eliminated by other means, such as the clerical insertion of distinctive symbols to designate components of name and address, as outlined in Section 4 below.

It should be noted that selection of an "optimum matching strategy" is heavily dependent upon the type(s) of matching system(s) considered and that the choice of type of matching system is a vital part of the determination of "optimum matching strategy."

3. Kind of Matching System

The paper by Winkler notes that matching systems can be manual or computerized and implies that preprocessing is largely unnecessary for manual matching systems. I think his suggestion that individuals can usually determine accurately whether a pair of name and address records is actually a match or nonmatch is somewhat optimistic. Individuals can make this determination (so can a computer system), but how accurately depends on the kind of system. The great advantage of a competent human matcher operating in a properly designed matching system is the use of judgmental flexibility, provided, of course, he or she has good judgment and the matching rules permit him (her) to use that judgment (and I have seen many sets of matching instructions which do not). The great disadvantage of a well-designed manual matching system with competent matchers is the human matcher's slowness and the inevitable drop in efficiency in operating in a system which requires examining large masses of records; and not in lack of clear decision rules, inconsistency of application of decision rules, and nonreproducibility of results. All of the latter do occur, but can be adequately controlled in a well-designed matching system (although it is not easy!). However, humans cannot match the forte of the computer--its speed in examining large masses of data.

The solution to this problem is to let the computer do what it does well and let humans do what they do well. That is, design a mixed computer-human system, in which the computer handles the large mass of cases which can be classified as positive links or positive nonlinks, on a mechanical, routine basis. Carefully trained and well-motivated humans could then try to match the remaining cases, using a "computer-interactive" system, where the human would specify a small class of possible matches and the computer would display the records in this class, until a positive link was found or there was adequate evidence that no such link existed.

4. Techniques of Preprocessing

Certain elements of preprocessing will unquestionably be valuable in any computerized matching system. In particular, it is important to develop some method so that the computer can quickly and accurately identify the various elements of the name and address: surname, house number, street name or number, first name, and the conventional prefixes and suffixes to name and address. If this involves elaborate manual rearrangement and keying of the name and address, substantial error is likely to be introduced, possibly as much as the preprocessing removes. The examples in the paper suggest that unaided computer formatting is also likely to introduce as much error as it removes. A solution may be something used in one of the earliest (1956) computerized matching systems, where clerks inserted a distinctive and computer-readable symbol in front of the components of name and address to be used in the matching; e.g., * before surname, # before house number, % before street name, $ before P. O. box number, @ before title, etc. After appropriate codes were placed in fixed fields, the symbols were deleted from the computer records.

DISCUSSION

Benjamin J. Tepping, Westat, Inc.

The papers by Kirkendall and Kelley contain much interesting material, with some of which I must take issue.

The Fellegi-Sunter model, on which these papers are based, recognizes that there are three possible outcomes, but (it seems to me) uses the wrong utility function. To simply minimize the probability of subjecting a case to clerical review conditional on bounds on the probabilities of erroneous matches and erroneous nonmatches ignores important facts:

(a) the value of an erroneous match is, in many (or perhaps most) applications, quite different from the value of an erroneous nonmatch;

(b) the cost and the probability of misclassification associated with the clerical review should be taken into consideration.

We do not necessarily want to minimize the number of clerical reviews. We do want to maximize the value of the record linkage operation. This implies that one must not only determine the costs of the various components of the operation, but must also set values on the possible outcomes. An illustration of this approach is the application of a theoretical model of record linkage to the Chandrasekar-Deming technique for estimating the number of vital events on the basis of data from two different sources. This was published in the Bureau of the Census Technical Notes No. 4, in 1971 [1].

It appears that neither author is aware of my paper [2] in JASA in 1968 in which is presented a model for the optimum linkage of records.

The authors treat the problem as an exercise in the testing of hypotheses. I think it is preferable to regard it as a problem of decision making, subject to a utility function which depends upon the state of nature. In these applications, the three possible decisions are to call the pair of records being compared a match or a nonmatch, or to make some kind of further investigation before deciding on a classification. That investigation may consist simply of subjecting the records to personal scrutiny or may involve seeking additional data. The utility function would specify a gain or loss for each of the possible decisions, conditional on whether the pair is in fact a match or a nonmatch.

Kirkendall's examples also ignore the problem of fixing the values of the probabilities of errors of the first and second kinds. Those probabilities should not be arbitrary. Any solution of the problem should depend upon evaluation of the loss or gain of alternative decisions as well as on the cost of nondecisions--e.g., resort to other means of arriving at a decision.

Kirkendall's first illustration assumes independence, both under H_0 and under H_1 . In the real world, this assumption may be far from true. For example, under either of the hypotheses H_0 or H_1 , an agreement on first name would increase the probability of an agreement on the item sex--two records both giving the first name as "Nancy" are not likely to indicate different sexes. Presumably the lack of independence could be treated as in her example of cancer patients, essentially by dividing the First Name item into two items: one for cases in which both records show the sex as male and one for cases in which both records show the sex as female. This comment also applies to Kelley's numerical example, in which independence of these components is assumed.

As is pointed out by Kelley, the literature that gives advice on the choice of blocking schemes is not extensive. Yet practical problems make blocking of the files being compared essential, and Kelley's work should contribute to the improvement of blocking designs. He does take account of costs, by considering both the decrease in operational costs, because blocking reduces the number of comparison pairs, and the increase in the probability of an erroneous nonmatch as a result of blocking. (I note, however, that he does not use the fact that the probability of an erroneous match decreases as a result of the blocking.) His numerical examples illustrate that the choice among competing admissible blocking schemes involves the implicit assignment of relative values to an increase in the probability of erroneous nonmatches and a decrease in the number of comparisons. In practice, no doubt, a similar implicit assignment of values to an erroneous match, an erroneous nonmatch and a case referred to personal review is made in order to fix the values of the parameters λ and μ of the Fellegi-Sunter model.

I think there is difficulty with the application of Kelley's Lemma 2 to the determination of a suitable blocking scheme even after dealing with the lack of independence of the components of the comparison vector. It seems that a choice must depend, among other things, on a knowledge of the probability, given that the pair is a match (or a nonmatch), that there is agreement between the units of the pair on specified components of the comparison vector. Estimates of such probabilities must ultimately depend upon extensive empirical investigations, although such estimates seem often to be made on the basis of assumed models.

REFEPENCES

[1] Tepping, B.J., "The application of a linkage model to the Chandrasekar-Deming technique for estimating vital events," U.S. Bureau of the Census, Technical Notes No. 4, Washington, D.C., 1971, pp. 11-16.

[2] Tepping, B. J., "A model for optimum linkage of records," Journal of the American Statistical Association, 63, 1968, pp. 1321-1332.

REJOINDER

William E. Winkler, Energy Information Administration

Eli Marks' comments provide a valuable perspective to the overall objectives of matching procedures.

Just as the Fellegi-Sunter matching procedure contains computerized (automatic designation of matches and nonmatches) and manual (review of records designated for further manual followup) components, so does preprocessing contain computerized (minor reformatting, spelling standardization, string comparison) and manual (keypunch/transcription, major reformatting) components.

The respective roles of the two components are best exemplified by Newcombe et al. (1983, 1959, 1962). Newcombe's view is that computer procedures should be developed for the most routine and repetitive tasks. As knowledge of the characteristics of address files and coding techniques increases, computerized procedures can replace greater proportions -- possibly all -- manual components.

It is my experience that reasonably designed manual procedures are difficult and expensive to implement. This is because of high turnover rates and the necessity of training and constantly supervising personnel performing manual processing. Computerized procedures can have the benefit of being more cost-effective, consistent, and reproducible.

Both Marks and I note that the Census Bureau's ZIPSTAN software -- which is designed for files of individuals -- induced minor errors in files of businesses. In Winkler (1985), I show that ZIPSTAN's identification of address subfields can yield substantial improvements in the discriminating power of the Fellegi-Sunter matching procedure.

The cost in using ZIPSTAN was a few days of my time installing it. The alternative would have been to do nothing or develop manual procedures, set up computer files suitable for manual review, train individuals in computer login and manual review procedures, and have the individuals perform the review. Marks notes, if identifying individual subfields of the name and address involves "elaborate manual rearrangement and keying ..., substantial error is likely to be introduced, possibly as much as preprocessing removes."

I strongly agree that our understanding of "matching tolerances" needs to be improved. The purpose of my discussion of string comparators was to show the limitations of tolerances such as SOUNDEX, particularly SOUNDEX abbreviations of surnames used as sort keys during the blocking stage of matching. For files of businesses, I show (Winkler, 1985) that individual sort keys are generally not suitable for creating blocks containing most matched pairs. My solution is to apply independently multiple sort keys.

String comparison metrics, such as Jaro's string comparator, can only be efficiently used during the discrimination stage because they involve the comparison of corresponding strings from pairs of records. In my view, they offer the best opportunity for developing tolerances. How such tolerances fit in the framework of the Fellegi-Sunter model needs to be described and quantified.

REFERENCES

Newcombe, H.B., Kennedy, J.M., Axford, S.J., and James, A.P. (1959), "Automatic Linkage of Vital Records," Science 130, 954-959.

Newcombe, H.B. and Kennedy, J.M. (1962), "Record Linkage," Communications of the ACM 5, 563-566.

Newcombe, H.B., Smith, M.E., Howe, G.R., Mingay, J., Strugnell, A., and Abbatt, J.D. (1983), "Reliability of Computerized Versus Manual Searches in a Study of the Health of Eldorado Uranium Workers," Comput. Biol. Med. 13, 157-169.

Winkler, W. E. (1985), "Exact Matching Lists of Businesses: Blocking, Subfield Identification, and Information Theory," Paper presented at the 1985 ASA Annual Meeting, Section on Survey Research Methods, August 4-8, 1985, Las Vegas, Nevada, (pp. 227-241 in this volume).

REJOINDER

R. Patrick Kelley, U.S. Bureau of the Census

Let me start my rejoinder by saying that I find Dr. Tepping's comments both interesting and helpful. The main criticism of my paper given by Dr. Tepping is my choice of the Fellegi-Sunter model as a basis for blocking research. As such, this exchange is simply another in a long debate over the handling of clerical costs and errors.

I have been aware of, and admired, Dr. Tepping's work on record linkage for quite some time. From a theoretical point of view, the utility theory approach is a fascinating one; however, clerical operations are hard to control and empirical investigations of clerical error rates and costs are data dependent. This makes estimates of the parameters in Dr. Tepping's model hard/expensive to obtain and highly variable.

Due to these facts, it is my opinion that the Fellegi-Sunter model provides the best general foundation for record linkage research and development. Methods which account for clerical costs should be used only after there have been several linkage projects run on data from the same source, using the same record linkage system.

Dr. Tepping also commented on the assumption of independence between comparison vector components, the difficulty of estimating, the difficulty of estimating model parameters, and the potential sensitivity of linkage error rates to errors in those parameter estimates. These comments are well placed, and I am continuing work on the blocking problem in an attempt to strengthen the results of this paper.

PROPERTIES OF THE SOCIAL SECURITY NUMBER RELEVANT TO ITS USE IN RECORD LINKAGES

Thomas B. Jabine, Consultant, Committee on National Statistics

Linkage of records from two data systems is aided greatly by the presence in both systems of the same numeric identifier, for example, the social security number (SSN) for persons or the employer identification number (EIN) for businesses. When matching variables for two records are compared, agreement on such numeric identifiers is usually given a large weight in deciding whether a true match exists.

Because of their importance for record linkage, it is important to have complete and current information on the relevant properties of each of these numeric identifiers. Such properties include: coverage, general structure and method of issuance, information content, and appropriate methods of validation. Properties relevant to sample selection using numeric identifiers are also of interest, since many record-linkage studies are based on a sample from one of the data systems.

This paper provides a description of the properties of the social security number (SSN) that are relevant to its use in record linkages. The description should be regarded as a first draft and readers are urged to suggest corrections and additions.

If this description of the SSN proves useful, it is suggested that the Administrative Records Subcommittee of the Federal Committee on Statistical Methodology make arrangements to: (1) prepare and disseminate descriptions, using the same format, of other commonly used numeric identifiers, such as the EIN and the unemployment insurance number, and (2) update the descriptions periodically and whenever significant changes occur.

Special thanks are due to Richard Wehrly of the Social Security Administration for providing information used in developing the SSN description. However, any errors are the sole responsibility of the author and readers are cautioned that the description of the SSN has not been officially reviewed by the Social Security Administration.

NUMERIC IDENTIFIER DESCRIPTION

1. Name of identifier

The social security number (SSN).

2. Administrative uses

SSNs were issued initially so that earnings of persons in jobs covered by the social security retirement program could be reported, by their employers, to the Social Security Administration (SSA) and credited to the persons accounts for subsequent use in determining benefit eligibility and payment amounts.

An early decision was made to use SSNs as identifiers in the State-operated unemployment insurance programs. No other significant uses developed until 1961 when the Internal Revenue Service, after discussions with SSA, decided to use the SSN as a taxpayer identification number. After implementation of this decision, other uses by Federal and State governments followed rapidly, and the SSN is now widely used as an identifier for workers, taxpayers, drivers, students, welfare beneficiaries, civil servants, servicemen, veterans, pensioners and others (HEW Secretary's Advisory Committee, 1973).

Legal justification for use of the SSN as an identifier by Federal agencies comes from Executive Order 9397, issued in 1943, which directed Federal agencies to use the SSN when establishing a new system of permanent account numbers. The Privacy Act of 1974 placed some restrictions on use of SSNs by Federal, State and local government agencies, but uses formally established prior to January 1, 1975 were not affected and these restrictions have had only a minor effect on widespread administrative use of the SSN by governments and private organizations (Privacy Protection Study Commission, 1977).

3. Coverage

a. Units.--SSNs are issued to persons.

b. Legal coverage provisions.--An SSN will be issued to any United States citizen upon application and presentation of acceptable evidence of identity. Foreign nationals legally present in the United States will be issued SSNs if legally entitled to work or if they have an acceptable "nonwork reason" for needing an SSN, e.g., the need for a taxpayer identification number.

All persons with Federally taxable income and their spouses are required to obtain SSNs for use as taxpayer identification numbers. SSNs are also required for many types of benefits and for other purposes: social security, driver's license, welfare benefits, voter registration, participation in scholastic aptitude testing programs, etc. For some of these, requirements vary by State.

c. Volume and characteristics of issuance to date.--SSNs were first issued in November 1936. By the end of 1975, over 235 million SSNs had been issued and there were an estimated 180 million living SSN holders (Social Security Administration, 1981b). As of the close of 1983, approximately 287,083,000 SSNs had been issued. It is estimated by SSA that there were 204,760,000 living SSN holders at the end of 1981. When SSN holders die, their SSNs are not reissued to other applicants.

The table in Attachment A shows the number of SSNs issued annually, by sex of applicant, through the end of 1979. Following the large number of issuances in the first 14 months (November 1936 to December 1937), the volume of annual issuances has fluctuated for a variety of reasons, with a tendency to increase in recent years as coverage of SSA benefit programs and the use of SSNs for non-SSA programs has expanded. Today most of the SSNs are issued to applicants under 20 years of age. In 1979, 62.8 percent of the SSNs were issued to persons under 15 and another 26.2 percent to persons between 15 and 19 (Social Security Administration, 1981b).

From time to time, surname counts based on the first six characters of the surname are made from SSA's account number files. Kilss and Tyler (1974) show the rankings of common surnames based on 1964 counts. Based on a 1974 tabulation, the ten most common surnames were:

Smith
Johnso(n)
Willia(ms)(mson)
Brown
Jones
Miller
Davis
Martin(ez)(son)
Anders(on)
Wilson

The letters in parentheses following some names are intended to show the more common surnames that have these first six characters.

d. Uniqueness, stability.--Until 1972, applicants for SSNs were not asked if they had already been issued numbers, nor were they asked for proof of identity. As a result many persons now have more than one SSN (Privacy Protection Study Commission, 1977). As of 1973, it was estimated that 4.2 million persons had two or more SSNs (HEW Secretary's Advisory Committee, 1973). More recent estimates are not available. Today, intentional issuance of multiple numbers to the same person is permitted only in exceptional circumstances, generally involving national security or the protection of the person in question.

In most cases where a person is known to have more than one SSN, SSA's computerized SSN files contain a record for each of his or her SSNs and cross references linking all of the SSNs.

Sometimes more than one person uses the same SSN. Some reasons why this happens are discussed in item 8b. Estimates of the frequency with which this occurs are not readily available, but it is believed to be much less prevalent than issuance of multiple numbers to the same person (HEW Secretary's Advisory Committee, 1973).

4. General structure and information content

The social security number has nine digits arranged as follows: 000-00-0000. The first three digits are called the area number, the next two are the group number, and the last four are the serial number. There are no check digits. The serial number provides no information about the person to whom an SSN has been assigned; however, the area and group numbers do contain a limited amount of information.

The area number, digits one to three of the SSN, carries some information either about the SSN holder's occupation or his or her place of residence at the time the number was issued. For the ranges of area numbers used to date, the information content is as follows:

(1) Area numbers 001 to 626. With a few exceptions, each of these area numbers has been assigned to a single State, one or more to a State. For most SSNs, the area number indicates only the SSN holder's State of residence at the time of issuance, as derived from the mailing address on the SSN application. For SSNs issued in the early days of social security, the area number indicated the specific SSA field office from which the number was issued, regardless of where the applicant lived.

(2) Area numbers 700-728. These numbers were assigned to railroad workers through 1963. Since then, railroad workers have been assigned SSNs with the same area numbers as other applicants.

The group number, digits four and five, in combination with the area number, provides a rough indication of when the SSN was issued. In particular, it is possible to tell whether an SSN was issued before or after another SSN having the same area number but a different group. Within an area number, the group numbers are always used in the following sequence:

- Odd numbers from 01 to 09
- Even numbers from 10 to 98
- Even numbers from 02 to 08
- Odd numbers from 11 to 99

The group number 00 has never been used. Only the first two sets of group numbers in the above sequence were used through 1965. Since then the third and fourth sets have been used with some area numbers. Current information on the last group number assigned for each area number can be obtained from SSA (see Section 9.a.).

5. Issuance procedures

All SSNs are issued by the Social Security Administration. Prior to July 1, 1963, the Railroad Retirement Board issued SSNs (in the 700 series) to all railroad employees.

A single application form, Form SS-5, Application for a Social Security Number Card, is used for initial applications, requests for replacements for lost cards and corrections, such as name changes. A copy of the application form is shown in Attachment B. Applications must be accompanied by evidence of age, identity and U.S. citizenship or lawful alien status. They may be submitted either in person or by mail, except that aliens and persons 18 or older making initial applications must apply in person.

Most SSN applications are submitted to SSA field offices. In 37 States, applications for new welfare applicants needing SSNs are developed by the State welfare agencies and submitted by the State directly to SSA's Office of Central Records Operations. SSA district offices sometimes make arrangements with schools for "mass enumerations" in which SSA and school officials collaborate in obtaining and reviewing applications from all students who wish to obtain SSNs.

The application forms (SS-5) and accompanying evidence submitted to district offices are screened for completeness and accuracy by district office personnel, who make further contacts with applicants when necessary. The SS-5 information is then keyed in the district office for direct transmission to SSA central operations.

The central processing of the applications consists of validation (which is essentially a matching operation) against existing SSN files, followed by appropriate actions. The exact nature of the validation depends on the type of application. For example, if an initial applicant alleges that he or she has not been issued an SSN previously, the purpose of the validation is to confirm that allegation. Validation procedures are discussed further in item 9b.

The final step depends on the results of the validation. The main possibilities are: assigning an SSN and mailing a card to a new applicant, mailing a replacement card to an applicant, correcting information (such as name) about the applicant in the SSN computerized files, or asking the field office to supply additional information.

When a new SSN is assigned, the next available number for the State from which the application was submitted is used. The sequence of availability proceeds from the lowest area number used in a given State through the highest area number for that State, using the same group number. For example, in New Hampshire, which has been assigned area codes 001, 002, and 003, the last available number in group 001-52 would be followed by the first available number in group 002-52, and the last available number in that group would be followed by the first available number in group 003-52.

6. Sampling properties

In theory, a probability sample could be selected using digital patterns based on any of the nine digits of the SSN or combinations thereof. However, consideration of the information content of the first five digits, as described in item 4, makes it clear that use of any of those digits should be avoided. It would be most inconvenient to select a sample that turned out to include only persons who were railroad workers at the time their SSNs were issued and had all been issued their SSNs not later than 1963!

The serial number part of the SSN, however, does not have this kind of problem and consequently is frequently used for digital sampling from a file of records that includes SSNs. Assuming a uniform distribution of 9,999 possible serial numbers (SSNs ending in 0000 have never been issued), it is possible to choose a digital sampling pattern that will approximate any desired sampling fraction. There are usually several alternatives. For example, to select a sample of approximately 5 percent (1 in 20) of the records, one could use

(1) 5 of the 100 possible combinations of the 8th and 9th digits;

(2) 50 of the 1,000 possible combinations of digits 7, 8 and 9;

(3) 500 of the 9,999 combinations of digits 6, 7, 8 and 9;

(4) 5 of the 100 possible combinations of the 7th and 8th digits

and so forth. The combinations of digits selected may be chosen at random with or without replacement (the latter would be preferable) or systematically with a random start. In the latter case, for exmple, we might choose the pair 73 at random and include with it the pairs 93, 13, 33 and 53.

The use of selected digits or combinations of digits for sampling is actually a form of cluster sampling. In the illustration used above, we could describe a population of records as consisting of 100 clusters, each consisting of all records with SSNs having a particular pair of 8th and 9th digits. Five of these clusters are selected by an appropriate probability sampling mechanism.

In practice, samples of this kind, especially when only the 8th and 9th digits are used, behave pretty much like random samples, chosen without replacement. In particular, reasonably accurate estimates of sampling error can be calculated as though the data were from a simple random sample.

In selecting samples based on the serial number portion of the SSN, the following points should be considered:

(1) The serial number 0000 is not used. The effect of this, which is quite small, on the expected sample size can easily be calculated.

(2) The digital patterns used for any particular sample determine only the expected sampling fraction or size. The sample size realized by using a particular set of digits or combination of digits will, in general, differ somewhat from its expected value. If precise control of sample size is important, this can be achieved by oversampling initially and then subsampling units at random or systematically from the initial sample.

(3) As discussed in item 3d, some persons have been issued more than one SSN. Such persons may have multiple chances of selection in a sample of persons obtained by selecting SSNs, depending on what record sets are being used. If the number of SSNs that each sample person has can be determined, appropriate adjustments can be made in estimates based on the sample. Because the phenomenon is infrequent, however, it is usually ignored in practice.

(4) Various studies (Hawkes and Harris, 1969; Page and Wright, 1979) have shown that the distributions of SSNs by ending digit in selected record sets is essentially uniform. However, studies conducted with various record sets in the late 1960s and early 1970s (Hawkes and Harris, 1969; Internal Revenue Service, 1973) showed a negative linear relationship between the ascending sequence of digits in positions 6 and 7 and the number of SSNs in these record sets having those digits. This probably resulted from the fact that, until 1972, SSNs in each area-group combination were issued consecutively by serial number, from 0001 to 9999. Since then, they have been issued in a randomized order, largely to avoid issuing consecutive numbers to persons with the same surname. Because of the new issuance procedure, one would expect this relationship to disappear gradually. However, to be on the safe side, it is recommended that: (1) digital sampling patterns use only the 8th and 9th digits whenever requirements can be met in that way, and (2) whenever multiple combinations of two or more digits are used, they should be selected systematically rather than at random from the range of possible combinations.

7. Links with other numeric identifiers

At the Federal level, there are two kinds of links between SSNs and employer identification numbers (EINs). For employees, the link occurs in the W-2/W-3 annual wage and tax reporting system (prior to 1978, reporting was quarterly). For many years SSA has used this link for statistical purposes, in the Continuous Work History Sample system, to add employer locations and industry data to records of earnings and demographic characteristics for sample persons. More recently, the Statistics of Income Division of IRS has used the same link to obtain employer industry codes to use as an aid in coding occupations reported by individual taxpayers on their returns.

The second link between SSNs and EINs applies to persons who operate businesses as sole proprietors. This link applies primarily to sole proprietors with employees; those with no employees are not, in general, required to obtain and use EINs. The link occurs in two ways: on income tax returns of sole proprietors, and on new applications for EINs. On income tax returns, the business schedules (C and F) call for entries of both the EIN (if the taxpayer has one) and the SSN. On EIN application forms (Form SS-4), applicants who are sole proprietors are asked to enter their SSNs.

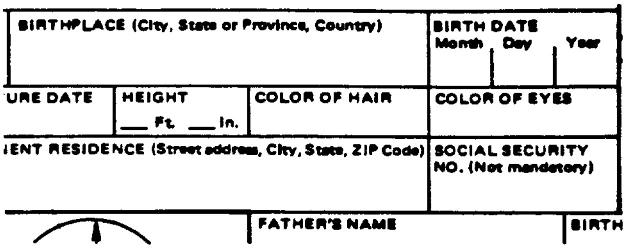
There are undoubtedly several links between the SSN and other numeric identifiers at the State and local levels. One obvious one is the link between SSNs and employer unemployment insurance (UI) identification numbers, which is necessary for the operation of the UI program. The precise nature of the linkage varies by State and, for the minority of States which operate under the "wage request" system, it may not exist in any readily accessible sense.

8. Reporting formats and problems

a. Formats.--Many different administrative and statistical forms include spaces for recording SSNs, either by the holders or by someone else completing the form. There is no standard format for this purpose. The particular format used may have some effect on the accuracy with which SSNs are entered on the forms and read from the forms for purposes of manual transcription or data entry.

Format features that vary include: width and height of the space provided for the number; separators used for the area, group, and serial numbers; use of boxes for individual digits; and the label used to indicate what should be entered. Some examples of these features appear below. All of them show the actual size of the entry space on the form.

Example 1. Department of State, Passport Application, Form SDP-11 (7-79)



BIRTHPLACE (City, State or Province, Country)

BIRTH DATE
Month Day Year

URE DATE

HEIGHT
___ Ft. ___ In.

COLOR OF HAIR

COLOR OF EYES

IENT RESIDENCE (Street address, City, State, ZIP Code)

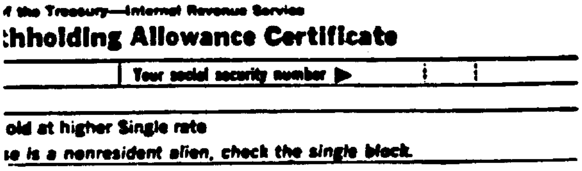
SOCIAL SECURITY NO. (Not mandatory)

FATHER'S NAME

BIRTH

Of several formats examined, this one provided the narrowest space for entering the SSN, with a width of 1 1/4 inches. Most others were in the range of 1 1/2 to 2 inches.

Example 2. Internal Revenue Service, Employee's Withholding Allowance Certificate, Form W-4 (10-79)



f the Treasury—Internal Revenue Service

:hholding Allowance Certificate

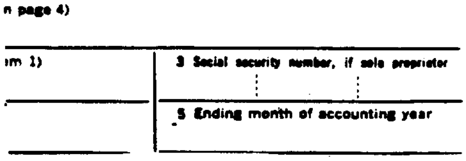
Your social security number ▶

old at higher Single rate

se is a nonresident alien, check the single block.

This format allowed the smallest vertical distance of those examined, 5/32 inch. It uses vertical dotted lines as separators for the three parts of the SSN.

Example 3. Internal Revenue Service, Application for Employer Identification Number, Form SS-4, (8-76).



n page 4)

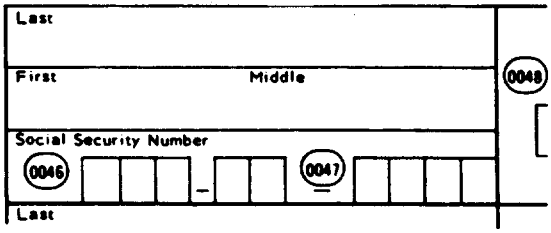
rm 1)

3 Social security number, if sole proprietor

5 Ending month of accounting year

This format also uses the dotted vertical lines as separators. In this case, the spaces for the three portions of the SSN are all the same length, 5/8 inch. Other forms using separators make the lengths of the three spaces roughly proportional to the number of digits to be entered, i.e., 3, 2, and 4.

Example 4. Bureau of the Census/Department of Health and Human Services, Income Survey Development Program, 1978 Research Panel-July Questionnaire, Form ISDP-403.



Last

First Middle

0048

Social Security Number

0046 □□□ - □□ 0047 - □□□□

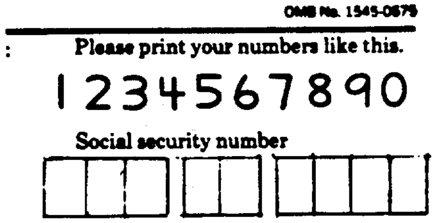
Last

This format illustrates the use of separate boxes for each digit of the SSN. The three parts of the SSN are separated by horizontal dashes. The circled numbers are source codes for data entry.

Example 5. Social Security Number Card (Original, Replacement or Correction), Form SS-5 (5-84) (see Attachment B).

This item is completed only for persons who already have SSNs and are applying for a replacement or correction. This format uses a box for each digit, with intervening spaces, and horizontal dashes to separate the three parts of the SSN. The wording of the item label reflects the fact that the form is sometimes completed by someone other than the "applicant."

Example 6. Internal Revenue Service, Form 1040 EZ Income Tax Return for Single Filers with no Dependents.



OMB No. 1545-0675

: Please print your numbers like this.

1234567890

Social security number

This format is used for handwritten entries by taxpayers that will be read automatically by optical character reading equipment. On the actual form, the boxes for the individual digits are in light blue. The boxes for the area, group and serial parts of the SSN are separated.

Example 4 above comes from a questionnaire that is completed by trained Census Bureau interviewers. The other examples are all from forms that are filled by members of the general public. No experimental research on alternative formats for recording SSNs has been identified. Some other research has suggested that the use of individual character separators may actually reduce legibility of entries (Wright, 1980).

b. Reporting and processing errors.--Most errors in SSNs in data files occur for two reasons: (1) the person completing the form or answering the questions gave an SSN for the wrong person, or (2) the SSN is for the right person, but it was reported, recorded, transcribed or keyed incorrectly.

The first type of error can occur, for example, when a widow reports the number under which she is receiving benefits, rather than her own. Another example is what SSA calls the "pocketbook number." The number 078-05-1120 appeared on a sample account number card contained in wallets sold nationwide in 1938. Several thousand people mistakenly reported this number to their employers as their own! By the 1970s there were over 20 different pocketbook numbers (HEW Secretary's Advisory Committee, 1973, p. 112).

People who lose their social security cards can apply for replacement cards bearing the SSN already issued to them. In cases where they are not able to give their SSN on the application, SSA must determine the correct SSN based on other identifying information. Occasionally a mismatch occurs and the person will be issued a replacement card bearing someone else's SSN.

The second type of error is usually an error in a single digit or a transposition of digits, types of errors that could be easily corrected if a check digit were used.

Cobleigh and Alvey (1974) describe errors detected when SSNs reported in the Current Population Survey were validated against Social Security Administration files. About three percent of the reported SSNs were clearly in error. Roughly two-thirds of these were found to have transposition or single-digit errors. Another one-sixth were SSNs belonging to other members of the same household, and the remainder could not be located in SSA's files.

9. Validation procedures

a. Intra-record validation.--When undertaking record linkages based on SSNs, it is usually desirable to start by identifying SSNs that are clearly invalid. A first step might be to look at the SSN itself and determine whether it is within the range of numbers issued to date. SSA will make available, on request, up-do-date information on the area numbers that have been issued so far and, for each of those numbers, the "highest" group number issued. "Highest" must be interpreted in terms of the standard sequence for use of group numbers within an area number, as explained in item 4 above.

Attachment C provides this information as of January 2, 1985. As of that date, the only area numbers used were those in the ranges 001 to 587, 589 to 595, 600 and 601, and 700 to 728. Also, group number 00 and serial number 0000 are never used. Current information on highest group numbers may be obtained from the director of the OASDI Statistics Division; Office of Research, Statistics and International Policy; Social Security Administration.

If records to be linked have information on date of birth or age, the SSN can be checked for consistency with age. The operating rule is that a person whose SSN was issued x years ago must be at least x years old. Since virtually all numbers issued through 1961 were issued to employed persons, only a few errors would be made by requiring that persons with numbers issued in this period be at least x + 15 years old. For SSNs issued from 1951 onwards, the SSA can provide fairly precise information about the years in which numbers with specific area-group combinations were issued (contact the source given in the preceding paragraph). For numbers issued prior to 1951, only rough estimates of issuance periods for area-group combinations are possible.

b. Validation against SSA records.--Validation is defined broadly here as a process in which SSN information for individuals from sources external to SSA records is checked against those records to determine its validity. Specifically, if the external record includes an SSN, it is desired to know whether the SSN is the correct one for that person and, if it is not correct, what the correct SSN, if any, is for that person. If the external record for a person has no SSN, it is desired to know whether that person has an SSN and, if so, what it is. This kind of validation requires matching external records to SSA records and should be thought of in that context.

Validation of SSN information is done routinely by SSA for program purposes. Somewhat less frequently it is undertaken for statistical purposes. Some examples of the latter are:

(1) Validation of SSNs collected in pretests for the 1970 Census of Population (Ono et al., 1968).

(2) Validation of SSNs collected in the March 1973 Current Population Survey, as a preparatory step before adding SSA and IRS administrative data to the survey records (covered in several reports and articles, e.g., Cobleigh and Alvey, 1974; Social Security Administration, 1981a).

(3) Validation of SSNs collected in panel surveys as part of the Income Survey Development Program (Kasprzyk, 1983).

(4) In various mortality followup studies, as a preparatory step before determining which members of an externally identified study population have died, according to SSA records.

Attachment D provides a summary description of SSA's current validation procedures for program operations. A combination of computerized and manual procedures is used, and unresolved cases are returned to district offices with an instruction to seek additional information from the applicant or claimant. The SSN files maintained by SSA are now fully computerized and a more sophisticated computer validation system is being developed.

A variety of validation procedures have been used in statistical applications; some of them are described in the references cited above.

The circumstances under which SSA will validate SSN information for administrative or statistical purposes are limited by law and by SSA regulations and policies. Anyone wishing to validate SSN information for statistical or research purposes should contact SSA's Office of Research, Statistics and International Policy.

10. Use as a matching variable

Arellano (n.d.) discusses use of the SSN in record linkages based on the model proposed by Fellegi and Sunter (1969). He recommends that the SSN not be used for blocking, because of the possibility that some individuals in the files to be linked may not have been issued SSNs. To use the SSN as a component of the comparison vector, Arellano recommends that the 9 digits of the SSN be partitioned into four elements on a 2,2,2,3 basis. He identifies 17 possible configurations of the SSN component of the comparison vector, covering the possible realizations of agreements and disagreements in the four elements, plus the case in which no SSN is available for one or both members of the comparison pair. He then suggests procedures for assigning conditional probabilities to these configurations for the matched and unmatched sets. These probabilities are based on assumptions about the kinds of errors that can occur in the matched set and on observed frequencies of realizations of the first three elements of the partitioned SSNs in the files to be linked (realizations of the fourth element are assumed to be uniformly distributed).

Rogot et al. (1983) report on linkages of records from the Census Bureau's Current Population Survey with the National Death Index, using each person's name, SSN and date of birth as key matching variables. Based on the results of an evaluation study in which "truth" (match or non-match) was based on a consensus of three raters using all available information for a set of "possible matches," they concluded that whenever SSNs agreed, it was appropriate to classify the pair of records as a positive link, provided there was agreement on sex. The use of probabilistic matching procedures was restricted to cases for which the SSNs did not agree or were missing on one or both records.

REFERENCES

Arellano, M.
(n.d.) Optimum utilization of the social security number for matching purposes. No further identification available.

Cobleigh, C. and Alvey, W.
1974 Validating reported social security numbers. American Statistical Association Proceedings, Social Statistics Section, 145-150.

Fellegi, I. and Sunter, A.
1969 A theory for record linkage. Journal of the American Statistical Association, 64(328); 1183-1210.

Hawkes, T. and Harris, R.
1969 An analysis of social security numbers in the SMI actuarial sample. Actuarial Note No. 62. Social Security Administration, U.S. Department of Health, Education, and Welfare.

HEW Secretary's Advisory Committee on Automated Personal Data Systems
1973 Records, Computers and the Rights of Citizens. Washington, D.C.: U.S. Department of Health, Education, and Welfare.
Chapt. VII. The Social Security Number as a Standard Universal Identifier
Chapt. VIII. Recommendations Regarding Use of the Social Security Number

Internal Revenue Service
1973 Evaluation of the randomness of the four ending digits of the social security numbers on the master file, 1971. Unpublished report by Mathematical Statistics Branch, Statistics Division. U.S. Department of the Treasury.

Kasprzyk, D.
1983 Social security number reporting, the use of administrative records, and the multiple frame design in the Income Survey Development Program. Pp. 123-144 in Technical, Conceptual and Administrative Lessons of the ISDP (M. David, ed.) Washington, D.C.: Social Science Research Council.

Kilss, B. and Tyler, B.
1974 Searching for missing social security numbers. American Statistical Association Proceedings, Social Statistics Section, 137-144.

Ono, M., Patterson, G. and Weitzman, M.
1968 The quality of reporting social security numbers in two surveys. American Statistical Association Proceedings, Social Statistics Section, 197-205.

Page, W. and Wright, G.
1979 A Statistical Study of the VA Annual Patient Census Sampling Procedure, 1975-1977, Controller Monograph Technical Series, No. 1. Veterans Administration.

Privacy Protection Study Commission
1977 Personal Privacy in an Information Society. Washington, D.C.: U.S. Government Printing Office.
Chapt. 16. The Social Security Number.

Rogot, E., Schwartz, S., O'Conor, K. & Olsen, C.
1983 The use of probabilistic methods in matching census samples to the National Death Index. Pp. 75-80 in Statistics of Income and Related Administrative Record Research: 1983, Internal Revenue Service.

Social Security Administration
1981a Methods of Estimation for the 1973 Exact Match Study. Studies from interagency data linkages, Report No. 10. Washington, D.C.: Department of Health and Human Services.

Social Security Administration
1981b Social security numbers issued, 1937-79. Research and Statistics Notes, No. 7, by F. Bamberger. Washington, D.C.: U.S. Department of Health and Human Services.

Wright, P.
1980 Strategy and tactics in the design of forms. Visible Language 14(2): 151-193.

ATTACHMENT A

Table 1.--Social Security Numbers Issued, By Sex of Applicants, 1937-79

(In thousands)

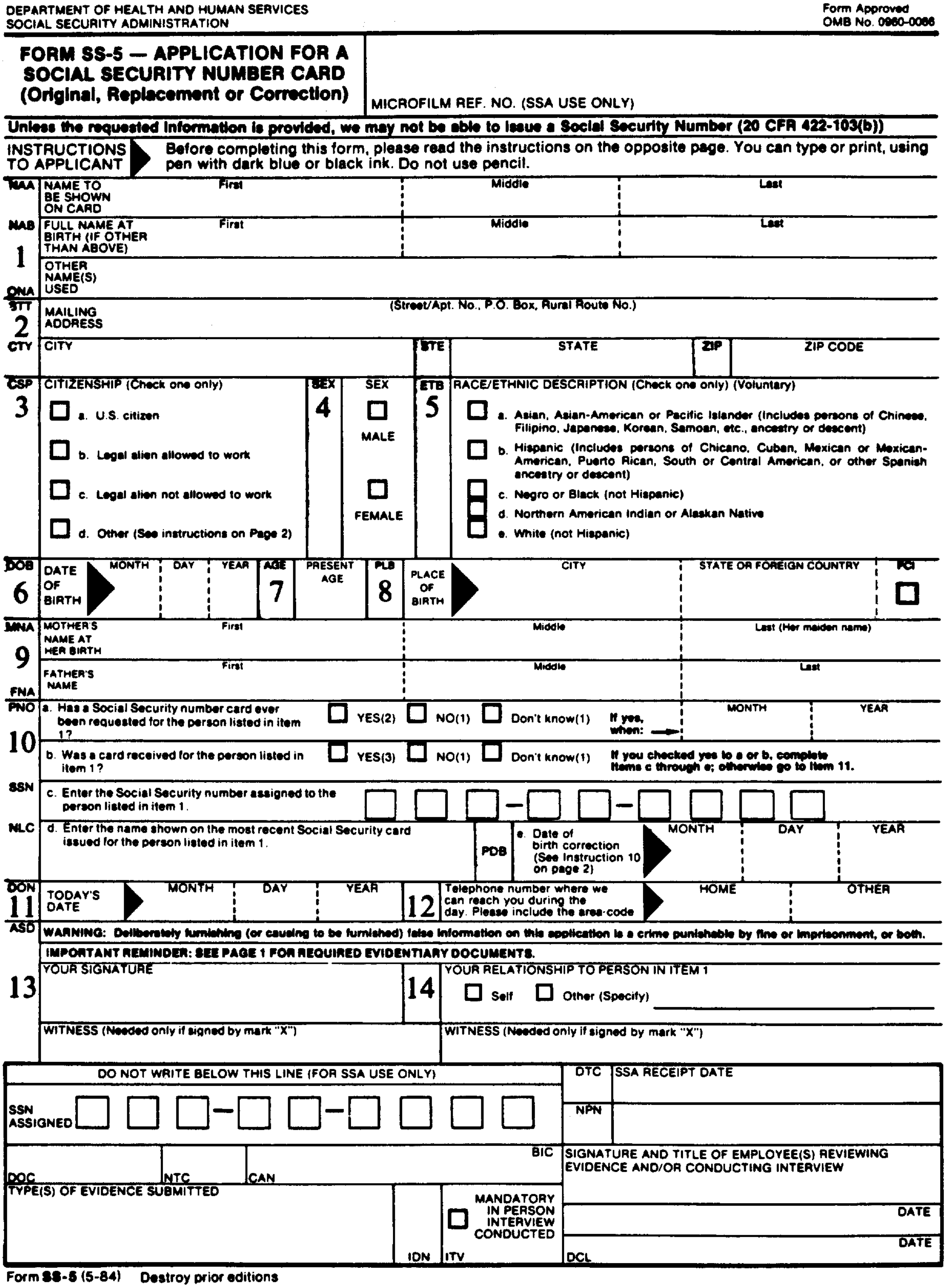
Year	Total	Male	Female
1937[1]	37,139	26,981	10,158
1938	6,304	4,010	2,294
1939	5,555	3,291	2,264
1940	5,227	3,080	2,147
1941	6,678	3,702	2,976
1942	7,637	3,547	4,090
1943	7,426	2,905	4,521
1944	4,537	1,830	2,707
1945	3,321	1,506	1,815
1946	3,022	1,432	1,590
1947	2,728	1,299	1,429
1948	2,720	1,305	1,415
1949	2,340	1,113	1,227
1950	2,891	1,406	1,485
1951	4,927	2,420	2,507
1952	4,363	2,292	2,071
1953	3,464	1,664	1,800
1954	2,743	1,299	1,444
1955	4,323	2,304	2,019
1956	4,376	2,391	1,985
1957	3,639	1,793	1,846
1958	2,920	1,384	1,536
1959	3,388	1,645	1,743
1960	3,415	1,663	1,752
1961	3,370	1,665	1,705
1962	4,519	2,109	2,410
1963	8,617	3,739	4,878
1964	5,623	2,707	2,916
1965	6,131	2,746	3,385
1966	6,506	2,894	3,612
1967	5,920	2,855	3,065
1968	5,862	2,856	3,006
1969	6,289	3,105	3,184
1970	6,132	3,004	3,128
1971	6,401	3,122	3,279
1972	9,564	3,948	5,616
1973	10,038	4,849	5,189
1974	7,998	3,950	4,048
1975	8,164	3,992	4,172
1976	9,043	4,507	4,536
1977	7,724	3,872	3,852
1978	5,260	2,682	2,578
1979	5,213	2,649	2,564

[1] Includes issuances in November and December 1936.

Source: Social Security Administration, 1981b.

ATTACHMENT B

Form SS-5.--Application for a Social Security Number Card



DEPARTMENT OF HEALTH AND HUMAN SERVICES
SOCIAL SECURITY ADMINISTRATION

Form Approved
OMB No. 0960-0066

FORM SS-5 — APPLICATION FOR A SOCIAL SECURITY NUMBER CARD (Original, Replacement or Correction)

MICROFILM REF. NO. (SSA USE ONLY)

Unless the requested information is provided, we may not be able to issue a Social Security Number (20 CFR 422-103(b))

INSTRUCTIONS TO APPLICANT ▶ Before completing this form, please read the instructions on the opposite page. You can type or print, using pen with dark blue or black ink. Do not use pencil.

NAA 1	NAME TO BE SHOWN ON CARD	First	Middle	Last
NAB	FULL NAME AT BIRTH (IF OTHER THAN ABOVE)	First	Middle	Last
ONA	OTHER NAME(S) USED			

STT 2	MAILING ADDRESS	(Street/Apt. No., P.O. Box, Rural Route No.)		
CTY	CITY	STE	STATE	ZIP — ZIP CODE

CSP 3 CITIZENSHIP (Check one only)

□ a. U.S. citizen

□ b. Legal alien allowed to work

□ c. Legal alien not allowed to work

□ d. Other (See instructions on Page 2)

SEX 4 SEX

□ MALE

□ FEMALE

ETB 5 RACE/ETHNIC DESCRIPTION (Check one only) (Voluntary)

□ a. Asian, Asian-American or Pacific Islander (Includes persons of Chinese, Filipino, Japanese, Korean, Samoan, etc., ancestry or descent)

□ b. Hispanic (Includes persons of Chicano, Cuban, Mexican or Mexican-American, Puerto Rican, South or Central American, or other Spanish ancestry or descent)

□ c. Negro or Black (not Hispanic)

□ d. Northern American Indian or Alaskan Native

□ e. White (not Hispanic)

DOB 6	DATE OF BIRTH ▶	MONTH	DAY	YEAR	AGE 7	PRESENT AGE	PLB 8	PLACE OF BIRTH ▶ CITY	STATE OR FOREIGN COUNTRY	FCI □

MNA 9	MOTHER'S NAME AT HER BIRTH	First	Middle	Last (Her maiden name)
FNA	FATHER'S NAME	First	Middle	Last

PNO 10 a. Has a Social Security number card ever been requested for the person listed in item 1? □ YES(2) □ NO(1) □ Don't know(1) If yes, when: → MONTH YEAR

b. Was a card received for the person listed in item 1? □ YES(3) □ NO(1) □ Don't know(1) **If you checked yes to a or b, complete items c through e; otherwise go to item 11.**

SSN c. Enter the Social Security number assigned to the person listed in item 1. □□□–□□–□□□□

NLC d. Enter the name shown on the most recent Social Security card issued for the person listed in item 1.

PDB e. Date of birth correction (See Instruction 10 on page 2) ▶ MONTH DAY YEAR

DON 11 TODAY'S DATE ▶ MONTH DAY YEAR

12 Telephone number where we can reach you during the day. Please include the area code ▶ HOME OTHER

ASD **WARNING: Deliberately furnishing (or causing to be furnished) false information on this application is a crime punishable by fine or imprisonment, or both.**

IMPORTANT REMINDER: SEE PAGE 1 FOR REQUIRED EVIDENTIARY DOCUMENTS.

13 YOUR SIGNATURE

14 YOUR RELATIONSHIP TO PERSON IN ITEM 1 □ Self □ Other (Specify) ____________

WITNESS (Needed only if signed by mark "X")

WITNESS (Needed only if signed by mark "X")

DO NOT WRITE BELOW THIS LINE (FOR SSA USE ONLY)

SSN ASSIGNED □□□–□□–□□□□

DTC SSA RECEIPT DATE

NPN

DOC NTC CAN BIC

SIGNATURE AND TITLE OF EMPLOYEE(S) REVIEWING EVIDENCE AND/OR CONDUCTING INTERVIEW

TYPE(S) OF EVIDENCE SUBMITTED

□ MANDATORY IN PERSON INTERVIEW CONDUCTED

IDN ITV

DATE

DATE

DCL

Form SS-5 (5-84) Destroy prior editions

ATTACHMENT C

Distribution of Social Security Numbers as of January 2, 1985: Highest Group Number Issued Within Each Area Number*

001	68
002	68
003	66
004	82
005	80
006	80
007	80
008	66
009	64
010	66
011	66
012	64
013	64
014	64
015	64
016	64
017	64
018	64
019	64
020	64
021	64
022	64
023	64
024	64
025	64
026	64
027	64
028	64
029	64
030	64
031	64
032	64
033	64
034	64
035	54
036	52
037	52
038	52
039	52
040	76
041	76
042	76
043	76
044	76
045	76
046	76
047	76
048	76
049	74
050	68
051	68
052	68
053	68
054	68
055	68
056	68
057	68
058	68
059	68
060	68
061	68
062	68
063	68
064	68
065	68
066	68
067	68
068	68
069	68
070	68
071	68
072	68
073	68
074	68
075	68
076	68
077	68
078	68
079	68
080	68
081	68
082	68
083	68
084	68
085	68
086	68
087	68
088	68
089	68
090	68
091	68
092	68
093	68
094	68
095	68
096	68
097	68
098	68
099	68
100	68
101	68
102	68
103	68
104	68
105	68
106	68
107	68
108	68
109	68
110	68
111	68
112	68
113	68
114	68
115	68
116	68
117	68
118	68
119	68
120	68
121	68
122	66
123	66
124	66
125	66
126	66
127	66
128	66
129	66
130	66
131	66
132	66
133	66
134	66
135	78
136	78
137	78
138	76
139	76
140	76
141	76
142	76
143	76
144	76
145	76
146	76
147	76
148	76
149	76
150	76
151	76
152	76
153	76
154	76
155	76
156	76
157	76
158	76
159	64
160	64
161	64
162	64
163	64
164	64
165	64
166	64
167	64
168	64
169	64
170	64
171	64
172	64
173	64
174	64
175	64
176	64
177	64
178	64
179	64
180	64
181	64
182	64
183	64
184	64
185	64
186	64
187	64
188	64
189	64
190	64
191	64
192	64
193	64
194	64
195	64
196	64
197	64
198	64
199	64
200	62
201	62
202	62
203	62
204	62
205	62
206	62
207	62
208	62
209	62
210	62
211	62
212	06
213	06
214	06
215	06
216	06
217	06
218	06
219	06
220	04
221	68
222	66
223	33
224	33
225	33
226	33
227	33
228	33
229	33
230	31
231	31
232	27
233	27
234	27
235	25
236	25
237	45
238	45
239	45
240	43
241	43
242	43
243	43
244	43
245	43
246	43
247	59
248	59
249	59
250	57
251	57
252	49
253	49
254	49
255	49
256	49
257	47
258	47
259	47
260	47
261	99
262	99
263	99
264	99
265	99
266	99
267	99
268	82
269	82
270	82
271	82
272	82
273	82
274	82
275	82
276	82
277	82
278	82
279	82
280	82
281	82
282	82
283	82
284	82
285	82
286	82
287	82
288	82
289	82
290	80
291	80
292	80
293	80
294	80
295	80
296	80
297	80
298	80
299	80
300	80
301	80
302	80
303	92
304	92
305	92
306	92
307	92
308	92
309	92
310	92
311	92
312	92
313	92
314	92
315	92
316	92
317	92
318	74
319	74
320	74
321	74
322	74
323	74
324	74
325	74
326	74
327	74
328	74
329	74
330	74
331	74
332	74
333	74
334	74
335	74
336	74
337	74
338	74
339	74
340	74
341	74
342	72
343	72
344	72
345	72
346	72
347	72
348	72
349	72
350	72
351	72
352	72
353	72
354	72
355	72
356	72
357	72
358	72
359	72
360	72
361	72
362	94
363	94
364	94
365	94
366	94
367	94
368	94
369	94
370	94
371	94
372	94
373	94
374	94
375	94
376	94
377	94
378	94
379	94
380	94
381	94
382	94
383	92
384	92
385	92
386	92
387	92
388	92
389	92
390	92
391	92
392	92
393	92
394	92
395	92
396	92
397	92
398	92
399	92
400	25
401	25
402	25
403	25
404	25
405	25
406	23
407	23
408	45
409	45
410	45
411	45
412	45
413	45
414	45
415	43
416	19
417	19
418	19
419	19
420	19
421	19
422	19
423	19
424	17
425	51
426	51
427	49
428	49
429	57
430	57
431	55
432	55
433	55
434	55
435	55
436	55
437	55
438	55
439	53
440	84
441	84
442	84
443	84
444	84
445	84
446	82
447	82
448	82
449	69
450	69
451	69
452	69
453	69
454	69
455	69
456	69
457	69
458	69
459	69
460	69
461	69
462	69
463	69
464	69
465	69
466	69
467	69
468	04
469	04
470	04
471	04

Distribution of Social Security Numbers as of January 2, 1985 (cont'd.)

472	04	495	88	518	11	541	11	564	81	587	49	610	00	706	18
473	04	496	88	519	11	542	11	565	81	588	00	611	00	707	18
474	02	497	88	520	04	543	11	566	81	589	30	612	00	708	18
475	02	498	88	521	43	544	11	567	81	590	30	613	00	709	18
476	02	499	88	522	43	545	83	568	81	591	30	614	00	710	18
477	02	500	88	523	43	546	83	569	81	592	30	615	00	711	18
478	06	501	04	524	43	547	83	570	81	593	30	616	00	712	18
479	06	502	02	525	53	548	83	571	81	594	28	617	00	713	18
480	06	503	04	526	99	549	83	572	81	595	28	618	00	714	18
481	06	504	04	527	99	550	81	573	81	596	00	619	00	715	18
482	06	505	13	528	49	551	81	574	76	597	00	620	00	716	18
483	06	506	13	529	49	552	81	575	27	598	00	621	00	717	18
484	04	507	11	530	08	553	81	576	27	599	00	622	00	718	18
485	04	508	11	531	96	554	81	577	11	600	16	623	00	719	18
486	90	509	88	532	96	555	81	578	08	601	14	624	00	720	18
487	90	510	88	533	96	556	81	579	08	602	00	625	00	721	18
488	90	511	88	534	96	557	81	580	19	603	00	626	00	722	18
489	88	512	88	535	94	558	81	581	99	604	00	700	18	723	18
490	88	513	88	536	94	559	81	582	99	605	00	701	18	724	28
491	88	514	86	537	94	560	81	583	99	606	00	702	18	725	18
492	88	515	86	538	94	561	81	584	97	607	00	703	18	726	18
493	88	516	04	539	94	562	81	585	51	608	00	704	18	727	10
494	88	517	04	540	11	563	81	586	78	609	00	705	18	728	14

*First three digits of the social security number are area numbers; second two digits are group numbers.

Group 00 is not a valid group -- it is for program purposes only.

ATTACHMENT D

Excerpt from
Validation and Screening Techniques for Social Security Numbers

VALIDATION OF SSN'S

Minimum information needed to validate an SSN is the person's name, sex, date of birth and the alleged SSN. Validation occurs only when the information on a current transaction exactly matches or can be reconciled with the information on the Alphident/Numident data bases or the microfilm subfiles of these systems. In certain circumstances, additional matching information is needed before validation can occur. If earnings are reported without an SSN or with an SSN or name that does not agree with these files and the correct SSN cannot be determined through internal screening operations, the employer or the worker is asked to furnish additional information to identify the record. The Internal Revenue Service (IRS) uses a similar system to validate SSN's of taxpayers.

MANUAL SCREENING OF DUPLICATE AND ORIGINAL SSN APPLICATIONS

The electronic screening operation to which every application is subjected is capable of processing roughly 85 percent of all applications input by field offices. Through a sophisticated series of screening grids, the computer makes a decision: is this applicant already represented in the Alphident data base? If the decision is yes, the previously assigned SSN is identified and a replacement card is prepared and mailed. If the decision is no, a number is assigned and a card is printed and mailed.

However, the decision-making capability of the system is deliberately limited because some applications have identifying information common to others or conditions exist which should receive a clerical review. These applications produce worksheets which are processed manually by OCRO.

Worksheets to be screened are checked against the Alphident Microfilm File and the Alphident Microfiche File, using the name and date of birth shown on the application. If an SSN is not located for the name and date of birth shown, another search is made using dates of birth somewhat different from the one given on the application. If an SSN is still not located, certain other variations are checked, including name at birth or on the signature line if different from the name in item 1; acceptable variations of common first names; dropping middle name shown; substituting different middle initials; substituting maiden surname for middle given name for married females; substituting initials only in place of complete given names; etc. Once a "possible" SSN is located, verification can be made immediately since full identifying information is available on the Alphident files. See RM 00204.020 for procedures for handling "UTL" and "Investigate" items.

THE ALPHIDENT MICROFILM AND MICROFICHE FILES

The electronic Alphident file is updated daily. If an SSN holder loses the social security card within the first days after it was issued, the number can be located and verified electronically.

The Alphident Microfilm File is an alphabetical file based on the Russell Soundex coding system. It contains essentially the same information as the electronic file.

Because the Alphident Microfilm File is updated only every 3 months, each week an accretion file is prepared on microfiche. This file contains all SSN assignments and corrections to our records processed during the preceding 12 weeks. This file is referred to when there is reason to believe that there was a recent SSN action for an individual.

Each record entry on both the Alphident Microfilm and the Alphident Microfiche Files consists of the following:

DATA	POSITIONS
Blank	1
Soundex	2-5
Blank	6
Applicant's Surname	7-27
Applicant's Given Name	28-43
Applicant's Middle Name	44-45
Month of Birth	56-57
Blank	58
Day of Birth	59-60
Blank	61
Century of Birth	62
Year of Birth	63-64
Blanks	65-66
SSN	67-77
Blank	78
Mother's Surname	79-91
Mother's Given Name	92-102
Mother's Given Initial	103
Blank	104
Sex/Race	105-106
Blank	107
Father's Surname	108-120
Father's Given Name	121-131
Father's Middle Initial	132
Blank	133
City/County of Birth	134-140
State/Country of Birth	141-142
Blanks	143-144
Form/Entry	145-146
Blanks	147-148
Reference Number	149-159
Blank	160

COMMON NAMES IN THE ALPHIDENT FILE

There are over 360 million records in the Alphident File, representing over 277 million SSN's assigned. Many of the names in the file are the same or are very similar. This is why it is extremely important to get complete and accurate identifying information on original applications and on requests for duplicate SSN cards. It is equally important to obtain information that is consistent with that on the original application. Applicants who have lost their original cards should be questioned closely to find out if any of the information on the current application is now different from that which they showed on their original application.

The latest tabulation of common surnames in the SSN file was made in 1974. Some examples of the number of times a common name could appear in Alphident are given below.

NAME	NUMBER OF ITEMS IN ALPHIDENT
Smith	2,382,509
Johnso(n)	1,807,263
Willia(ms)(mson)	1,568,939
Brown	1,362,910
Jones	1,331,205
Miller	1,131,861
Davis	1,047,848
Martin(ez)(son)	1,046,297
Anders(on)	825,648
Wilson	787,825

THE RUSSELL SOUNDEX CODE

By using the Russell Soundex Code system, searching for possible SSN's on the Alphident film and fiche in OCRO is accomplished quickly.

Here are the basic rules for using the Soundex Code.

Use the first letter of the surname, then code the remaining letters as follows:

LETTERS	CODE SYMBOLS
BPFV	1
CGJKQSXZ	2
DT	3
L	4
MN	5
R	6

Vowels are not coded, nor are the letters W, H, and Y. Two successive letters with the same code numbers are coded only once.

Example:
"Mack" is coded M-200. The "a" is not coded since it is a vowel. "c" falls under code symbol 2. "k" also falls under code symbol 2, but is not used since two successive letters with the same code sumbol are coded only once. Since the complete Soundex Code must consist of the first letter of the name followed by three numbers, we add enough zeros to complete the 3-digit code.

Here are some other examples:

1. Snyder - S-536
2. Way - W-000
3. Bear - B-600
4. Brown - B-650

LIMITATIONS IN OCRO SCREENING FOR SSN's

When an applicant has indicated a previous SSN in item 10 of the SS-5 and the correct number cannot be found in the electronic or OCRO screening operations, the data are returned via form SSA-4310 to the district office. This is because studies show that many such applicants are mistaken in stating they previously applied for a number, and it is not worthwhile spending additional time on the case unless different information can be found. When the district office receives a form SSA-4310 from OCRO, it should recontact the applicant for any different information that may be useful in screening. See RM 00204.020 A.1. Take appropriate action, but do not return the SSA-4310 to OCRO.

Upon recontacting the applicant, the district office may discover that a married woman obtained her original SSN under a first husband's name, but is now applying for the duplicate in her second husband's name; that a man who calls himself "Winslow" obtained his number earlier in life as "Buddy;" or that Mr. Kline's record was set up originally under "Cline." There is also a possibility that the applicant may be able to locate the previously issued SSN on an old pay stub or by asking a present or a past employer. This new information may enable OCRO to locate the original SSN. If the applicant is unable to give any information different from what was previously given and is unable to locate the alleged number, the district office has no other choice but to request assignment of an original SSN. However, this should be done only as a last resort, particularly if the person has earnings under the original number which might not be credited when the SSN holder applies for benefits.

These facts point up the need for obtaining the most accurate information possible during the initial interview with the applicant, whether it be for an original or duplicate SSN card; otherwise, multiple numbers may result. Any reasonable assistance should be extended to the applicant to help find out definitely what the alleged prior SSN is. (See RM 00202.025 I.10.)

Source: "The Social Security Number," Program Operations Manual System, Part I, Chapter 00201.000, Section 00201.015, Social Security Administration.

EXACT MATCHING LISTS OF BUSINESSES: BLOCKING, SUBFIELD IDENTIFICATION, AND INFORMATION THEORY

William E. Winkler, Energy Information Administration

1. INTRODUCTION

The purpose of this paper is to present an evaluation of matching strategies for name and address files of businesses. In evaluating matching methods, we wish to minimize erroneous matches and nonmatches and the amount of manual review.

This work and previous work by various authors (Newcombe, Kennedy, Axford, and James, 1959; Newcombe and Kennedy, 1962; Newcombe, Smith, Howe, Mingay, Strugnell, and Abbatt, 1983; Coulter, 1977; Coulter and Mergerson, 1977; Rogot, Schwartz, O'Conor, and Olsen, 1983; Kelley, 1985) rely on matching strategies based on a theory of record linkage formalized by Fellegi and Sunter (1969) and first considered by Newcombe et al. (1959). The Fellegi-Sunter model provides an optimal means of obtaining weights associated with the quality of a match for pairs of records. Linked pairs (designated matches) and nonlinked pairs (designated nonmatches) receive high and low weights, respectively. Pairs designated for further manual followup receive weights between the sets of high and low weights.

Early work by Newcombe et al. (1959, 1962) showed the potential improvement (lower rates of erroneous matches and nonmatches and of manual followup) when weights were computed using surname and date of birth in comparison to when weights were computed using surname only. Coulter (1977) provided an example of the decrease in discriminating power as the probability of identifiers (such as surnames, first names, middle names, and place names) being misreported (transcribed inaccurately) and/or pairs of identifiers associated with individuals being different but accurately reported increases.

While the applied work referenced above involved files of individuals only, this paper provides an evaluation involving files of businesses. Matching using files of businesses is different from matching files of individuals because business files lack universally available and locatable identifiers such as surnames.

Matching consists of two stages. In the blocking stage, sort keys, such as SOUNDEX abbreviation of surname, are defined and used to create a subset of all pairs of records from files A and B that are to be merged. Records having the same sort key are in the same block and are considered during further review. Records outside blocks are designated as nonmatches. In the discrimination stage, surnames and other identifying characteristics are used in assigning a weight to each pair of records identified during the blocking stage.

With the exception of Newcombe et al. (1959, 1962), little work has been performed in evaluating how many erroneous nonmatches arise due to a given blocking strategy. The chief reason that little work has been performed is that identifying erroneous nonmatches due to blocking and accurately estimating error rates is difficult (Fellegi and Sunter, 1969; Winkler, 1984a,b).

The key to identifying difficulties in blocking files of businesses is having a data base in which all matches are identified and which is representative of problems in many business files. In section 2, the construction of such a data base from 11 Energy Information Administration (EIA) and 47 State and industry files is described. Section 2 also contains a summary of the Fellegi-Sunter model and the criteria used in evaluating competing matching strategies.

Section 3 is divided into two parts. The first part contains results obtained by multiple blocking strategies using a procedure in which the numbers of erroneous nonmatches and matches are minimized under a predetermined bound on the number of pairs to be passed on to the discrimination stage (for related work see Kelley, 1985). The results are related to results obtained during the discrimination stage and build on earlier work of Winkler (1984a, 1984b).

In the second part, the main results of the discrimination stage are presented. The effects of improved spelling standardization procedures and identification of additional comparative subfields are highlighted. Although the deleterious effect of poor spelling standardization is covered by the Fellegi-Sunter theory and presented in the simulation results of Coulter (1977), no concrete examples have previously been presented.

The second part also contains results on the variation of cutoff weights and misclassification and nonclassification rates during the discrimination stage. The results are based on small samples used for calibration and obtained using multiple imputation (Rubin, 1978; Herzog and Rubin, 1983) and bootstrap imputation (Efron, 1979; Efron and Gong, 1983). Fellegi and Sunter (1969, p. 1191) indicate that results based on samples are unreliable.

Finally, the second part presents results addressing the strong independence assumptions necessary under the Fellegi-Sunter model and conditioning techniques that can be used in improving matching performance in some situations when direct application of the Fellegi-Sunter model yields high misclassification and/or nonclassification rates. The investigation of independence uses the hierarchical approach of contingency table analysis (Bishop, Fienberg, and Holland, 1975). The conditioning argument uses a steepest ascent approach (Cochran and Cox, 1957).

Section 4 contains a summary and further discussion of the results and problems for future research.

2. EMPIRICAL DATA BASE, METHODS, AND EVALUATION CRITERIA

This paper´s approach to developing more effective matching strategies involves:

1. constructing an empirical data base for testing procedures;
2. employing the Fellegi-Sunter model of record linkage;
3. defining evaluation criteria; and
4. refining procedures in response to empirical results.

A suitable data base should have all duplicates identified and connected to their respective parents (records used for mailing purposes) and present problems that are representative of similar data files (in this case, files of businesses). The identification of all duplicates allows determination of erroneous nonmatches during the blocking stage. Evaluation criteria should be such that they are suitable for adoption by others performing research in matching methodologies.

2.1. Creation of a Suitable Empirical Data Base

The empirical data base consists of 66,000 records of sellers of petroleum products. It was constructed from 11 EIA lists and 47 State and industry lists containing 176,000 records. Easily identified duplicates having essentially similar NAME and ADDRESS fields were deleted when the melded file was reduced from 176,000 to 66,000 records.

The data base contains 54,850 records identified as headquarters or parents (records used for mailing purposes); 3,050 records identified as duplicates (records having names and addresses similar to their parents´); and 8,511 records identified as associates (records such as subsidiaries and branches that have names and/or addresses different from their parents´).

Duplicates were identified primarily through elementary computer-assisted techniques (see Winkler, 1984a); associates were identified through surveying and call-backs. Our evaluation will only consider how well various strategies perform in matching duplicates with headquarters. The presence of unidentified associates, however, can cause falsely higher error rates (see section 2.3.1).

2.1.1. General Applicability of Results

Procedures developed for dealing with problems in the main empirical data base would be generally applicable to most EIA systems because the data base:

1. is larger than any other master frame file in EIA;
2. is involved with retail sales-- such frames are often more difficult to work with than files of individuals or files of headquarter addresses of large corporations; and
3. had greater formatting and spelling standardization difficulties-- it was constructed from many more sources than any other EIA frame.

Because the main empirical date base is constructed from many different lists and contains many records associated with retailers, results should be representative of the difficulties encountered with similarly constructed, non-energy files of businesses.

2.1.2. Improved Spelling Standardization

The original spelling standardization software contained two basic loops. The first replaced most punctuation with blanks and deleted multiple blanks within a field. The second used lookup tables to replace a given spelling of a word with a standardized spelling or abbreviation. Blanks were generally used to delimit words within fields.

Spelling standarization software was updated in two ways. First, the logic of the processing was enhanced to cause changes in character strings that are not easily updated because they contain embedded punctuation or blanks. For instance, "´S" is replaced by "S" and "MC NEELY" by "MCNEELY."

Second, standardization tables were updated with a very large number of spelling variations of words such as ´COMPANY,´ ´DISTRIBUTOR,´ ´SERVICE,´ and ´CORPORATION.´ The key to systematically identifying such spelling variations was a program that created an alphabetic listing and frequency count of every word in a prespecified field such as NAME or STREET ADDRESS. As more than 90 percent of keypunch errors occur after the first character (see e.g., Pollock and Zamora, 1984), most spelling variations of commonly occurring words in the empirical data base have probably been identified.

2.1.3. Identification of Subfields

The identification of subfields was done in two stages. In the first, ZIPSTAN software (U.S. Dept. of Commerce, 1978b) was used to process the STREET ADDRESS field. Although the Census Bureau uses a UNIVAC computer system, we were able to obtain an unsupported version of ZIPSTAN that had been created for use on IBM systems.

The basic idea of ZIPSTAN was to identify key subfields of the STREET ADDRESS field for files of individuals. Although ZIPSTAN assumes that the street address begins with a numeric word, which is the usual situation in the files of individuals for which ZIPSTAN was designed, it is able to process other types of street address subfields that typically occur in files of establishments or businesses.

Although ZIPSTAN provided warning messages for 18 percent of the 66,410 records in the empirical data base, it was still helpful for most cases. Warning messages consisted of ´MISSING STATE NAMES´ (records associated with non-US postal addresses), ´PLACE NAMES CONVERTED´ (minor conversion of the city field), ´STREET NAMES CONVERTED´ (minor conversion of the street name), ´SYNTAX CONVERSION´ (conversion of unacceptable patterns of word characteristics), and ´POST OFFICE BOXES´ (containing PO BOX).

The following examples show some representative EIA records before and after ZIPSTAN processing.

Before ZIPSTAN

1. EXCH ST
2. HWY 17 S
3. 1435 BANK OF THE
4. 2837 ROE BLVD
5. MAIN & ELM STS
6. CORNER OF MAIN & ELM
7. 100 N COURT SQ
8. 100 COURT SQ SUITE 167
9. 2589 WILLIAMS DR APT 6
10. 15 RAILROAD AVE
11. 2ND AVE HWY 10 W
12. MAIN ST
13. 184 N DU PONT PKWY
14. 1230 16TH ST
15. BOX 480

After ZIPSTAN

No.	House No.	Pre-fixes		Street Name	Suf-fixes		Unit
		1	2		1	2	
1.				EXCH	ST		
2.		HW		17TH	S		
3.	1435			BANK OF THE			
4.	2837			ROE	BL		
5.				MAIN ELM STS			
6.				CORNER OF MAIN ELM			
7.	100	N		COURT	SQ		
8.	100	CT	SQ	*** NO NAME ***			RM 167
9.	2589			WILLIAMS	DR		AP 6
10.	15			RAILROAD	AV		
11.				2ND	AV	HW	10
12.				MAIN	ST		
13.	184	N		DU PONT	PW		
14.	1230			16TH	ST		
15.	480			*PO BOX*			

ZIPSTAN is able to identify accurately subfields in 13 of 15 cases. The two exceptions are cases 2 and 8. In case 2, 'HWY' is moved to a prefix position and '17' is placed in the STREET NAME position. In case 8, 'COURT,' the street name, is placed in a prefix location.

Although ZIPSTAN accurately identifies the subfields associated with intersections (cases 5, 6, and 11), such identification may not allow accurate delineation of duplicates in comparisons of various lists. Some lists may contain STREET ADDRESSes in the following forms, none of which can be readily comparable with the forms in examples 5, 6, and 11.

5. 34 Main St
5. Elm and Main Streets
11. Hwy 10 W
11. 7456 Richmond Hwy

In the second stage of subfield identification, the following words in the NAME field were identified:

KEYWORD1	Largest word in NAME field
KEYWORD2	2nd largest word in NAME field (ties broken by alpha sort)
CON	Concatenation of initials

The above three subfields were used for comparison purposes because the NAME field in lists of businesses generally does not contain words such as SURNAME and FIRST NAME that are present in files of individuals. Based on a sample of 1000 records, an upper bound of 27 percent at the 95 percent confidence level is placed on the number of records containing a word that could be identified as SURNAME.

The identification of SURNAMEs was not performed for three reasons: (1) it is difficult to develop software that accurately identifies records that contain SURNAME (see U.S. Dept. of Agriculture, 1979); (2) it is difficult develop software to identify SURNAMES within the NAME field (e.g., PAUL ROBERT or ROBERT PAUL- which is the SURNAME?); and (3) the small number of records to be compared and containing surnames was not sufficient to justify such a development effort.

The following provides examples of legitimate variations associated with NAME field of one company:

J K Smith Co
Smith Jonathon K
Smith Fuel Service Co
J K Smith Exxon Fuel Service
J K S Fuel

Fellegi and Sunter (1969, pp. 1193-1194) provide an explicit theoretical model for how much such legitimate spelling variations decrease the accuracy with which matches and nonmatches are delineated. Coulter (1977) provides an empirical example of the decrease based on a simulation.

Identifying and comparing the largest words in the NAME field are only performed after spelling standardization and/or abbreviation so that the chance of designating large words with little distinguishing power is minimized.

For instance, if a character string such as 'DISTRIBUTOR' appeared in the name field, it would likely be the longest word. Replacing the various spellings of 'DISTRIBUTOR' with an abbreviation such as 'DSTR' either allows it to be deleted so that it is not considered by the keyword-identification program or allows longer words with possibly more distinguishing power to be identified.

Although methods of identifying subfields might be considered results, we are primarily concerned with how their identification affects the efficacy of various matching procedures. Consequently, the identification can be considered a preprocessing step (see e.g., Winkler, 1985) that is used in creating the data base used in evaluations.

2.1.4. Completeness of Identification of Duplicates

It is likely that few, if any, additional erroneous nonmatches of duplicates are present in the empirical data base for three reasons. First, no additional duplicates were identified in the set of headquarters records during a manual review of all 1,500 records in a random sample of 3-digit ZIP codes. Second, no additional duplicates were identified during a review of a sample of 20 pages (each containing 60 records) in a listing that was ordered alphabetically using the NAME field. Third, no additional duplicates were identified during the discrimination stage (section 3.2).

Without further manual followup, it is impossible to determine how many unidentified associate records are in the set of headquarters records. It is unlikely that surveying and callbacks--because they were first-time efforts--would have been able to identify them all.

Even if more associates are identified, the results of matching duplicates against headquarters will not be seriously affected. The main effect of identifying more associates will be to lower the estimated rates of erroneous matches. Some duplicates are now matched to headquarters that are not identified as their parent and that are actually associates of the duplicates' parents. Each such match is presently counted as an erroneous match.

2.2. Methods

2.2.1. The Formal Probabilistic Model

The Fellegi-Sunter model (1969) uses an information-theoretic approach embodying principles first used in practice by Newcombe (Newcombe et al., 1959). For a review of existing techniques and their relationship to classical information theory see Kirkendall (1985).

In the Fellegi-Sunter model, agreements on characteristics such as SURNAME or ZIP code are assumed to be more common among truly matched pairs than among erroneously matched or unblocked pairs. In practice, specific binit weights of agreement (or disagreement) are computed by,

$$W = \log_2 A/B$$

where

A= the proportion of a particular agreement (or disagreement) defined as specifically as one wishes among matched pairs, and

B= the corresponding proportion of the same agreement (or disagreement) among pairs that are rejected as matches.

The following table will help us to understand more specifically the computation of weights.

Table 1: Counts of True State of Affairs

Specified Characteristic	Match	Nonmatch
Agree	a	b
Disagree	c	d

If we wish to compute the weight associated with agreement on a specified characteristic, then we take A=a/(a+c) and B=b/(b+d); for disagreement, we take A=c/(a+c) and B=d/(b+d).

For each detailed comparison of a pair of records, the weights for appropriate agreements and disagreements are added together, and the total weight, TWT, is used to indicate the degree of assurance that the pair relates to the same entity. The procedure assumes that weights associated with individual agreements or disagreements are uncorrelated with each other (at least conditionally, see e.g., Fellegi and Sunter, 1969, p. 1190).

Cutoffs UPPER and LOWER are chosen (using empirical knowledge or educated guesses) and the following decision rule is used:

If TWT > UPPER, then designate pair as a match.

If LOWER <= TWT <= UPPER, then hold for manual review.

If TWT < LOWER, then designate pair as a nonmatch.

Given fixed upper bounds on the percentages of erroneous nonmatches having TWT < LOWER and of erroneous matches having TWT > UPPER, Fellegi and Sunter (1969, p. 1187) show that their procedure is optimal in the sense that it minimizes the size of the manual review region.

In some cases, either looking at disjoint subsets of the set of blocked pairs and/or increasing or decreasing individual weights used in computing the total weight, TWT, can improve the efficacy of the above decision rule. For instance, among a set of records that are blocked into pairs using the first six characters of the STREET field, individual weights associated with agreements and disagreements on characteristics of the NAME field might be increased and decreased, respectively.

A procedure that uses individual weights, that have been varied in order to achieve greater accuracy in the set of pairs designated as matches and nonmatches and/or a reduction in the set of records held for manual review, will be referred to as a modified information-theoretic procedure. An unmodified procedure will be referred to as the basic information-theoretic procedure.

2.2.2. Specific Weight Computation

In addition to individual weights computed using the subfields HOUSE NUMBER, PREFIX, STREET NAME, SUFFIX, UNIT DESIGNATOR, KEYWORD1, KEYWORD2, and CO given in section 2.1.3, the following subfields were used in computing individual weights:

Field	Subfield Columns	Designated as
NAME	1-4,5-10,11-20,21-30	N1,N2,N3,N4
STREET	1-6,7-15,16-30	S1,S2,S3
ZIP	1-3,4-5	Z1,Z2
CITY	1-5,6-10,11-15	C1,C2,C3
STATE	1-2	
TELEPHONE	1-3,4-6,7-10	T1,T2,T3
WL-NAME 1/	1-4,5-10,11-20,21-30	W1,W2,W3,W4

1/ Sort words in NAME field by decreasing order of wordlength. Break ties with alpha sort.

Generally, corresponding subfields were used in computing individual weights. The exceptions were comparisons of the first and second keywords (section 2.1.3) in the NAME field.

It is important to note that if any weight associated with a given SORT KEY, say TELEPHONE, used in blocking is computed only for records within the subset of pairs having the SORT KEY agreeing, then the comparison has no discriminating power and the resulting weight is zero. If, however, a weight is computed for a comparison of a SORT KEY within a subset of pairs which do not all agree on the SORT KEY, then the weight could be nonzero. Also, it is intuitive that some of the comparisons, say of the above defined subfields of the NAME and KEYWORDs (section 2.1.3) may not be independent.

2.2.3. Variances

As the truth and falsehood of matches in the set of blocked pairs were known for the evaluation files, estimated error rates and their variances were obtained using multiple samples.

The basic procedure was to draw samples of equal size, compute cutoff weights using each sample (based on at most 2 percent of nonmatches being classified as matches and at most 3 percent of matches being classified as nonmatches), use each pair of cutoff weights on the entire data base to determine overall error rates, and compute the variances of the cutoff weights and the overall error rates over the set of samples.

The multiple imputation procedure of Rubin (1978) has been used for evaluating the effects of different methods of imputing for missing data but is applicable in our situation. Multiple imputation entails obtaining several estimates using different samples and then computing the mean and variance over samples. In using Rubin´s procedure, we sample without replacement.

The key difference from Efron´s bootstrap is that sampling is performed with replacement. Our application corresponds almost exactly to the first example in the paper of Efron and Gong (1983).

2.2.4. The Independence Assumption

Fellegi and Sunter (1969, pp. 1189-90) state that the independence assumption for the comparisons of information contained in different subfields is crucial to their theory but that the independence assumption may not be crucial in practice. They note that obtaining total weights having a probabilistic interpretation only necessitates that comparisons be conditionally independent. The conditioning must be consistent with the way total weights are computed.

There are several practical difficulties with testing their independence assumption. First, it must be tested separately for matches and nonmatches. Newcombe and Kennedy (1962) provide a method of approximating the weights for nonmatches and show that accurately approximating the weights for matches is difficult. The chief reason is that the number of nonmatches is close to the number of pairs in the cross product of two files A and B while matches represent a relatively small subset (of all pairs) having specific characteristics.

Second, the weights of nonmatches and matches may vary substantially depending on what blocking criteria are used. If, say, four independent criteria are used, then it might be necessary to examine as many as 15 (2**4-1) mutually exclusive subsets of the set of blocked pairs (see sections 3.1 and 3.2).

Third, the collection of the information necessary for contingency table analyses is difficult because we have no strong control over sampling design (Bishop, Fienberg, and Holland, 1975, pp. 36-39). Even with moderately large samples, some of the subsets determined by blocking criteria may be too small for adequate analysis of the conditional independence of two variables given two or more variables because of the number of marginal constraints that are zero (see section 3.2.8).

Fourth, if many different subfields and/or different means of comparing them are considered (we will consider 30; Newcombe and Kennedy, (1962, p. 566), considered 200), then modelling the conditional relationships using contingency table techniques (Bishop, Fienberg, and Holland, 1975) can be cumbersome.

Even if dependencies occur, it may be possible to vary weights associated with individual comparisons (i.e., steepest ascent, see e.g., Cochran and Cox, 1957, pp. 357-369) to determine whether the efficacy of the overall weighting procedures can be improved. Our specific steepest ascent method generally involved choosing a few individual weights in disjoint subsets determined by blocking criteria (sections 3.1 and 3.2) and varying them by +/- 0.5.

It is important to note that modifications to individual weights may be heavily dependent on the subsets determined by the blocking criteria.

2.3. Criteria for Evaluation

2.3.1. Type I and II Errors

A Type I error is an erroneous nonmatch and a Type II error is an erroneous match. The Type I error rate is U/D*100 where U is the number of erroneous nonmatches and D is the number of matches. The Type II error rate is F/M*100 where M is the number of pairs designated as matches and F is the number of erroneous matches.

As duplicates unmatched during the blocking stage are considerably more difficult to identify than false matches during the discrimination stage, the primary emphasis in developing a new strategy was minimizing Type I errors during the blocking stage before minimizing Type II and Type I errors during the discrimination stage.

It is important to note that if a pair of files has no erroneous nonmatches, then any matching strategy applied will yield either no pairs during the blocking stage or a Type I error rate of 0 percent and a Type II error rate of 100 percent. Because the empirical data base is relatively free of duplicates (as a result of reducing the empirical database from 176,000 to 66,000 records), application of any matching strategy will produce relatively high Type I error rates during the blocking stage.

As we are primarily concerned with evaluating methodologies for accurately matching pairs that are not readily matched using elementary comparisons (e.g., having major portions of key fields agreeing exactly), the data base of 66,000 records is more suitable for use than the original set of 176,000 records.

2.3.2. Overall Rate of Duplication

The number of erroneous nonmatches as a percentage of the total number of records in a file is also an important evaluation criteria. We define the overall rate of duplication as Q/(X+Q)*100 where Q is the number of erroneous nonmatches and X is the number of parent records.

This additional evaluation criteria is important because the Type II error rate criteria will not provide a measure of how free of duplicates a file is. The Type II error rate does not work well because, as the number of matches, D, in a file decreases, the Type I error rate (U/D*100, where U is the number of erroneous nonmatches) will necessarily increase.

In the analysis of the empirical data base, D is held constant so that the comparative advantages of various strategies can be assessed using Type I error rates. The overall rate of duplication will not work well for these comparative evaluations because it is too dependent on the number of parent records, X, which does not change. That is, if U1 and U2 are the numbers of erroneous nonmatches under two matching strategies and U1<U2<<X, then U1/(U1+X) and U2/(U2+X) are approximately equal.

2.3.3. Amount of Manual Review

The amount of manual review is a critical feature in any matching procedure because manual review is both time-consuming and expensive. If one procedure requires one half as much manual review as another, yields Type I error rates that are only somewhat higher than the other, and yields similar rates of erroneous nonmatches (section 2.3.2), then there is strong justification for adopting the procedure requiring less manual review.

3. RESULTS USING THE EMPIRICAL DATA BASE

Results of the empirical analyses for the blocking stage and the discrimination stage are presented in sections 3.1 and 3.2 respectively.

3.1. Comparison of Sets of Blocking Strategies

The following five criteria were used for blocking files into sets of linked pairs used in the discrimination stage. The set of five criteria were developed by comparing a large number of criteria. If the upper bound on the overall rate of erroneous matches during the blocking stage is set at 65 percent, then this set of five gave the largest overall reduction in erroneous nonmatches (see Winkler, 1984a).

BLOCKING CRITERIA

1. 3 digits ZIP, 4 characters NAME
2. 5 digits ZIP, 6 characters STREET
3. 10 digits TELEPHONE
4. Word length sort NAME field, then use 1. *
5. 10 characters NAME

* This criterion also has a deletion stage which prevents matching on commonly occurring words such as 'OIL,' 'FUEL,' 'CORP,' and 'DISTRIBUTOR.'

3.1.1. Type I and II Error Rates by Individual Blocking Criteria

Table 2 presents counts and rates of matches, erroneous matches, and erroneous nonmatches for each of the five matching criteria given above.

As we can see, no single criterion provides a significant reduction in the rate of erroneous nonmatches. The best is criterion 4 (wordlength sort) which leaves 702 (23 percent) duplicates unlinked. The reason criteron 4 works best is that the NAME field does not have subfields (generally words) that are in fixed order or in fixed locations. Consequently, criterion 4 links NAME fields from headquarters and duplicates having the following form:

John K Smith
Smith J K Co

Criterion 3 (TELEPHONE) provides the lowest rate 8.7 percent (186/(186+1952)) of erroneous matches and the second best rate 34.7 percent (1057/3050) of erroneous nonmatches. Criterion 5 (10 characters of the NAME) provides both the worst rate of erroneous matches, 58.6 percent (1259/1259+889)), and the worst rate of erroneous nonmatches, 63.3 percent (1932/3050).

Table 2: Rates of Matches, Erroneous Matches, and Erroneous Nonmatches by Blocking Criteria

Criterion	Link with Correct Parent 1/	Link with Wrong Parent	Not Linked 2/	Actual Number of Matches
1	1460 (66.8)	727	1387 (45.5)	3050
2	1894 (82.5)	401	1073 (35.2)	3050
3	1952 (91.3)	186	1057 (34.7)	3050
4	2261 (80.3)	555	702 (23.0)	3050
5	763 (14.4)	4534	1902 (62.4)	3050

1/ Type II error rates are in parentheses.
2/ Type I error rates are in parentheses.

3.1.2. Comparison of Sets of Criteria

In comparing subsets of the five blocking criteria, the primary concern is in reducing the number of erroneous nonmatches. The number of matches and erroneous matches in the set of pairs created in the blocking stage is dealt with primarily during the discrimination stage.

The comparison takes the form of considering the incremental reduction in the number of erroneous nonmatches as each individual criteria is added. Although criteria 3 and 4 perform best on the empirical data base, they are considered later than criteria 1 and 2.

Criteria 1 and 2 are applicable to all EIA files because all of them have identified NAME and ADDRESS fields. As many non-EIA source lists used in updating do not contain telephone numbers, criterion 3 is not applicable to them. As a number of EIA lists have consistently formatted NAME fields, criterion 4 will yield little, if any, incremental reductions in the number of erroneous matches during the blocking stage.

Table 3: Incremental Decrease in Erroneous Nonmatches and Incremental Increase in Matches and Erroneous Matches by Sets of Blocking Criteria

Set of Criteria Used	Rate of Erroneous Nonmatches	Erroneous Nonmatches/ Incremental Decrease	Matches/ Incremental Increase	Erroneous Matches/ Incremental Increase
1	45.5	1387/ NA	1460/ NA	727/ NA
1,2	15.1	460/927	2495/1035	1109/ 289
1,2,3	3.7	112/348	2908/ 413	1233/ 124
1,2,3,4	1.3	39/ 73	2991/ 83	1494/ 261
1,2,3,4,5	0.7	22/ 17	3007/ 16	5857/4363

NA- not applicable.

3.1.3. The Preferred Set of Blocking Criteria

The preferred set of blocking criteria are criteria 1, 2, 3, and 4. Criterion 5 (10 characters of the NAME) was considered because it yielded the greatest reduction in erroneous nonmatches of any fifth blocking criteria while keeping the overall percentage of erroneous matches below 65 percent.

Criterion 5, however, is not suitable for inclusion because it incrementally adds 16 matches and 4363 erroneous matches while reducing the number of erroneous nonmatches from 39 to 22. As the discrimination stage (section 3.2) delineates matches and nonmatches with an error rate of 3 percent and 99.6 (4363/4379) of the incrementally-added pairs are false, inclusion of criterion 5 would yield an overall increase in the number of erroneous nonmatches.

Blocking 3050 duplicates with 54,850 parents using the preferred set of blocking criteria yielded 4485 pairs (2991 matches and 1494 nonmatches) for consideration during the discrimination stage.

It is important to note that the 39 matches not identified during the blocking stage are never again considered. Erroneous matches created during the blocking stage are considered during the discrimination stage and still can be correctly designated. These reasons led to our emphasis on minimization of Type I errors during the blocking stage prior to minimization of Type I and II errors during the blocking stage.

3.2. Discrimination

The discrimination stage was divided into two parts: (1) a part in which 2240 pairs were designated as matches using an ad hoc decision rule and (2) a discrimination stage in which the remaining 2245 pairs were designated as either matches, erroneous matches, or candidates for manual review.

The ad hoc decision rule generally consisted of designating those pairs as matches that had been connected by two or more blocking criteria. The exceptions were records connected by 1 and 4, only (NAME and WL-NAME), and 2 and 3, only (STREET and TELEPHONE). Slightly more than 98 percent of the 2240 records designated as matches were actually matches.

Prior to use in the information-theoretic discrimination procedure, the 2245 remaining pairs were further divided into four mutually exclusive classes using the preferred blocking criteria (section 3.1.3):

Class 1 (1021 records): Linked by 1, only, and by 1 and 4, only.
Class 2 (624 records): Linked by 2, only, and by 2 and 3, only.
Class 3 (256 records): Linked by 3, only.
Class 4 (344 records): Linked by 4, only.

3.2.1. Overall Results

Table 4 presents a summary of results obtained during the discrimination stage. It shows that 2148 (96 percent) of 2245 records are classified as matches or nonmatches and that only 3 percent (68/2148) of the classified records are misclassified. Results are based on using the entire data set for calibration (i.e., obtaining cutoff weights) and evaluation. Variance results (section 3.2.6) based on 25 different samples used for calibration yield cutoff weights and error rates that are consistent with results in Table 4.

Two observations are that the cutoff weights vary substantially across classes and that 100 percent of the records in classes 2 and 4 can be classified. The varying cutoff weights indicate that cutoff weights may vary with different types of address lists. Thus, new calibration information may be needed for each new file encounted. Calibration information is based on knowing the actual truth and falsehood of matches within a representative set of blocked pairs.

Table 4: Results from Using a Modified Information-Theoretic Model for Delineating Matches and Erroneous Matches (3 Percent Overall Misclassification Rate)

Class	Cutoff Weights		Misclassed as		Total Classed as		Total Classed	Total Records
	LOWER	UPPER	Non-Match	Match	Non-Match	Match		
1	4.5	7.5	28	8	692	274	966	1021
2	2.5	2.5	5	3	379	245	624	624
3	-0.5	4.5	5	6	104	110	214	256
4	8.5	8.5	9	4	266	78	344	344
Totals			47	21	1441	707	2148	2245

The largest group of misclassified records are those erroneous matches that have the same address and phone number as the headquarters' records. For example:

(a) Apex Oil 222 Columbia St NE Salem
OR 97303 503/588-0455
Jones Co 222 Columbia St N E Salem
OR 97303 503/588-0455
(b) A A Oil Main St Smallsville TX
77103 713/643-2121
Smith J K Co Main St Smallsville TX
77103 713/643-2121

Example (a) represents two different companies located in the same office building. Example (b) represents two different fuel oil dealers, one of which has gone out-of-business.

Misclassified matches (erroneous nonmatches) generally had typographical differences or missing data in a number of subfields, as in the examples below:

```
(c)  Smith Oil          W 31st St N Church St
      Hardsburg        PA 18207     713/643-2121
     Smith J K          N Church St
      Hardsburg        PA 18207     missing
(d)  Mcneely R          3312-14 Harris Ave
      MPLS             MN 55246     612/929-6677
     R Mcden Neely      3312 Harris Ave
      St Louis Par     MN 55246     612/929-6677
```

Example (c) has a minor variation in the NAME field, a major variation in the STREET field, and a missing TELEPHONE field. Example (d) has major variations in the NAME field and CITY fields and a minor variation in the STREET field.

3.2.2. Improvement Due to New Spelling Standardization

The improvement due to the new spelling standardization was quite minor as the results in Figures 1 and 2 show. Figures 1 and 2 represent plots of the numbers of matches and nonmatches against total weight using the early and new spelling standardizations, respectively.

The results are only shown for Class 2 (section 3.2 and section 3.1.3) because records blocked using STREET ADDRESS only or STREET ADDRESS and TELEPHONE only are intuitively among the most difficult to work with (see examples in section 3.2.1). Both figures will be compared with other figures corresponding to Class 2 that appear in sections 3.2.2, 3.2.3, and 3.2.4. Although characteristic results for other classes will be mentioned, no graphs will be presented for them.

Figures 1 and 2 show the classic patterns in matches and nonmatches (Newcombe et al., 1959; Newcombe et al., 1983; Rogot et al., 1983). In both figures, the curves of matches almost entirely overlap with the curves of nonmatches. As the distinguishing power of the weighting scheme improves, the curves move apart.

FIGURE 1: Total Weight Versus Counts of Matches and Nonmatches Prior to New Spelling Standardization Prior to Identification of Subfields

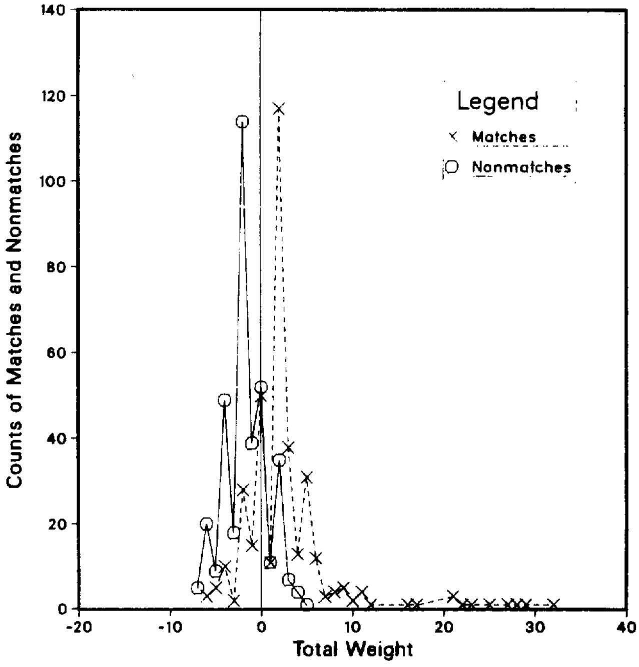
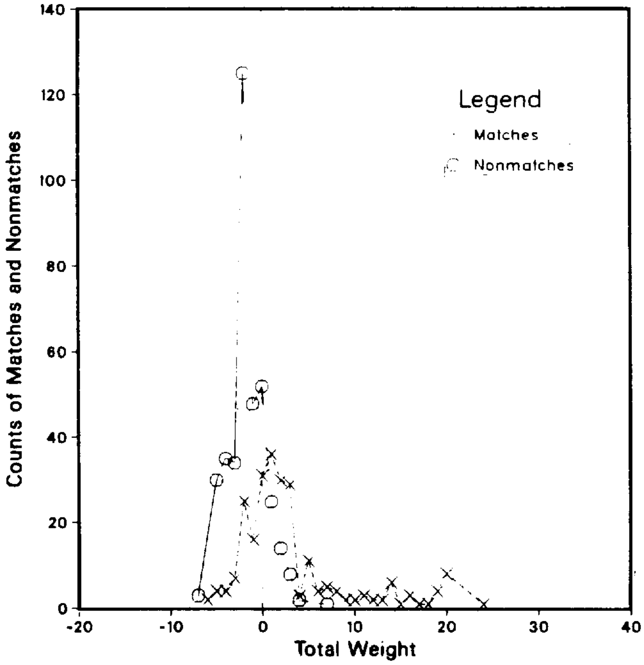


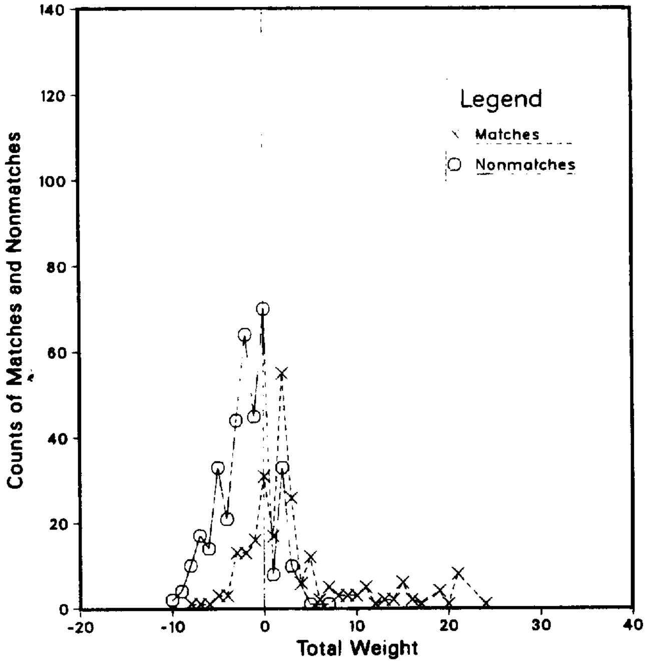
FIGURE 2: Total Weight Versus Counts of Matches and Nonmatches After New Spelling Standardization Prior to Identification of Subfields



3.2.3. Improvement Due to Address Subfield Identification

Figure 3 is a plot of the numbers of matches

FIGURE 3: Total Weight Versus Counts of Matches and Nonmatches After New Spelling Standardization Address Subfield Identification



and nonmatches against total weight when the new spelling standardization and address subfield identification (section 2.1.3) is used. Comparison with Figure 2 shows that the subfield identification yields a moderate improvement (i.e., the curves of matches and nonmatches overlap less.)

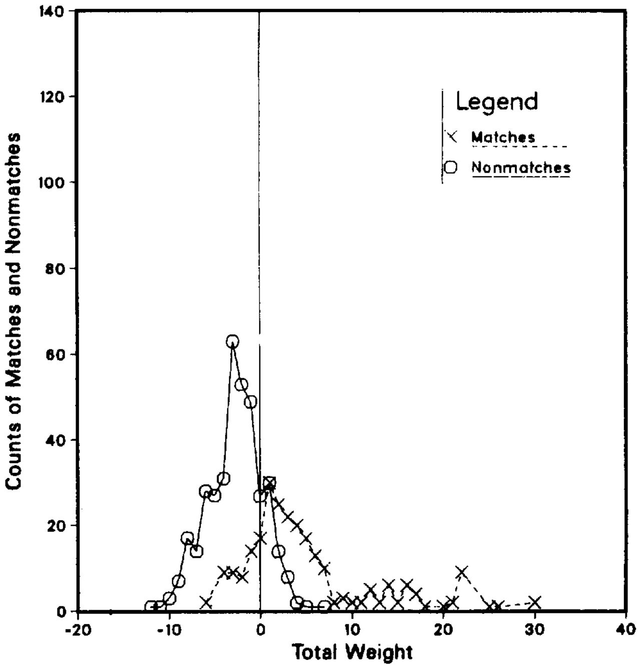
Although not shown in this paper, examination of similar sets of plots for other classes, particularly those blocked using the NAME field, show less improvement when additional weights obtained using the ADDRESS subfields are used.

3.2.4. Improvement Due to Name Subfield Identification

Figure 4 is a plot of the numbers of matches and nonmatches against total weight when the new spelling standardization and name and address subfield identification are used (see section 2.1.3 for a list of the subfields). Comparison with Figure 3 shows that the NAME subfield identification yields little, if any, improvement.

Although not shown in this paper, examination of similar sets of plots for other classes, particularly those blocked using the NAME field, show greater improvement when additional weights obtained using the NAME subfields are used.

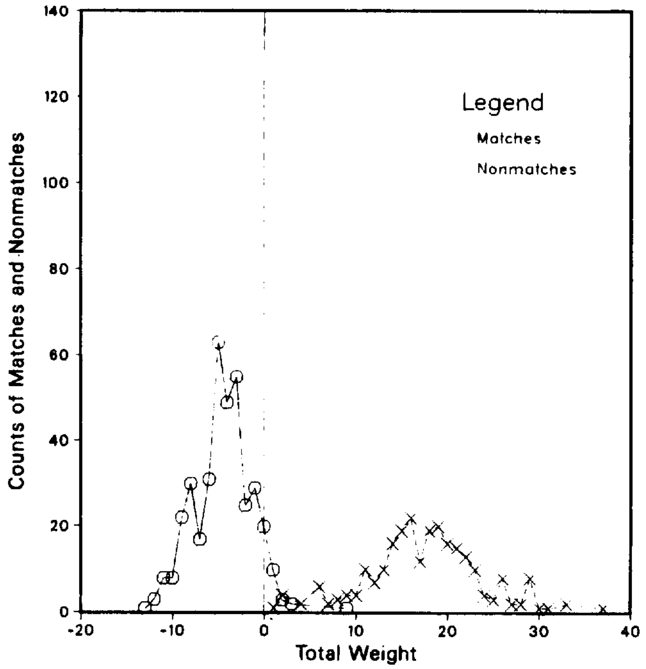
FIGURE 4: Total Weight Versus Counts of Matches and Nonmatches After New Spelling Standardization Name and Address Subfield Identification



3.2.5. Improvement Due to Conditioning

Figure 5 is a plot of the numbers of matches and nonmatches against total weight when a special conditioning (see section 2.2 and section 3.2.8) procedure in addition to the new spelling standardization and name and address subfield identification is used. Comparison with Figure 4 shows that the conditioning yields a substantial improvement in Class 2. Other classes (not shown) show slight improvements.

FIGURE 5: Total Weight Versus Counts of Matches and Nonmatches Name and Address Subfield Identification Conditioning



Comparison of Figure 5 with Figures 1 or 2 show the significant improvements obtained using the modified information-theoretic model that includes all enhancements.

Table 5 shows the results from using the basic information-theoretic model that are comparable to the results in Table 4. The only difference is that a modified information-theoretic procedure is used in obtaining Table 4 results. Overall comparison shows that the modified information-theoretic procedure performs better than the basic information-theoretic procedure.

Specifically, comparison of the two tables shows that the total number of records classified rises from 1526 (out of 2245) to 2148 while the overall misclassification rate falls from 5 percent to 3 percent.

Comparison of Tables 4 and 5 also shows that the main difference in the modified and basic procedures is that the modified procedure allows classification of all 624 records in class 2 while the basic procedure allows classification of only 215.

Table 5: Results from Using an Information-Theoretic Model for Delineating Matches and Erroneous Matches (5 Percent Overall Misclassification Rate)

Class	Cutoff Weights		Misclassed as		Total Classed as		Total Classed	Total Records
	LOWER	UPPER	Non-Match	Match	Non-Match	Match		
1	0.5	6.5	39	14	674	264	938	1021
2	-4.5	3.5	2	4	100	115	215	624
3	-4.5	6.5	2	1	55	42	97	256
4	2.5	11.5	11	2	254	46	300	344
Totals			54	21	1055	471	1526	2245

3.2.6. Variances

Tables 6, 7, and 8 present estimates and their coefficients of variation obtained using 25 calibration samples and Rubin's multiple imputation technique. For each calibration sample, the sample sizes in Classes 1, 2, 3, and 4 were 240, 200, 120, and 160, respectively. Cutoff weights and misclassification rates were obtained for each sample. Estimates are the average cutoff weights and average misclassification rates over 25 replications (samples). Variances of the estimates are over 25 replications.

Overall, the results indicate that the estimated cutoff weights and misclassification rates vary significantly from calibration sample to calibration sample. The variances are functions of both the sample sizes on each replication and the number of replications. When the number of replications was held at 25 and the sample sizes decreased to 120, 100, 80, and 90 for the four classes, estimated coefficients of variation over 25 replications were approximately 30 percent higher on the average for misclassified matches and about the same for misclassified nonmatches.

The fact that the coefficients of variation decrease substantially as sample sizes increase indicates that calibration samples should be as large as possible. As the total number of records considered in these analyses was quite small, taking substantially larger samples was not practicable.

Examination of Table 6 shows that the estimated coefficients of variation associated with the cutoff weights using the modified information-theoretic procedure range from 15.3 percent to 99.5 percent; and from 14.3 percent to 115.4 percent with the basic information-theoretic procedure. The cutoff weights are consistent with the cutoff weights given in Table 4 and Table 5. Results in Tables 4 and 5 were obtained using the entire data set instead of samples.

Examination of Tables 7 and 8 show that the misclassification and nonclassification rates can vary significantly. Coefficients of variation of the estimated misclassification rates for the modified information-theoretic procedure vary from 33.2 to 109.9; for the basic procedure from 33.8 to 112.9.

Table 6: Estimated Cutoff Weights and Their Variances
25 Replications, With and Without Conditioning

Class	Status 1/	Estimated Cutoff Weights		Variance of Estimated Cutoff Weights		CVs of Estimated Cutoff Weights	
		LOWER	UPPER	LOWER	UPPER	LOWER	UPPER
1	C	2.66	7.72	7.02	2.05	99.5	18.5
2	C	1.44	1.44	0.62	0.62	54.9	54.9
3	C	-3.39	5.82	8.74	2.08	87.2	24.8
4	C	6.89	1.92	1.11	7.57	15.3	23.1
1	WC	-1.92	8.05	4.90	1.50	115.4	15.2
2	WC	-5.04	4.56	0.52	1.41	14.3	26.1
3	WC	-6.38	6.82	1.46	1.66	18.9	18.9
4	WC	1.71	12.13	3.11	7.56	102.9	22.7

1/ C-Conditioning, WC-Without Conditioning.

Table 7: Estimated Counts and Rates of Misclassification
and Nonclassification
25 Replications, With and Without Conditioning

Class	Status 1/	Total Records	Misclassed as		Not Classed	Correctly Classed as		Proportion Misclassed as	
			Match	Non-Match		Match	Non-Match	Match	Non-Match
1	C	1021	10.4	27.4	75.2	260.7	647.2	.038	.041
2	C	624	9.7	3.0	0.0	244.0	367.3	.038	.008
3	C	256	3.0	3.5	94.2	85.2	70.0	.034	.048
4	C	344	1.4	10.2	23.5	54.3	254.6	.026	.039
Total		2245	24.5	44.1	192.9	644.2	1338.1	.037	.032
1	WC	1021	8.9	26.2	145.4	237.1	603.3	.036	.042
2	WC	624	3.8	3.9	450.6	89.4	76.3	.040	.048
3	WC	256	1.6	2.3	178.8	38.1	35.1	.041	.062
4	WC	344	1.3	9.6	57.7	38.8	236.6	.032	039
Total		2245	15.6	42.0	832.5	403.4	951.3	.037	.042

1/ C-Conditioning, WC-Without Conditioning.

Comparison of the modified and basic weighting procedures shows that the modified procedure is able to classify accurately significantly more records, particularly in classes 2 and 4, than the basic procedure. The results are consistent with those presented in Tables 4 and 5.

Results obtained using Efron's bootstrap imputation with 25, 100, 200, and 500 replications are consistent with the results in Tables 6, 7 and 8.

3.2.7. Overall Rate of Duplication

The overall rate of duplication (section 2.3.2) is 0.19 percent (100*102/(54850+102)) where the number of headquarters records is 54,850 and an estimated upper bound on the number of erroneous nonmatches is 102).

The estimated upper bound, 102, on the number of erroneous nonmatches is the number of matches

Table 8: Coefficients of Variation of Estimated Counts of Misclassification and Nonclassification 1/

25 Replications With and Without Conditioning

Class	Status 2/	Total Records	Misclassed as		Not Classed
			Match	Non-Match	
1	C	1021	69.5	47.4	54.7
2	C	624	64.6	81.1	0.0
3	C	256	96.6	84.1	40.9
4	C	344	109.9	33.2	60.8
1	WC	1021	62.3	42.3	34.0
2	WC	624	112.9	96.2	9.0
3	WC	256	106.9	65.5	8.1
4	WC	344	99.6	33.8	34.3

1/ Units are percentages.
2/ C-Conditioning, WC-Without Conditioning.

that are unblocked plus an upper bound on the the number that are erroneously classified as nonmatches during the discrimination stage. Thirty-nine records (section 3.1.2) are unblocked using the preferred set of blocking criteria.

The estimated upper bound consists of the sum of the estimated upper bounds on the numbers of automatically erroneously matched records in classes 1-4 and an estimate of the number of matches that are misclassified during manual review. The upper bounds at the 95 percent confidence level in classes 1-4 (using the estimates in Tables 7 and 8) are 24.9, 22.2, 8.9, and 4.5, respectively.

We assume that two percent of the estimated 124.3 matches in the estimated set of 192.9 records (see Tables 7 and 8) will be misclassed during manual review. This yields that 2.5 matches will be misclassed as nonmatches.

Thus, the upper bound is 102 (=39+24.9+22.2+8.9+4.5+2.5).

3.2.8. The Independence Assumption

Independence of comparisons does not hold. This is shown by the significant variation of the lower and upper cutoff weights across Classes 1 thru 4 in Tables 4, 5 and 6. If the comparisons were independent, then individual weights and cutoffs for the total weights would be reasonably consistent across classes. Individual weights (not shown) vary more than the cutoff weights across classes.

Independence of interactions within classes is illustrated by Tables 9 and 10. They show the two-way independence of the interactions of some of the subfields given in section 2.1.3 for subfields that are generally not connected and somewhat connected with blocking characteristics respectively. The variables used in the comparisons were defined in sections 2.1.3 and 2.2.2, respectively.

Table 9: Independence of Two-Way Interactions for Selected Subfields that are Generally Not Connected with Blocking Characteristics, By Class 1/

Class	K11/H	K22/H	K11/SN	K22/SN
1	yes	yes	no 2/	no 2/
2	NA	NA	yes	yes
3	no 4/	no 3/	no 2/	yes
4	yes	yes	yes	yes

NA- not applicable because one of two variables is basically the same as a blocking characteristic due to small sample size.

1/ Kii is the comparison of KEYWORDi with KEYWORDi, for i=1, 2; H is comparison of HOUSE NUMBER with HOUSE NUMBER; and SN is the comparison of STREET NAME with STREET NAME.

2/ Independent when H is included in a 3-way contingency table analysis.

3/ Independent when K11 is included.

4/ Independent when K22 is included.

Table 10: Independence of Two-Way Interactions for Selected Subfields that are Somewhat Connected with Blocking Characteristics, By Class

Class	W1/S1	W1/S2	W1/S3	W2/S1	W2/S2	W2/S3	W3/S1	W3/S2	W3/S3
1	yes	yes	yes	yes	yes	yes	yes	yes	yes
2	NA	yes	yes	NA	yes	yes	NA	yes	yes
3	no 1/	no 2/	no 3/	no 4/	no 2/	no 1/	no 5/	no 2/	no 1/
4	NA	NA	NA	yes	yes	no 1/	no 1/	no 2/	no 1/
A 6/	no	no	yes	yes	yes	yes	yes	yes	yes

NA- not applicable because one of two variables is used as a blocking characteristic.

1/ Independent when S2 is included in a 3-way contingency table analysis.

2/ Independent when S1 is included.

3/ Independent when W2 is included.

4/ Independent when W3 is included.

5/ Independent when S3 is included.

6/ Aggregate of Classes 1-4.

The Fellegi-Sunter model (1969, pp. 1189-1190) does not require full independence of interactions. It only requires that interactions be conditionally independent.

In over half the entries in Tables 9 and 10, the two-way interactions are independent unconditionally at the 95 percent confidence level and the hierarchical principle (Bishop, Fienberg, and Holland, 1975) assures that all such two-way interactions are always conditionally independent. In all cases in which two-way interactions are not unconditionally independent, a third variable was found so that the two-way interactions were independent at the 95 percent confidence level given the third variable.

It is important to note two points. First, some of the interaction of variables (not presented in the tables) such as H and S1 or W1 and K11 are often not independent unconditionally and it seems likely that they will generally not be independent conditionally. Second, building a precise model, by mutually exclusive class, in which only the minimal set of variables necessary for effective discrimination is included, and which precisely models the conditional relationships, is likely to be difficult and heavily dependent on the empirical data base used.

What we attempted to do in our approach was to find a superset of the minimal set of variables needed for effective discrimination; apply them all in creating the weights for each class; perform minimal modification in the basic procedures for creating the weights; and show that the failure of the independence assumption is not too crucial.

4. CONCLUSIONS AND FUTURE WORK

This section contains a brief summary of the results of this paper, a discussion of how the results relate to previous applied work and existing theory, and a set of problems for future research.

4.1. Summary

The results of this paper imply that the keys to delineating matches and nonmatches accurately are: (1) good spelling standardization and (2) accurate identification of corresponding subfields. They also imply that the independence assumption, required by the information-theoretic model of Fellegi and Sunter (1969), is not critical in practical applications of the type performed in this paper.

A key advantage of the Fellegi-Sunter approach is that it lends itself to incremental improvements, as knowledge of both file properties and data manipulation techniques (via software) increase.

4.2. Further Discussion of Results

4.2.1. Independent Application of Multiple Blocking Criteria

Newcombe et al. (1962, pp. 563-564) provide an example of applying multiple blocking criteria independently. They blocked first on surname and then on maiden name in files of individuals used for epidemiological research. In their study of a special sample of 3560 matches (linkages in their terminology), 98.4 percent (3504) were obtained using SOUNDEX coding of surname and an additional 1.4 percent (to a total 99.8 percent) were obtained using SOUNDEX coding of maiden surname. The increase in the total number of pairs considered for review when the second blocking criterion was used was 100 percent.

The results of section 3.1 show that, within the set of criteria considered, no single blocking criterion can yield a subset of pairs containing 80 percent of matches and no two can yield subsets containing 90 percent. The work of Winkler (1984a,b) provides a considerably more exhaustive study of blocking criteria and shows how the set of criteria used in this study work reasonably well on two additional sets of files.

Kelley (1985) provides a theoretical foundation for the simultaneous consideration of several subfields which is consistent with the Fellegi-Sunter model. In hypothetical examples, he shows how best to apply simultaneously first name, surname, and sex as blocking criteria. Section 3.1 results show that criterion 5, 10 characters of the NAME, does not perform well (62.4 percent of matches are not blocked and only 14.4 percent of the blocked pairs are matches) while criterion 1, 3 digits of the ZIP and 4 characters of the NAME, performs considerably better (45.5 percent of matches unblocked and 66.8 percent of the blocked pairs are matches). Thus, our results serve as partial corroboration of Kelley's results.

It seems likely that independent application of multiple blocking criteria such as done in this paper will be necessary to identify matches in other files of businesses. This is primarily due to lack of identifiers such as surnames.

4.2.2. Spelling Standardization

The comparison of Figures 1 and 2 in section 3.2.2 showed that improved spelling standardization of commonly occurring words did not yield any dramatic improvement in the ability to distinguish matches and nonmatches. Results for other classes (not shown) were similar. The results, however, may not be representative because the files had already been standardized using a somewhat more elementary set of tables. It is possible that improvements could be more dramatic when results using totally unstandardized files are compared with results using well standardized files.

Additionally, consistent spelling of commonly occurring words can allow their identification; thus, making it easier to identify other subfields having greater distinguishing power.

4.2.3. Subfield Identification

Section 3.2 results (particularly Figures 2-4) showed improvements in the Fellegi-Sunter weighting procedure's ability to delineate accurately matches and nonmatches and reduce the size of the manual review region. The improvements were due to the identification of subfields in the NAME and STREET fields using ZIPSTAN and KEYWORD software, respectively.

The improvements using ZIPSTAN in classes 1 and 4 (not shown) were quite substantial. They were, however, not as dramatic as the improvements in classes 2 and 3 when conditioning procedures were used.

The results basically show us that it may be possible to delineate and compare subfields (particularly within the NAME field) that yield greater distinguishing power. In particular, if such comparable subfields are distinguished, then string comparator metrics (see e.g., Winkler, 1985) which allow assignment of weights of partial agreement between strings (rather than just 1-agree and 0-disagree) could be used to deal with subfields containing minor keypunch/transcription errors.

4.2.4. Independence, Conditioning, and Steepest Ascent

The results in section 3.2 (particularly subsections 3.2.1 and 3.2.8) show that the comparisons of characteristics of various subfields are generally not independent. Fellegi and Sunter (1969, p. 1191) indicate that their weighting scheme may work well in practice even when the independence assumption is not met.

In an early analysis (not shown), weights were computed uniformly over all pairs within the set of blocked pairs, rather than separately in the four subclasses. Analyses similar to those in section 3.2 (particularly, using figures like Figures 1-5) showed that weights computed uniformly did not have as much distinguishing power. In particular, the curves of nonmatches and matches never moved as far apart as the curves moved apart in Figure 5. Results (not shown) for other classes used in this paper were quite similar to those in Figures 1-5.

We can conclude that, at least in our example, dependence of comparisons leads to less discriminating power. We should note, however, that a large number of comparisons were performed, some of which are likely not to be independent conditionally. It may be possible that subsets of the comparisons (they are likely to vary significantly from class to class) may be created in which the comparisons are conditionally independent. For such subsets, however, it is not clear whether the overall discriminating power will increase.

It is important to note that, for those procedures in which only one blocking criterion is used (such as blocking on SOUNDEX abbreviation of surname in files of individuals), it may be possible to compute weights uniformly over the entire set of blocked pairs. The four classes which we considered were created using the preferred set of four blocking criteria. Thus, our weight creation scheme is conditional on the set of blocking criteria.

The conditioning arguments in this paper consisted primarily of the subdivision of the set of blocked pairs into four classes based on the four blocking criteria and steepest ascent methods of weight variation. Both procedures are cumbersome to apply, the second particularly so. It may be possible to produce some algorithm for conditioning or some other method which allows a systematic approach to conditioning. Bishop, Fienberg, and Holland (1975, Chapter 11) provide a useful discussion of the difficulties with some of the measures of association that have been developed.

4.2.5. Legitimate Representation Differences and Keypunch/Transcription Error

Fellegi and Sunter (1969, pp. 1193-1194) provided a specific model which incorporates error rates associated with legitimate representation differences of the same entity (see e.g., the name variations in section 2.1.3) and/or keypunch/transcription error. Their results (see also Coulter, 1977; Kirkendall, 1985) show that, in the presence of such errors, agreement weights remain approximately the same as agreement weights in the absence of such errors, while disagreement weights (which are generally negative) increase. The results have substantial intuitive appeal.

Review of figures like Figures 1-5 for classes 1, 3, and 4 (not shown) and examination of pairs that are either misclassified or not classified in all 4 classes indicate that keypunch error plays a substantially greater role in classes 1 and 3 than in classes 2 and 4. The results are consistent with Table 4 results in which all records in classes 2 and 4 are classified (none held for manual review) while a moderate number of records in classes 1 and 3 (55 of 1021 and 42 of 256, respectively) are held for manual review.

A partial explanation of the differences is that classes 1 and 3 contain a moderate number of pairs of records having substantial variations in the NAME and/or STREET fields while classes 2 and 4 do not. In class 1, many keypunch errors occur after the first four characters of the NAME. Being able to block on TELEPHONE (class 3), allows significant reduction in the number of erroneous nonmatched because so many keypunch/transcriptions can occur in the NAME and STREET fields (see also Winkler, 1984a).

An additional series of steepest ascent variations were performed in classes 1 and 3. In all cases, the distinguishing power remained constant or became slightly worse. In some cases, graphs such as given by Figure 5 contained curves of nonmatches and matches for which the humps moved apart but for which the manual review region remained constant or increased in height. Thus, it seems unlikely that more conditioning in the form presented in this paper will improve procedures. Rather, it seems likely that improvements will depend more on better identification and comparison of subfields.

4.2.6. Adaptability of the Fellegi-Sunter Procedures

Newcombe et al. (1959, 1962) first showed that the basic weighting procedure as presented in Fellegi and Sunter (1969) could be improved by adapting it to make use of additional comparative information. Figures 1-5 in this paper illustrate successive improvements which can be obtained using spelling standardization, additional comparisons of subfields of the NAME and STREET fields, and conditioning arguments.

Further improvements seem likely. They can be obtained using techniques that are already available. For instance, Statistics Canada (1982) has developed sophisticated methods of delineating subfields within the NAME field for use on the Canadian Business Register. Identifying subfields as Statistics Canada has done could allow a number of less sophisticated comparisons (such as first four characters and next six characters of the NAME field) to be dropped and discriminating power to increase. ZIPSTAN software (U.S. Dept. of Commerce, 1978b) yielded subfields of the STREET field which provided increased discriminating power.

Use of frequency counts of the occurrence of substrings (e.g., Zabrinsky occurs less often and has more distinguishing power than Smith) could be incorporated in matching lists of businesses. Presently, such matching using frequency counts is applied to lists of individuals (e.g., U.S. Dept. of Agriculture, 1979; U.S. Dept. of Commerce, 1978a). The theoretical justification for procedures using frequency-based matching are explicitly described by Fellegi and Sunter (1969, pp. 1193-1194).

Use of frequency-based matching involves use of lookup tables for obtaining weights associated with individual comparisons. Such lookups can be performed efficiently using K-D trees (Friedman, Bentley, and Finkel, 1977). EIA presently uses K-D trees for search of lookup tables during spelling standardization.

String comparator metrics (see e.g., Winkler, 1985) allowing comparison of strings containing minor keypunch errors could also be used in adapting the weighting procedures.

4.3. Problems Remaining

Effective evaluation of the efficacy of various matching procedures requires having a representative data base in which matches and nonmatches have been identified and tracked. Such data bases can be created during list updating projects and are necessary if incremental improvements in procedures are to be made (see e.g., Coulter and Mergerson, 1977; Smith et al., 1983).

Effective evaluation also requires having common terminology and measures that allow rough comparison of results obtained using significantly different data bases and/or methodologies. The results of this paper and others (see e.g., Newcombe et al., 1983; Rogot et al., 1983) suggest a number of avenues for future research that can be incorporated into existing procedures in a straightforward manner.

4.3.1. Error Rates

Various authors (see e.g., Newcombe et al., 1983; Rogot et al., 1983) have presented the rates of erroneous matches and nonmatches during the discrimination stage but generally do not mention the rates of erroneous nonmatches that remain unlinked during the blocking stage. As the Fellegi-Sunter model explicitly provides measures of the Type I and Type II error rates, it seems natural to extend investigation of such rates to both blocking and discrimination stages.

The results of this paper imply that error rates occurring during both stages must be investigated simultaneously. For instance, during early stages of the work at EIA no effective methods existed for accurately delineating matches and nonmatches during the discrimination stage. As more effective methods of delineating matches and nonmatches during the discrimination stage are developed, it seems likely that additional blocking criteria (such as criterion 5 in section 3.1) may be adopted without increasing the rate of erroneous nonmatches.

Other measures, such as the overall rate of duplication given in this paper (see also Winkler, 1984a,b), may provide additional insight into how well a specific application is performed and provide additional information comparable with other applications.

Type I error rates based on samples (see e.g., Winkler, 1984a,b) have been shown to yield coefficients of variations of approximately 100 percent even with samples as large as 1800. Although Fellegi and Sunter (1969) indicate that estimating error rates based on samples yields high variances, they did not provide an example showing the magnitude of the problem. There may be better methods for obtaining such error rates and their variances when samples are used.

4.3.2. General Applicability of Linkage Mechanisms

Winkler (1984a,b) showed that the preferred set of blocking criteria are reasonably applicable to two other data bases having different characteristics from the empirical data base that was used for analyses in this paper. In those papers, however, blocking criteria were investigated independent of the discrimination stage.

Investigations of the efficacy of different blocking strategies when both blocking and discrimination stages are considered simultaneously are necessary. The investigations should be performed on files with significantly different characteristics.

For instance, is the use of an abbreviation method such as SOUNDEX (e.g., Bourne and Ford, 1961) or NYSIIS (e.g., Lynch and Arends, 1977) abbreviation of SURNAME the only way to block files of individuals? If so, why are such blocking methods effective in reducing the rate of erroneous nonmatches? What methods were investigated and why were they rejected? Should files of individuals be blocked several different ways using significantly different blocking criteria?

4.3.3. String Comparators

If corresponding strings such as SURNAME are identified, then it is possible to define distance or weighting functions that compare nonidentical strings. Such weighting functions (see e.g. Winkler, 1985, pp. 12-16) can be derived using abbreviation methods such as SOUNDEX (e.g., Bourne and Ford, 1961), using the Damerau-Levenstein metric (e.g., Hall and Dowling, 1980, pp. 388-390), or the string comparator of Jaro (e.g., U.S. Dept of Commerce, 1978a, pp. 83-101).

Each of the methods is intended to allow comparison of strings in which minor typographical differences occur. What are the relative merits of different weighting functions? Are there any better algorithms for string comparison?

4.3.4. Tracking True and False Matches

In linking pairs of records in lists of businesses, many erroneous matches will have similar NAMEs and/or STREET ADDRESSes. Matches may have different NAMEs and/or STREET ADDRESSes (e.g., subsidiaries, successors). Delineation of most such matches and nonmatches can require manual followup which is both time-consuming and expensive.

If matches and nonmatches are tracked properly and the weighting methodology for delineating matches and nonmatches is reasonably effective, then many nonmatches that have similar NAMES and STREET ADDRESSes to previous nonmatches or matches having different NAMES and/or STREET ADDRESSes from their true parents will not require manual review.

To determine if it is cost-effective to track matches and nonmatches, research is needed to show:

1. how classes of matches and nonmatches of records linked using various blocking criteria should be set up to allow tracking;
2. how effective weighting schemes should be determined that allow maximum use of the tracking system;
3. how pairs newly linked during an update should be compared within equivalence classes and across equivalence (a record can be linked truly once and falsely many times);
4. how updating using the results of 1, 2, and 3 should be performed; and
5. how the results of the updating should be evaluated.

ACKNOWLEDGMENT

The author wishes to thank Dr. Yvonne M. Bishop for a number of helpful comments and conversations.

REFERENCES

Bishop, Y. M. M., Fienberg, S. E., and Holland, P. W. (1975), Discrete Multivariate Analysis, MIT Press, Cambridge, MA.

Bourne, C. P., and Ford, D. F. (1961), "A Study of Methods for Systematically Abbreviating English Words and Names," J. ACM, 8, 538-552.

Coulter, R.W. (1977), "An Application of a Theory for Record Linkage," Technical Report from the Statistical Reporting Service of the U.S. Dept. of Agriculture.

Coulter, R.W. and Mergerson, J.W. (1977), "An Application of a Record Linkage Theory in Constructing a List Sampling Frame," Technical Report from the Statistical Reporting Service of the U.S. Dept. of Agriculture.

Cochran, W.G. and Cox, G.M. (1957) Experimental Designs, J. Wiley and Sons, New York.

Efron, B. (1979), "Bootstrap Methods: Another Look at the Jackknife," Ann. Stat., 7, 1-26.

Efron, B. and Gong, G. (1983), "A Leisurely Look at the Bootstrap, the Jackknife, and Cross-Validation," The American Statistician, 37, 36-48.

Fellegi, I. P., and Sunter, A. B. (1969), "A Theory for Record Linkage," JASA, 40, 1183-1210.

Friedman, J. H., Bentley, J. L., and Finkel, R. A. (1977), "An Algorithm for Finding Best Matches in Logarithmic Expected Time," ACM Trans. Math. Software, 3, 209-226.

Hall, P. A. V. and Dowling, G. R. (1980), "Approximate String Matching," Computing Surveys, 12, 381-402.

Herzog, T. and Rubin, D. (1983), "Using Multiple Imputations to Handle Nonresponse," in Incomplete Data in Sample Surveys, Volume 2: Theory and Bibliographies, edited by Madow, W.G., Olkin, I., and Rubin, D.B. Academic Press, New York, 210-245.

Kelley, R. P. (1985), "Advances in Record Linkage Methodology: A Method for Determining the Best Blocking Strategy," Invited paper presented at the Workshop on Exact Matching Methodologies in Rosslyn, VA, on May 9-10, 1985.

Kirkendall, N. (1985). "Weights in Computer Matching: Applications and an Information Theoretic Point of View," Record Linkage Techniques--1985, Internal Revenue Service.

Lynch, B.T. and Arends, W.L. (1977), "Selection of a Surname Coding Procedure for the SRS Record Linkage System," Technical Report from the Statistical Reporting Service of the U.S. Dept. of Agriculture.

Newcombe, H.B., Kennedy, J.M., Axford, S.J., and James, A.P. (1959), "Automatic Linkage of Vital Records," Science, 130, 954-959.

Newcombe, H.B. and Kennedy, J.M. (1962), "Record Linkage," Communications of the ACM, 5, 563-566.

Newcombe, H.B., Smith, M.E., Howe, G.R., Mingay, J., Strugnell, A., and Abbatt, J.D. (1983), "Reliability of Computerized Versus Manual Searches in a Study of the Health of Eldorado Uranium Workers," Comput. Biol. Med., 13, 157-169.

Pollock, J. and Zamora, A. (1984), "Automatic Spelling Correction in Scientific and Scholarly Text," Communications of the ACM, 27, 358-368.

Rubin, D. (1978), "Multiple Imputations in Sample Surveys--A Phenomenological Bayesian Approach to Nonresponse," ASA 1978 Proceedings of the Section on Survey Research Methods, 20-28.

Rogot, E., Schwartz, S., O'Conor, K., and Olsen, C. (1983), "The Use of Probabilistic Methods in Matching Census Samples to the National Death Index." ASA 1983 Proceedings of the Section on Survey Research Methods, 319-324.

Smith, M., Newcombe, H.B., and Dewar, R. (1983), "Automated Nationwide Death Clearance of Provincial Cancer Registry Files--The Alberta Cancer Registry Study," ASA 1983 Proceedings of the Section on Survey Research Methods, 300-305.

Statistics Canada/ Systems Development Division (1982), "Record Linkage Software."

U. S. Department of Agriculture/ Statistical Reporting Service (1979), "List Frame Development: Procedures and Software."

U. S. Department of Commerce, Bureau of the Census/Agriculture Division (1981), "Record Linkage for Development of the 1978 Census of Agriculture Mailing List."

U. S. Department of Commerce, Bureau of the Census/Survey Research Division (1978a), "UNIMATCH: A Record Linkage System."

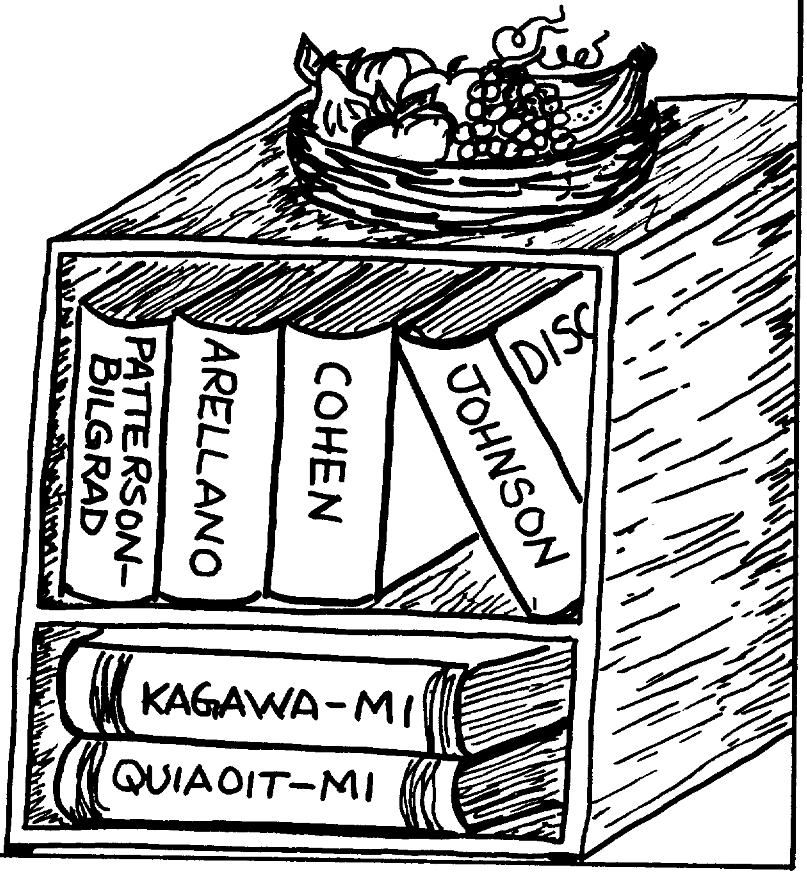
U. S. Department of Commerce, Bureau of the Census/Survey Research Division (1978b), "ZIPSTAN: Generalized Address Standardizer."

Winkler, W. E. (1984a), "Issues in Developing Frame Matching Procedures: Exact Matching Using Elementary Techniques." Presented to the ASA Energy Statistics Committee in April 1984.

Winkler, W. E. (1984b), "Exact Matching Using Elementary Techniques." ASA 1984 Proceedings of the Section on Survey Research Methods, 237-242.

Winkler, W. E. (1985), "Preprocessing of Lists and String Comparison," Record Linkage Techniques--1985, Internal Revenue Service.

Section IV: Application Case Studies I



THE NATIONAL DEATH INDEX EXPERIENCE; 1981-1985

John E. Patterson and Robert Bilgrad, National Center for Health Statistics

The National Death Index (NDI) is a central, computerized index to the death certificates filed in each State vital statistics office. This computer file contains a standard set of identifying information for each person dying in the U.S., beginning with 1979. The NDI was established to assist health and medical investigators in determining whether persons in their studies may have died, and if so, to provide the names of the States in which those deaths occurred, the dates of death, and the corresponding death certificate numbers. The NDI user can then obtain copies of death certificates from the appropriate State offices.

The NDI became operational in November 1981. As of March 31, 1985, the NDI file contained 10.3 million death records for the five-year period 1979-1983. A total of 168 NDI file searches have been performed, involving 2,352,001 records submitted by 99 NDI users. This report gives a brief overview of the NDI users and their research activities, and describes recent evaluations and planned revisions of the NDI matching criteria. Procedures for using the NDI are also presented.

1. OVERVIEW OF NDI USERS

The NDI has been used in a variety of health and medical research projects which rely on the successful ascertainment of the vital status of their study subjects. The research projects of the 99 NDI users have been grouped into five broad research categories in Table 1. These categories are (1) exposure cohorts, involving studies of the effects of being exposed to potential risk factors in the workplace, the environment, or as a result of diagnostic or therapeutic procedures; (2) disease cohorts, involving followup of persons diagnosed as having cancer or other diseases; (3) life style/risk factors, involving studies of the effects of activities such as smoking or drug abuse; (4) clinical trials, primarily involving studies of the potentially beneficial effects of various therapies for specific diseases; and (5) general population cohorts, involving followup of survey participants not selected on the basis of a specific diagnosis or exposure to risk factors.

Forty percent of the NDI users are conducting occupational studies involving followup of rosters of employees to determine whether there have been any harmful effects resulting from their exposures to potentially harmful substances. Most of these studies are being performed by the National Institute for Occupational Safety and Health as well as by oil and chemical companies. Another 28 percent of the NDI users are involved in followup activities on cohorts of persons diagnosed as having cancer or other diseases.

Table 1 also shows the types of organizations using the NDI. It should be noted that while Federal agencies account for only 18 percent of the NDI users, the Federal government is actually providing the funding support for about three-fourths of the studies being performed by universities and consulting firms.

Many of the NDI users are either following cohorts of under 2,500 persons or use the NDI only to check on those study subjects which are considered lost to followup. Almost three-fourths of the users have submitted fewer than 10,000 names. The fewest records submitted for an NDI file search were 7. The largest volume of records was submitted by the Census Bureau for the National Longitudinal Mortality Study being supported by the National Heart, Lung and Blood Institute. Thus far, this study has involved the submission of a test file of 225,875 Census Bureau records and the main study file of 994,195 records. The study's methodology involves a search of the NDI file every two years. The second NDI search for the main study is scheduled for around July 1985 and will involve approximately 1.2 million Census Bureau records.

2. COMPLETENESS AND QUALITY OF NDI AND USER DATA

The effectiveness of the NDI matching process is dependent on the following three factors: (1) the completeness and quality of the death certificate data submitted to the National Center for Health Statistics (NCHS) by the State vital statistics offices for use in creating the NDI file, (2) the completeness and quality of the data provided by the NDI user, and (3) the effectiveness of the NDI matching criteria.

The completeness of the NDI file is probably well in excess of 99 percent. Data on virtually all deaths occurring from 1979 to 1983 have been submitted by the fifty States, the District of Columbia, New York City, Puerto Rico and the Virgin Islands. The NDI file now contains 10.3 million records. Table 2 shows that the completeness of data for most data items exceeds 97 percent except for middle initial (71.7 percent), father's surname (86.2 percent), and social security number (91.0 percent). Although 9.0 percent of the records do not contain social security numbers (as shown in Table 3), only 6.0 percent of the records for persons 22 years and older do not contain such numbers. As might be expected, death records for females have higher percentages of social security numbers not reported than records for males.

It is very difficult to assess the quality of the data on the NDI file, but we have reason to believe that it is probably quite good. The quality of the NDI data is most affected by how the death record information is reported to and recorded by funeral directors. The death certificate is a legal document which must be filed in the State where the death occurs. Most States continually encourage funeral directors to make every effort to obtain accurate information from the person making the funeral arrangements. Funeral directors have a strong incentive for obtaining and accurately recording good identifying information on each decedent. Their clients would not be pleased if errors appeared on the certificate, since this would very likely delay settlement of claims for life insurance and other survivor benefits. All States perform 100 percent verification of the coding and keying of their records. NCHS also performs various quality control checks as the States' data are received.

The completeness and quality of data submitted by NDI users, on the other hand, vary greatly depending on how the data were collected. The complete and accurate collection of the NDI data items listed in Table 2 will, of course, enhance the effectiveness of any subsequent searches of the NDI file. This table summarizes the overall completeness of the data submitted by NDI users; however, the completeness of each data item varies greatly among the different users, especially for such items as middle initial, social security number, State of residence and State of birth.

Because of the newness of the NDI program, many users did not or could not insure the collection of all of the NDI data items. NCHS strongly encourages investigators who are or will be planning studies to make every possible effort to collect all of the NDI data items, even if the investigators do not have specific plans to conduct a followup of study subjects to ascertain their vital status. Once a study is completed, the same or other health investigators may decide that future followup of the study group may indeed be very useful. Internally, NCHS has instituted a policy requiring each new survey to collect all of the NDI data items, regardless of whether the survey staff or others in NCHS plan to use the NDI to followup on the survey participants in the future.

3. RECENT REVISIONS IN THE NDI MATCHING CRITERIA

When the NDI retrieval program was first designed and implemented, a fairly simple set of seven matching criteria was developed (1) to use most effectively the principal identifiers on the death record; (2) to satisfy the needs of the majority of potential users; (3) to make searches against the NDI very routine, eliminating the need for special programming for individual users; and (4) to take into account the policy concerns of the States. These concerns were very significant and had a major impact on the development of the initial matching criteria. Many States felt that the NDI users should be required to provide a fairly substantial body of identifying information for their subjects. They should not accept matching solely on the basis of social security numbers, for example. A number of States were also concerned about probabilistic matching. They felt that their regulations would prevent them from searching their files on a probabilistic basis, and they did not believe that they could delegate authority to NCHS to do what they would not be permitted to do themselves.

For an NDI record to qualify as a possible match with a given user record, under the initial matching criteria, at least one of the following seven combinations of data items must agree on both records:

1. Social security number, first name.
2. Social security number, last name.
3. Social security number, father's surname.
4. If the subject is female: social security number, last name (user's record) and father's surname (NDI record).
5. Month and exact year of birth, first and last name.
6. Month and exact year of birth, first name, father's surname.
7. If the subject is female: month and exact year of birth, first name, last name (user's record) and father's surname (NDI record).

Nine evaluations of the effectiveness of the above matching criteria have been performed by NCHS and by several NDI users. The results are summarized in Table 4. Each of these evaluations involved study files of known decedents which were searched against the NDI file. In those evaluations where social security numbers were available for a large proportion of decedents, the resulting percentages of true matches (user records which were correctly identified as deceased) ranged from 92.1 percent to 98.4 percent. The differences in these percentages are attributed primarily to differences in the quality of the users' data sets. Three evaluations showed that, without the benefit of any social security numbers true matches amounted to only 79.7 percent [8], 80.0 percent [10], and 81.9 percent [9], primarily because of discrepancies in year of birth and names. However, two other users apparently had much better data on dates of birth and names because they achieved true matches of 91.1 percent [1] and 96.5 percent [3] without the benefit of social security numbers.

Most of our advisers and users have stressed that our first efforts to improve our matching criteria should be to maximize the number of true matches, even if this means a significant increase in the false matches which may be generated as a by-product. Our users have generally found that nearly all false matches can be eliminated easily by simply reviewing the output of the NDI search. This is especially true for small studies. For very large studies computerized processing of the NDI output is necessary to identify true matches and to isolate questionable matches which deserve closer inspection. Several users have developed their own computerized algorithms for this purpose.

As a result of the evaluations mentioned above, NCHS is planning to add five new matching criteria to the initial seven. The five additional matching criteria are listed below and are numbered 8 through 12 to distinguish them from the initial seven. A possible NDI record match would be generated if any of these combinations of data items agree on an NDI and a user record.

8. Month and ± 1 year of birth, first and last name.
9. Month and ± 1 year of birth, first and middle initials, last name.

10. Month and exact year of birth, first and middle initials, last name.
11. Month and day of birth, first and last name.
12. Month and day of birth, first and middle initials, last name.

Our evaluations have shown that by also permitting matches on month and day of birth and on month and ± 1 year of birth, the percentage of true matches generated can be increased significantly. One of the NCHS evaluations mentioned previously, involving a cancer registry file containing social security numbers on 85.9 percent of its 2,598 records, showed an increase in true matches from 92.1 percent to 96.2 percent with the addition of the five new matching criteria [8]. The increase in matching effectiveness is greatest, however, for study files having very few or no social security numbers. Another NCHS evaluation involved a file without social security numbers for 607 decedents in the NCHS National Health and Nutrition Examination Study. This evaluation showed an increase in true matches from 81.9 percent to 89.5 percent [9].

The initial retrieval program permitted first names, last names and fathers' surnames to match on the basis of either their exact spelling or Soundex codes. Evaluations showed that the use of Soundex codes often generated agreements on names which were dissimilar, however, causing a number of unnecessary false positives to be generated, while adding very little to the number of true positives. With the planned implementation of the revised matching criteria, the use of Soundex codes will be eliminated. Phonetic matching will be performed only on last names and fathers' surnames and will be based on NYSIIS codes (New York State Identification and Intelligence System). The NYSIIS coding system which will be used was first modified abd tested by the U.S. Department of Agriculture [11] and was subsequently adopted for use in Statistics Canada's Mortality Data Base. The computer program which assigns the modified NYSIIS codes was obtained by NCHS from Statistics Canada.

4. USING THE NDI

As mentioned above, health investigators planning to use the NDI are encouraged to collect as many of the NDI data items as possible and to insure that the data are of good quality. To become an NDI user, health investigators must first complete and submit an NDI application form. Each form is reviewed by the advisers to the NDI program to insure that (1) the proposed use of the NDI is solely for statistical purposes in medical or health research and (2) the applicant provides adequate assurances that the identifying death record information obtained from the NDI and from the State vital statistics offices will be kept confidential and will be used only for the proposed study.

Once the applicant is notified that the application is approved, the NDI user may then submit records for an NDI file search. The user must submit records on a magnetic tape which conforms with the NCHS tape specifications, file format requirements, and coding instructions. Users planning to submit under 300 records have the option of using NCHS coding sheets. The results of an NDI file search are sent to the user (along with the user's data) within three weeks after the user's records are received by the NCHS computer facility.

The user must assess the quality of each possible NDI record match listed and determine which NDI matches are worthy of further investigation. A sample of the planned revision of the NDI Retrieval Report is presented in Table 5. The Retrieval Report lists all user records involved in a match with one or more NDI records. The State of death, death certificate number and date of death are listed for each possible match, along with an indication of which data items are in agreement. Two changes in this report should further assist NDI users in evaluating the quality of possible matches. First, the revised Retrieval Report will show which digits of the social security numbers are in agreement. The current report merely indicates whether or not there was an agreement on the entire social security number. Second, the new report will indicate the extent to which the years of birth disagree; e.g., +1 year, -1 year, -15 years, etc. The current report simply indicates whether or not there is exact agreement on the year of birth.

The user must decide which, if any, of the NDI records are true matches and then obtain copies of the death certificate from the appropriate State vital statistics offices. Most users are interested in obtaining the cause of death from the death certificate. Some users also conduct death record followback activities to the hospitals, physicians, next-of-kin, and/or other persons or establishments indicated on the death certificates. Other users simply obtain copies of certificates to assist in confirming whether a questionable match is actually the person in the study.

Once an application is approved, requests for repeat searches of the NDI file (for additional years of death or for different study subjects) do not need to go through the formal review and approval process again, as long as the information provided in the initial application remains essentially the same. Death records for a particular calendar year are added to the NDI file annually, approximately 12-14 months after the end of that calendar year. Records for deaths occurring in 1984 are scheduled to be added to the NDI file around February 1986.

5. ADDITIONAL REFERENCES CONCERNING THE NDI

In addition to the NDI users' articles and studies cited above, several other articles have been written describing the experience of NDI users [12-15]. There have also been articles written regarding the potential use of the NDI for various studies [16-18]. Finally, papers have been written in which birth certificates from the NCHS 1980 National Natality Survey were searched against the NDI to produce infant mortality rates [19-22]. Copies of these four unpublished papers can be obtained from NCHS [23].

Persons interested in receiving copies of the NDI User's Manual [24] and an NDI Application

Form should write or call:

NATIONAL DEATH INDEX
Division of Vital Statistics
National Center for Health Statistics
3700 East West Highway, Room 1-44
Hyattsville, Maryland 20782
Telephone: (301) 436-8951

NOTES AND REFERENCES

[1] Wentworth, Deborah N., et al., "An Evaluation of the Social Security Administration Master Beneficiary Record File and the National Death Index in the Ascertainment of Vital Status," American Journal of Public Health, Vol. 73, No. 11, November 1983, pages 1270-1274.

[2] Acquavella, J.F., Donaleski, D., and Hanis, N.M., "An Analysis of Mortality Follow-up through the National Death Index for a Cohort of Refinery and Petrochemical Workers," (Accepted for publication in the American Journal of Industrial Medicine, 1985).

[3] Stampher, Meir J., et al., "Test of the National Death Index," American Journal of Epidemiology, Vol. 119, No. 5., 1984, pages 837-839.

[4] Data are based on preliminary, unpublished results of an evaluation of the NDI matching criteria performed by Nancy Fink, Johns Hopkins School of Hygiene and Public Health, 1983.

[5] Lubitz, J. and Pine, P., "Initial Findings: Development and Use of a Linked Medicare-NCHS Mortality File," (Read before the 112th Annual American Public Health Association Meeting, Anaheim, California, November 11, 1984).

[6] Curb, J. David, et al., "Ascertaining Vital Status through the National Death Index and the Social Security Administration," American Journal of Epidemiology, Vol. 121, No. 5, 1985, pages 754-766.

[7] Davis, Kathryn B., et al., "A Test of the National Death Index Using the Coronary Artery Surgery Study (CASS)," (Accepted for publication by Controlled Clinical Trials, 1985).

[8] Patterson, John E., "Evaluation of the Matching Effectiveness of the National Death Index," American Statistical Association Proceedings of the Social Statistics Section, 1983, pages 1-10.

[9] Data are based on unpublished results of an evaluation of both the initial and revised NDI matching criteria performed by Helen Barbano, Division of Analysis, National Center for Health Statistics, 1984.

[10] Rogot, Eugene, et al., "On the Feasibility of Linking Census Samples to the National Death Index for Epidemiologic Studies: A Progress Report," American Journal of Public Health, Vol. 73, No. 11, November 1983, pages 1265-1269.

[11] Lynch, Billy T. and Arends, William L., Selection of a Surname Coding Procedure for the SRS Record Linkage System, Statistical Reporting Service, U.S. Department of Agriculture, February 1977.

[12] Arellano, Max G., et al., "The California Automated Mortality Linkage System (CAMLIS)," American Journal of Public Health, Vol. 74, No. 12, December 1984, pages 1324-1329.

[13] Lubitz, James, "The Cost of Cancer and the Medicare Hospice Benefit," Proceedings of the 19th National Meeting of the Public Health Conference on Records and Statistics, August 1983, pages 145-147.

[14] Nelson, Nancy A. and Van Peenen, P.D.F., "RE: 'Test of the National Death Index'" (letter to the editor), American Journal of Epidemiology, Vol. 121, 1985, page 626, (with reply from the first author, Meir Stampher).

[15] Rogot, Eugene, et al., "The Use of Probabilistic Methods in Matching Census Samples to the National Death Index," Statistics of Income and Related Administrative Record Research: 1983, Statistics of Income Division, Internal Revenue Service, October 1983.

[16] Beebe, Gilbert, "Record Linkage Systems - Canada vs. the United States," American Journal of Public Health, Vol. 70, December 1980, pages 1246-1247.

[17] Edlavitch, S.A., Feinleib, M. and Anello, C., "A Potential Use of the National Death Index for Postmarketing Drug Surveillance," Journal of the American Medical Association, Vol. 253, No. 9, March 1, 1985, pages 1292-1295.

[18] MacMahon, Brian, "The National Death Index," (editorial), American Journal of Public Health, Vol. 73, No. 11, November 1983, pages 1247-1248.

[19] Placek, Paul, et al., "Methodology for the '1980 National Natality Survey/National Death Index Match' Project," 1984 Proceedings of the American Statistical Association, Section on Survey Research Methods.

[20] Keppel, Kenneth G., et al., "Infant Mortality Rates Based on Linked Records from the 1980 National Natality Survey," (Read before the 1985 Annual Meeting of the Population Association of America, Boston, Massachusetts, March 28-30, 1985).

[21] Kessel, Samuel S., et al., "Fetal, Perinatal, and Neonatal Death Rates According to Hospital, Maternal, and Infant Characteristics: United States, 1980," (presented at the 1985 Annual Meeting of the American Statistical Association, Las Vegas, Nevada, August 5-8, 1985).

[22] Placek, Paul J., "Record Linkage Methodologies in the 1980 National Natality Survey (NNS) and 1980 National Fetal Mortality Survey (NFMS)," (presented at the meetings of the International Statistical Institute, Amsterdam, The Netherlands, August 12-22, 1985).

[23] Copies of references [19], [20], [21], and [22] can be obtained by writing to: Natality Statistics Branch, Division of Vital Statistics, National Center for Health Statistics, 3700 East-West Highway, Room 1-44, Hyattsville, Maryland 20782.

[24] User's Manual: The National Death Index, U.S. Department of Health and Human Services, National Center for Health Statistics, DHHS publication number (PHS) 81-1148, September 1981.

Table 1

NATIONAL DEATH INDEX (NDI) USERS AND RECORD VOLUMES

NDI User Characteristics	Users		NDI Searches		User Records	
	Number	Percent	Number	Percent	Number	Percent
Types of Research:						
Total	99	100.0	168	100.0	2,352,001	100.0
Exposure cohorts						
Occupational	40	40.4	57	33.9	636,752	27.1
Environmental	5	5.1	18	10.7	78,824	3.4
Diagnostic/therapeutic	2	2.0	3	1.8	7,566	0.3
Disease cohorts						
Cancer registries	13	13.1	16	9.5	38,002	1.6
Other	15	15.2	18	10.7	42,120	1.8
Life style/risk factors	9	9.1	14	8.3	116,875	5.0
Clinical trials	9	9.1	14	8.3	86,333	3.7
General population cohorts	6	6.1	28	16.7	1,345,529	57.2
Types of NDI Users:						
Total	99	100.0	168	100.0	2,352,001	100.0
Federal Government	18	18.2	62	36.9	1,516,313	64.5
State Government	4	4.0	6	3.6	45,056	1.9
University	28	28.3	37	22.0	327,060	13.9
Private Industry	13	13.1	17	10.1	221,942	9.4
Hospital	19	19.2	22	13.1	63,120	2.7
Consulting firm	17	17.2	24	14.3	178,510	7.6
Record Volume:						
Total	99	100.0	168	100.0	2,352,001	100.0
Under 2,500	42	42.4	45	26.8	29,259	1.2
2,500 - 9,999	29	29.3	38	22.6	165,711	7.1
10,000 - 24,999	12	12.1	31	18.5	225,466	9.6
25,000 - 99,999	13	13.1	33	19.7	513,014	21.8
100,000 - 499,999	2	2.0	7	4.2	424,356	18.0
500,000+	1	1.0	14	8.3	994,195	42.3

Table 2

NUMBER OF RECORDS AND PERCENT COMPLETENESS
OF NATIONAL DEATH INDEX (NDI) AND USER DATA ITEMS

Data Items	NDI File	User Files
No. of Records--------	10,290,730	1,131,931*
Percent Complete:		
Last Name-----------	99.9	99.9
First Name----------	99.9	99.7
Middle Initial------	71.7	73.4
Social Security No.-	91.0	84.2
Birth Month---------	98.8	95.7
Birth Day-----------	98.7	87.9
Birth Year----------	99.4	97.0
Father's Surname----	86.2	8.9
Sex-----------------	99.9	92.6
Race----------------	97.9	53.1
Marital Status------	99.4	17.9
State of Residence--	99.9	44.2
State of Birth------	99.5	18.6
Age at Death--------	99.9	10.6

* The total number of user records shown excludes 1,220,070 records associated with the National Longitudinal Mortality Study, sponsored by the National Heart, Lung and Blood Institute and involving both the Census Bureau and the National Center for Health Statistics. This large volume of records was eliminated from this table to give a more realistic presentation of the completeness of the data items submitted by the other 98 NDI users.

Table 3

REPORTING OF SOCIAL SECURITY NUMBER ON NATIONAL DEATH INDEX (NDI) RECORDS; BY SEX AND AGE AT DEATH

Age at Death	Number of NDI Records			Percent not Reported WITHIN Age/Sex Group		
	Both Sexes*	Male	Female	Both Sexes*	Male	Female
All Ages--	10,289,958	5,536,778	4,753,180	9.0	7.8	10.3
0-16------	356,704	208,377	148,327	88.6	87.4	90.3
17-21-----	126,475	95,242	31,233	17.8	16.9	20.6
22-59-----	1,965,257	1,279,175	686,082	8.4	7.2	10.6
60+-------	7,841,522	3,953,984	3,887,538	5.3	3.6	7.2

* The record counts and percentages do not include 772 records for which sex was not reported.

Table 4

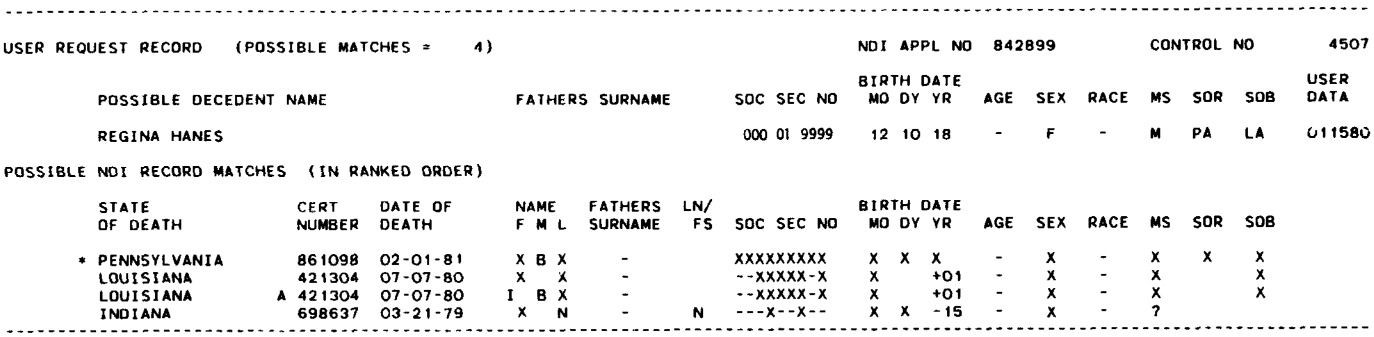
EVALUATIONS OF THE EFFECTIVENESS OF THE NATIONAL DEATH INDEX (NDI)
MATCHING CRITERIA USING RECORDS OF KNOWN DECEDENTS

NDI Users and User Studies*	Known Decedents	True Matches	Percent True Matches
University of Minnesota **School of Public Health**			
(Multiple Risk Factor Intervention Trial (MRFIT) for coronary heart disease) [1]....	191	188	98.4
Exxon Corporation **Research & Environmental Health Division**			
(Mortality study update of Exxon workers) [2]....	1,449	1,407	97.1
Harvard Medical School			
(Nurses health study) [3].......................	346	334	96.5
Johns Hopkins School of **Hygiene and Public Health**			
(Health effects of low-level radiation in shipyard -workers) [4]........................	8,947	8,485	94.8
Health Care Financing Administration			
(Use and costs of Medicare services by cause of death) [5]..........................	69,631	65,000	93.3
University of Texas at Houston **School of Public Health**			
(Hypertension Detection and Follow-up Program post trial survey) [6]..................	1,154	1,074	93.1
University of Washington			
(Coronary Artery Surgery Study) [7].............	370	344	93.0
National Center for Health Statistics **Division of Vital Statistics**			
(Evaluation of NDI using cancer registry records) [8]			
INITIAL matching criteria:......................	2,598	2,394	92.1
Using Social Security Number (SSN)..........	2,231	1,874	84.0
Using birth month/year......................	2,596	2,069	79.7
NEW matching criteria..........................	2,598	2,500	96.2
Using SSN..................................	2,231	1,874	84.0
Using birth month/day <u>or</u> birth month/±1 year.	2,596	2,351	90.6
National Center for Health Statistics **Division of Analysis**			
(First National Health and Nutrition Examination Survey epidemiologic follow-up) [9]			
INITIAL matching criteria (without SSN).........	607	497	81.9
NEW matching criteria (without SSN)............	607	543	89.5

* Numbers in brackets refer to studies cited in the NOTES and REFERENCES Section.

Table 5

RETRIEVAL REPORT -- REVISED
(All the information in this example is hypothetical.)



```
------------------------------------------------------------------------------------------------------------------------------------

USER REQUEST RECORD   (POSSIBLE MATCHES =    4)                                 NDI APPL NO  842899         CONTROL NO         4507

                                                                                  BIRTH DATE                                          USER
         POSSIBLE DECEDENT NAME                  FATHERS SURNAME       SOC SEC NO  MO DY YR   AGE  SEX  RACE  MS  SOR  SOB   DATA

         REGINA HANES                                                  000 01 9999  12 10 18   -    F    -    M   PA   LA    011580

POSSIBLE NDI RECORD MATCHES  (IN RANKED ORDER)

         STATE              CERT    DATE OF      NAME   FATHERS  LN/              BIRTH DATE
         OF DEATH           NUMBER  DEATH        F M L  SURNAME   FS  SOC SEC NO   MO DY YR   AGE  SEX  RACE  MS  SOR  SOB

       * PENNSYLVANIA       861098  02-01-81     X B X     -          XXXXXXXXX    X  X  X     -    X    -    X   X    X
         LOUISIANA          421304  07-07-80     X   X     -          --XXXXX-X    X    +01    -    X    -    X        X
         LOUISIANA        A 421304  07-07-80    I  B X     -          --XXXXX-X    X    +01    -    X    -    X        X
         INDIANA            698637  03-21-79     X   N     -       N  ---X--X--    X  X -15    -    X    -    ?
------------------------------------------------------------------------------------------------------------------------------------
```

COLUMN HEADING ABBREVIATIONS:

LN/FS = Last name on user record compared to father's surname on National Death Index (NDI) record.

MS = Marital status

SOR = State of residence

SOB = State of birth

SYMBOLS USED WITHIN THE TABLE:

* = All items provided on user record matched exactly with NDI record.

Blank = User and NDI data items did not match.

X = User and NDI data items matched exactly.

- = Data item not provided by user.
For SSN: specific digits did not match.
For LN/FS: comparison was not attempted.

SYMBOLS (CONTINUED):

? = Insufficient information on NDI record.

A = Alias NDI record.

I = Only first initial of first name matched.

N = Names matched only on NYSIIS codes.

B = Middle initials not provided on either record. This occurrence is treated as a match on middle initial.

+01 = Birth year on the NDI record is one year more than the year on the user record.

-01 = Birth year on the NDI record is one year less than the year on the user record.

-15 = Difference between the two years of birth. (The two-digit birth year on the user record is subtracted from the two-digit birth year on the NDI record. Note: No distinction is made to accomodate birth years in the 1800's versus birth years in the 1900's.)

AN IMPLEMENTATION OF A TWO-POPULATION FELLEGI-SUNTER PROBABILITY LINKAGE MODEL

Max G. Arellano, University of California, San Francisco

I. INTRODUCTION

The California Automated Mortality Linkage System (CAMLIS) has been in operation at the University of California, San Francisco, since the fall of 1981. It was organized under the sponsorship of the Department of Epidemiology and International Health to facilitate the clearance of study population files submitted by qualified investigators against mortality files for the State of California.

The linkage of two independently generated data files has long been thought to be the exclusive province of highly trained clerks because of the need to process the discrepancies which frequently occur between sets of identifying information for the same person on the two files.

A computerized approach to the record linkage problem can adopt either deterministic or probabilistic decision criteria. Deterministic linkage criteria require the formulation of a 'match key' to establish the relationship between records on the two files to be linked. This match key functions on an 'either or' basis, i.e., if an identical value of the match key is found on both files, the records with the identical values are said to be matched. Otherwise, the records are said to be unmatched. In order to perform its required function with minimal error, this match key must possess as many of the characteristics of a unique identifier as possible. Match keys can be constructed from any conceivable combination of last name, first name, sex, social security number, birth date (or portions thereof), or any other identifying items present on the file. Although it is not a true unique identifier, the ready availability of the social security number has led to its widespread use as the match key of choice in deterministic linkage applications.

Probabilistic linkage criteria are based on a linkage weight calculated for each pairwise comparison between records on the two files to be linked; these linkage weights are the sum of component weights calculated for each item of identification contained on the two files. The component weights are functions of occurrence probabilities and of the reliability of the data items. Probabilistic decision criteria provide an attractive alternative to deterministic linkage criteria as a means of computerizing the record linkage activity primarily because: 1) they assign weights in a manner that is consistent with our own human intuition and 2) they can accommodate partial agreements. On the debit side: 1) they require the estimation of many parameters, some of which are inestimable, 2) they are much more difficult to program and 3) they are more costly to use.

Our decision to adopt probabilistic decision criteria was based primarily on our conviction, based on a careful analysis of the available information, that the requirements of investigators in the health and medical care research fields could not be met solely by deterministic linkage criteria. Our experience over the last four years has served to confirm the validity of that decision.

II. THE FELLEGI-SUNTER WEIGHTING ALGORITHM

The Fellegi-Sunter [1] weighting algorithm requires the estimation of two probability distribution functions:

If we let,

p_{jA} = P(Occurrence of the jth configuration in population A)

p_{jB} = P(Occurrence of the jth configuration in population B)

$p_{jA\cap B}$ = P(Occurrence of the jth configuration in $A\cap B$)

$w(\gamma_j)$ = Probability linkage weight for the jth agreement configuration

$m(\gamma_j)$ = P(Occurrence of the jth agreement configuration|the record pairs are associated with members of the matched set)

$$= P(\gamma_j | (a,b) < M)$$

$$= p_{jA\cap B}(1-e_A)(1-e_B)(1-e_T)$$

$u(\gamma_j)$ = P(Occurrence of the jth agreement configuration|the record pairs are associated with members of the unmatched set)

$$= P(\gamma_j | (a,b) < U)$$

$$= p_{jA}p_{jB}$$

Then, $w(\gamma_j) = \log[m(\gamma_j)/u(\gamma_j)]$

Among the obvious difficulties encountered in the implementation of this model are:

(A) It does not address the problem of estimating the e or e_T terms. We generally refer to these as the "component error probabilities."

(B) The $p_{A\cap B}$ term requires information which can only be obtained when the linkage has been completed in a satisfactory manner, if then.

If the populations represented by the files that are being linked can be regarded as samples drawn from the same population, i.e., the "one-population" model, some simplifications can be introduced into the above expressions:

$$m(\gamma_j) = p_j(1-e)^2(1-e_T)$$

$$u(\gamma_j) = p_j^2$$

$$w(\gamma_j) = \log[m(\gamma_j)/u(\gamma_j)]$$

$$= \log[p_j^{-1}(1-e)^2(1-e_T)]$$

Moreover, if the data are being collected continuously, as is generally the case under the circumstances to which the one-population model is applicable, procedures can readily be developed to iteratively obtain "good" estimates of the component error probabilities. This is, unfortunately, not the case for situations to which the two-population model would generally be applied. For one thing, if the populations being linked do not overlap, the $p_{A\cap B}$ term is meaningless. The model also requires estimates of component error probabilities specific to the files that are being linked.

Prior information on the record-pairs that correspond to the intersection of the two populations is obviously desirable, if not absolutely necessary, before probability linkage can be initiated. However, since this is precisely the information we are attempting to obtain by means of probability linkage, if it can be obtained by other means, one may legitimately question the need for probability linkage.

In this paper I will describe the approach that has been adopted by the CAMLIS project to the problem of implementing a two-population Fellegi-Sunter model.

III. THE CAMLIS IMPLEMENTATION OF THE TWO-POPULATION FELLEGI-SUNTER MODEL

Central Concepts

The CAMLIS approach is based on the following central concepts:

(A) A two-stage linkage process, consisting of a deterministic first stage (primarily based on the social security number) followed by a probabilistic second stage, is necessary to achieve the desired performance characteristics. This strategy has several benefits:

(1) Each stage is capable of detecting valid linkages which will escape detection by the other stage.

(2) Since deterministic linkage is carried out first, the correctly matched records which it produces can be used to derive estimates of the component error probabilities required by probability linkage.

(B) A phonetic name encoding algorithm with superior operating characteristics must be used to form the basic comparison groups for probability linkage to minimize the number of pairwise record comparisons that must be carried out. We chose to adopt a modified version of the New York State Identification and Intelligence System (NYSIIS) phonetic coding system for this purpose. It is doubtful if CAMLIS could be operated on a cost-effective basis without the use of a phonetic name coding system with the superior performance characteristics of NYSIIS.

(C) A modification of the weighting algorithm for the two-population Fellegi-Sunter model is necessary to compensate for the inestimable parameters.

(D) Component error probabilities can be estimated from the "matched set" produced by first stage or deterministic linkage.

In this presentation, I will focus primarily on points (C) and (D) above, i.e., on our approach to the estimation of the parameters required by the two-population Fellegi-Sunter weighting algorithm.

The Estimation of Relative Frequency Parameters

In CAMLIS applications, a user file, which we denote as file L_A, is linked to a California State mortality file, which we denote as file L_B. Since the characteristics of most user files are significantly different from those of the California mortality file, the two-population model is obviously called for. However, many of the parameters required by the two-population model, e.g., $p_{A\cap B}$ and e_A, are inestimable. We therefore carefully scrutinized the expressions for the two probability distribution functions to determine whether a simplification was possible. We first made the observation that the characteristics of the user file are always subsets of the characteristics of the mortality file; we also observed that, for those components that are independent of mortality, $p_A \sim p_{A\cap B}$. These observations resulted in the elimination of the p_A term from the weighting algorithm and served to justify the use of relative frequencies derived only from the mortality files. Since these relative frequencies can change over time, files have been developed which contain the necessary relative frequencies at five-year intervals; CAMLIS procedures retrieve them as necessary.

The component for which the assumption is not tenable is birth year; an entirely different approach to weight computation for the birth year component has, therefore, been developed.

The Estimation of Component Error Probabilities

Within the context of a mortality clearance system, it is not possible to derive separate estimates of component error probabilities for files L_A and L_B; there is just not enough information available. We therefore made the simplifying assumption that the corresponding component error probabilities in the two files were identical, i.e., we assume that:

$$e = e_A = e_B$$

Estimates of e and e_T are derived from the matched record-pairs produced by first stage deterministic linkage. To eliminate spurious matches, we require a high concordance among the identifying elements on the two files that are not incorporated into the match key.

The basic algorithm that we utilize to calculate agreement configuration weights is therefore:

$$m(\gamma_j) = p_{jA}(1-e)^2(1-e_T)$$

$$u(\gamma_j) = p_{jA}p_{jB}$$

$$w(\gamma_j) = \log[m(\gamma_j)/u(\gamma_j)]$$

$$= \log[p_{jB}^{-1}(1-e)^2(1-e_T)]$$

IV. CONCLUSION

The Fellegi-Sunter model requires an assumption regarding the independence of the components of the comparison vector; this assumption is frequently a major concern in linkage applications. It is not my intention to minimize the importance of this assumption. The real concern, however, must be the extent to which violations of this assumption affect the results produced by the model.

(A) The components of the comparison vector should be carefully chosen. Only one of several highly dependent components should be incorporated into the model.

(B) Although it is possible to correct for the effect of dependence, for moderately dependent components, these efforts are hardly ever worth the small gain in precision that can be realized.

(C) We have done a great deal of difference analysis. Our conclusion is that the estimated component error probabilities and relative frequencies must differ considerably from the appropriate values to significantly affect the computed weights.

(D) For matches that achieve a linkage weight significantly greater than the upper threshold value, a bias in the weight is obviously of no consequence. Similarly, for matches that achieve a linkage weight significantly below the lower threshold value, a bias in the weight is also of no consequence. The vast majority of record-pairs achieve either very low or very high linkage weights.

(E) Record-pairs which achieve a linkage weight between the lower and upper threshold values are subject to manual review. Since record-pairs fall into this category because they either contain ambiguous or sparse identifying information, it is extremely doubtful whether they would differ significantly if the weights were computed according to a more precise model. In any case, comparable results could be obtained by redefining the upper and lower threshold values.

The major advantage of probability linkage is that it permits a meaningful ranking of matched record-pairs. The ranking makes it possible to focus review efforts on the comparisons which have been assigned borderline weights. It can readily be shown that the gain achieved by verifying the probability linkage decisions above a certain threshold value and below a certain threshold value is negligible.

Our experience with the Fellegi-Sunter probability linkage criteria has been uniformly favorable. It is our considered opinion, however, that probabilistic linkage and deterministic linkage are best utilized as complimentary procedures and that both are necessary to achieve optimum results.

REFERENCES

[1] Fellegi, I., and Sunter, A. (1969) "A Theory for Record Linkage," Journal of the American Statistical Association, 64, 1183-1210.

DERIVING LABOR TURNOVER RATES FROM ADMINISTRATIVE RECORDS

Malcolm S. Cohen, University of Michigan

U.S. nonagricultural establishments will hire workers new to their firms an estimated 64 million times during 1985. These hiring transactions probably will involve only 12-16 million workers who changed their primary jobs.

An econometric model was constructed using administrative records from Social Security files, and estimates of new hires were made by industry, state, age, race, and sex. When this study was done, Social Security records were available only through the mid-1970s. Wage records used in the administration of the unemployment insurance system were available in sixteen states to verify the accuracy of the econometric estimates. Because wage records were available only for sixteen states, and because of differences in state laws and data processing procedures, wage records could not be used for obtaining national estimates.

Organizationally, this paper is divided into two main sections. In the first, the methodology employed is described. The second presents examples of the various results, as well as some general comments about the usefulness of these administrative records.

METHODOLOGY

Social Security data from a one-percent sample of a continuous work history file for the period 1971-76 were used to construct labor turnover measures. Instructions for using the methodology were given to three government agencies, who then did the matching and provided tabulations for different years. These agencies were the New York Department of Labor, the Social Security Administration, and the Bureau of Economic Analysis. The provisions of the 1976 tax reform act require the Internal Revenue Service to screen the data for possible confidentiality disclosures prior to release. All analyses of Social Security records were from tabulations provided by the government agencies. No Social Security data were released on individual workers or firms.

Employee records were matched with employer records. If a worker's identification number appeared in a firm's file in a given quarter, but did not appear in the file in the previous quarter, the worker was classified as an accession to the firm [1]. If a worker classified as an accession did not work for the firm for the prior four quarters, that worker was classified as a new hire. The decision to use four quarters as a determining factor was somewhat arbitrary. That period of time was chosen because it was long enough to identify workers who return to a firm seasonally, although it would not exclude workers who may have worked for a firm sometime in the more distant past. The higher degree of accuracy that might be attained by matching records several years back, however, was not considered great enough to justify the substantial increase in cost of matching data for more than four quarters [2].

It is also possible to generate other turnover measures using the pattern of employment within the firm. For example, if a worker is present in a given quarter and absent in the next quarter, this is a separation. If a worker is a new hire who continues to work for a period of, say, an additional two quarters, this is a permanent new hire. If a worker is an accession (not employed in previous quarter) who did work for the firm sometime in the previous four quarters, this is a recall. If a worker is an accession and separation in the same quarter, this is a short-term accession. Various turnover measures were developed based on these definitions.

Data were constructed for new hires from quarterly Social Security records from the second quarter of 1972 to the second quarter of 1975. A special tabulation for 1975-76 was used for special analyses but not included in the quarterly analyses used to generate current estimates.

A model was developed to predict new hires. The model's derivation begins with a tautology:

$$(1)\quad \Delta E = NH + Recalls - Quits - Layoffs - OS$$

where ΔE is change in employment; NH is new hires; and OS is other separations.

From this we obtain:

$$(2)\quad NH = \Delta E - Z$$

where Z = Recalls − Quits − Layoffs − OS

To obtain rates, both series were divided by E. It was assumed that the unemployment rate would be a good proxy for Z. It was assumed that there was a negative correlation between Z and the unemployment rate.

When the equation was estimated, data from the Bureau of Labor Statistics (BLS) 790 series were used for employment, and data from the monthly Current Population Survey were used for unemployment rates and seasonal dummy variables. The final equation was:

$$(3)\quad NHR_t = \alpha_0 + \alpha_1 \%\Delta E_t + \alpha_2 UR_{t-1} + \\ + \alpha_3 S_1 + \alpha_4 S_2 + \alpha_5 S_3 + \alpha_6 D + E_1$$

where NHR is the new hire rate; %ΔE is the percentage change in BLS 790 employment; UR is the unemployment rate; S_1,S_2 and S_3 are seasonal dummies for the first three quarters of the year; D is 1 in the first quarter of 1974; and E_1 is a random term.

The dummy variable was used because of a data error in the first quarter of 1974 in the data provided. The coefficient α_1 is expected to be positive, while α_2 is predicted to be negative. The equations were estimated for each state with a total of thirteen observations. The results of the model for fiscal 1975 were simulated to determine goodness of fit.

Figure 1 provides the %ΔE and UR_{t-1} parameters, the proportion of variation explained by the model (R^2), actual new hire rate, and percent error in the forecast for all 50 states. All parameters significant at the .05 level are indicated by an asterisk.

One of the difficulties with this model is that data for the dependent variable cannot be obtained from Social Security data beyond 1977 on a quarterly basis. Only annual new hire rates can be computed. These can only be obtained by special arrangements with the Internal Revenue Service and the Social Security Administration. To verify the model in selected states, however, wage records were obtained using similar concepts for workers covered by unemployment insurance. These data can be generated quarterly on a current basis in wage records states. Over 40 states are wage records states. Special arrangements must be made, however, in each state to obtain these data. The arrangements require considerable data processing to match workers and firms over at least four quarters.

Our estimates were compared with the wage records data in sixteen states. The results of the comparisons are shown in Figure 2. The errors are generally relatively small except in Florida. Here, however, the Florida data provided were probably more prone to error than our estimates. The significantly lower reported new hires in Florida probably represents an undercount in the state's processing. The state used a different processing methodology than the other states.

We simulated our model and obtained new hire estimates for 1975-85 [3].

RESULTS

Figure 3 shows the predicted number of new hires from 1975 through 1985 using our model. Figure 4 illustrates the five states with the largest number of new hires. These states accounted for 40% of all new hires in the United States. Converting the new hires into rates, Figure 5 shows the parts of the United States with the highest and lowest rates. The highest rates are west of the Mississippi. A prominent exception is Florida.

It is also possible to compare new hire rates by industry. Figures 6 and 7 show the industries with the highest and lowest rates, respectively.

In 1985 it is unlikely that social services would be among the high new hire rate industries. This reflects changes in government priorities over the decade. It is probable, however, that the other industries are high and low turnover industries in 1985.

Individuals versus Transactions

One of the difficulties in interpreting our measures is reconciling the incredibly high turnover (e.g., 80% in 1985) with our knowledge of how often workers change jobs. The number of turnover transactions include instances where one worker changed jobs more than once, so the total does not reflect the actual number of workers who changed jobs. Thus, when turnover is expressed as a percentage of employment, the result should not be interpreted as the percentage of workers who changed jobs. To gain some insight into reconciling this apparent dilemma, we developed some special tabulations from 1975-76 Social Security files. First we computed an annualized 84% new hire rate for 1976 by multiplying the rate obtained in the second quarter of 1976 by 4. This is certainly comparable to the rates we had been obtaining for other years. A different analysis was carried out where workers were assigned to their primary jobs, where they earned the most money during 1976. Only 18% of the workers were new hires in their primary jobs, based on the second quarter of 1976. Some of these workers could have accounted for several new hire transactions. Similarly, workers who were not new hires in their primary jobs could be new hires in secondary jobs. Thus, we estimated that of the 64 million new hires, about 14 million workers were new hires in their primary jobs. In another quarter we estimated a ratio which would suggest that slightly under 16 million workers were new hires in their primary jobs. An estimate of 12-16 million seemed appropriate due to the limited number of quarters on which we could base our ratio.

Another comparison we made with our special tabulation was the average number of employers for whom employees worked in different industries. We assigned workers to the employer from whom they received the majority of their earnings and tabulated the number of different employers. Four nonagricultural industries--heavy construction contractors, water transportation, eating and drinking places, and motion pictures--had an average of two or more employers per' worker. Water transportation (longshore) averaged 2.5 employers per worker. The industries with an average of l.25 or fewer employers (with at least 100,000 persons in the industry) included: primary metals, communications, and public utilities.

Areas for Further Research

The information obtained from Social Security records and state unemployment insurance records represent about the only currently comprehensive source of labor turnover data. Our model permits obtaining current estimates from these data. It would be useful to tabulate annual Social Security files to determine labor turnover from more recent Social Security files. It would also be useful to forecast the turnover rates by industry, age, and sex. The 1975-76 special tabulations by person and transaction provide detailed characteristics by state, SMSA, industry, age, wage class, sex, and race. Additional analyses of these data remain to be carried out, as well as additional analyses of separations and short-term new hires. Finally, more efficient forecast estimates can be made by combining cross-section and time-series turnover data.

NOTES AND REFERENCES

[1] A worker's identification number appears in the file if the worker had wages greater than zero in a given quarter.

[2] Using California wage records from the Unemployment Insurance system, the California Employment Development Division did a test of how many fewer new hires there would be if seven quarters were used as a cut-off instead of four, and found only about 2% fewer new hires. (Glen Siebert, Employment Service Potential: Indicators of Labor Market Activity, pp. 48-9. Sacramento, CA: Employment Development Department, 1977.)

[3] For a more complete description of the simulation methodology, see Malcolm S. Cohen and Arthur R. Schwartz, "A New Hires Model for the Private Nonfarm Economy," Economic Outlook for 1984, Department of Economics, University of Michigan, Ann Arbor, 1984.

Figure 1. New Hire Rates by State, Fiscal 1975, % Error, R^2, Selected Coefficients

State	1975 New Hire Rate	1975 % Error	R^2	%E	URLAG
Alabama	19.1	−.3	.943	51.94	−1.59*
Alaska	42.0	5.2	.941	165.85*	2.34
Arizona	24.9	.3	.978	148.44*	−1.65*
Arkansas	22.4	.5	.966	48.90	−2.24*
California	23.4	−.9	.930	87.91	−1.21
Colorado	28.3	1.6	.951	97.29*	−2.75*
Connecticut	15.0	.9	.984	97.25*	−1.03*
Delaware	15.9	−.6	.828	−64.85	−3.25*
D.C.	20.8	−3.5	.822	89.90	−1.49
Florida	26.3	−1.3	.973	178.70*	−2.29*
Georgia	20.2	−1.1	.982	118.40*	−2.24*
Hawaii	20.9	.9	.819	122.97	−.85
Idaho	26.3	.1	.898	68.38	−.52
Illinois	16.8	.1	.988	111.27*	−1.19*
Indiana	15.5	−.9	.992	83.49*	−1.56*
Iowa	18.7	3.0	.951	25.81	−1.61*
Kansas	23.1	3.1	.944	63.74	−1.33
Kentucky	17.7	−1.2	.980	107.40*	−1.07*
Louisiana	26.3	1.7	.890	−15.77	−1.37
Maine	18.2	−3.0	.943	105.95	−.88
Maryland	18.2	−.1	.982	162.01*	−.71
Massachusetts	16.5	−1.9	.976	126.06*	−.81*
Michigan	14.5	−4.1	.935	73.53*	−1.48*
Minnesota	17.3	−.1	.958	62.99	−1.30*
Mississippi	19.5	.2	.938	96.48*	−1.36
Missouri	18.2	.4	.989	99.74*	−1.13*
Montana	23.5	−1.3	.959	191.26*	−.20
Nebraska	20.6	1.7	.971	74.97	−.86
Nevada	33.2	−.5	.975	165.36*	−1.42*
New Hampshire	17.5	−2.0	.917	135.78*	−2.02*
New Jersey	17.1	−.1	.978	121.49*	−1.20*
New Mexico	28.3	−2.3	.916	103.08	−1.61*
New York	15.7	−1.7	.959	109.77*	−1.16*
N. Carolina	16.9	−1.5	.970	112.58*	−2.03*
N. Dakota	22.2	2.0	.902	229.05*	.72
Ohio	15.0	−.3	.996	91.35*	−1.33*
Oklahoma	24.8	−.1	.944	131.44	−1.08
Oregon	23.3	1.1	.925	103.60	−1.22
Pennsylvania	13.9	.2	.980	134.31*	−.96*
Rhode Island	17.8	−1.9	.960	72.75*	−1.84*
S. Carolina	17.6	−1.9	.918	69.73*	−1.77*
S. Dakota	19.9	−2.4	.968	133.96*	−.50
Tennessee	18.1	−.6	.978	93.82*	−1.38*
Texas	27.1	−.3	.977	34.35	−1.57*
Utah	23.9	.0	.967	109.57	−1.20
Vermont	18.0	.9	.821	161.11	−.18
Virginia	18.0	−.2	.970	107.94*	−1.66*
Washington	22.4	.7	.953	141.46*	−.07
W. Virginia	15.7	−2.3	.964	145.31*	−.27
Wisconsin	14.8	−.1	.988	72.78*	−1.39*
Wyoming	33.4	4.4	.899	21.54	−1.22

%E = percentage change in employment
URLAG = unemployment rate in previous quarter
* = coefficient significant at the .05 level
N = 13 for each state

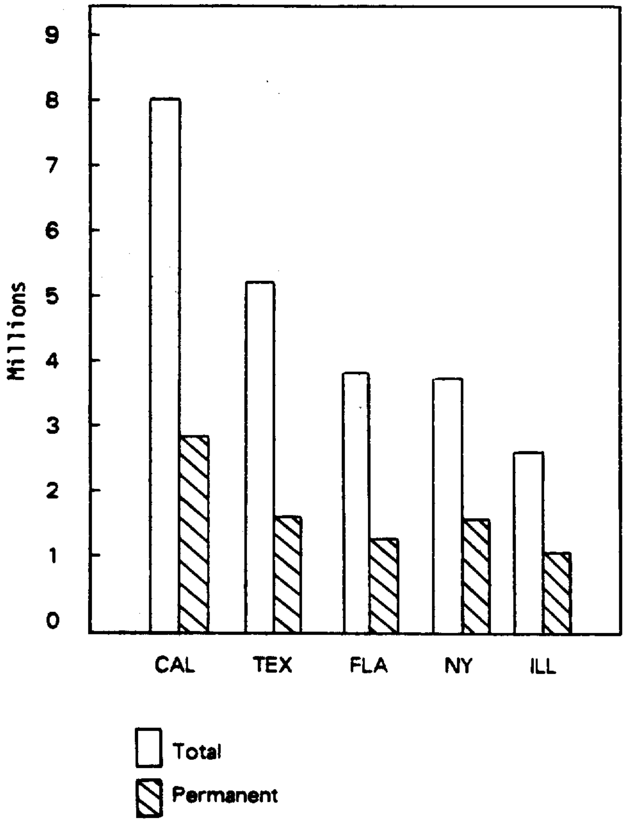
Figure 2. Comparison of New Hire Forecasts with Actual New Hire Data

State	Period	New Hires Reported by State Employment Agencies	Predicted New Hires	% Difference
Arkansas	Fiscal 1979	583,990	603,500	+3.34
Pennsylvania	Fiscal 1976	2,051,553	2,147,100	+4.66
South Dakota	Fiscal 1979	177,433	155,800	-12.19
	Fiscal 1980	142,795	137,500	-3.70
	Fiscal 1981	134,109	142,900	+6.57
Idaho	Fiscal 1976	238,989	241,000	+0.84
California	Fiscal 1976	6,142,625	5,796,000	-5.64
	Fiscal 1977	6,625,804	6,506,800	-1.80
	Fiscal 1978	7,523,644	7,640,400	+1.55
	Fiscal 1979	8,366,534	8,226,400	-1.67
North Dakota	Fiscal 1976	147,081	144,300	-1.88
North Carolina	1979 - 4th Q.	392,663	370,300	-5.71
Nevada	Fiscal 1976	309,100	298,300	-3.48
	Fiscal 1979	452,679	476,800	+5.32
	Fiscal 1980	464,348	466,600	+0.48
	Fiscal 1981	438,880	477,600	+8.95
South Carolina	1979 - 1st-3rd Q.	611,324	627,700	+2.68
	1981 2nd-4th Q.	550,619	522,900	-5.03
Maine	Fiscal 1978	263,175	268,900	+2.17
Illinois	1979 3rd-4th Q.	1,436,475	1,593,500	+10.93
New Mexico	Fiscal 1979	410,927	412,000	+0.26
	Fiscal 1980	378,288	386,200	+2.10
Missouri	1979 -3rd-4th Q.	718,946	670,400	-6.75
	Calendar 1981	1,073,311	1,204,900	+12.26
Iowa	Fiscal 1981	587,016	582,500	-0.77
Mississippi	1981 4th Q.	101,921	107,400	+5.40
Florida	Calendar 1980	2,673,019	3,790,500	+41.81
	Calendar 1981	2,918,487	3,729,700	+27.80

Figure 3. Number of New Hires in the Private Nonfarm Economy by State
(annual totals in thousands)

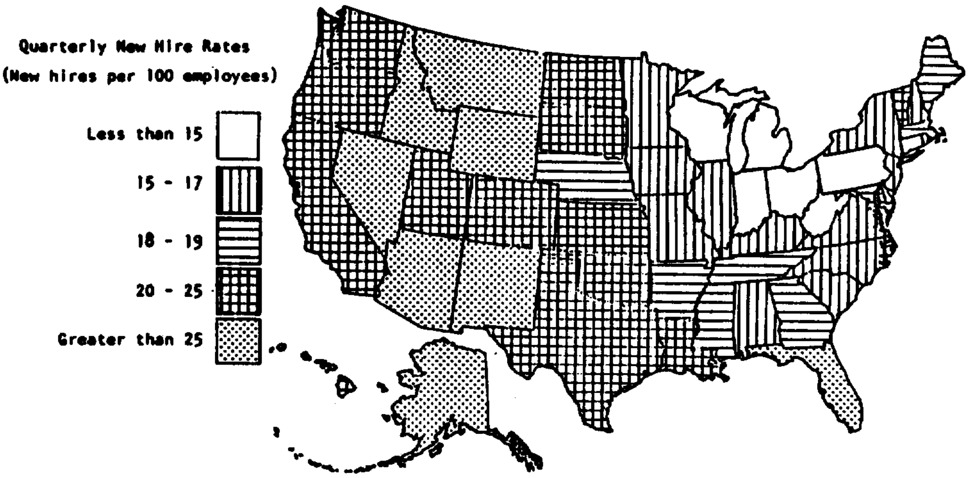
State	1975	1976	1977	1978	1979	1980	1981	1982	1983	1984	1985
Alabama	610.9	693.3	765.6	878.3	895.3	808.1	770.0	620.8	598.7	744.9	829.9
Alaska	178.8	112.3	92.8	102.6	109.2	111.7	126.8	136.2	171.3	162.5	172.3
Arizona	514.5	624.0	735.7	903.1	1005.0	891.5	889.8	730.0	789.2	921.3	1019.3
Arkansas	378.8	443.0	498.7	580.6	606.2	534.0	503.7	392.4	364.2	472.8	541.5
California	5219.2	6059.2	6811.6	7838.4	8294.0	7770.4	7760.8	6700.8	6743.6	8001.6	8532.8
Colorado	656.3	813.2	951.6	1148.8	1242.7	1140.8	1097.3	888.8	913.8	1190.4	1371.7
Connecticut	530.9	653.1	719.2	829.0	852.6	795.4	757.8	642.0	654.6	781.2	862.5
D.C.	183.9	164.2	174.4	196.3	205.8	187.0	180.8	147.8	151.3	174.2	186.5
Delaware	126.4	182.4	203.2	243.4	259.2	220.3	208.1	140.3	138.5	183.7	211.3
Florida	2006.5	2567.0	2968.8	3614.6	3884.4	3790.5	3729.7	3104.6	2983.2	3778.2	4162.8
Georgia	1031.3	1226.7	1388.1	1667.6	1720.3	1547.1	1437.8	1143.7	1223.9	1464.5	1634.2
Hawaii	179.6	205.6	219.9	263.0	268.2	254.5	238.6	213.0	221.4	265.2	278.5
Idaho	207.4	250.0	261.4	289.9	282.7	260.9	260.3	247.2	274.5	290.2	303.1
Illinois	2195.1	2718.2	2813.5	3178.2	3172.0	2826.4	2639.7	2074.3	2241.8	2632.6	2761.8
Indiana	876.1	1090.2	1192.6	1393.0	1351.4	1115.6	1078.8	799.8	844.7	1062.0	1153.4
Iowa	485.4	547.4	602.0	691.4	718.5	633.8	580.4	453.0	405.5	518.4	596.0
Kansas	507.6	564.0	612.7	694.4	726.4	655.0	646.4	532.3	521.9	623.2	680.3
Kentucky	563.8	647.9	733.1	826.1	785.0	689.9	659.4	541.6	633.7	687.9	739.6
Louisiana	868.4	1019.4	1092.7	1239.0	1308.0	1298.7	1295.5	1167.1	1084.5	1240.2	1369.5
Maine	201.8	241.0	247.6	278.0	277.4	261.7	247.4	222.8	237.4	267.8	272.5
Maryland	774.1	859.4	981:1	1111.2	1077.0	1008.1	976.8	872.3	927.6	1007.8	1053.0
Massachusetts	1181.1	1416.4	1535.9	1707.1	1768.3	1697.5	1652.0	1422.0	1525.1	1717.6	1812.5
Michigan	1377.7	1639.4	1862.6	2110.8	2036.3	1676.4	1574.2	1177.4	1292.1	1574.6	1728.0
Minnesota	713.2	823.4	905.2	1065.5	1123.6	993.8	950.0	766.8	766.2	958.9	1074.3
Mississippi	403.6	460.4	517.6	566.2	577.7	507.2	501.5	406.8	429.4	514.0	558.6
Missouri	947.7	1096.5	1201.8	1347.0	1366.0	1184.6	1204.9	1019.6	959.4	1158.4	1251.6
Montana	167.2	213.5	208.6	241.4	219.7	201.2	224.6	183.8	200.5	241.8	249.0
Nebraska	318.7	366.4	381.3	418.9	441.6	399.7	394.3	336.2	317.7	381.7	406.1
Nevada	255.1	319.4	380.8	473.4	488.6	459.4	470.8	400.0	442.6	575.5	637.0
New Hampshire	153.5	208.2	236.8	277.1	292.2	255.4	253.0	183.3	202.5	256.6	278.6
New Jersey	1396.3	1648.9	1793.3	2038.4	2069.0	1917.8	1878.5	1572.4	1628.0	1887.0	2041.0
New Mexico	269.6	312.5	358.1	399.3	414.6	378.9	384.8	334.4	338.0	407.4	454.4
New York	3211.6	3568.6	3809.7	4285.6	4391.2	4072.8	4015.6	3356.4	3296.5	3719.8	4014.8
North Carolina	1035.5	1225.3	1377.6	1622.5	1707.0	1741.9	1378.6	1027.8	1072.5	1366.1	1512.6
North Dakota	134.8	140.5	138.7	160.2	161.8	141.1	159.9	145.9	171.8	185.3	190.0
Ohio	1702.0	2077.8	2324.2	2632.2	2622.1	2204.4	2152.4	1653.7	1595.4	2062.7	2247.7
Oklahoma	632.1	715.4	775.4	904.6	933.1	942.4	972.8	820.1	836.3	975.2	1072.2
Oregon	562.8	661.6	744.8	839.5	884.7	750.3	697.0	592.8	625.4	748.8	813.8
Pennsylvania	1864.6	2214.4	2330.8	2717.4	2651.9	2285.8	2289.0	1628.0	1840.2	2118.2	2271.6
Rhode Island	187.4	223.7	244.6	279.3	286.8	258.4	244.0	188.8	188.5	236.0	265.1
South Carolina	501.3	594.0	649.8	764.0	797.9	719.9	684.2	536.7	508.5	654.4	725.6
South Dakota	119.1	139.9	147.9	159.8	155.1	134.2	141.2	122.7	141.0	155.8	163.8
Tennessee	838.4	975.2	1091.6	1231.3	1223.3	1079.2	1072.2	869.2	850.8	1063.3	1135.4
Texas	3369.7	3886.5	4228.4	4909.2	5327.6	5266.0	5359.2	4772.0	4376.8	5132.4	5760.4
Utah	287.7	337.3	366.9	425.1	434.8	395.6	402.6	350.8	359.1	434.4	475.6
Vermont	96.3	113.2	123.7	136.9	133.7	125.3	126.0	123.0	143.8	149.4	152.7
Virginia	848.0	1014.2	1124.8	1308.3	1378.2	1234.8	1151.6	953.5	912.0	1172.3	1303.1
Washington	828.6	952.7	1034.8	1174.8	1212.2	1085.9	1072.0	1026.9	1156.8	1313.9	1364.2
West Virginia	267.3	288.8	307.0	337.4	353.1	324.6	299.3	259.6	237.7	272.8	294.8
Wisconsin	702.2	830.5	922.1	1075.1	1121.9	932.2	898.1	683.0	667.8	856.0	973.9
Wyoming	123.9	147.2	166.3	193.0	209.8	210.6	212.2	192.7	189.0	222.8	235.9
U.S. Total	42794.9	50296.0	55356.0	63768.0	65824.0	60108.0	58904.0	48876.0	49396.0	58984.0	64196.0

Figure 4. States with the Highest Number of New Hires, 1984



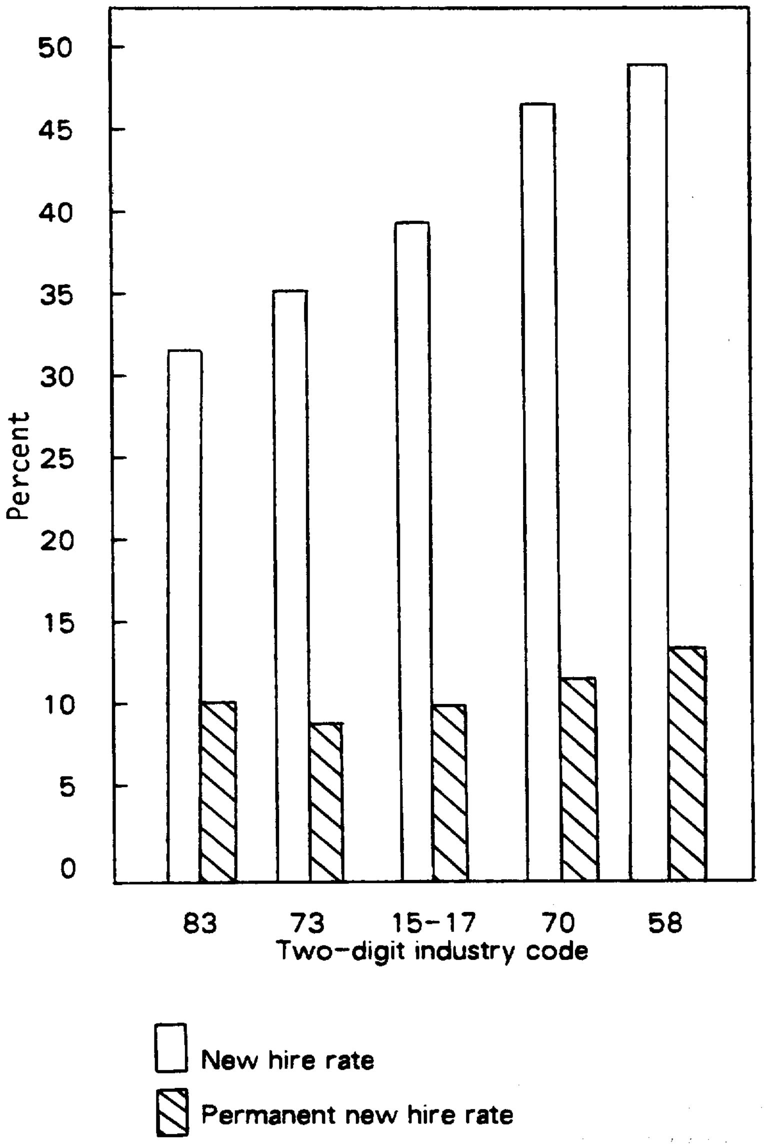
Source: Institute of Labor and Industrial Relations, University of Michigan, November 1983

Figure 5. Projected Quarterly New Hire Rates, 1984.



SOURCE: Institute of Labor and Industrial Relations, University of Michigan. November 1983.

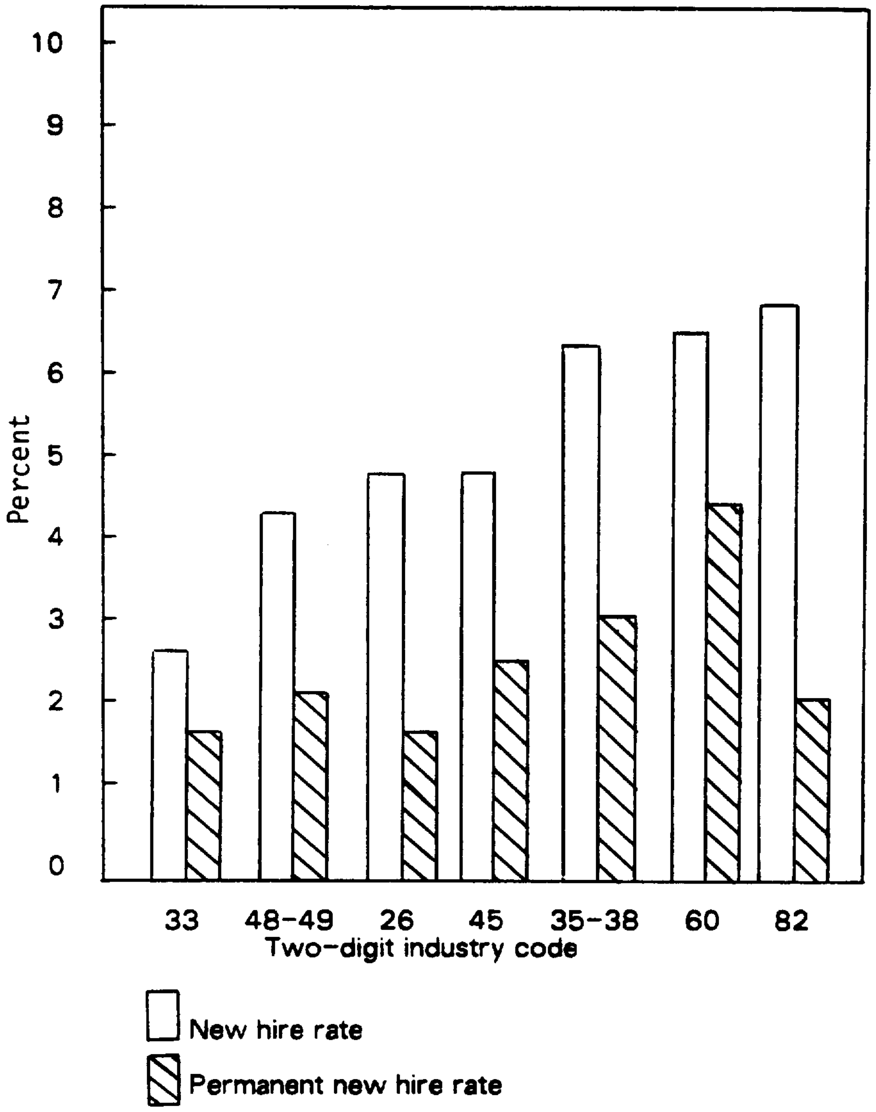
Figure 6. Industries with Highest New Hire Rates, 1975 2nd Quarter



83	- Social services
73	- Business services
15-17	- Construction
70	- Hotel and other lodging
58	- Eating and drinking establishments

Source: Institute of Labor and Industrial Relations, University of Michigan

Figure 7. Industries with Lowest New Hire Rates, 1975 2nd Quarter



33	- Primary metal manufacturing
48-49	- Communications and public utilities
26	- Paper manufacturing
45	- Air transportation
35-38	- Machinery + transportation + instrument manufacturing
60	- Banking
82	- Educational services

Source: Institute of Labor and Industrial Relations, University of Michigan

DISCUSSION

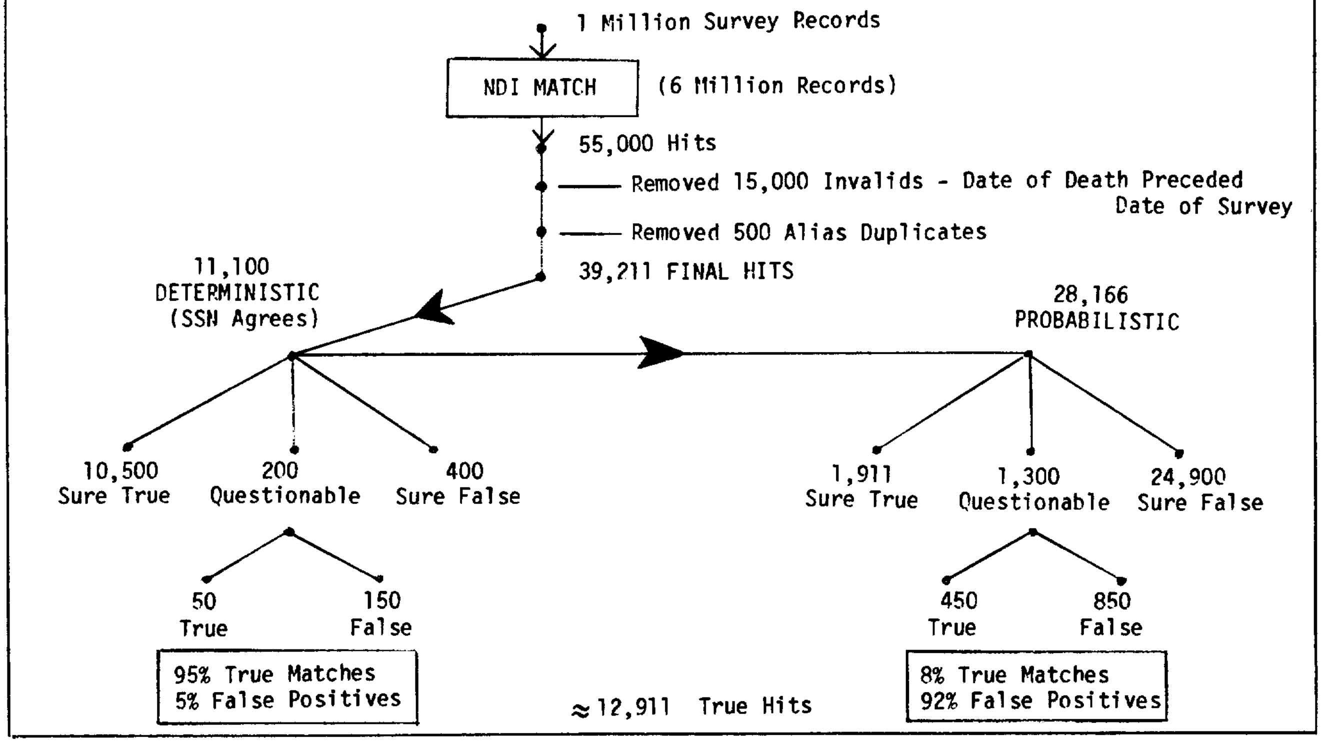
Norman J. Johnson, U.S. Bureau of the Census

I would like to present my discussion of these three papers in terms of points which we have encountered in an application of matching from our project. I have been working on developing the data base for The National Longitudinal Mortality Study (NLMS). This study is being conducted jointly by the National Heart, Lung, and Blood Institute, the National Center for Health Statistics and the U.S. Census Bureau. The primary objectives of the NLMS are to analyze socioeconomic, demographic and occupational differentials in mortality within the United States. A major interest of our analysis will be to compare survival rates of different subsets of the cohorts.

The study population consists of eight cohorts of selected Census samples. Deaths in this population are identified through periodic matching to the National Death Index (NDI), the index discussed in the first paper by Mr. Patterson. As pointed out in that presentation, in terms of number of records submitted for matching, our project is a major user of the National Death Index. The National Longitudinal Mortality Study currently consists of approximately 1 million records from eight cohorts. One match has been made to the NDI, which at the time consisted of approximately 6 million records. We intend to conduct follow-up matches approximately every two years.

The process we used to obtain the final matched records was completed in two steps. First, our files were matched to the NDI using the NCHS criteria. Then, an extensive screening was made of the resulting match using some of the methodologies discussed in presentations given earlier in these sessions to determine the final true match status. This second step involved both computer and manual matching. Our approach in the computer matching phase was similar to that used in the CAMLIS project of Mr. Arellano, the presenter of the second paper of this section. A link was made deterministically for all matches in which there was an exact agreement on social security number. Records not matched deterministically were then matched probabilistically using a modified Newcombe model. Weights for this model were estimated from a subsample of records from the NCHS match which had been reviewed manually to establish correct match status. Three categories of records from the probabilistic match resulted: true, false and questionable matches. Questionable matches were decided on the basis of a manual review. This process and the final results have been schematically diagrammed in Figure 1. From the initial one million records, approximately 12,900 links occurred. The information in the figure also indicates the substantial difference in the true match rate between the deterministic and the probabilistic steps.

Figure 1. -- National Longitudinal Mortality Study



PATTERSON AND BILGRAD

As I mentioned in my introduction, our project is a major user of the National Death Index. Deaths in our cohorts are determined by linking our records to records in this Index. The NDI matching algorithm is, in a sense, deterministic. It uses combinations of five major variables in seven criteria to determine a link. These criteria are soon to be expanded to twelve. A link is made if any one of the seven criteria is satisfied. As other studies continue to match using this index, the NDI may wish to incorporate some probabilistic components into their matching procedure based on the experience of their users. Results from our project may be helpful in this regard.

Five major categories of users were summarized in the presentation. The major users identified are in health-related fields. In many health studies, analysis is done by comparing survival of cohorts, as is the case in our study. Rare events are often of interest and small counts may be greatly affected by match rates. For this reason, in our study, we feel that matching algorithms should put emphasis on detecting true matches, with willingness to manually review more questionable matches, in order to rule out false positives. The additional criteria made available in the new NCHS matching algorithm are a step in the right direction. The expanded criteria will generate more true links as well as more false positives.

The paper presents results of studies to measure the improvements in the match rate to the NDI due to the replacement of the Soundex Code for matching of names by the NYSIIS code. If the NCHS studies of the effects of this change are true, that is, 18 percent fewer true matches and 31 percent fewer false matches could be expected, then, in view of the comments which I made earlier, the Soundex Code would be preferable to us.

ARELLANO

I will focus my discussion on the three points mentioned in the conclusion section of the paper. The paper deals with the use of the Fellegi-Sunter approach in the CAMLIS project to link user files to death certificates from the state of California. The first point discussed concerns the potential for making estimates of error terms in the Fellegi-Sunter model. The estimation of error terms is a major difficulty encountered in application of the theory. In some applications, making simplifying assumptions is the only way to obtain estimates of errors. The similarity of the CAMLIS study and the National Longitudinal Mortality Study may enable us to exchange estimated parameter values once they are obtained.

The conclusion on the robustness of error probability estimates is important and potentially very useful. This quality of the estimates would allow the use of approximate values without great risk of poor matching results and permit a more frequent borrowing of parameter values from other studies. A nice collection of results in the literature demonstrating this robustness would be very useful.

The third point covered in the conclusion deals with the effects of bias. We have observed a positive bias in our scoring algorithm. It would be helpful for us to know if the CAMLIS project has identified any consistent bias in their procedure. If so, what explanation do they have for it?

COHEN

The findings of this particular study are based on the results of a match of two files performed by a Government agency. The match was based on an apparently deterministic match procedure using a certain identification number. The provider of such match results should advise clients of error rates and nonmatch results of similar studies. Error rates of such matches should be required as part of publications and presentations in order to give the reader a chance to determine if any biases have resulted due to the matching procedure. This is similar to documenting which computer and software were used when publishing papers based on computer simulation. In this paper, matching determines the study and data base. What is the error rate in the identification number in both files? Errors in deterministic match variables are more important than in probabilistic match variables. The paper does compare the finding of this study with those of other sources to demonstrate that the match was effective.

The question of what impact effective matching algorithms have on the confidentiality of person records was mentioned in the paper. The law provides specific statements on this subject. Some confidentiality problems were discussed in an earlier session. By linking data from several sources, individual records can be identified more easily. In the case of data collection at the Census Bureau, there is an additional concern. The Bureau is a passive collector of data. Cooperation of the respondent is of crucial importance in obtaining reliable information. As the public becomes aware of our ability to link records from several Governmental agencies, response rates to our questionnaires may decrease, become biased, and possibly inaccurate due to the fear of person-record identification. This is in spite of the potential to provide more beneficial information than would exist without the linked records.

ON MATCHING WITH PERSONAL NAMES

J. T. Kagawa, Cancer Research Center of Hawaii
M.P. Mi, University of Hawaii, Honolulu

In the record linkage process, personal names are important matching criteria for comparing documents to identify information belonging to the same individual or family. The discriminating power of the surname, given name, and middle name for linkage varies depending on the frequencies of various possible configurations in the population. Although the total number of possible configurations of personal names is extremely large, the distribution of these configurations are not uniform.

Due to the many people of different nationalities in Hawaii, the name structure has become very diverse and therefore, offers a good opportunity to study the name configurations that are available in the population. Migratory waves of contract laborers and others seeking new opportunities introduced many new names to Hawaii. Often times, names written in Chinese or Japanese characters had to be phonetically translated and anglicized by immigration officers who had little or no knowledge of these languages. This process created further heterogeneity and inconsistencies within names. It is not uncommon to find two or more different names derived from the same character or to find that one surname was actually derived from two completely different characters. Names were also shortened or modified if they were too difficult to pronounce.

In an attempt to develop an optimal strategic approach for computerized linkage of various documentary sources, studies are being conducted to elucidate the variation in personal names in the population. Some pertinent questions to be answered are: 1) how many possible configurations for surname, given name, and middle initials there are in each racial group? 2) how are these configurations distributed in the population? and 3) is there any evidence of time trends in these distributions or name patterns? Preliminary results from the analysis of the 1942-43 Hawaii Population Registration are presented in this report.

MATERIALS AND METHODS

The Population Registration was conducted in Hawaii during 1942-1943 under martial law. There were a total of 439,601 residents registered and fingerprinted. Eight major racial groups were selected including Caucasian, Hawaiian, Portuguese, Chinese, Filipino, Japanese, Puerto Rican, and Korean. The description of each of these racial groups in Hawaii was given previously by Adams (1937), and Lind (1955).

Recorded configurations for surname, given name and middle intials were tabulated separately by sex and race directly from the 1942-1943 population. For each of the eight racial groups, the name configurations were grouped into four types based on the relative frequency in the registration file. The first type was for unique configurations. The next type was for configurations with a relative frequency less than 0.1 percent. The third type was for configurations of fairly frequent appearance equal to or greater than 0.1 percent but less that 1 percent. Lastly, any configuration with a relative frequency of 1 percent or greater was considered in the fourth group. Since the number of configurations was tabulated directly from the data, which were subject to errors in reporting and recording, possible errors could have been included. Errors could have occurred by insertion, substitution, deletion, and switching of one or more alphabetic letters and such an alteration could or could not be a valid configuration. It was therefore assumed for this analysis that most errors are made accidently, presumably at random, and the altered configuration should be unique.

The relative frequency for each of the configurations for surname, first name, and middle initials was calculated. The relative frequency of the ith configuration is $p_i=m_i/M$, where M is the total number of individuals in the population and m_i the number of individuals having the ith configuration. The probability that two individuals randomly sampled from the population would match on the ith configuration is p_i^2. This also approximates the probability of a chance match for the ith configuration when two documentary sources of vital events from the population are brought together for linkage. The sum of these probabilities over all configurations, that is Σp_i^2, is the probability of a chance match on any configuration for a given criterion. Therefore, the greater the total probability, the less discriminating is the linkage criterion among individuals.

RESULTS AND DISCUSSION

Table 1 gives the number of males and females in each racial group. These groups represented 83 percent of the total population in 1942. The Japanese group was the largest, accounting for 37 percent, and larger than any other two groups combined. The Caucasian group ranked second, followed by the Filipino, Portuguese, Chinese, Hawaiian, Puerto Rican, and Korean. These groups and outcrosses among these groups have contributed to the ethnic diversity of Hawaii's present population.

The surname distributions are shown in Table 2. Data on females were not used because of the possible inclusion of their married surname. The total number of surnames varied greatly from one race to another. There were only 241 configurations in the Korean group, while the Filipino group had approximately 60 times more configurations. There were no common names in the Filipino group based on the relative frequency of 1 percent or greater. There were a total of only five common names representing only a very small proportion of individuals in the Caucasian, Hawaiian, and Japanese groups. Conversely, a large number of individuals shared more than 12 common names in the Korean and Chinese groups. The total probability of chance match also differed markedly among the eight racial groups. The probability of match between two individuals randomly selected from the population was approximately 6 in 10,000 for the Filipinos as compared to the estimate of 850 in 10,000 for the Koreans. In the Korean group, about one-half of the subpopulation shared four common surnames, namely: Kim (22.4%), Lee (15.2%), Park (6.8%), and Chung (4.5%). A high probability equal to 293 in 10,000 was also found for the Chinese group. There were 25 common surnames shared by 68 percent of the Chinese population. The most common Chinese surnames being Wong (8.1%), Lee (6.3%), Chung (5.2%), Ching (5.1%), and Chang (5.1%).

The distribution of the given name for each racial group is shown in Table 3. The ratio of the number of surname configurations to the number of given names varied from race to race. For the Caucasian, Portuguese, and Hawaiian groups, there were a greater number of surname configurations than given names. This relationship was completely reversed for the Chinese and Koreans. The Japanese and Puerto Rican groups had approximately the same number of surnames and given names. As shown in the table, there were very few common given names. However, these common names accounted collectively for a significant portion of each of the subpopulations. For males, the percentage of the population sharing common names was 65 for the Portuguese, 62 for the Hawaiian, 49 for the Puerto Rican, and 46 for the Caucasian. Among the females, the percentage estimates were lower, varying from 25 to 43. In the Chinese, Japanese, and Korean groups the common given names for males and females were of Western origin. Yoshiko, being a common given name of Japanese origin among the Japanese females was the only exception. As shown with surnames, the probability of chance match for the given name as a matching criterion also varied from race to race. The highest value was 323 in 10,000 for the Portuguese males and the lowest was 33 in 10,000 for the Japanese females. The Portuguese and Hawaiians showed the highest probabilities of chance match for both the male and female given names.

The possibility of time trends of selecting given names was also tested based on the 1942 population file. The recorded given names were tabulated by sex and age for each of the eight racial groups. The age groups were 0-19, 20-49 and 50-99. Except for native Hawaiians, individuals with birth years between 1843-1892 were mainly those who immigrated to the islands. The other two age groups were comprised of a mixture of later arriving immigrants and individuals born in Hawaii. A given name was determined popular if the relative frequency was 1.0 percent or greater of the total number of individuals in each race. The distributions based on age groups also showed variations among the different racial groups.

The majority of the given names of the oldest age groups were the names from their native country. With the influence of Western culture, the given names of the younger age groups showed the trend towards adopting the popular English names of the times. It was also observed that the names in the 20-49 age group of the Japanese continued to be largely Japanese. Although still of Japanese origin, the names were quite distinguishable from those of the older generation. Also the selection of Spanish names for the Filipino group prevailed over the three age groups. The popular English male given names among the racial groups remained unchanged throughout the years. The popular female names showed more distinctive periods of rise and decline, which may be attributed to the influence of literary characters and famous people.

Two middle initials were recorded for individuals registered in the 1942 population file. The middle initials distributions are shown in Table 4. The blank configuration represented 44 percent in the males and 37 percent in the females of the eight racial groups analyzed. The blank response indicated either missing information or a valid configuration. Many immigrants to Hawaii from China, Japan, and Korea did not have middle names. Out of the total possible configurations, the Chinese had the largest number of different combinations for both males and females. Middle initials for the Chinese and Korean groups, mostly comprised of double initials, generated a large number of possible configurations. The frequency of uncommon middle initials was reflected in the lower probability of chance match for both of these groups. The frequencies of common middle initials were high in the remaining racial groups.

The observed variations in name patterns among the different racial groups in Hawaii provides a unique testing ground for the study of record linkage methodology. The analysis of the 1942 Hawaii Population Registration file showed that the distributions of the configurations for surnames, given names, and middle initials were definitely nonuniform. Personal names for the different racial groups maintained varying degrees of discriminating power. A study is being planned to analyze the name structure of the present Hawaii population. There has undoubtedly been many more new names introduced into the population.

REFERENCES

1. Adams, R. 1937. Interracial Marriage in Hawaii. New York: MacMillan. pp. 353.
2. Lind, A.W. 1955. Hawaii's People. Honolulu: University of Hawaii Press. pp 121.

Table 1. Size of Subpopulations

Sex	Racial Groups[1]							
	CAU	PTG	HAW	CHI	FIL	JAP	POR	KOR
No. individuals								
Males	34566	15790	7752	16118	40323	84298	4372	3786
Females	25988	15886	7321	12426	10946	78669	3385	2738

[1] CAU = Caucasian; PTG = Portuguese; HAW = Hawaiian; CHI = Chinese; FIL = Filipino; JAP = Japanese; POR = Puerto Rican; KOR = Korean.

Table 2.--Distribution of Surnames by Racial Groups

Sex / Type[2]	Racial Groups[1]							
	CAU	PTG	HAW	CHI	FIL	JAP	POR	KOR
				Number of Configurations				
Males								
Unique	8548	866	896	240	8960	1111	553	101
Rare	4658	546	943	205	5341	3831	199	48
Fair	79	167	231	76	73	192	157	74
Common	1	16	1	25	0	3	15	18
All	13286	1595	2071	546	14374	5137	924	241
			Σp_i					
Males								
Common	0.01	0.29	0.01	0.69	0.00	0.03	0.32	0.72
Other	0.99	0.71	0.99	0.31	1.00	0.97	0.68	0.28
			$\Sigma p_i^2 \times 10^{-2}$					
Males								
All	0.07	0.83	0.15	2.93	0.06	0.20	1.20	8.50

[1] See Table 1.

[2] Unique = single count in the population; Rare = 0.01% - 0.09%; Fair = 0.10% - 0.99%; Common = 1% or greater.

Table 3.--Distribution of Given Names by Racial Groups

Sex / Type[2]	Racial Groups[1]							
	CAU	PTG	HAW	CHI	FIL	JAP	POR	KOR
				Number of Configurations				
Males								
Unique	1512	432	619	3798	2971	4883	467	1664
Rare	905	239	217	1054	1266	3795	168	253
Fair	113	81	71	99	219	153	98	86
Common	20	23	21	15	7	9	22	14
All	2550	775	928	4966	4463	8840	755	2017
Females								
Unique	1866	723	680	2030	1486	1963	393	730
Rare	869	412	235	570	656	1882	108	99
Fair	165	136	116	137	206	228	138	147
Common	14	15	19	17	5	4	18	13
All	2914	1286	1050	2754	2353	4077	657	989
				Σp_i				
Males								
Common	0.46	0.65	0.62	0.23	0.13	0.13	0.49	0.20
Others	0.54	0.35	0.38	0.77	0.87	0.87	0.51	0.80
Females								
Common	0.25	0.32	0.43	0.24	0.09	0.04	0.36	0.23
Others	0.75	0.68	0.57	0.76	0.91	0.96	0.64	0.77
				$\Sigma p_i^2 \times 10^{-2}$				
Males, all types	1.69	3.23	2.82	0.51	0.49	0.40	1.96	0.43
Females, all types	0.77	1.80	1.59	0.57	0.40	0.33	1.39	0.71

[1] CAU = Caucasian; PTG = Portuguese; HAW = Hawaiian; CHI = Chinese; FIL = Filipino; JAP = Japanese; POR = Puerto Rican; KOR = Korean.

[2] Unique = single count in the population; Rare = 0.01% - 0.09%; Fair = 0.10% - 0.99%; Common = 1% or greater.

Table 4.--Distribution of Middle Initials by Racial Groups

Sex / Type[2]	Racial Groups[1]							
	CAU	PTG	HAW	CHI	FIL	JAP	POR	KOR
				Number of Configurations				
Males								
Unique	122	64	50	72	96	52	15	73
Rare	134	22	22	219	24	8	2	59
Fair	1	4	13	120	7	10	7	92
Commmon	20	17	11	8	17	11	16	5
All	277	107	96	419	144	81	40	229
Females								
Unique	118	84	47	91	96	80	18	73
Rare	107	59	37	179	31	78	2	29
Fair	3	7	16	137	7	11	8	89
Common	20	15	9	18	17	12	14	20
All	248	165	109	425	151	181	42	211
				Σp_i				
Males								
Blanks	0.17	0.39	0.38	0.46	0.34	0.60	0.54	0.61
Common	0.81	0.58	0.55	0.10	0.63	0.36	0.43	0.06
Others	0.02	0.03	0.07	0.44	0.03	0.04	0.03	0.33
Females								
Blanks	0.14	0.30	0.23	0.20	0.39	0.49	0.43	0.31
Commmon	0.83	0.64	0.70	0.32	0.57	0.45	0.52	0.39
Others	0.03	0.06	0.07	0.48	0.04	0.06	0.05	0.30
				$\Sigma p_i^2 \times 10^{-2}$				
Males								
Blanks	2.83	15.35	14.67	21.16	11.57	35.36	28.60	37.13
Common & Others	4.12	2.35	10.46	0.28	2.92	1.60	1.54	0.19
All	6.95	17.70	25.13	21.44	14.49	36.96	30.14	37.32
Females								
Blanks	1.81	9.12	5.25	3.81	15.34	23.79	18.30	9.89
Common & Others	5.25	3.54	14.88	0.96	2.36	2.12	2.69	1.02
All	7.06	12.66	20.13	4.77	17.70	25.91	20.99	10.91

[1] CAU = Caucasian; PTG = Portuguese; HAW = Hawaiian; CHI = Chinese; FIL = Filipino; JAP = Japanese; POR = Puerto Rican; KOR = Korean.

[2] Unique = single count in the population; Rare = 0.01% - 0.09%; Fair = 0.10% - 0.99%; Common = 1% or greater.

SURNAME BLOCKING FOR RECORD LINKAGE

F. Quiaoit, Cancer Research Center of Hawaii, and
M.P. Mi, University of Hawaii, Honolulu

In the linkage between two documentary sources, each record from one source is compared with all the records in the other source. For one-file linkage involving a single source, each record is compared with all other records except itself. In either case, the number of such pair-wise comparisons becomes extremely large even if the size of the documentary source is moderate. The fact that only a small fraction of these comparisons are meaningful emphasizes the need for the grouping of records based on one or more selected items of identifying information. This is known as blocking. Once blocks are formed, the comparison of records is only made between the two corresponding blocks for two-file linkage or within the block for one-file linkage.

In principle, any identifier may be used as a blocking criterion. Surname is often selected for this purpose. Blocking may be made on the whole or part of the surname configuration. The use of a phonetic code on the surname for blocking has become popular in many applications. The objective of the present study was to evaluate the performance of several blocking methods based on prevalent name patterns in various racial groups in a multi-ethnic population, and to test the effects of blocking on linked pairs in which one or both records had known reporting or recording errors in the surname field.

MATERIALS AND METHODS

Data on surnames from the complete 1942-43 Population Registration in Hawaii were used. There were a total of 439,601 individuals registered and fingerprinted under martial law. Eight major racial groups were selected including Caucasian, Portuguese, Hawaiian, Chinese, Filipino, Japanese, Puerto Rican, and Korean. All recorded surname configurations for male subjects were analyzed in the present study. Two methods, namely: the New York State Identification and Intelligence System (NYSIIS) and the Russell's Soundex system were chosen to pre-code surnames phonetically. Under each method, records were blocked with the same code. These two systems were compared specifically to the other five methods of blocking, namely, by the whole surname, first character of surname, first two, three, or four characters of surname, respectively. Criteria such as the total number of blocks formed, distribution of block size, and surname information in matching were used for evaluation.

A set of known linked record pairs was obtained from the linkage project between the 1942 Population Registration file and the death file (1942-79) in Hawaii. It consisted of all male subjects aged 60 and over in the 1942 population who died during the 38-year period from 1942 to 1979. A total of 11,367 linked pairs were established by computer as well as by manual search (Mi et al., 1983). Pairs, in which recorded surname and first name were switched, were excluded. There were 672 pairs with various error conditions in surname. The concordance rate of each method, which is the percentage of record pairs that were properly placed in the same block regardless of these errors, was used for comparison.

RESULTS AND DISCUSSION

The number of male subjects in the 1942 Population Registration is shown for each racial group in Table 1. The total number of recorded configurations for surname varied greatly among racial groups ranging from only 241 in the Korean group to 14,374 among the Filipino. The average number of individuals possessing the same surname varied from 2.6 for the Caucasian group to 29.5 for Chinese men. The value for each racial group was also the average block size when blocking was based on the whole surname of twelve characters. Most of the surname configurations were unique, having only a single representation in the population. These unique configurations included rare spelling variations, and errors in reporting and recording. When a part of the surname was used for blocking, records having the same leading characters in their surname fields were grouped together. As shown in Table 1, the number of blocks increased from an initial maximum of 26, based on the first character of the surname, to several hundreds or thousands using more leading characters for blocking. However, the magnitude of increase was not linear for each additional character used, and varied from one race to another. The distribution of blocks by size also changed. When the whole surname was used for blocking, most blocks were small with 10 or less records. If blocking was based on the first character of surname, the block size increased tremendously. If more leading characters were used, the number of records in each block decreased as expected. The performance of the first four characters of surname for blocking was comparable to the NYSIIS and Soundex method in the percentage distribution of blocks by size in all groups except the Chinese and Koreans. The NYSIIS and Soundex method produced a much higher percentage of large blocks of over 50 records in the Chinese and Korean groups. This was because almost all the Chinese and Korean surnames were five characters or less in length.

It should be emphasized that block size is an important consideration in the choice of a blocking method for linkage. Since the number of pair-wise comparisons is equal to the product of the size of two corresponding blocks in two-file linkage and to the product of the block size and block size minus one in one-file linkage, a larger block size will greatly affect the cost of a linkage.

The other criterion which deserves attention is the loss of surname information in matching by blocking. Suppose that there is no blocking and the whole documentary source or file is used as a giant block for pair-wise comparison. The amount of information provided by surname in matching is approximately $1 - \Sigma p_i{}^2$ where p_i is the relative frequency of the i^{th} surname configuration and $\Sigma p_i = 1$. The squared term represents the probability of chance match on the i^{th} configuration. When summed over all configurations, the squared term gives the total probability of chance match in surname. The exact probability of chance match is $1 - \Sigma p_i p_i'$ in the two file linkage where p_i' is the relative frequency of the ith configuration in the second source. If all individuals have the same surname, that is, $p_i = 1$, every record pair must agree on surname and the total probability of chance match reaches the maximum of 1. Under this special condition, surname clearly provides no information. On the other hand, if each individual record has a different surname, the probability of chance match is minimal and the amount of information provided by surname reaches the maximum. When blocking is made based on surname (a part or whole), the newly structured block consists of records of one or more surnames, each with the relative frequency of p_{ij}, the j^{th} surname within the i^{th} block. The relative frequency of the i^{th} block is q_i, and the probability of chance match for records with the i^{th} blocking criterion is $q_i{}^2$. The probability of chance match on surname within newly structured blocks is $\Sigma\Sigma p_{ij}{}^2/\Sigma q_i{}^2$, and the amount of information of surname in matching is estimated by $1 - \Sigma\Sigma p_{ij}{}^2/\Sigma q_i{}^2$. Suppose that the whole surname is used for blocking. Because each block is characterized by a different surname, obviously $\Sigma\Sigma p_{ij}{}^2/\Sigma q_i{}^2 = 1$, therefore surname is no longer informative and provides no discrimination among records within any block in which pair-wise comparisons are made.

The average and maximum number of surnames per block and the estimates of surname information in matching under various blocking methods are given in Table 2. When blocking is based on the first character, the amount of surname information was generally high except for the Korean group. The probability of chance match on surname was estimated to be 0.085, the highest among the eight racial groups studied (Kagawa and Mi, 1985). The amount of information decreased rapidly, particularly in the Chinese group, as the number of leading characters for blocking increased. When blocking is based on the NYSIIS and Soundex codes, the amount of information was close to those estimates derived from the blocking based on the first four characters in several racial groups. These phonetic coding methods seemed to be desirable especially for the Chinese and Korean groups, but not for the Japanese. The concordant rate was defined as the percentage of total pairs in which both members were blocked concordantly by a given method. Table 3 gives the estimates of the concordant rate for the four selected methods. The rate over all racial groups was 56.7, 43.9, 56.4, and 64.9 percent, respectively, for blocking based on the first three characters, first four characters, NYSIIS code, and Soundex code of surname. Both NYSIIS and Soundex methods consistently produced a high concordant rate in all racial groups. Because Chinese and Korean surnames are generally short (composed of three to five characters), errors would have to occur in the first few characters. It was anticipated that blocking based on the first three and four characters would not be highly desirable. Among the 672 linked pairs, 176 linked pairs were found to be concordant by all four methods. Erroneous conditions at the end of the surname were not detected even by the modified NYSIIS system. There were 87, 106, 98, 86 and 119 record pairs in which errors occurred in the first, second, third, fourth, and between the fifth and eighth positions, respectively. Therefore, it may be concluded that in a population where spelling variations or errors in reporting and recording usually occur after the fourth position of the surname, these four methods would perform equally well for blocking. Otherwise, NYSIIS and Soundex should be more promising than methods which are based on the use of leading characters.

REFERENCES

Mi, M.P., J.T. Kagawa, and M.E. Earle. 1983. An operational approach to record linkage. Meth. Inform. Med. 22:77-82.

Kagawa, J.T. and M.P. Mi. 1985. On matching with personal names, pp. 269-273 in this volume.

Table 1. Block Characteristics by Methods

Item	Racial Groups[1]							
	CAU	PTG	HAW	CHI	FIL	JAP	PUR	KOR
Number of Male Subjects	34566	15970	7752	16118	40323	84298	4372	3786
Blocking by Complete Surname								
Number of Blocks	13286	1595	2071	546	14374	5137	924	241
Block Size Distribution, %								
1 - 10	96.7	85.1	93.4	77.5	96.6	73.8	92.3	80.1
11 - 50	3.0	10.5	6.4	14.6	3.0	19.9	6.5	13.7
51 - 100	0.2	2.6	0.1	2.0	0.2	3.1	0.8	4.6
101 - 500	0.1	1.6	0.0	5.5	0.1	3.1	0.4	0.8
501 - 1000	0.0	0.2	0.0	1.1	0.0	0.2	0.0	0.8
> 1000	0.0	0.0	0.0	0.4	0.0	0.0	0.0	0.0
Average Size	3	10	4	30	3	16	5	16
Maximum Size	397	550	97	1313	289	1022	288	848
Blocking by First Character of Surname								
Number of Blocks	26	26	23	24	26	25	24	22
Block Size Distribution, %								
1 - 10	3.9	11.5	17.4	12.5	3.9	16.0	8.3	31.8
11 - 50	3.9	19.2	26.1	12.5	3.9	4.0	25.0	27.3
51 - 100	3.9	3.9	21.7	0.0	3.9	8.0	8.3	9.1
101 - 500	15.4	15.4	17.4	45.8	23.1	12.0	50.0	18.2
501 - 1000	15.4	23.1	13.0	16.7	15.4	8.0	8.3	9.1
> 1000	57.7	26.9	4.4	12.5	50.0	52.0	0.0	4.6
Average Size	1329	614	337	672	1551	3372	182	172
Maximum Size	3474	1922	4214	4157	4539	11229	811	1055
Blocking by First 2 Characters of Surname								
Number of Blocks	280	155	142	113	232	178	144	82
Block Size Distribution, %								
1 - 10	34.3	36.1	62.0	39.8	35.8	32.6	58.3	65.9
11 - 50	21.8	26.4	24.7	27.4	17.2	18.0	24.3	15.9
51 - 100	10.0	12.3	4.2	8.0	12.1	10.1	9.7	12.2
101 - 500	28.6	18.7	7.8	18.6	26.3	18.5	7.6	2.4
501 - 1000	5.0	5.8	0.7	3.5	4.7	6.7	0.0	3.7
> 1000	0.4	0.7	0.7	2.7	3.9	14.0	0.0	0.0
Average Size	123	103	54	143	174	474	30	46
Maximum Size	1008	1128	2869	4153	2809	6321	422	872

See note at the end of the table.

Table 1. Block Characteristics by Methods (Continued)

Item	Racial Groups[1]							
	CAU	PTG	HAW	CHI	FIL	JAP	PUR	KOR
Blocking by First 3 Characters of Surname								
Number of Blocks	2212	655	491	354	1880	835	471	179
Block Size Distribution, %								
1 - 10	68.6	68.8	75.6	68.1	66.5	50.1	84.1	77.1
11 - 50	24.5	19.1	18.3	19.5	23.7	24.9	12.3	14.5
51 - 100	3.8	6.6	3.1	3.1	4.9	7.3	2.3	5.6
101 - 500	3.1	4.9	3.1	6.8	4.6	12.7	1.3	1.7
501 - 1000	0.0	0.6	0.0	2.5	0.2	2.9	0.0	1.1
> 1000	0.0	0.0	0.0	0.9	0.0	2.2	0.0	0.0
Average Size	16	24	16	46	21	101	9	21
Maximum Size	471	575	487	1378	740	3879	300	849
Blocking by First 4 Characters of Surname								
Number of Blocks	6941	1112	974	490	5719	1818	709	229
Block Size Distribution, %								
1 - 10	90.6	79.9	82.3	75.9	85.9	61.1	89.0	79.0
11 - 50	8.2	13.1	15.4	13.9	11.9	24.5	9.0	14.9
51 - 100	0.9	4.1	1.4	2.7	1.5	5.9	1.4	4.4
101 - 500	0.3	2.6	0.8	5.9	0.6	6.9	0.6	0.9
501 - 1000	0.0	0.3	0.0	1.0	0.0	0.7	0.0	0.9
> 1000	0.0	0.0	0.0	0.6	0.0	0.8	0.0	0.0
Average Size	5	14	9	33	7	46	6	17
Maximum Size	401	554	255	1322	422	3838	300	848
Blocking by NYSIIS								
Number of Blocks	7293	1025	631	209	6526	1922	649	89
Block Size Distribution, %								
1 - 10	91.7	79.4	80.0	71.8	87.6	55.8	88.4	68.5
11 - 50	7.1	12.5	13.8	12.4	10.7	26.4	9.2	14.6
51 - 100	0.8	4.6	4.3	3.3	1.2	6.8	1.5	10.1
101 - 500	0.4	3.2	1.9	7.7	0.6	10.0	0.8	4.5
501 - 1000	0.0	0.3	0.0	2.9	0.0	0.8	0.0	2.3
> 1000	0.0	0.0	0.0	1.9	0.0	0.2	0.0	0.0
Average Size	5	16	13	77	6	44	7	43
Maximum Size	414	586	406	2311	366	1114	300	965

See note at the end of the table.

Table 1. Block Characteristics by Methods (Continued)

Item	Racial Groups[1]							
	CAU	PTG	HAW	CHI	FIL	JAP	PUR	KOR
				Blocking by Soundex				
Number of Blocks	2864	813	441	161	2779	948	555	86
Block Size								
Distribution, %								
1 - 10	72.9	73.8	77.1	60.9	66.8	43.1	85.8	62.8
11 - 50	22.1	16.0	15.7	16.2	26.8	26.9	11.5	16.3
51 - 100	3.6	5.8	3.6	4.4	4.8	9.5	1.6	12.8
101 - 500	1.5	4.1	3.0	13.0	1.6	15.5	1.1	5.8
501 - 1000	0.0	0.4	0.7	3.7	0.0	4.3	0.0	2.3
> 1000	0.0	0.0	0.0	1.9	0.0	0.6	0.0	0.0
Average Size	12	20	18	100	15	89	8	44
Maximum Size	449	587	774	2275	352	1395	300	885

[1] CAU = Caucasian; PTG = Portuguese; HAW = Hawaiian; CHI = Chinese; FIL = Filipino; JAP = Japanese; PUR = Puerto Rican; KOR = Korean.

Table 2. Surname Characteristics within Blocks

Blocking Criterion	Racial Groups[1]							
	CAU	PTG	HAW	CHI	FIL	JAP	PUR	KOR
	Average Number of Surnames Per Block							
First character	511	61	90	23	553	206	39	11
First 2-characters	48	10	15	5	62	29	6	3
First 3-characters	6	2	4	2	8	6	2	2
First 4-characters	2	1	2	1	3	3	1	1
NYSIIS	2	2	3	3	2	3	1	1
Soundex	5	2	5	3	5	5	2	2
	Maximum Number of Surnames Per Block							
First character	1407	184	961	73	1553	834	113	31
First 2-characters	352	100	632	53	962	376	48	22
First 3-characters	178	31	118	12	269	210	23	23
First 4-characters	37	10	60	8	117	89	10	10
NYSIIS	51	13	71	39	52	70	9	
Soundex	68	16	136	24	74	71	15	15
	Surname Information in Matching							
First character	0.99	0.89	0.99	0.81	0.99	0.98	0.86	0.47
First 2-characters	0.94	0.70	0.99	0.70	0.97	0.94	0.63	0.29
First 3-characters	0.75	0.32	0.93	0.20	0.85	0.84	0.34	0.08
First 4-characters	0.40	0.14	0.78	0.07	0.57	0.79	0.18	0.02
NYSIIS	0.48	0.17	0.90	0.57	0.46	0.43	0.20	0.25
Soundex	0.64	0.20	0.95	0.54	0.61	0.64	0.27	0.14

[1] CAU = Caucasian; PTG = Portuguese; HAW = Hawaiian; CHI = Chinese; FIL = Filipino; JAP = Japanese; PUR = Puerto Rican; KOR = Korean.

Table 3. Concordant Rate of Blocking

Blocking Method	Racial Groups[1]								
	Total	CAU	HAW	CHI	FIL	JAP	PUR	KOR	OTH
	Number of Linked Pairs with Errors in Surname								
	672	167	77	28	78	222	54	10	36
	Concordant Rate (%)								
First 3-characters	56.7	56.3	62.3	32.1	48.7	54.5	79.6	50.0	63.9
First 4-characters	43.9	50.3	52.0	14.3	32.1	41.4	59.3	20.0	44.4
NYSIIS	56.4	60.5	57.1	57.1	59.0	51.4	70.4	40.0	44.4
Soundex	64.9	66.5	53.3	71.4	71.8	65.3	75.9	50.0	44.4

[1] CAU = Caucasian; PTG = Portuguese; HAW = Hawaiian; CHI = Chinese; FIL = Filipino; JAP = Japanese; PUR = Puerto Rican; KOR = Korean; OTH = All Others.

Section V: Application Case Studies II

1979 SOLE PROPRIETORSHIP EMPLOYMENT AND PAYROLL: PROCESSING METHODOLOGY

Nick Greenia, Internal Revenue Service

I. BACKGROUND

As the result of an interagency agreement between the Internal Revenue Service (IRS) and the Small Business Administration (SBA), IRS Statistics of Income (SOI) Division is augmenting its tabulations of business financial data (income statement, and balance sheet, when possible) with two additional data items, payroll and employment, from employment tax returns, Form 941 and Form 943. Employment is also to be used as an additional table classifier. The Small Business Administration (SBA) expects that the tabulations will prove useful in the continuing development of its Small Business Data Base in fulfillment of its Congressional mandate (P.L. 96-302 Title IV) to evaluate public policy and economic trends that affect small businesses without thereby placing any additional data collection burden on small businesses [1].

To produce these enhanced data, SOI is linking its perfected [2] sample files of business information and tax records for corporations (Form 1120 series), partnerships (Form 1065), and sole proprietorships (Schedules C, F, or Form 4835 appended to Form 1040) to their respective Forms 941 (Employer's Quarterly Federal Tax Return) and/or Forms 943 (Employer's Annual Tax Return for Agricultural Employees) in order to abstract employment and payroll from the latter two types of records. The linkage is effected through the Employer Identification Number (EIN).

These studies commence with Tax Year 1979 and will be repeated for all three types of business entity for Tax Year 1982 to coincide with the Economic Censuses. Thereafter, they will be undertaken annually for corporations and quinquennially for partnerships and sole proprietorships [3].

For the Tax Year 1979 Sole Proprietorship Employment and Payroll Study, the process entailed attempting to (a) link the 108,335 business Schedules C and F and Forms 4835 appended to Forms 1040 on the SOI Individual sample file to possible counterpart employment and payroll records in the population files of some 5 million Forms 941 and 943 for all types of business entity; (b) resolve multiple matches and mismatches for matched sole proprietorship/ employment and payroll records; and (c) reweight for false unmatched sole proprietorship records.

II. SOURCE FILES

Each of the business employment and payroll studies will add employment and payroll data to the financial data already available from the IRS SOI business statistics series by matching SOI sample files of business income and tax returns with the corresponding quarterly or annual Employer's Tax Returns reporting Federal income tax withheld and Social Security (FICA) taxes (Forms 941 and Forms 943).

Processing for the 1979 Sole Proprietorship Study consisted of linking by EIN sole proprietorship business records associated with the SOI-perfected Tax Year 1979 Form 1040 sample file [4] to Census-perfected extracts of their corresponding Form 941 (Employer's Quarterly Federal Tax Return) and Form 943 (Employer's Annual Tax Return for Agricultural Employees) records. Sole proprietorship business records were appended to the sole proprietor's Form 1040 and for this study were one of the following three types:

(1) Schedule C (Profit or Loss from Business or Profession),
(2) Schedule F (Farm Income and Expenses), and
(3) Form 4835 (Farm Pental Income and Expenses and Summary of Gross Income from Farming or Fishing).

File extracts containing EIN, payroll, and employment were provided by Census for the population of some 5 million Forms 941 and 943 (Census deleted Form 943 employment due to its unreliability as a consequence of the March 12 reporting requirement, seasonality of farm employment, and exclusion of certain employee groups not under Social Security) for Calendar Years 1978, 1979, and 1980. The Census-perfected extracts of Form 941 and Form 943 data were themselves derived from tape extracts originally produced on a contractual basis by IRS (initial processor of the complete data set for tax administration purposes) as authorized by Internal Revenue Code section 6103 for Census as part of Census' ongoing effort to update annually its Standard Statistical Establishment List (SSEL).

Generally, problems of access to data were minor for SOI since all source documents were IRS-related and originally filed with IRS. While data access posed little difficulty for SOI, however, SBA could receive only tabulations of aggregated data--no files of microdata records--due to the restrictions IRS places on the disclosure of confidential taxpayer data under sections 7213 and 7431 of the Internal Revenue Code.

III. MATCH/MERGE METHODOLOGY

Foremost among the challenges presented by the 1979 Sole Proprietorship Study were those relating to the matching variable itself, the EIN, and the sole proprietorship's filing period. Each of these factors directly affected linking procedures and strategies regarding the Form 941 and Form 943 data.

While the EIN was a required entry for a Form 4835 if Form 943 was filed, it was required for a Schedule C or Schedule F if the sole proprietor had a Keogh plan (self-employed deferred compensation plan) or was required to file an employment (Form 941 or Form 943), excise, or alcohol, tobacco, and firearms tax return. Matters were complicated for Schedule C and Schedule F, however, by the Keogh plan provision as follows. Prior to 1978, employers maintaining Keogh plans were required to have an EIN in order to complete Form 5500-K (Annual Return/Report of Employee Pension Benefit Plan for Sole Proprietorships and Partnerships with Fewer than 100 Participants and At Least One Owner-Employee), even if the only participants were owner-employees (sole proprietors and certain partners). In 1978 and 1979, owner-employee Keogh plans without common-law employee participants (i.e., with only owner-employee participants) were no longer required to file Form 5500-K, but Schedule C and Schedule F instructions for EIN completion still read as described above; that is, Keogh plans without common-law employees were not excluded explicitly. Of the more than 650,000 Forms 5500-K filed for Plan Year 1977, some 450,000 were for plans without common-law employees. Therefore, while it is unclear what the impact of such a situation was for 1979 Schedules C and F, it is apparent that the potential for problems in the 1979 Sole Proprietorship Employment and Payroll Study (false matches to Forms 941 and Forms 943) was considerable.

The EIN potential problem was compounded by the fact that while sole proprietorship Forms 941 and 943 were processed by IRS and posted by EIN to the IRS Business Master File (the computer data storage system from which the original Form 941 and Form 943 file extracts were produced for Census processing/perfection), the sole proprietorship records (Schedules C and F, Form 4835) were processed with the appropriate Forms 1040 and posted to the IRS Individual Master File (IMF) by the Form 1040's Social Security Number (SSN). Little testing or perfection was performed for the sole proprietorship's EIN, and thus, the potential for false matches as well as false non-matches--due to incorrect and even missing EIN's on the IMF side--was significant.

If the sole proprietorship's EIN posed a problem for the link operation, so did its filing or accounting period. Since (a) no such item existed on the business records themselves (it was abstracted from the one Form 1040 to which multiple sole proprietorship records could be appended), (b) a Form 1040 whose accounting period ended in other than December was presumed to have a full-year fiscal accounting period, and (c) 98.6 percent of the 92,694,302 Forms 1040 processed for Tax Year 1979 had Calendar Year 1979 accounting periods, SOI decided that part-year records and other possibly out-of-scope records (e.g., certain prior-year returns) would not be excluded from processing. Instead, the assumption was made that all sole proprietorship records should be treated as full-year calendar 1979 accounting period records. Accordingly, significant savings of both time and money were realized by disregarding the accounting period from the SOI Form 1040 sample file and using only the 1979/1980 Census Form 941/943 file for this study (instead of both the 1978/1979 and 1979/1980 files, as was done for the 1979 Partnership Employment and Payroll Study).

Since EIN generally was required as an entry on the business schedule only in the event of payroll taxes (Forms 941 and 943) or a Keogh plan, EIN-linkages could be contemplated for just a subset of the sole proprietorship sample. In fact, of the 108,335 Schedules C and F and Forms 4835 on the SOI Sole Proprietorship sample file, only 31,008 had EIN's and, therefore, could be viewed as potential initial matches with the Forms 941 and 943. By type of record, the sample counts were the following.

Form 4835:	40
Schedule F:	2,612
Schedule C:	28,356

IV. PROBLEMS AND RESOLUTIONS

Of the 31,008 records with EIN's (see Figure 1), 24,153 matched on EIN with Forms 941 and/or Forms 943 on the 1979/80 Census extract (EIN was unique for each Form 941 or Form 943 but could have been shared by a Form 941 and a Form 943). Of these 24,153 matches, 4,503 were multiple matches, meaning an SOI sole proprietorship record matched to a Form 941 or Form 943 matching either another SOI sole proprietorship record, an SOI partnership record, or an SOI corporation record. Of the inter-business entity (instead of intra-business entity) multiple matches, 117 were for sole proprietorships matching Forms 941/943 with records on either the SOI Partnership sample file or the SOI Corporation sample file. Consequences would have been dire indeed had all these multiple matches not been individually reviewed (an operation to be treated as obligatory, given the size of the largest possible sole proprietorship weight--over 2,000--and the simply astronomical amounts of payroll, hundreds of millions of dollars per Form 941 for a number of cases, reported for what were probably large corporations).

Figure 1. 1979 Sole Proprietorship Employment and Payroll
Preliminary Unweighted Processing Counts
(Pre-Reweighting)

Item	Number of Businesses (Schedule C and F, Form 4835)
Statistics of Income	
Sample..................	108,335
Without EIN...........	77,327
With EIN..............	31,008
Initially matched on EIN to 1979/80 Form 941 and/or Form 943............	24,153
Initially unmatched on EIN to 1979/80 Form 941 and/or Form 943...........	6,855

All multiple matches were manually reviewed using one-line record listings containing the following data items: EIN; sole proprietorship industry code; sole proprietorship code (to distinguish between Schedules C and F and Form 4835); Form 1120/1065 code (to identify inter-business multiple matches, but only those from SOI sample files); sole proprietorship business receipts, business deductions, and proxy payroll (salaries and wages plus cost of labor); Form 941 calendar 1979 payroll; Form 941 calendar 1980 payroll; Form 943 calendar 1979 payroll; and Form 943 calendar 1980 payroll.

At least two factors (other than the questionability of the sole proprietorship's EIN) are responsible for exacerbating the multiple match (as well as the false non-match) situation. The first is the sole proprietorship/corporation "connection" and helps explain at least some of the sole proprietorship/corporation multiple matches and mismatches. Apparently, sometimes a corporation such as a large department store will subcontract work to a sole proprietorship, say, for appliance repair or upholstery cleaning, and the sole proprietorship will incorrectly report the corporation's EIN instead of its own. The second factor concerns multiple sole proprietorships run by the same sole proprietor, even in different business activities. The sole proprietor might legitimately file several different business returns--each with the same EIN (when EIN is necessary)--and either one Form 941 or Form 943 for all businesses or one for each (also using only one EIN). Regardless, IRS would end up processing several business returns but only one consolidated (by either the proprietor or IRS) Form 941/943 containing all employment and payroll data for the sole proprietor. This latter consideration turned out to be quite significant due to the high number of "multiple matches" which were of this variety.

Resolution of multiple matches was accomplished first by "transcribing to unmatched status" sole proprietorship records with non-zero proxy payroll (the sum of salaries and wages plus cost of labor) which matched to a Form 941 or Form 943 whose payroll was egregiously greater than the sole proprietorship's proxy payroll (often sole proprietorship/corporation matches probably). Second, the assumption was made that for purposes of this processing stage, records with zero proxy payroll generally should become unmatched records. Finally, within each group of both like SSN's and EIN's (to ensure that "like" sole proprietorships also belonged to the same sole proprietor or Form 1040), the remaining matches of sole proprietorship records with non-zero proxy payroll were "perfected" by reapportioning the Form 941/943 payroll and employment data among the sole proprietorship records based on their share of the like group's total proxy payroll. When possible, this reapportionment scheme was applied according to the type of sole proprietorship record best corresponding to the Form 941 or Form 943. For example, if a Form 941 and a Form 943 matched a Schedule C and a Schedule F, the Form 941 data were accorded to the Schedule C and those of the Form 943 to the Schedule F. If a Form 941 or a Form 943 matched both a Schedule C and a Schedule F, the Form 941 or Form 943 was reapportioned among both schedules.

Comparison listings were used after resolution to ensure that all problem matches had, in fact, been remedied. Subsequent to multiple match processing, the final stage in mismatch or false match testing was performed: scrutiny and resolution of matches in which Form 941 or Form 943 payroll exceeded the business record payroll or proxy payroll by at least $1,000 (see Figure 2). Manual review of one-line listings for these records identified only 45 matches worth retaining; the remainder were dispatched to unmatched status via an algorithm which required Form 941/943 payroll to be strictly less (no tolerance) than the sole proprietorship's business deductions (business deductions was chosen in case proxy payroll had been reported or was "hidden" in deduction items other than cost of labor and slaries and wages) in order for the match to be kept. (The tolerance was dropped for this resolution process due to the large weights observed for a number of sole proprietorships and also because business deductions was sometimes zero.) Comparison listings were again used to verify that no anomalies slipped through processing [5].

Figure 2. 1979 Sole Proprietorship Employment and Payroll
Unweighted Match-Processing Counts
(Pre-Reweighting)

Category	Sole Proprietorship Records		
	Initial EIN Matches to Form 941/943	Retained as Match	Rejected as Match
TOTAL........	24,153	22,279	1,874
Multiple business record matches.....	4,503	3,612	891
Form 941/943 payroll exceeded business deductions by $1,000*....	737	45	692
Records with zero 1979 Form 941/943 employment and payroll*	291	0	291
Other matches.....	18,622	18,622	0

* NOTE: Matched records meeting this condition but resolved as unmatched during other processing stage are excluded from this count.

The intent underlying both multiple match and mismatch processing was that only matches with almost certain probabilities of being "good" were to remain as matches. That is, the assumption was that possibly marginal matches were to be treated during these processing phases as "truly false" matches. The goal was to produce a solid reweighting base of good matches so that reweighting for false non-matches based on their characteristics would be as accurate as possible. It was thought that any marginal cases would be more suitably accounted for later by those characteristics which allied them more closely with either true matches or true non-matches as a result of reweighting analysis.

V. REWEIGHTING

On a weighted basis, only 11.1 percent of the 12,329,982 sole proprietorships in the SOI 1979 population matched a Form 941/943 after resolution of multiple matches and mismatches. Since 82.3 percent of sole proprietorships did not have an EIN and only 7.4 percent of all unmatched records had EIN's, however, this statistic is not as discouraging as it might first appear. In fact, the match rate was 63.0 percent when only records with EIN's are considered.

Final problem adjustments consisted of reweighting for false non-matches [6], based on analytical tables of matched and unmatched frequencies classified by industry, Form 1040 adjusted gross income, business receipts, and proxy payroll (cost of labor plus salaries and wages). Unmatched frequencies were further broken down according to whether sole proprietorship records were with or without EIN, since imputation factors might differ considerably for these two sets.

Reweighting was more significant in terms of impact for the 1979 Sole Proprietorship Study than the 1979 Partnership Employment and Payroll Study [7] largely due to the sole proprietorship EIN problem (the EIN's potential absence and other complications as discussed above) and the distribution of unmatched proxy payroll. Of the $42.4 billion reported as proxy payroll by all sole proprietorship records (matched and unmatched), only $28.8 billion or 67.9 percent was accounted for by matched records. If proxy payroll is a good indicator of "true matchability" (97.7 percent of matched records also reported proxy payroll), it seemed that a significant portion of true matches remained to be "found," given that 27.6 percent of unmatched records with EIN's and 22.2 percent of unmatched records without EIN's also reported proxy payroll. Of course, to the extent that proxy payroll consists of contract labor or other "non-true" payroll components, it might not be such a good indicator for certain sole proprietorships--especially for proprietorships filing Schedules F but not required to file Form 943 for employees not under Social Security (see Data Limitations below). Imputation for "missing" data rather than reweighting for false non-matches might be more the issue then.

Reweighting was based upon a file of data defined differently in terms of matched and unmatched status from that of the 1979 Partnership Employment and Payroll Study. For the 1979 Partnership Study, a matched record was defined, primarily for reasons of simplicity and expediency (it was also the first of the business employment and payroll studies to be undertaken and, consequently, the first to encounter new obstacles and the attendant deadlines and cost restrictions in surmounting them), as any Form 1065 matching on EIN with a 1978, 1979, or 1980 Form 941 or Form 943 containing either employment or payroll for 1978, 1979, or 1980. This definition unfortunately allowed into tabulations some records with both zero employment and zero payroll for 1979, since they contained data for either 1978 or 1980. While this definition is being discontinued for future business employment and payroll studies, it also was not used for the 1979 Sole Proprietorship Study, even though a file containing two years (1979 and 1980) of Census Form 941/943 data was used for matching purposes. In fact, only records matching on EIN to a 1979 Form 941 or Form 943 containing employment or payroll data are considered matches--and these criteria must have been met even after multiple match and mismatch problem resolution. That is, records initially "matched" but later transformed to unmatched status as a result of resolution processing are not considered matched for reweighting and table purposes.

VI. DATA LIMITATIONS

Following are qualifications necessary to better understand the data in terms of conceptual limitations posed by slightly different terminologies employed across return forms as well as differences in data reporting requirements:

(a) Sole proprietorship proxy payroll was defined as the sum of salaries and wages plus cost of labor in order to be consistent with the definition of proxy payroll used for the 1979 Partnership Employment and Payroll Study. While this item was used primarily for purposes of comparison with Form 941/943 payroll during multiple match and mismatch processing, definitional differences between these two versions of payroll also warrant aggregate comparisons to ascertain what effect not only actual but also perceived differences had on the data.

Salaries and wages and cost of labor were available from Schedule C as the items wages (form instructions required the reporting of both salaries and wages) and cost of labor but from Schedule F and Form 4835 only as the item labor hired. All of these items should have excluded compensation of the proprietor, but since the Sole Proprietorship Study required gross payroll, they included amounts deducted for jobs or WIN credits.

Overstatement of proxy payroll may have occurred due to inclusion of payments for contract labor, such as certain janitorial, secretarial, or agricultural employees not reportable on Forms 941/943 but deducted on the business schedule, probably under cost of labor. On the other hand, understatement of payroll may have occurred if payroll were reported as commissions, legal and professional fees, repairs, other costs of sales and operations, or other business deductions. Additionally, for certain businesses in the Retail and Services industry groups, tip income would have been reportable on Form 941 but not claimed as a deduction on the Schedule C. Finally, a definition of payroll conforming more closely to the concept of total compensation might also contain contributions to both pension and profit-sharing plans and employee benefit programs (such as health and prepaid dental insurance), though the proprietor's contributions to the latter were not specifically excluded by Schedule C instructions.

(b) For payroll, Form 941 appears to have required as reportable compensation virtually what was required in the counterpart Form W-2 and Form W-3 items; i.e., income which was taxable but not necessarily tax "withholdable." Form 943 required the reporting of all taxable cash wages to employees subject to FICA taxes, but excluded the value of non-cash items, such as food and lodging--potentially significant components of compensation for agricultural employees and also reportable on Schedule F as a deduction under labor hired. A further limitation was that reportable taxable wages were only required for workers under Social Security (thus, excluding many non-resident alien agricultural workers) and were not to exceed the FICA maximum, a little more than $22,000 for 1979 and for purposes of this study probably not too detrimental.

In addition to taxable wages, Form 941 required the reporting of all tips and other compensation to employees even if income or FICA taxes were not withheld and specifically excluded only annuities, supplemental unemployment compensation benefits, and gambling winnings--even if income taxes were withheld on these.

(c) While the Form 941/943 March 12 reporting date for employment was an obvious data limitation, it was exacerbated by the possibility of employment double-counting due to employees who worked two or more jobs with different employers filing different employment tax returns.

(d) While testing was conducted to identify possible mismatches in which Form 941/943 payroll was abnormally high, none was attempted (primarily due to time and other cost constraints) for possible false matches or mismatches in which it was too low. For the 1982 study, it might be possible to establish acceptable ranges for payroll/proxy payroll ratios by industry, geography, and certain size classes, but any such operation should be excessively circumspect, given "hidden" proxy payroll, as well as the problem with EIN's previously discussed. (For other recommended enhancements, see also section 10, Greenia, Nick, Match Group Case Study #00002, "1979 Sole Proprietorship Employment and Payroll.")

ACKNOWLEDGMENTS

For their thoughtful review of material in this report, the author thanks Tom Jabine (National Academy of Sciences), Carol Utter (Bureau of Labor Statistics), and Doug Sater (Bureau of the Census). Appreciation is also extended to Wendy Alvey and Beth Kilss for their help in editing the manuscript and to Dawn Nester and Rodney Turner for typing its many drafts.

NOTES AND REFERENCES

[1] For further information on the Small Business Data Base, see Kirchhoff, Bruce A. and Hirschberg, David A., "Small Business Data Base: Progress and Potential," 1981 Proceedings: American Statistical Association, Section on Survey Research Methods; Hirschberg, David A. and Phillips, Bruce, "Using Financial Statement Data to Evaluate the Status of Small Business," 1982 Proceedings: American Statistical Association, Section on Survey Research Methods; and Rose, Paul and Taylor, Linda, "Size of Employment in Statistics of Income: A New Classifier," 1982 Proceedings: American Statistical Association, Section on Survey Research Methods.

[2] File perfection essentially consisted of testing and resolving obvious math errors as well as data inconsistencies in each file record. Errors could have been made by the taxpayer or during a data processing stage.

[3] A more comprehensive treatment of small business employment and payroll, forthcoming from David A. Hirschberg and Bruce Phillips of SBA, will folow the conclusion of the Tax Year 1979 corporation and sole proprietorship studies. Final tabulations for these two studies were provided to SBA in July 1985.

[4] For a detailed account of the sampling scheme involved in selecting this sample, as well as other information--including tabulations--concerning this file, see Statistics of Income--1979/80, Sole Proprietorship Returns.

[5] For more details on the false match resolution phase, see Problems and Resolutions, Greenia, Nick, Match Group Case Study #00002, "1979 Sole Proprietorship Employment and Payroll."

[6] For a complete discussion of the reweighting process, including its assumptions, see Day, Charles, "Imputation Methodology, 1979 Forms 1040/941/943 Link Study," June 1985. This unpublished report is available upon request by writing to Director, Statistics of Income Division, D:R:S, Internal Revenue Service, 1111 Constitution Avenue, N.W., Washington, DC 20224.

[7] See Greenia, Nick, Match Group Case Study #00006, "1979 Partnership Employment and Payroll."

THE DEVELOPMENT OF THE MASTER ESTABLISHMENT LIST

David Hirschberg, Small Business Administration

As part of its data base developmental effort, the Office of Advocacy, Small Business Administration (SBA), has developed a Master Establishment List (MEL) with over 8.1 million businesses. In creating the list, two commercially available lists were merged. The first, the Dun's Market Identifier file, contained over 4.6 million records; the second, the Market Data Retrieval file--a "yellow-page" listing--contained over 7 million records.

The MEL provides direct statistics on the number and geographic distribution of America's small businesses. It also facilitates communication with the small business sector and is a vital tool for conducting surveys and mailings to selected industrial sectors regarding governmental policy.

This paper describes the development of the Master Establishment List. First, some background is provided on existing small business files. Then the MEL is discussed, some of its uses are described and some on-going validation efforts are mentioned. The paper concludes by raising some of the policy implications of concern to SBA.

BACKGROUND

Although major progress has been made, the small business sector remains poorly documented in the Federal statistical system. Most existing Federal statistical data and administrative record sources are not adequate for assessing the impact on small business in a variety of policy analysis and decision-making areas. It is interesting to note that of the 124 pages of statistical tables appearing in the Economic Report of the President, 1985, only one is relevant to small business activity, "Business Formation and Business Failures, 1940-84." [1] (The source of this business formation and business failure data is Dun and Bradstreet.) Two other sources of information on business formation are the Bureau of Economic Analysis and the Internal Revenue Service. However, there are obvious problems in using their data as well. For example, the Index of Net Business Formation, published by the Bureau of Economic Analysis, is 114.8 for 1983 (with 1967 = 100). This growth level is sharply at variance with the number of business tax returns reported by IRS, as shown below. Furthermore, the number of enterprises has increased from 3.3 million in 1976 to 4.4 million in 1982.

The Small Business Administration, Office of Advocacy's Small Business Data Base was designed to provide more reliable information on the scope and contribution of the small business sector. This data base is drawn from commercially available data and places little additional paperwork burden on the business community. It permits the maintenance of confidentiality and provides policy-relevant data.

The first project, which is now complete, was the development of the United States Establishment and Enterprise Microdata (USEEM) files for 1976, 1978, 1980 and 1982. These files are based on Dun and Bradstreet's Market Identifier (DMI) files, which are collected for credit and insurance purposes. They have been edited, cleaned and reformatted, and are the basic centerpiece of the Small Business Data Base.

These four files contain information on business organizations that reported business activity in any one year. Each record which identifies an establishment has the following information: (1) Dun's number--this is a number assigned by Dun and Bradstreet that uniquely identifies each establishment and can be used to merge with prior-year files; (2) geographic location -- city, county, SMSA, state and zip code; (3) year business started; (4) number of employees; (5) annual sales volume; (6) Standard Industrial Classification (SIC) code; (7) parent and headquarter's city and state; (8) Dun's number of parent and ultimate parent; (9) subsidiary indicator; (10) status indicator -- single location, headquarters, establishment or branch; and, (11) manufacturing indicator -- indicates if manufacturing takes place at the location.

Table 1. IRS Business Tax Returns by Legal Form of Organization (in millions)

Year	Total	Proprietorships	Partnerships	Corporations
1967	8.5	6.1	.9	1.5
1976	11.3	8.1	1.1	2.1
1982	14.6	10.2	1.5	2.9

Source: Statistics of Income Division, IRS.

The USEEM files now contain data for the estimated 8 million business establishments which existed during the period 1976-82. For each year, annual files include approximately 5 million records. These records provide estimated employment and industry classification for establishments and firms, the start date (age), organizational status and geographic data for each firm.

These USEEM files have been linked into a longitudinal sample file, the United States Establishment Longitudinal Microdata File (USELM), enabling researchers to follow the same establishments over time. This is a primary and necessary requirement to address policy-relevant research issues. The 1984 files are currently being developed; they will later be merged with the USELM 1976-82 files.

The second project involves working with Dun and Bradstreet's raw financial statement file (FINSTAT). The FINSTAT file contains about 150,000 financial statements for 1975, but for the past few years the number has increased to almost 500,000 per year. To preserve the confidentiality of cooperating companies, all identifying information has been removed by Dun and Bradstreet. Although the file includes the major U.S. corporations, approximately 95 percent of the firms have fewer than 100 employees and 74 percent have fewer than 20 employees. By comparing these data with other sources, we are beginning to resolve the question of how well these data represent the small business community.

Finally, a major effort is underway to have data available on small business from the various statistical and administrative agencies of the Federal Government. Together with the Internal Revenue Service (IRS), for example, the Small Business Administration is supporting an effort to link IRS' business Statistics of Income files for partnerships, proprietorships and corporations with that agency's tax reports of employment and payrolls. This overcomes a significant shortcoming in the IRS files. As rich as they are for analytical purposes, there is no employment reported on business tax returns. Other projects include organizing the IRS Corporate Source Book [2] into machine-readable form and examining disclosure and confidentiality issues, particularly as they relate to business data from IRS and Census sources, so as to develop disclosure strategies for the release of microdata (data on individual firms).

THE MASTER ESTABLISHMENT LIST (MEL)

A universe list of firms and establishments is the core element of a statistical program. The Bureau of the Census uses the annual IRS business tax returns, combined with employer withholding/social security reports and multi-establishment company surveys, to develop their list of businesses with employees--the Standard Statistical Establishment List (SSEL). Multi-establishment companies of the Company Organization Survey enable the SSEL data to provide linkage between establishments and their parent firms. The total number of establishments in the SSEL in 1977 was approximately 4.3 million, compared with the 15.6 million business tax returns. Most of this difference is made up of firms without employees.

The Bureau of Labor Statistics (BLS) also prepares lists of establishments or, more correctly, tax units. Administrative records from each of the State unemployment insurance systems are compiled annually. Linkages between the establishments and their enterprises are not available. Other agencies have developed lists to meet their needs as well. An example is the Post Office/Survey Research Center Sample of Nonhousehold Mailers.

Unfortunately, Advocacy cannot use the Census, IRS, or BLS lists as the basis of its sampling frame. By law, the information in these sources cannot be disclosed. Therefore, Advocacy undertook to develop a Master Establishment List based on merging two publicly available private sources: (1) the Dun and Bradstreet's Market Identifier (DMI) file and (2) a "yellow-page" listing from Market Data Retrieval, Inc. (MDR) for the year 1981. The MDR file is compiled from 9 million entries, including duplicates, in the nation's telephone directory yellow pages. The MDR covers many of the establishments in the DMI file and also many small establishments and persons who do not have credit ratings.

Merging the DMI and MDR files involved a considerable effort, given the enormous size of these files and the absence of unique identifiers. [3] About 3.5 million unduplicated records in the MDR file were identified as not having a matching record in the DMI file. The resulting MEL file contains a total of 8.1 million firms and establishments for 1981. [4]

The coverage of the MEL is important. It is useful to compare with comparable tabulations of employment from the Census Bureau's County Business Patterns (CBP). Table 2 does this for the DMI components of the Master Establishment List.

The first two columns of Table 2 list the number of establishments identified in the DMI and CBP. As mentioned previously, the DMI file covers all establishments with Dun and Bradstreet credit ratings. This includes a small number of establishments with no employees, as well as an undetermined number of small establishments with employees. In contrast, the CBP includes only establishments with employees. Given these coverage differences, it is noteworthy that there is a basic similarity in the total number of establishments.

Several reasons exist for the differences by industry, but they are difficult to quantify. Discrepancies may result from differences in industrial classification between the DMI and the CBP. The extent to which the DMI file includes firms with no employees, as well as establishments which are no longer in business, is not known.

Given these classification and coverage problems, the employment estimates are remarkably similar at the major industry division level, as shown in Table 3. Total employment in the DMI file is 6 percent less than that of BLS and 2 percent more than that of CBP. For mining, contract construction, manufacturing, and services, the DMI reports slightly more employment

Table 2. Establishment Counts by Major Industry Division: Dun's Market Identifier (DMI) and County Business Patterns (CBP), 1981

(Establishments in Thousands)

Industry	DMI	CBP	Ratio DMI/CBP
All Industries, Total	4,635	4,587	1.01
Agriculture, Forestry & Fishery	120	804	.15
Mining	42	359	.12
Construction	612	626	.98
Manufacturing	441	336	1.31
Transportation, Communications & Public Utilities	182	162	1.12
Wholesale Trade & Retail Trade	1,846	1,887	.98
Finance, Insurance & Real Estate	372	387	.96
Services	1,019	1,445	.71

Note: Components may not add to total due to rounding.
Source: Tabulations from the DMI and County Business Patterns, U.S. Bureau of the Census (selected years).

Table 3. Employment by Major Industry Division: Dun's Market Identifier (DMI), Bureau of Labor Statistics (BLS) and County Business Patterns (CBP), 1981

(Employment in Millions)

Industry	DMI	BLS	CBP	Ratio		
				CBP/DMI	BLS/DMI	BLS/CBP
All Industries, Total	74.7	75.1	74.8	1.001	1.005	1.004
Agriculture, Forestry & Fishery	.8	NA	.3	.38	NA	NA
Mining	1.3	1.1	1.1	.85	.85	1.00
Construction	5.9	4.2	4.3	.73	.71	.98
Manufacturing	21.2	20.2	20.4	.96	.95	.99
Transportation, Communications, & Public Utilities	4.1	5.2	4.6	1.12	1.27	1.13
Wholesale Trade & Retail Trade	16.7	21.6	20.3	1.22	1.29	1.06
Finance, Insurance & Real Estate	4.6	5.2	5.4	1.17	1.15	.98
Services	19.0	18.6	17.9	.94	.98	1.04

Note: Components may not add to total due to rounding.
Source: Preliminary Report on the Development of the Master Establishment List, 1982, Social and Scientific Systems, Inc.

Table 4. Dun's Market Identifier (DMI) and Market Data Retrieval (MDR) Files as Components of the Master Establishment List, 1981

Number of Establishments in Thousands

Industry	DMI	MDR	MEL	Ratio MDR/DMI
All Industries, Total	4,635	3,488	8,123	.75
Agriculture, Forestry & Fishery	120	49	169	.40
Mining	42	10	52	.25
Construction	612	215	828	.35
Manufacturing	442	82	524	.19
Transportation, Communications & Public Utilities	182	84	267	.46
Wholesale Trade & Retail Trade	1,846	1,054	2,900	.57
Finance, Insurance & Real Estate	372	407	779	1.09
Services	1,019	1,577	2,595	1.54

Note: Components may not add to total due to rounding.
Source: Preliminary Report on the Development of the Master Establishment List, 1982, Social and Scientific Systems, Inc.

than the CBP or BLS files. However, there is significant undercoverage for wholesale and retail trade; transportation, communications and public utilities; and finance, insurance and real estate.

Unfortunately, employment is not available from the MDR file, but the number of establishments added to the DMI file is shown in Table 4. It was apparent from the detailed industry tabulations that the added MDR firms were mostly professionals, such as doctors and lawyers, as well as taxi operators, truckers, insurance agents, and real estate brokers -- businesses that generally do not use credit. These sectors are basic to small business activity and it is important that they be included in lists of small businesses.

In contrast to the 15 million tax returns filed with IRS, the Master Establishment List contains 8.1 million firms and establishments. It does not follow that there is a deficiency in the MEL. Inspection of the sales distribution reported in IRS' proprietorship files suggests that they include persons with other occupations and do not truly reflect full-time business activity. Of the 12.7 million proprietorship reports in 1980, almost half have business receipts below $5,000.

The analysis of the DMI file and the business units added by the MDR file indicate that, for most purposes for which the file will be used, the MEL is representative of the full-time business population with employees.

USES OF THE MEL

The Master Establishment List has been used for a variety of purposes. Users studying specific problems relating to small business have requested that the Small Business Administration make specialized tabulations from the MEL, draw samples based on those tabulations, and provide mailing lists for the sample cases. In some cases the requests have asked for firms by industry and size for a specific State or designated SMSAs or even particular counties. Although some users have been concerned with the broad spectrum of business units, other users' interests have been highly specialized.

An example of the use of the MEL to create a specialized data base was its use in analyzing the proposed legislation on enterprise zones. Because the establishments in the MEL have addresses, it is possible to examine the existing location of business activity in central cities and non-central cities in relation to the proposed enterprise zones. Some measure of the magnitude of potential costs and benefits of the legislation can be obtained by analyzing projected changes in business activity and employment.

In another application, using a three percent sample of the MEL's businesses, an Ownership Characteristics Survey was initiated in January of 1984. It asked respondents for the legal form of ownership as well as for the sex, race and veterans status of the business owner.

Summary results are available in the "Report of the President on the State of Small Business, 1985." [5]

VALIDATION EFFORTS

The exact matching of the 4.6 million DMI records and the 9 million MDR records to produce 8.1 million Master Establishment List records was considerably more successful than might have been expected, and the resulting MEL file has had wide use. As the tabulations of MEL show, the DMI data were augmented in precisely those areas where it was known that coverage was incomplete (i.e., services and trade). Although there are undoubtedly additional small businesses that are without Dun's credit ratings and are not listed in the yellow pages, it is not clear that further efforts to extend the MEL would be worthwhile.

Validation studies have been carried out analyzing the MEL's coverage, consistency, and completeness. One such study involved matching the establishments in the area samples of the University of Michigan's Survey Research Center with the establishments listed in source areas in the Master Establishment List. Another study is comparing State unemployment insurance data with DMI files.

The former study revealed important differences in the MEL list and the list compiled by Michigan. However, recent research has indicated that these lists are subject to obsolescence. Turnover is about one percent a month; therefore, if lists compiled for different time periods are compared, a large number of nonmatches should be expected. This and other experience has shown that a large proportion of nonmatches occurs when business lists are matched using different sources of information. [6]

In the latter study, unemployment insurance microdata files and DMI files were matched for a recent time period for Texas and Pennsylvania. When the comparisons are completed, they will yield information of considerable value in evaluating the DMI file. It can be noted that only about 40 percent of the firms in the files were matched.

FEDERAL POLICY IMPLICATIONS

Using the January 1985 DMI and MDR files, an updated MEL is being created. We are asking support from the various statistical agencies to provide resources to continue this effort, to improve its quality and help make it generally available to the statistical community.

There is a clear need throughout the Federal establishment for a consistent and reliable business universe frame for a variety of research and sampling purposes. Each Federal agency now operates its own system, virtually oblivious to the activities and requirements of others. Employment differences between systems are explained as due to classification, reporting and coverage procedures. In this time of considerable budgetary restraint, cooperation in the development of databases such as the MEL is absolutely necessary.

ACKNOWLEDGMENTS

The development of FINDIT and the actual computer matching were carried out by Social and Scientific Systems, Inc., a Maryland based 8(a) firm. The contributions of Dennis Abels, Wendy Alvey, Sam George, and Pal Khera are gratefully acknowledged.

NOTES AND REFERENCES

[1] Council of Economic Advisors. (1985) Economic Report of the President, 1985, table B-91, p. 337.

[2] The Internal Revenue Service's Statistics of Income Division produces a Corporate Source Book annually, which presents detailed income and balance sheet data classified by industry and size of total assets. For more information, contact the Corporation Returns Analysis Section, Statistics of Income Division (D:R:S:C) at IRS.

[3] For a detailed description of the methodology and computer algorithm, see "File Matching Utilizing Automated Heuristic Techniques (FINDIT)," by Social and Scientific Systems, Inc., Bethesda, MD, January 1983.

[4] See "Preliminary Reports on the Development of the Master Establishment List," by Social and Scientific Systems, Inc., Bethesda, MD.

[5] Small Business Administration. (1985) The State of Small Business: A Report of the President, 1985.

[6] Converse, Muriel and Heeringa, Steven G. (1984) "An Evaluation of the Accuracy and Current Utility of the 1981 Master Establishment List (MEL)," Institute for Social Research, University of Michigan, Ann Arbor, MI.

ENHANCING DATA FROM THE SURVEY OF INCOME AND PROGRAM PARTICIPATION WITH DATA FROM ECONOMIC CENSUSES AND SURVEYS--A BRIEF DISCUSSION OF MATCHING METHODOLOGY

Douglas K. Sater, Bureau of the Census

This discussion involves the enhancement of data from the Survey of Income and Program Participation (SIPP) with data from economic censuses and surveys. This is a pilot project and is still in the development stages. This discussion focuses on the matching methodology, problems, and problem resolution.

I. INTRODUCTION

The Survey of Income and Program and Participation is a new Census Bureau Survey designed to collect a host of information on the social, demographic, and economic situation of the nation's individuals and families.

The data will be extremely valuable to labor market analysis, but they have one major shortcoming--they do not include characteristics of the employer for which the sample persons worked. This gap can be bridged by the addition of information on employers that is collected in the economic censuses.

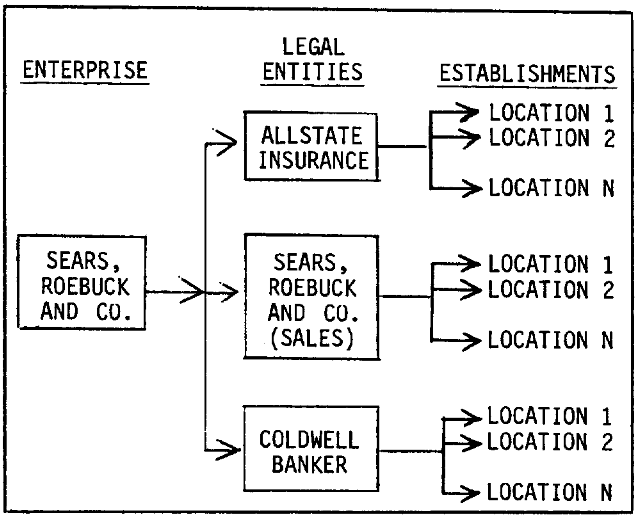
The addition of economic data to the SIPP will enable researchers to obtain improved estimates of the impact of economic and institutional forces which have been intensively studied but are only partially understood or measured. Some of the areas in which the matched file can yield new insights are: the relationship between capital and wage rates, structural unemployment, the transition from a goods to a service economy, unions and the labor market, productivity analysis and numerous other studies. For some of the studies, data at the establishment level are appropriate, and for others, enterprise level data are needed.

II. DEFINITIONS

An establishment is defined as a single physical location where business is conducted or where services or industrial operations are performed. Where separate activities are performed at a single physical location, each activity is treated as a separate establishment. The legal entity is an organizational unit which is assigned an employer identification number (EIN) by the IRS for tax reporting purposes. The legal entity represented by the EIN may comprise one or more establishments. The enterprise is the entire economic unit consisting of one or more establishments or legal entities under common ownership or control. The following figure (Figure 1) shows a partial example of these definitions.

We will be conducting the matching activity for about 20,000 persons in Wave 6 of the SIPP -- the first annual "round-up." In addition to the demographic and economic information, the Wave 6 questionnaire also asks for the employer name, address, and employer identification number for up to three employers.

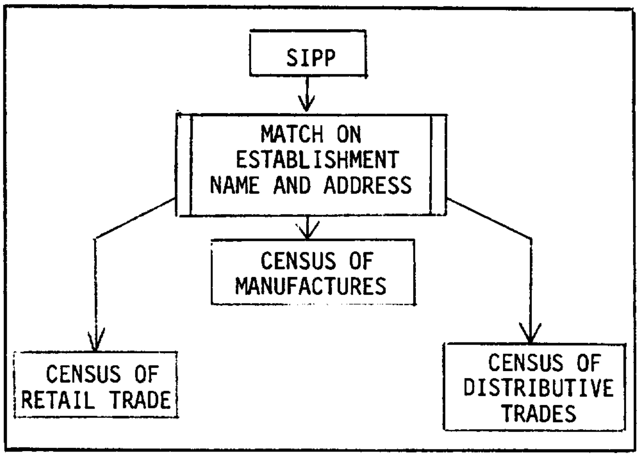
Figure 1.--A Partial Example of Basic Definitions



The first step in this process was to examine the available economic data sources. The Census Bureau conducts numerous economic censuses and surveys, such as the Census of Manufactures, which contain the needed economic data. For linkage purposes, the economic census records also contain a census file number (CFN) which uniquely identifies the establishment. They also contain the establishment name and the establishment address, but they do not contain the EIN.

The first option would be to match the SIPP directly to each economic census needed. (Figure 2 shows a simplified diagram with

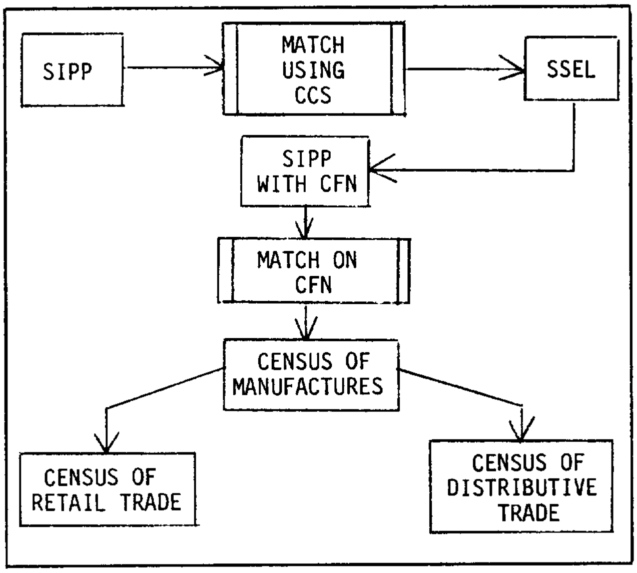
Figure 2.--Simplified Diagram of Direct Match to Three Economic Censuses


SIPP
MATCH ON ESTABLISHMENT NAME AND ADDRESS
CENSUS OF MANUFACTURES
CENSUS OF RETAIL TRADE
CENSUS OF DISTRIBUTIVE TRADES

only 3 possible economic data sources.) This would involve numerous matches on employer names and addresses. Since we are only trying to match about 20,000 cases, the development and testing of programs and the sorting of the economic files were more than we wanted to tackle in this pilot project. Further, the economic censuses do not cover all establishments. That is, they do not cover some "out-of-scope" establishments nor do they cover small establishments. Since about half of all establishments have less than 5 employees, this is a serious shortfall for our purposes.

A more attractive approach would be to conduct the match through an intermediate data set and program system, namely the Standard Statistical Establishment List (SSEL) and the Census Control System or CCS (Figure 3). The SSEL is a centralized multipurpose computerized name and address file of all known employer firms and nonemployer agricultural firms. (This includes the out-of-scope and small establishments as well as establishments covered by the economic census.) The CCS is an interactive random access name search program and series of files derived from the SSEL. It contains the establishment name and address, the EIN and the census file number. The file also contains selected search keys: ZIP Code from the address, a name search key and the EIN. Further, these files also contain selected data such as the number of employees and the annual payroll. In essence, the CCS is a computer assisted manual search program, and it seems to fit our needs quite nicely. Thus, the approach taken is to use the CCS to match to the SSEL to pick up the CFN and selected bits of data. The CFN will then be used to match to the economic censuses. The CFN has another nice property, it allows us to match at the establishment or the enterprise level.

Figure 3.--Simplified Diagram of Match to Three Economic Censuses Using the SSEL and the CCS



The CCS operates in two basic modes:

1. In the EIN mode, one provides the system with the EIN and it returns an abbreviated SSEL record for that EIN.
2. In the name search mode, one provides the system with the name. The system compresses the name, selects the search key, locates the block of records corresponding to this name key, and returns all records in this block. Additional screening is performed based on other data (such as ZIP Code) if it's provided to the system. The selection of the correct record is then done manually.

   For multi-establishment enterprises, located in either the EIN or the name search mode, a second search is done which lists all establishments within the legal entity or enterprise, as appropriate. The selection of the correct establishment record is then done manually.

A hypothetical example would be as follows: Suppose one wanted to locate American Art Supplies, 1235 Main Street, 20735. We would provide the system with "American Art Supplies, 20735".

It would return, for example, the following three records from the Block:

1. American Art Supplies
2. American Fabricaters
3. American Farm Products

We then select record (1) and it provides a second listing containing, for example, the following two records:

1. American Art Supplies-Hqt.
   1235 Main Street.
2. American Art Supplies-Sales
   425 Canal Street.

We then extract the CFN associated with record 1. This is an oversimplification of the system but it gives a general idea of the process.

To make the process as efficient as possible, a stage-by-stage process has been designed which maximizes the amount of computer work and minimizes the amount of manual review. For example, well-considered sorting of the SIPP file can greatly speed the process. That is, assembling the same employer names into groups will allow one search for many records with the same name. Employers of 250 or more employees account for less than 1 percent of all employers, but account for 31 percent of all employees.

III. MATCHING PROBLEMS

There are numerous problems with name matching. First, there are reported name variations due to abbreviations, misspellings, etc. For a household interview survey, such as the SIPP, there are several things that must occur to get a correct name spelling. The interviewer must hear the response and spell the name when filling in the form. The data keyer must be able to read the written entry and key the name. This, in itself is more than ample opportunity for the introduction of errors. Plus, there are errors introduced through phonetic problems. Names such as KROEHLER, BEALLS FLORIST, BURROUGHS, and PFEIFFER BREWERY would pose such problems.

Also, the SSEL, as good as it is, does contain some typographic errors. At any rate, most of these cases are expected to be resolved through the computer assisted manual search process using the reported address and "judgement." For example, if we are trying to locate "KRAYLER, 75 Ely Street, Binghamton, N.Y." we might decide that this is really "Kroehler Manufacturing Co. of Binghamton." We are referring to this process of decision as "judgement" because some degree of uncertainty may exist. If the level of uncertainty seems excessive, the case will be referred for further review. However, care must be exercised in the implementation of "judgement." It implies a lack of uniformity and nonempirical matching criterion.

Another problem is the reported name variations for franchises and "Doing Business As" vs. legal name. As an example, an establishment may be commonly known as "Wendy's," but in actuality, it is a franchise using the Wendy's name and whose legal name is John Smith Enterprises. The match process does not have, in its design, an a priori process to resolve these problems, but the professional review process may be able to identify and resolve such cases.

A potential problem is the presence of mailing address on the SSEL rather than the physical address. Although every effort is made to obtain the physical address for the SSEL file, there are occurrences where the address on the SSEL is the address of the lawyer, accountant, or the administrative office. Depending on the particular circumstances, the problems may be solved or may be intractable.

Also, multiple establishment names on SSEL records may cause problems.
These are occurrences of different establishments having the same name. A hypothetical example would be as follows:

Clinton Aluminum (Hdqts.)
1235 Main Street
Clinton Aluminum (Mfg)
751 Ash Street
Clinton Aluminum (Sales)
755 Ash Street

This, in itself, poses no major problems, unless the address is not reported in the SIPP. Thus, the first question is whether there is sufficient name detail reported in the SIPP to match such a case without address? That is, are division or group names reported in the SIPP? Given the amount of space on the form, I think not. A typical SIPP entry for this example would simply be "Clinton Aluminum." In this event, other matching criteria need to be implemented. If each establishment is in a different part of the country, the selection of the establishment within the same SMSA as the SIPP respondent's may be a reasonable criterion. Another possibility would be to use the SIPP respondent's occupation. For example, if the occupation were salesman, a reasonable criterion would be to assign the case to Clinton Aluminum - Sales Division.

Suppose, in the Clinton Aluminum example, we have located the correct legal entity, but cannot match to the correct establishment. This case should not be hastily written off as a nonmatch. We already know alot about it. We know the enterprise, the legal entity, and we know that it is one of three establishments. It seems that a conditional allocation process will maximize the amount of information. There are several ideas for performing this allocation. One approach would be to use an average value for all three establishments. Another would be to randomly assign the case to one of the three establishments or to do the assignment according to a probability function based on employment size. The probability of correct match is that dependent on the probability function and, for mismatches, data utility is dependent on the degree of homogeneity of the three establishments. In the Clinton Aluminum example, suppose that all three establishments are the same size. Then the chance of a correct match is one in three. In this same example, the wage structure and degree of unionization, etc. are likely to be quite different between the establishments. Thus, a mismatch will distort the data. In a case such as Wendy's or McDonald's, such data distortion would be minimal.

I have not considered this allocation process in depth, but will in the next few months. At any rate, I will need to assign two sets of flags to keep track of what was done and how well the record was matched. The first will identify the type of match. The second will apply to allocated matches and will provide an assessment of the probability of correct match.

IV. PRE-TEST RESULTS

A small-scale familiarization test of this computer-assisted manual search process using the Census Control System was conducted. The sample was comprised of 166 employer names reported in the Waves 1 and 2 of the 1984 SIPP. These cases were drawn from a sample of Primary Sampling Units (PSU). These PSU's were not scientifically sampled, but were arbitrarily chosen to include (1) a variety of PSU's (by size and region), and (2) a variety of manufacturers. Because this is not a scientific sample and only manufacturers are included, the results cannot be generalized and are included only as an approximate indicator. The purpose of this exercise was primarily educational; that is, to see how the process works with real data.

Waves 1 and 2 asked for the name of the employer for which the person worked during the reference period. Although the employer address and Employer Identification Number were not collected in these waves, we tried to obtain the employer addresses for these cases from a variety of reference materials, such as the Major Employer Lists from the 1980 census, telephone directories, and Standard and Poor's Index of Corporations. Table 1 shows the different levels of employer information and the proportion of cases at each of these levels. Table 2 shows selected results of this test.

Table 1.--Results of Address Search Operation

Item	No.	PCT
Total........................	166	100.0
With Corp. Hdqts...........	94	56.6
No Corp. Hdqts.............	72	43.4
With Estab. Address..........	72	43.4
With Corp. Hdqts...........	44	26.9
No Corp. Hdqts.............	28	16.9
No Estab. Address............	94	56.6
With Corp. Hdqts...........	50	30.1
No Corp. Hdqts.............	44	26.5

Even though an establishment address was found for only 43 percent of the cases, the employer name in the SIPP was matched to the correct enterprise 78 percent of the time. The similar match rate is 78 percent for legal entities and 51 percent for establishments. For those cases where there was an establishment address, the match rates are: 88 percent for enterprises, 88 percent for legal entities, and 81 percent for establishments. (Note that the lines "Matched to Enterprise" and "Matched to Legal Entity" are not equivalent. As an example, if a person reported he/she worked for Sears, Roebuck and Company, the person can be matched to the enterprise, but not to the legal entity. That is, which of the following would be the correct legal entity: Allstate Insurance, Coldwell Banker & Co., Dean Witter Financial Services, or Sears Merchandise group? As it turns out in this very small-scale test, we did not encounter any cases of this type. Hence, the number matched to legal entity is 130 and the number matched to enterprise is 130.)

1. Type 1 -- These nonmatches represent cases where there were more than one establishment with the same name all at different addresses. If the address was reported in the SIPP, we would have been able to match these cases. Thirty-one of the 46 nonmatch cases were Type 1's.

Table 2.--Results of Matching Test

SIPP-SSEL Match Status	Total		With Establishment Address		No Establishment Address	
	Number	Percent	Number	Percent	Number	Percent
Total......................................	166	100.0	72	100.0	94	100.0
Matched to Enterprise......................	130	78.3	63	87.5	67	71.3
Matched to Legal Entity (EIN).............	130	78.3	63	87.5	67	71.3
Matched to Establishment.................	84	50.6	58	80.6	26	27.7
Uniquely Identified by Name.............	75	45.2	49	68.1	26	27.7
Uniquely Identified by Name & Address...	9	5.4	9	12.5	X	X
Not Matched to Establishment............	46	27.7	5	6.9	41	43.6
Type 1..................................	31	18.7	X	X	31	33.0
Type 2..................................	9	5.4	5	6.9	4	4.3
Type 3..................................	6	3.6	0	.0	6	.0
Type 4..................................	0	0	0	.0	0	.0
Not Matched to Legal Entity (EIN).........	36	21.7	9	12.5	27	28.7
Not Matched to Enterprise.................	36	21.7	9	12.5	27	28.7

X -- Data cell does not apply.

Type 1 -- These nonmatches represent cases where more than one establishment was found in the SSEL, all at different addresses (but part of the same company) and the company name matched the name reported in the SIPP.

Type 2 -- These nonmatch cases represent more than one establishment at the same address in the SSEL; that is, we would need more information than just the address (such as plant or division name or SIPP occupation) to identify the correct establishment.

Type 3 -- These are cases where the SSEL contains mixed types of entries, some Type 1 and some Type 2.

Type 4 -- These are cases where we could not identify any establishments in the enterprise by name. There were no Type 4's in the test.

(See text for more details on the definitions of the nonmatch types 1-4.)

2. Type 2--These are cases where there are more than one establishment with the same name and at the same address that is, we need more information than just the name and address (such as plant or division name or SIPP occupation). Nine of the 46 nonmatch cases were of this type.

3. Type 3--These are cases where the SSEL contains mixed types of entries, some Type 1 and some Type 2.

4. Type 4--These are cases where we could not identify any establishments within the enterprise by name. There were no Type 4's in the test.

There were 36 cases for which we could not locate the enterprise on the first pass. A large part of this is due to the lack of address for these cases. For the 16 of these, the location was apparently outside the search area we tried (PSU of SIPP respondents address). An address reported in the SIPP will permit us to match most of these. Also, we were able to locate an additional 12 through further research. These were, in general, very small companies. The remaining 8 are, as yet, unresolved. Given the nature of this test, these results were most encouraging.

The 130 SIPP-SSEL matched cass were also matched to the Census of Manufacturers (CM). Of these, 100 matched exactly 26 matched to the enterprise, but the establishment was non-manufacturing and not in the CM, 3 very small and out-of-scope for the CM, and the remaining case was a true nonmatch.

V. OTHER ISSUES

There are a number of other issues to be faced in this project, some of which are:

1. Adjustment for nonmatches--allocation or reweighting. Nonmatch rates will be significantly different between large and small employers. Since much of the analysis will be affected by this, some sort of allocation or reweighting will be necessary.

2. Development of match status flags and probability of correct match status.

3. Development of a process of computing match error rates.

4. Errors in EIN's.

5. Differences in reference periods between the Economic Censuses, SSEL, and the SIPP.

6. Suppression issues in data releases.

We will be investigating these issues in the next few months as work on this pilot project progresses.

BIBLIOGRAPHY

1. Sater, Douglas K., "Enhancing Data from the Survey of Income and Program Participation with Data from Economic Censuses and Surveys," Unpublished paper, July, 1985.

2. Haber, Sheldon E., et al., "Matching Economic Data to the Survey of Income and Program Participation: A Pilot Study," American Statistical Association Proceedings, Social Statistics Section, 1984.

3. U.S. Department of Commerce, Bureau of the Census, The Standard Statistical Establishment List Program, Technical Paper 44, January, 1979.

4. Kasprzyk, Daniel, and Roger A. Herriot, "The Survey of Income and Program Participation," American Statistical Association Proceedings, Social Statistics Section, 1984.

5. U.S. Department of Commerce, Office of Federal Statistical Policy and Standards, Statistical Policy Working Paper 2-- Report on Statistical Disclosure and Disclosure Avoidance Techniques, May 1978.

6. Kasprzyk, Daniel, "Social Security Number Reporting, the Uses of Administrative Records and the Multiple Frame Design in the Income Survey Development Program," Technical, Conceptual and Administrative Lessons of the Income Survey Development Program, Social Science Research Council, New York, New York.

DISCUSSION

Joseph Steinberg, Survey Design, Inc.

INTRODUCTION

The three papers presented illustrate three of a number of varying objectives of exact matching:

(1) addition of data from second file to host file for the same IRS business tax unit;
(2) construction of a more comprehensive frame by combining files; and
(3) addition of variables on establishment economic data to data for individuals in the Survey of Income and Program Participation (SIPP).

This discussion primarily comments on earlier drafts of these papers.

These papers describe the files used and how the matching was done in fine detail. I leave it to those more expert to comment on these matters; I will not try to comment on that.

PERSPECTIVE OF COMMENTS

The point of view taken in preparing these comments was:

(1) How does the quality (or likely results) of the exact matching conform to statistical standards used to judge a statistical study or to judge completeness of a frame?
(2) After reading or listening to the paper, what is known about factors (and their magnitudes) affecting the nonsampling error component of the results?
(3) What additional information should be made available to judge the nonsampling error?
(4) What more (should) might possibly be tried to reduce the nonsampling errors?
(5) Further, if a sample reinterview program is considered useful in measuring coverage and content (net and gross) differences in a sample survey or census, why not use a sample reinterview program for evaluation and calibration in matching studies?
(6) Is the matching approach optimal or is it better to collect data through a survey process?

In view of the review approach, you will see that this discussion provides some comments and a series of ouestions for the presenters.

GREENIA

Nick Greenia has an interesting problem, even though both files come from IRS forms. The supplementary forms for individuals (C, F, and 4835), which are of interest, may not show the EIN or, if EIN is shown, it may be incorrect. What is known (if anything) about false nonmatches or false matches as a result (since only the 1979/1980 files of the Forms 941/943 were used, and not 1978/1979)? What is known about the false nonmatch rate which resulted?

It is interesting to observe that many identifier systems have similar problems -- here it is the "sole propietorship/corporation connection" re the EIN. There used to be (and may still be) the problem in the SSN: multiple people gave an identical SSN as a result of the purchase of a wallet that had a valid SSN on a specimen identification card.

I noted that matched cases were dropped when the 941/943 payroll was greater than the sole proprietorship's business deductions. Was any effort made to contact any sole proprietorship when this was found? Wouldn't it be of interest to know, for a small sample, at least, under what circumstances this situation arose? May not treating such cases as unmatched eliminate an important class of novel situations? Why do you think, Nick, that reweighting overcomes the problem?

Given the assertion in the paper "... that a significant portion of true matches remained to be found ..." (Section V), would the analytic objectives be served if the tabulations of "matched" data are based on not much more than the original set of matches? Would the nonsampling error of the results be too large?

I have often wondered whether information on the Forms W-3 was available on any accessible file. Since the Form 941 employment is only for employees for the pay period ending March 12, would a more useful source of employment and payroll be:

(1) the number of statements--counts of Forms W-2 and
(2) total payroll for the year from the summary W-3 process?

Incidentally, if any of these ouestions suggest a need for contact with a business (as re 941/943 payroll greater than business deductions), a statistical study (perhaps conducted by a third party) should be considered the vehicle, with results available to IRS only in tabulations (screened for disclosure problems). Consider, a statistical reinterview program may be a useful means for evaluating overall quality and not just for special issues.

HIRSCHBEPG

Now I turn to Dave Hirschberg's paper. In the paper, I found the interesting points:

(1) that the Master Establishment List (MEL) is unique in its representativeness of small businesses of all size categories, and
(2) that the total number of businesses included in the MEL exceed more than half of the population or universe of all (small and large) businesses reporting to the Internal Revenue Service.

My question is: How complete is MEL? The tables show the relation of the Duns Market Identifiers (or DMI) to County Business Patterns. How do the distributions of MEL compare with some standard? And, by Standard Industrial Classification (SIC code) and employment size?

At another point, the author indicates that businesses not represented in the MEL are mostly smaller businesses or individuals that might be located in their homes or who, due to limited activities, would not appear in the credit markets nor advertise in the yellow pages.

In view of this, what problems are there in the Small Business Administration (SBA) use of MEL? Also, what is known about the rate of inclusions in MEL files of firms no longer in existence (given the slowness of purge of the DMI and Market Data Retrieval, Inc's "yellow-page" listings)? What is the duplication rate still in the file? (One source paper says "... hopefully relatively few.") Further, what is known of the proportion of false matches -- discards from one file or the other that really didn't match? This is not to suggest that "Findit" as a match program has any discernible problems -- at least to my knowledge.

Now, I turn to another matter. This project, creation of MEL, was initiated since there was essentially no single file available to SBA which satisfied its needs--and it is understandable why various agencies have statutes (Census) or regulations which require confidentiality of frames, privacy being deemed more important than government-wide efficiency.

What is the confidentiality status of MEL? Does SBA have a regulation which prohibits disclosure? What are any other possible public uses - could another government agency, say, Department of Energy, or could a research firm doing a study for a government agency have access? At what price? How does this compare to your costs?

On another matter -- what improvements in file completeness would there be from access to the UI files in the 25 states willing to share their files? Has anyone explored the possibility that uniform files for these 25 states may exist in a Federal agency's hands -- the Bureau of Labor Statistics (BLS)? And what cooperation can be worked out between SBA and BLS, given written agreement by these 25 states to permit SBA access?

The paper recognizes that data collection is "non-rigorous" and, therefore, employment, and possibly SIC codes, too, may be inaccurate. What, if anything, can be said about the effects of possible inaccuracies on the use of subsets of MEL as survey frames? Consider the value of a sample reinterview program to check on quality.

Finally, the paper mentions that some checks were planned, e.g., MEL vs. University of Michigan, Survey Research Center's sample of their nonhousehold establishment list. Are there any results of such checks available? What do they show about the completeness of MEL?

SATER

Now concerning Doug Sater's paper; first, I turn to the SIPP information collection to be used for the match. Has Census considered the desirability of expanding the questions being asked (name of employer, address, employer identification number)? Perhaps, in addition to address (or, if not available), they could consider getting nearest street intersection; asking for telephone number at place of employment -- for possible use, when no EIN is given, for calling the employer; or, if no address, calling to establish an address?

Also, re SIPP-collected data -- what steps are taken to assure that SIPP-collected EIN is consistent with SIPP-collected information on employer name and address?

The paper discusses a prospective matching project, and it is interesting to read about the decision process that leads to the decision concerning the source file and matching method. It will be interesting to hear, in the future, what actually took place: the degree of manual effort and the various costs. Incidentally, what is the relative budget planned for this matching activity compared to the SIPP data collection phase? It would be interesting to know, both here and in other matching projects, about relative budgets for matching vs. data collection of source surveys.

In view of the author's contention that they expect to obtain (in the SIPP) valid EINs about 40 percent of the time and that there is a need to use a variety of methods to try to determine the EIN in the remainder, how will the match validity be tested? (The paper says error measurement will be the subject of future development. And evaluation strategies will be the subject of future development.) What about considering a sample reinterview program as part of the evaluation strategy?

The paper describes a small scale familiarization test. Admittedly, it was not a true test, since address and EIN had not been collected in the nonprobability set of units used for the test.

How secure are you, Doug, in the rates of exact matching cited in the paper? Do you have plans for another, truer, test, using a subsample of the SIPP that you plan to use, before mounting the full-scale matching project? Suppose the results are not as good as in the small-scale familiarization test; what if the results suggest a 60-70 percent match rate. Would you recommend the project move forward?

The paper notes that adjustments are planned for matching problems. What order of magnitude of matching problems do you believe are likely to occur, for which allocation or reweighting is the preferred solution? What do you anticipate will be the net effect on the level of nonsampling error in some principal result?

REJOINDER

Nick Greenia, Internal Revenue Service

The discussant's observations are, of course, most appreciated and exhibit a grasp of the Sole Proprietorship Link Study's fundamental problem: as a first time study, it had to cope with how much was simply unknown.

The decision to employ the 1979/1980 file of Form 941/943 records and omit the 1978/1979 file as well as the fiscal filing period possibility was due to two factors: higher processing costs and the 1979 calendar filing period assumption. Higher costs of additional linkage processing for files not originally designed for the link studies per se (i.e., the SOI-perfected sole proprietorship sample file and the Census-perfected Form 941/943 population file) were deemed unwarranted primarily because (a) for Tax Year 1979 some 99% of all Forms 1040 had calendar year 1979 filing periods and (b) of those which had fiscal or non-1979 filing periods, many were probably filed for members of partnerships.

Other than what is known of false matches obtained from match processing as well as the increase in aggregate data resulting from reweighting for false non-matches (increases of 16% for number of businesses, 10% for payroll, and 11% for employment), nothing is known of this processing decision's direct impact on false matches and non-matches. Probably it had little impact since match problems in general were thought to be attributable primarily to the Employer Identification Number (or lack of it) on the sole proprietorship's business schedule. The second Sole Proprietorship Link Study (Tax Year 1982) is expected to benefit from the 1979 experience in this regard primarily because such tradeoff decisions as necessitated for the 1979 Study will be precluded by the 1982 sample file format design.

No sole proprietorships were contacted during the study's match processing phase primarily due to resource constraints. Although the payroll/deductions discrepancy was designed to catch "hidden payroll" on the business return, the 1982 study probably will compare payroll to proxy payroll. This change is suggested by the 1979 experience which has led us to believe that hidden payroll is less of a potential problem than the overstating of proxy payroll--primarily due to its inclusion of contract labor payments as well as payroll not reportable on Form 941/943 for certain employee classes. Again though, it is important to err on the conservative side (particularly when examining the payroll/deductions relationship) by building a sound match base, due to the large weights on some sample business records. Reweighting is thought to overcome potential problems of omission by compensating for any marginal matches missed through groups of solid match records with similar characteristics. Further, it was a desirable step in order to provide the Small Business Administration (SBA) with as full a data set as possible to meet SBA's own analytic needs.

The discussant's suggestion to replace the Form 941 file with W-3 file counterpart information (total conpensation for payroll, number of W-2's attached as an employment proxy) would be desirable if control problems currently confronting the W-2/W-3 tapes--annually provided to IRS by SSA for the Combined Annual Wage Reporting Agreement Form 941/943 reconciliation effort--could be overcome. SSA is planning to overhaul its current computer processing system in 1987, which might be a more appropriate time to reconsider such an approach. In the meantime, however, it might be worthwhile to pursue this idea with the thought of supplementing Form 943 information--weakened in the past by reporting qualifications as well as the general problem of reporting employment only for the March 12 pay period.

REJOINDER

David Hirschberg, Small Business Administration

Joseph Steinberg's questions regarding the Master Establishment List's (MEL) quality and conformity to statistical standards lie at the heart of the matter, once the major issue of mechanically merging files is solved.

Limited opportunity exists here for full discussion of the quality issues raised by Joseph Steinberg. However, there are several studies and reports which provide the interested researcher with such information. Discussions of the overall quality of the Dun's DMI file can be found in "D&B, DMI: Data User Conference." [1] Another publication of interest includes, "A Comparison of Employment Data From Several Sources: County Business Patterns, UI and Brooking's USEEM," by Candee Harris. [2] That report provides a fairly extensive examination by industry of the small business population.

Generally the nonsampling errors which are of concern can be examined from the information presented. The impact of the matching on the overall quality of the MEL is more complicated. From a statistical point of view, little is known about how completely the "yellow pages" cover the universe of business.

Definitive efforts to evaluate the Master Establishment List are hampered by the lack of uniform numerical identifiers in the various systems. Even when numerical identifiers, such as Federal employer identification numbers, are available, the matching of files from different systems is not a straightforward task, as Nick Greenia has pointed out in his paper. [3]

A great deal of work is needed in this area, and access to administrative records from State and Federal agencies is necessary. In addition, a requirement exists to more carefully define a small business for statistical purposes.

The overall documentation of the Small Business Data Base work can be found in the appendices to the "State of Small Business: A Report of the President" for each year beginning with 1982. [4] A more comprehensive guide to information relating to specific issues can be found in "The Development of the Small Business Data Base of the U.S. Small Business Administration: A Working Bibliography" by Bruce D. Phillips. [5] Most of these publications are available from the Office of Advocacy. Methodological and quality issues raised by Steinberg are directly addressed. Steinberg also raised the issue of the MEL's confidentiality status. This is now under discussion with the firms producing the files, and a formal statement on this issue should be forthcoming.

As mentioned previously, the inability to match files of business firms, along with a large turnover rate, plagues any attempt to develop independent verification of the MEL. The University of Michigan Survey Research Center report, although vigorous in its approach, was not able to overcome these problems. [6] When differences between the two files occurred, it was difficult to determine precisely what the problem was.

One final comment with regard to the State unemployment insurance data is in order. The potential use of these files was explored with the States and the Bureau of Labor Statistics; because of confidentiality provisions, access could not be provided. Although a few States did decide to make their files available for research purposes, the cost involved in integrating them into the MEL precluded their use.

REFERENCES

[1] Advisory Commission on Intergovernmental Relations. (1979) "D&B, DMI: Data Users Conference," Washington, D.C. (mimeographed).

[2] Harris, Candee S. (1981) "A Comparison of Employment Data for Several Business Data Sources: County Business Patterns on Unemployment Insurance, and the Brookings U.S. Establishment and Enterprise Microdata File," The Brookings Institution, Washington, D.C.

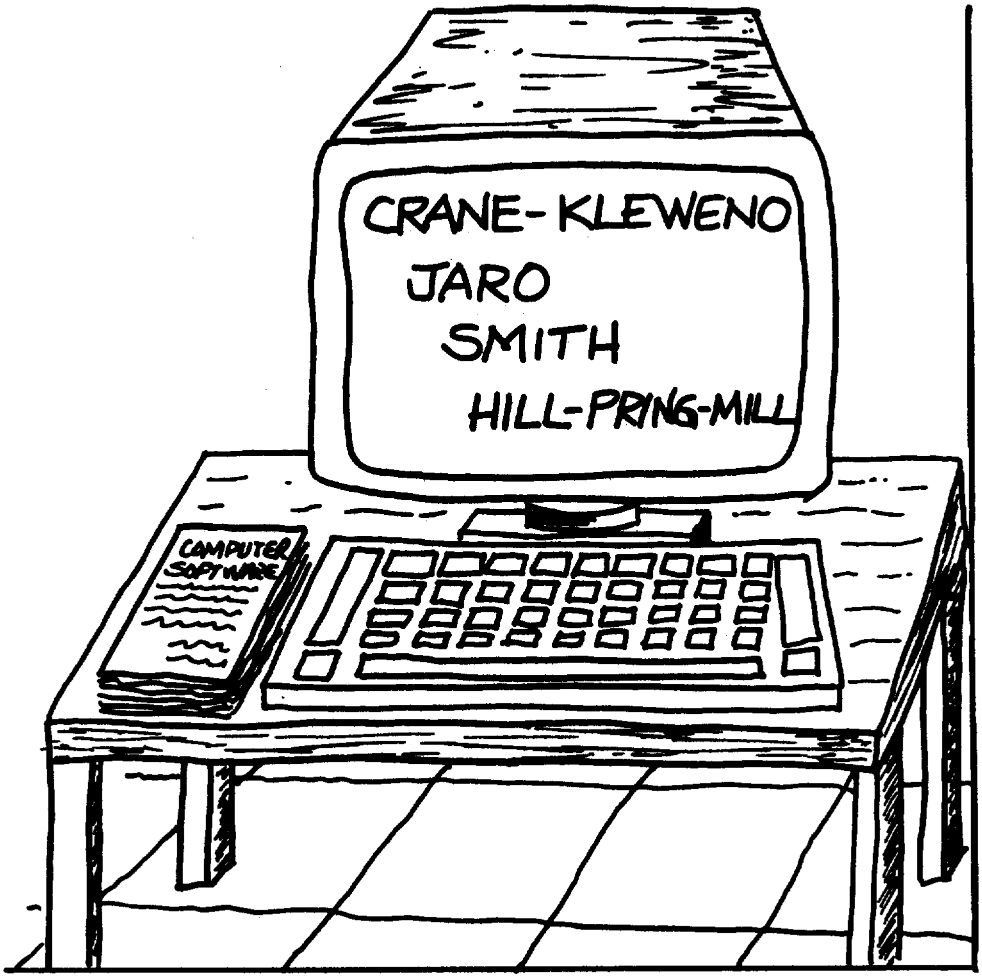
[3] Greenia, Nick (1985) "1979 Sole Proprietorship Employment and Payroll: Processing Methodology," Record Linkage Techniques--1985, Internal Revenue Service, Washington, D.C.

[4] Small Business Administration. (Annually, 1982 to 1985) The State of Small Business: A Report of the President.

[5] Phillips, Bruce D. (1985) "The Development of the Small Business Data Base of the U.S. Small Business Administration: A Working Bibliography," Record Linkage Techniques--1985, Internal Revenue Service, Washington, D.C.

[6] Converse, Muriel and Heeringa, Steven G. (1984) "An Evaluation of the Accuracy and Current Utility of the 1981 Master Establishment List (MEL)," Institute for Social Research, University of Michigan, Ann Arbor, MI.

Section VI: Computer Software



PROJECT LINK-LINK: AN INTERACTIVE DATABASE OF ADMINISTRATIVE RECORD LINKAGE STUDIES

Jane L. Crane, National Center for Education Statistics
Douglas G. Kleweno, U.S. Department of Agriculture

Much information exists on linkage studies using administrative records and, in some cases, survey data. A database called LINK-LINK illustrates the electronic retrieval of linkage study information. This paper is a guide for a prospective user of LINK-LINK. It will briefly describe the database and potential uses of the system, explain how one searches the database for general or specific linkage project information, outline procedures for obtaining copies of the database and address the future direction of the project.

The database is the end-product of a pilot study by the statistical policy committee formed from the Matching Group of the Administrative Records Subcommittee, a standing committee of the Federal Committee on Statistical Methodology. The committee encourages use of the database and solicits comments and suggestions from all users.

A DESCRIPTION OF LINK-LINK

LINK-LINK is an interactive information database devoted to administrative record and survey data linkage studies. The initial database contains 30 studies which were selected for complexity, originality, and diversity of record linkages. Appendix A provides a list of these studies by title.

Information for each study in the database was obtained using a self-administered questionnaire. The questionnaire, designed by the statistical policy committee, was completed for each linkage study. Respondents for the pilot study were contacted by telephone and letter before receiving the questionnaire. After the information was collected, it was edited for clarity and completeness and then it was keyed into the database.

The database is comprised of a series of menu-type prompts to direct the inquirer during the interactive information search. The menu allows the user to choose the search category from a list that appears on the screen. There is considerable flexibility in the database because of a variety of search categories. In addition, the prompts also allow selection of a particular area of user interest.

LINK-LINK was written using a dBASEIII software program. The database, which was developed on an IBM PC/XT personal computer, is on a 5¼" floppy disk.

Equipment requirements for LINK-LINK include: an IBM PC/XT or any other fully compatible personal computer with the MS-DOS or PC-DOS Version 2.0 or greater operating system; a minimum of 256K bytes of memory; two 360K floppy disk drives or one 360K floppy disk drive and a hard disk drive; and a printer with at least an 80 column capacity.

Objectives for Developing LINK-LINK

The primary objectives of the database are as follows:

1) inform and educate data users about record linkage activities;
2) identify and describe major record linkage data files;
3) illustrate procedures to meet confidentiality requirements associated with a particular record file;
4) demonstrate linkage methodology including software limitations, data quality concerns and linkage solutions; and
5) identify a knowledgeable contact person for further linkage information.

Type of Information Available

Each study in the database can be referenced to obtain a broad spectrum of linkage study information including: the linkage purpose; linkage methodology including software used; linkage data files; methods used to meet legal requirements for matching; type of dissemination of the linked data; names of cooperating institutions and their contact person; and titles of supporting linkage publications. A more detailed description of the database contents is given in Tables 2 and 3.

Potential Uses of LINK-LINK

LINK-LINK is a reference source for people seeking information on record linkage studies involving administrative records and/or survey data. The database is a useful tool to:

1) identify new and significant linkage programs using administrative records and survey data, or discover the most recent research activity involving linkage of records;
2) identify the potential uses of linkages involving administrative and/or survey data records;
3) identify the complexity and limitations of data linkages as dictated by public policy;
4) keep abreast of research in administrative record and survey data linkages and avoid redundancy of research efforts; and
5) use as a basis for additional research.

LINK-LINK'S MAIN MENUS

There are two main menus which provide the user with a large selection of information to investigate record linkage studies contained in

LINK-LINK. It is possible to search the database to identify all linkage studies for a certain characteristics such as the linkage purpose or linkage method. It is also possible to search a specific project for detailed study information. The logic of the system flow is from general categories to specific study detail.

Table 1 shows the system's two main menus with the initial user selection categories. Based on the user's interest, the appropriate menu selector value is entered.

Main Menu I is an exploratory menu to give the user a listing of linkage studies by general category. Main Menu II provides detailed data specific to a study in the database. A series of submenus direct the user to the appropriate information of interest within the main menu.

Main Menu I

The user, upon entering the database, keys "do explore" to display the Main Menu I selection categories. As the user responds to additional menu prompts, the search for information narrows until a list of record linkage studies is identified. The format for the list of studies is a five-digit database reference number, a project title, and a brief statement of the study description. The listing is displayed on the computer monitor and is also routed to a printer for hard copy.

Table 2 provides a brief description of the Main Menu I selection categories. For example, to obtain a list of linkage studies used for the construction of a sampling frame, the user keys a "1" in the Main Menu I and a "1" in the submenu. The end point of the Main Menu I is a list of database linkage studies satisfying the conditions as defined by the user in one or more menus.

At the end point of a path search in Main Menu I, the system prompts the user 1) to return for further exploring using major categories in the Main Menu I; 2) to request specific information for one or more studies listed using Main Menu II; or 3) to leave the system entirely with a series of "0" or quit prompts.

Main Menu II

Main Menu II provides the user access to detailed information on a specific linkage study. The user must know the five-digit database reference number which is provided when the listing of studies is printed at the end of Main Menu I. Only one study can be searched at a time. The user can request information on additional studies by entering each reference number as requested. All information displayed on the monitor is again routed to the printer for hard copy.

Table 1: Main Menu Selection Categories in LINK-LINK

Menu	Selector	Category
MAIN MENU I	(1)	Identification of Linkage Purpose
	(2)	Restrictions on Access of Files for Linkage Purposes
	(3)	Linkage Methods and Related Issues
	(4)	Data Files Used in Linkages
	(5)	Subjects and Respondents on Files
	(6)	Title and Short Description of Linkage Project
	(7)	Type of Dissemination
	(8)	Documentation of Linkage Studies by Title and Author
MAIN MENU II	(1)	Access to Files for Linkage Purposes
	(2)	Linkage Methodology
	(3)	Data File Description
	(4)	Titles/Authors of Written Documentation
	(5)	Contact Person for Study Information

Table 2: Description of Selection Categories for Main Menu I

Selector Category	Description of Contents
1. Identification of Linkage Purpose	Ten linkage purposes are identified. The user selects a category for a list of studies.
2. Restrictions on Access to Files	A submenu with two options are available to the user to identify general study safeguards: 1) studies where access to linkage records is permitted when respondent permission is obtained, and 2) studies where agency policy or legal authority restricts disclosure (general or specific statutes).
3. Linkage Methods	Four options in the submenu permit the user to investigate how database study files were linked: 1) software used for data preparation; 2) software used for matching; 3) data quality problems; and 4) linkage problems. Each submenu prompts the user to select a category of interest.
4. Data File Used in Linkages	Datasets used in all linkage studies contained in the database are listed. Number and title of a study are listed first, followed by the dataset(s).
5. Subjects and Respondents	Four general categories of subject/respondent interest are available.
6. Title and Description	List of linkage studies with database reference number, title, and study description is available.
7. Type of Dissemination	Four dissemination categories in the submenu are available for the user to obtain a list of linkage studies: 1) realeased in aggregate form; 2) public use microdata file; 3) restricted use microdata file; and 4) no dissemination.
8. Documentation of Linkage Studies	List of linkage studies with any published documentation by author, title, and date is available.

The user will generally access Main Menu II after exploring for information in Main Menu I. The user simply enters Main Menu II with the five-digit database reference number for which additional information is requested. Table 3 describes the five selection categories available.

It is possible, if the database reference is known, to skip Main Menu I and go directly to Main Menu II by keying "do lnktomn2." This command will place you at the beginning of Main Menu II where you will be asked to select from the categories identified in Table 3.

THE FUTURE OF LINK-LINK

At this time, the future of LINK-LINK is uncertain. The Matching Group of the Administrative Records Subcommittee is searching for an individual or Agency to assume responsibility for the database. Because the current version of LINK-LINK is a pilot effort still in the development stage, an evaluation of the database design is in order. In addition, the mechanics for updating current linkage studies and adding new studies to the database must be addressed. It is also necessary to support users who request a copy of the database.

Copies of the LINK-LINK database may be obtained by mail. Send two formatted floppy disks for each copy of the database requested and a pre-addressed mailer to return the disks.

Specifications for the floppy disks are:

5¼" flexible disk
Double Sided
Double Density
40 tracks

Send correspondence and floppy disks to:

Fritz Scheuren, Ph.D.
Chairperson, Administrative Records
Subcommittee, Federal Committee on
Statistical Methodology
c/o Statistics of Income Division
Internal Revenue Service D.R.S
1111 Constitution Avenue, N.W.
Washington, DC 20224

ACKNOWLEDGMENTS

The success of this project is due to the contributions of members of the Matching Group of the Administrative Records Subcommittee. Group members were: Jane L. Crane, chairperson, National Center for Education Statistics; Mary Bentz and Beth Kilss, Internal Revenue Service; Douglas G. Kleweno, U.S. Department of Agriculture; Wendel Thompson, U.S. Department of Energy, and Tom Reilly, Bureau of the Census. Fritz Scheuren, Internal Revenue Service, provided leadership as the chairperson of the standing subcommittee. Thomas B. Jabine, Consultant, Committee on National Statistics, made many valuable contributions including the identification of the studies for inclusion in the LINK-LINK database.

Table 3: Description of Selection Categories for Main Menu II

Selector Category	Description of Contents
1. Access to Files for Linkage Purpose	Specific information on: parties to the transaction; incentives; how legal requirements were met; how records were obtained; procedures to protect identifiable records during linkages; type of dissemination, if any; and steps taken to prevent disclosure after records have been linked.
2. Linkage Methodology	Specific study information on: software used to prepare data files and to link records; problems in data quality; and problems encountered during the linkage process are listed.
3. Data File Description	Specific linkage study data set names and key variables are listed from each data set.
4. Titles/Authors of Documentation	References of publications by title, author, and date for specific linkage study are provided.
5. Contact Person for Study Information	Specific linkage study resource person including individual's title, employer, address and telephone number are identified.

APPENDIX A: DATABASE STUDIES BY TITLE

Tax Year 1979 Sole Proprietorship Employment and Payroll

Residential Energy Consumption Survey

Developing A Sampling Frame Of Petroleum Sellers

IRS/Census Direct Match Study

Tax Year 1979 Partnership Employment and Payroll

Employer Reporting Unit Match Study (ERUMS)

SRS/ASCS Data Exchange

Intergenerational Wealth Study

Enhancing Data From the SIPP With Economic Data

IRS 1979 Occupational Coding Study

Linked IRS-SSA Data File

Updating of the SSEL

IRS 1982 Estate Collation Study

Deriving Labor Turnover Rates From Admin Records for U.S. and 30 States

Mail List Development for 1982 Census Of Agriculture

High School and Beyond--Third Follow-up Student Financial Aid Record Component

National Health and Nutrition Exam Survey, Epidemiologic Follow-up Study

Census/IRS Link Study

1982 Partnership Employment and Payroll Link Study

1982 Sole Proprietorship Employment and Payroll Link Study

Continuous Wage/Benefit History Project

IRS Mortality Statistics Study

Current Population Survey/ National Death Index Match Study.

Forward Trace Study

Continuous Work History Sample System

Wage and Tax Statement Extract

Information Returns Program Match

IRS/SSA/DOD Match

Special Frame Study

Master Employment List-Unemployment Insurance Records of Texas and Pennsylvania

CURRENT RECORD LINKAGE RESEARCH

Matthew Jaro, U.S. Bureau of the Census

This paper discusses problems involved in the design and implementation of record linkage algorithms for file matching under conditions of uncertainty. Current research activities in this area are summarized, along with a brief survey of some underlying theoretical considerations. This paper stresses techniques that might be used for obtaining confidence in the match decision and algorithm validation. The research being conducted for the 1985 pretest in Tampa, Florida is discussed.

1. SUMMARY OF RESEARCH ACTIVITIES

Record linkage is the process of examining two computer files and locating pairs of records (one from each file) that agree (not necessarily exactly) on some combination of identifiers (or fields). For the Census Bureau this process is typically executed on two files containing individual names, addresses and demographic characteristics. Specifically, record linkage is important for census undercount determination, address list compilation and general census evaluation.

Record linkage research is focused on the development of an algorithm and accompanying manual procedures that will accomplish the above goals in a statistically justifiable manner. To this end the following major activities must be initiated:

A. development of a statistical foundation for the record linkage process;

B. construction of a data base that can be used for calibration, validation and testing of the characteristics of the linkage process;

C. development of methods to obtain information on the discriminating power of the various identifiers and their associated error rates (discriminating power is a measure of an identifier's usefulness in predicting true match pairs); and

D. design and implementation of computer algorithms to perform the actual linking.

The results of this research will be:

A. more accurate undercount determination and coverage analysis;

B. reduction of costly clerical procedures by use of automated methods;

C. a statistically valid process which can replace previous ad hoc techniques; and

D. algorithms that will be useful for overcoverage determination and address list compilation.

2. AREAS OF INVESTIGATION

There are several areas of investigation that must be pursued in order to design and implement a successful matching system. These areas are currently the focus of attention for the Record Linkage Research Staff.

2.1 Blocking and Other Search Restricting Techniques

The set of records that will be examined to find a match for a given record is called a block. Obviously, if an entire file were searched for a match for each record, the probability of finding a true match would be highest, since no records are excluded from consideration. However, the cost of such a process would be prohibitive. As we restrict our search, we exclude records and increase the probability that the "true match" record would be excluded--but the cost of searching decreases.

The ideal blocking identifier would be one which nearly always agrees in "true match" record pairs but nearly always disagrees between pairs which are not valid matches. This ideal blocking identifier must have a large enough number of possible values to insure that the file will be partitioned into many (and therefore smaller) blocks. R. Patrick Kelley of our staff has developed a method for computing an optimal blocking strategy, considering the tradeoffs of computation cost against errors introduced by restricting the search for matches. See [4].

2.2 Weights

The terms "identifier" or "component" represent fields on a computer file (and are used interchangeably). Typical components are street name, street type (e.g., Street, Avenue, etc.) surname, given name, etc. The discriminating power of a component (or identifier) is a measure of how useful that component is in predicting a match. Consider a component such as surname. Common values of surname (such as "Smith") have greater chances of accidental agreement than do rare values (such as "Humperdinck"). Consequently, the frequency of occurrence of a particular value of an identifier is one determinant of the weight or importance of that value as an indicator of matched or unmatched records. Another determinant of the weight is the error rate associated with the value of that component. High error rates diminish the predictive usefulness of an identifier or its values.

Fellegi and Sunter, in [1], presented a general theory of record linkage, including discussions of weight calculations and the development of optimal decision rules. Their basic idea for weighting is summarized below.

The two files (A and B) to be linked consist of a number of components (identifiers) in common. Consider all possible pairs of records. A particular pair is either truly a matched pair (an element in the set M of all matched pairs) or an unmatched pair (an element in the set U of all unmatched pairs).

For all pairs (p) and each component (or component-value state) i let:

m_i= Pr (component agrees | p ε M)
u_i= Pr (component agrees | p ε U).

Weight for the ith component = $\log_2 (m_i/u_i)$.

The above computation would be the same if we were considering specific values of components (such as "Smith" or "Humperdinck") rather than the component as a whole (surname). Similar weights can be computed for disagreements. m_i is computed by examining all matched pairs; u_i is computed by examining all unmatched pairs. For the two files A and B,

$$\{U\} = \{A \times B\} - \{M\} .$$

Since the cartesian product A x B is $O(n^2)$ and M is O(n) (where n is the number of records in the smaller file), then { U } is much greater than { M } and the u_i can be computed by taking the frequency counts of the components in both files.

The calculation of m requires a prelinked set of records M. This fact presents the greatest practical difficulty because of the large sample size necessary, the cost of producing such samples and the inherent error in manual processes.

Fellegi and Sunter, in [1], suggest a method of weight calculation that does not require prelinked pairs. It uses an assumption of the statistical independence of the components and requires the solution of a non-linear system of equations. We plan to investigate the use of this method, which to our knowledge has never been tested.

Another method of weight calculation that we will consider is that of iterative refinement. We propose this method to avoid the construction of costly samples. If there were no errors in a given component, the value "m" for that component would be 1 and the weight for the component could be calculated from the frequency of occurrence of the component value states.

These initial weights can be refined as follows: Whenever a record pair disagrees on a component, that pair would be presented to an operator by the matching program. The operator can then make a decision as to whether the pair is a match or not. This places the pair in either the set M or U and the weights can now be updated (since m is now less than 1 -- because of the detected error -- if this pair is placed in {M}).

The program can obtain information regarding the error rates of each component in this manner, updating the probability as records are processed. The operator supplies the "truth" regarding each record in question (does this pair belong to set { M } or to set { U } ?). This teaches the program to make similar decisions to those of the operator.

The operator can set the level of errors that will control the display of candidate record pairs. In this way, records can be matched automatically despite small errors in components. As confidence is gained, the thresholds for manual intervention can be moved. After all records have been processed, the entire file can be rematched using the new weights and the process can be continued until consecutive iterations produce small differences.

An investigation into this technique is required to determine whether such iterations will converge to a stable set of weights and to determine the amount of bias introduced by such estimation techniques.

A third method of weight calculation that might be explored would involve automatically making the "M" or "U" decisions, instead of relying on human operators. This would be accomplished by considering pairs of records 'that match on all fields except a specified number. Those pairs could be assigned a match status if the composite weight (Σw_i) for the pair was sufficiently greater than the cut-off threshold. The distance from the cut-off would leave room for weight estimation error without effecting the "M" or "U" decision, and hence, the "M" decision could be made automatically with some degree of confidence. These cases would be used to tabulate the error rate probabilities.

Since the cut-off threshold for a match decision is dependent upon the weights of each field, this threshold would move as weights are revised. The effect of this concomitant variation on the weight estimation must be investigated.

2.3 Composite Weights

If the components are assumed to be statistically independent, then the composite weight is equal to the sum of the individual component weights. Adding the weights is equivalent to multiplying the conditional probabilities. Weights for disagreements can be computed similarly to weights for agreement. Disagreements are generally given negative weights, whereas agreements receive positive weights.

We know that some dependencies exist (such as sex and given name) but the extent to which dependence changes the matching decision rules must be analyzed. For example, "Robert" is principally a male given name, but "Stacy" could be either male or female. Such dependencies could have an effect on the probabilities of agreement given unmatched pairs. If the errors in the fields are dependent, then the probabilities of agreement given matched pairs could change. The disagreement weights would also change proportionally.

We are currently designing simulation experiments to study the effect of covariance on the decision results. It is hoped that a regression analysis will provide information concerning this relationship after a number of runs with differing covariance configurations.

2.4 Error Rates

If a plot were to be made of numbers of observations versus composite weight, a bi-modal distribution would result. Since most pairs are elements of { U } , the disagreement mode is much larger than that for agreement.

For each pair, one of three decisions is made. The pair is said to match if the weight is greater than a threshold μ, or not to match if the weight is less than a second, lower threshold λ . Pairs having weights between these thresholds are classed in the "don't know" category. These pairs must be followed-up using a computer-assisted manual approach.

Once the thresholds are set, bounds on the probabilities of false matches and false nonmatches can be computed by integrating the portions of the distribution tails lying beyond the threshold values. By tabulating weights of candidate pairs, the matcher program could provide information on the error rates associated with the component values. These error rates are useful for verification. The success of this technique will depend upon our ability to fit a curve to the observed tails of each mode in order to perform the integration.

2.5 Component Values

The matcher algorithm will use a table of weights derived from investigations on weight methodologies (see 2.2). One weight would be associated with each predetermined component or identifier value. The algorithm would store the most frequent values of components from tables prepared by other programs and component values not in this list would be given a relatively high weight. Thus, popular names (which have low discriminating power) would receive lower weights than comparatively rare ones, without requiring the construction of exhaustive lexicons. Value tables would only be used if successful results could not be obtained by considering a component to have a single weight.

The weight tables for the program will include expected frequencies of occurrence of component values, error rate information and number of records processed for past data. Information from the current data could be used to update the weight tables as the program gains experience matching.

2.6 Bayesian Adjustment

In addition to keeping records of expected frequencies (based on earlier observed frequencies), the program will also keep observed frequencies of a block for a specific file. If there is much deviation between observed and expected frequencies, temporary modification to the weights can be considered. For example, in a Spanish-speaking area, the name "GONZALEZ" might occur relatively more frequently than it does on the average for the United States.

Missing data values could also result in the reduction of discriminating power of a field within a block.

We have incorporated a Bayesian adjustment technique into our experimental matcher. We have assumed a Beta prior distribution and are investigating parameter estimation techniques for this distribution.

2.7 Distance Metrics

Simple agreement/disagreement patterns of component pairs are not adequate for character strings and numeric data. We are investigating prorating the weight on the basis of degree of agreement.

A number of character-string comparison routines for component values which do not agree completely are available, including the routine designed by Jaro and Corbett, which has been used for 12 years in the UNIMATCH system [3]. Through the use of such a routine, words can be matched despite spelling errors. The UNIMATCH algorithm is an information-theoretic comparator which takes into account phonetic errors, transpositions of characters and random insertion, replacement and deletion of characters. These approaches will be tested in the matcher algorithm.

2.8 Assignment

After blocking, the program uses the various techniques described above to construct a composite weight for each pair in the block. These weights are stored in a cost matrix and the assignments can be made by solving the problem:

$$\text{Maximize} \qquad Z = \sum_{i=1}^{n} \sum_{j=1}^{n} C_{ij} X_{ij}$$

$$\text{Subject to}$$

$$\sum_{j=1}^{n} X_{ij} = 1 \qquad i=1,2,\dots,n$$

$$\sum_{i=1}^{n} X_{ij} = 1 \qquad j=1,2,\dots,n$$

where C_{ij} is the cost (weight) of matching record i with record j. X is an indicator variable. The matrix is made square by the use of dummy weights.

This problem is the linear sum assignment problem, which is a degenerate transportation problem that can be solved efficiently using only additions and subtractions. Once an optional assignment set is obtained, the Fellegi-Sunter decision procedure is applied to determine whether an assignment represents a match, a clerical review case or a non-match.

3. MATCHER IMPLEMENTATION PLANS

An experimental program has been implemented that incorporates the techniques discussed in this paper so that controlled tests can be conducted without undue difficulty. This program is operating on an IBM Personal Computer.

For production matching it is anticipated that not more than two passes will be required to match nearly all records not requiring professional review. Records failing to match on blocking components in the first pass would have a second chance to match on different blocking components during a second pass. By selecting two high discrimination/low error rate sets for blocking, the probability of intersecting errors is minimized. The high discrimination/low error rate property for a component means there is a high probability that the component can accurately predict a matching record pair. By using two such components, the chance of a successful match is relatively good, since errors on both components would be required to reject a record.

We plan to utilize experience gained by Statistics Canada (the Generalized Iterative Record Linkage System [2]) and others in our investigation into the problems of record linkage. It is our intent to have an operational program for use with the 1985 Census pretest. One of the most important applications will be coverage evaluation for the Decennial Census.

REFERENCES

[1] Fellegi, I.P. and Sunter, A.B., A Theory for Record Linkage, Journal of the American Statistical Association, Vol 64, 1969 pp 1183-1210

[2] Generalized Iterative Record Linkage Systems (GIRLS), Institutional & Agriculture, Survey Methods Division, Statistics Canada, Internal Documentation, Oct. 1978

[3] Jaro, Matthew. UNIMATCH - A Computer System for Generalized Record Linkage Under Conditions of Uncertainty. Spring Joint Computer Conference, 1972, AFIPS -- Conference Proceedings, Vol 40, 1972, pp. 523-530

[4] Kelley, R. Patrick. Blocking Considerations for Record Linkage Under Conditions of Uncertainty. Proceedings of the American Statistical Association, Social Statistics Section, Philadelpia, 1984, pp. 602-605. (Sections 3 & 4 were prepared with the assistance of D. Childers.)

RECORD-KEEPING AND DATA PREPARATION PRACTICES TO FACILITATE RECORD LINKAGE

Martha Smith, Statistics Canada

Lack of adequate personal (or "entity") identifying information and appropriate documentation on what is contained in historical files can be major stumbling blocks in carrying out long-term follow-up studies. Over the past few years, considerable experience has been gained in the use of existing administrative (e.g., industrial employee, mortality, hospital, cancer, marriage, birth) survey and census data files for record linkage studies in Canada [1-3].

The purpose of this paper is to give some practical pointers for agencies and individuals involved in implementing future linkage projects, particularly those where large historical files are being used, and where no unique identity numbers are available. Specific examples will be given here which relate to occupational and environmental health studies, but many of the record linkage problems and their solutions apply also to other areas of statistical research.

Organizationally, the present paper is divided into six main sections. The first section gives the main results and conclusions. The second section outlines the kinds of data files required for occupational and environmental health studies. The third section describes the role that various broad categories of records can play in the linkage process. The fourth section gives examples of the practical problems in the preparation of existing files for linkage, along with the methods and some of the software developed to cope with these problems. The fifth section deals with the probabilistic matching technique and the art of designing an efficient linkage operation. The last section makes recommendations for future record keeping and data preparation practices to facilitate record linkage.

I. MAIN RESULTS AND CONCLUSIONS

A generalized record linkage system has been developed based on the concepts of probability and the use of 'weighted' record comparisons [4-7]. The probabilistic methods developed have several desirable features:

- records can be linked which lack unique numerical identity numbers;
- records are able to link despite discrepancies which may exist between identifying particulars;
- 'weights' can be assigned for agreement, disagreement, and partial agreement; and
- the technique discriminates between rare and common values of a given identifier.

On the basis of fairly extensive experience with computerized record linkage of a probabilistic kind, using the generalized iterative record linkage system (GIRLS), it seems unlikely that the technology and the software will be major limiting factors in the future. The major costs, which can limit the application of the approach, are often likely to be associated with the need to do data entry for additional identifiers in a standard fashion, if these have not already been captured in machine readable form. For historic data files, lack of appropriate documentation and standard data entry rules can cause problems. Some software has been developed to aid in the preprocessing of such files. It is therefore recommended that if the files are to be used for record linkage, sufficient identifying items be captured at the time of the initial data entry. Compromises whereby the amount of identifying information is restricted in order to reduce costs will be reflected in reduced accuracy of the linkages, and of the kinds of uses that can be made of the files.

Certain files may serve in the role of intermediate files that facilitate the linkage of other files.

Procedures to evaluate the quality of the linkage should be planned early. For example, it may be possible to incorporate known alive cases in a mortality search; to carry out independent manual follow-up on a sample of the file and compare with the computer results; or to carry out an alive follow-up to complement the death search.

Improvement of present data sources and the development of new sources would seem to be necessary if further demands for occupational and environmental health statistics are to be met. A checklist of data items to be collected has been described elsewhere [3-4].

Collaboration and co-operation among individuals and agencies are often required to complete studies. Suitable communication networks among investigators must be established, particularly if there is a long geographic distance between the interested groups.

II. KEY ELEMENTS IN A TYPICAL FOLLOW-UP STUDY THE KINDS OF DATA FILES REQUIRED

Certain general principles shape whatever epidemiological studies for long-term health effects are undertaken and influence the nature of the procedures for data gathering and analysis. The data gathering could include examining data systems already available which could facilitate the study. The requirements for identifying information are similar whether one is looking for changes to the exposed individual, or for inherited changes affecting the offspring from such individuals.

The key elements for data collection that should be included in any such study are described in [4]. A typical follow-up study often requires some knowledge of work histories, dose histories, health outcomes and the personal identification of the individuals involved. The software available must be capable of bringing all the various relevant files together at appropriate times.

The kinds of linkages involved may be a series of internal linkages to identify data pertaining to the same individual (e.g., to create individual work histories) as well as two-file linkages (e.g.

to match a work record against a death record). The matching techniques can use individual identification numbers (e.g., Social Insurance Numbers), probabilistic matching techniques, or a combination of the two.

III. THE FUNCTIONS OF BROAD CATEGORIES OF SOURCE RECORDS

The kinds of source records required for studies of delayed health effects may serve one or more of four possible functions in the follow-up process.

First, such records may identify an individual as belonging to an "at risk" population (or to a "control" population with which the other is to be compared). In this case they are referred to as "starting point" records which initiate the follow-up process.

Alternatively they may identify an end effect, such as cancer or death in an individual who is a member of a study population, in which case they are referred to as "endpoint" records. One example of an endpoint file is the Canadian Mortality Data Base consisting of the records of all deaths in that country dating back to 1950. Follow-up thus will consist of using a file of starting-point records to search a file for potential end-point records, and of linking those records from the two files which relate to the same individuals.

The third possible function of a record file is that of an intermediate file which facilitates the searching and the linkage process. For example, where a starting-point record carries the maiden surname of a woman who later married, and the endpoint record contains her married surname, the search of the endpoint file may be more productive and accurate where reference can be made to another file, such as a marriage file or the Social Insurance Number Index which contains both of these names.

The fourth function of record files is as a source of the detailed statistical variables required for the analysis. For example, linkage may be required to bring together individual work histories, dose histories and smoking histories.

In considering the possible uses of various available files, all four functions must be kept in mind.

IV. PREPARING THE INPUT FILES

Prior to linkage of any kind, the records being used need to be brought into the formats that are required for making the necessary comparisons, and into the sequences that are appropriate for the linkages. The quality of the identifiers needed for linkage may also be tested by looking for blank fields and for values of the identifiers that are not permitted (such as day of birth = 32). If data collection and data entry have not been done with record linkage initially in mind, this phase can be quite time consuming and costly.

We have found the Statistical Analysis System (SAS) very helpful at this stage, and as a routine we systematically scrutinize the values of fields in files to be used in linkage. These are compared with any available documentation regarding coded values and their meanings. One can check how many fields have non-missing values, valid values, ranges, codesets, or invalid characters or values.

Whereas blank fields can only be filled from other sources, fields which have unacceptable values may sometimes be corrected.

One may wish to create a new field for each record to indicate the "availability" and validity of fields on the same record. For example the value "120112001" could indicate "present and with the valid code range" (1), "present, but with an invalid code" (2), or "absent" (0). A SAS distribution of this word facilitates one's assessment of the likelihood that one will be able to link the files.

It is necessary to obtain copies of the forms of the original source documents, the record layouts and any file documentation, along with detailed information regarding how the administrative system works.

Some problems one may expect to encounter have to do with the quality of the records, and some methods which have been used to deal with the problems are as follows:

(1) Lack of a standard format - particularly for the name and address fields

If name fields have been entered in string format and if a variety of delimiters have been used to separate surnames from forenames, it may be necessary to put the values of the fields into a standard fixed format. It is particularly difficult to separate the components in a name field if blanks have been used as the delimiter. A simple NAMESCAN routine has been developed, which changes all alphabetic characters to "A" and leaves all other characters intact. A SAS distribution can then be made to look at the various patterns on the file.

When standardizing name fields, titles should be put in a separate field e.g., Mrs, Jr, Sr. Two-part surnames can be concatenated (SMITH-JONES to SMITHJONES) and retained along with alternate entries for SMITH and for JONES, special characters may be eliminated (O'CONNOR to OCONNOR) and prefixes concatenated (VAN DYK to VANDYK). A prefix list is shown in Table 1. Geographic and disease codes will usually have changed over time. It may be necessary to recode fields so that all records share a common system of codes, or to use ranges of codes that are comparable.

Table 1. --List of Surname Prefixes

BON	DI	LE	O
D	DO	LES	ST
DA	DU	LI	STE
DE	EL	LU	VAN
DEL	FITZ	LOS	VANDEN
DEN	L	M	VANDER
DER	LA	MAC	VON
DES	LAS	MC	VONDER

(2) Spelling errors

To get around spelling errors in surnames, a phonetic encoding scheme can be used. We currently use the modified New York State Identification and Intelligence System (NYSIIS) surname code [8]. In the 1950-79 Mortality Data Base file, there were about 200,000 unique surnames which mapped into about 40,000 NYSIIS codes. Based on evaluation studies of earlier linkage projects, we are currently considering making modifications to this coding scheme based on some of the phonetics involved with Canadian names (particularly French names).

(3) Incomplete files
Due to the rules regarding cutoff dates for preparation of statistics from certain files, one may find that records are missing due to late registrations. If the files are assigned numbers in an orderly fashion, a sequence and continuity check of the numerically sorted file can be carried out, missing gaps listed, as well as the first and last record numbers of the files. We have done this for the Mortality Data Base file. Where exposure data files have been maintained separately from the Master Identification file, some utilites can be used to match files for "orphan" records i.e. an exposure record with no corresponding record on the master identification file or vice versa.

(4) Missing identifiers
These can be assessed from SAS output of individual fields, as well as using the availability word for a number of variables. It is advisable to split a field into its component parts - for example, for birth date use year, month and day. Sometimes sex code has been found missing from files.A list of all forenames appearing on the Mortality Data Base has been created. This has been used to impute a sex code e.g., 1=male only, 2=female only, 3=either male or female forename. Sex code is required so that appropriate weights can be assigned for forenames in the frequency weighting.

(5) Lack of documentation of old historical files
Here we have found SAS output very helpful, and created documentation regarding the contents of each field.

(6) Possible correlated data items
Certain data fields may be correlated, therefore caution has to be taken when assigning weights to these items e.g., birth place of father, mother, and a child. In certain instances the information relates to identical items (e.g., an address and postal code); in other cases it may reflect multiple wrong guesses (e.g., a birthdate being incorrectly reported).

(7) Duplicate records not properly identified
It is important that for a two-file linkage, all records that are known in advance to relate to the same individual be properly identified. This is to ensure that any groups to which either record of such a pair may belong can be combined by the linkage system. Typical examples are records relating to women who have both a maiden name and a married surname. One is unlikely to want to discard one record and keep the other, because there may be records on the other file that relate to either surname. A field can be added to the record to contain a value of 1,2,3 etc. to indicate whether this is the first, second or third "duplicate" entry for this record. If no duplicate exists, the value of the field can be set to zero. Such duplicate records must all be assigned the same unique number (in the GIRLS system this is referred to as the SEQUENCE number).

If an intermediary file is used, alternative entries can be put in with different versions of the identifying information. These may be either entries from both files separately or in hybrid form (i.e. certain items from one file and other items from the other file).

(8) An internal linkage should have been done first
Any file that is going to be used for a two-file linkage, should first be examined to determine whether an internal linkage is required to bring together all records which refer to the same individual (or entity). If one is uncertain about whether there are duplicates, sometimes a fairly inexpensive first check may be to sort the file by surname, first forename, and birth date and to create a microfiche copy of the file for visual examination. A great deal of work in a two-file linkage can be saved by first unduplicating in this fashion the two files that are to be linked.

(9) Length of data fields
If two fields are to be compared, the lengths of the data fields need to be compatible. For example, as a standard, we encode ten letters of the surname into the NYSIIS code. If the number of letters in one file is less than ten characters, problems can arise when the codes are compared. It is therefore advisable to use a surname field that is ten characters or greater. If special characters were originally used, the data entry of the field should be large enough to allow for the elimination of these special characters in the preprocessing.

(10) Separating out values where the same field was used for more than one purpose
As an example, the same field on some files may be used for maiden as for alias surname. One may wish to try to separate out the two types of surnames that have been entered, so that during the linkage step appropriate rules can be used.

(11) Several unique numbering systems used over time
In certain files, several numbers may have been used over time to refer to the same individual. In administrative sytems, there may be a rather different problem; one often needs to clarify whether such numbers have ever been reassigned to other individuals.

In certain cases, one may wish to chain all the various numbers that were used by the same person over time and use this as a pocket identifier within which a probabilistic match could be made.

V. PROBABILISTIC RECORD LINKAGE TECHNIQUES

The Basic Principle

There are three major difficulties to be overcome in order to achieve efficient record linkage. The personal identifying items are often inadequate to discriminate between the person to whom a record truly refers, and other persons in the population who have similar names. A second difficulty arises because when people report personal identifiers they frequently make mistakes. The third difficulty arises because of the large volume of records involved in record linkage. Some related difficulties include the setting of appropriate threshold values for acceptance and rejection of linkages, deciding how most efficiently to carry out a multi-step operation, deciding on the number of partial agreements to use and the selection of pocket identifiers.

The objective of the Generalized Iterative Record Linkage System was to make it possible for computer procedures to efficiently carry out the data processing involved in the probabilistic matching of data files, and to do so easily for a wide variety of diverse data requests. The GIRLS system has involved optimizing four major tasks: (1) the search operation, (2) the decision-making step, (3) the grouping of records, and (4) the retrieval of information.

In the searching step, the sequencing information is used as a means of avoiding the many unprofitable pairings that would have to be examined if every record initiating a search were compared with every other record in the file being searched. Generally for searches of the Mortality Data Base, comparison pairs are created only where both the sex and the phonetically coded form of the surname agree.

For other applications, the sequencing may be by one of several systems of numerical identifier or by phonetically coded surname. Regardless of the means by which the record pairs are brought together, the next step will be a detailed comparison of the remaining identifiers. This is necessary even where the numeric identifiers agree, because such identifiers are occasionally used improperly by persons to whom they do not belong, and sometimes even by a relative of the rightful owner who has the same surname.

At the present time, a test is being made to provide a measure of the usefulness of employing personal identifiers from the Social Insurance Number (SIN) index file to supplement those from the work records, for the purposes of carrying out automated death searches. Not only are the names, birth dates and such more likely to be recorded on the SIN record, they are also more likely to be complete, and as well they will frequently include the mother's maiden surname, which carries considerable discriminating power and is quite unlikely to be available from any work record.

In the decision-making step, each of the remaining identifiers is compared in turn, wherever it is represented on both members of the comparison pair of records.

The odds associated with any specified outcome from the comparison of any identifier are:

$$\text{odds} = \frac{\text{freq of specific outcome in linked pairs}}{\text{freq of specific outcome in unlinked pairs}}$$

This applies equally to agreements, disagreements and to any degree of similarity or dissimilarity no matter how it is defined (as long as both definitions are identical above and below the line).

When pairs are sorted in descending order of total weight, a point is reached at which the record pairs should be judged unlinkable or borderline. To calculate where this threshold should be, two further values are required to be weighted for a two-file linkage. These are: (1) the likelihood that the individual is represented in the file being searched, so that there is a potential for linkage, and (2) the size of the file being searched, since the opportunity for fortuitous agreement increases in proportion to the file size.

The logarithms of both of these values will be negative. When added in with the weight from the identifier comparisons, the resultant sum is known as the "absolute total weight".

$$W^* = W + \log_2 \frac{Na(L)}{Na} + \log_2 \frac{1}{Nb}$$

where,
W^* = $\log_2$ of the absolute odds in favour of a correct linkage;
W = $\log_2$ of the relative odds in favour of a correct linkage = $w1 + w2 + w3$ .... where these are each logs to the base 2 of the odds ratios for the successive identifier comparison outcomes;
Na and Na(L) are respectively, the total number of records in the file initiating the searches and the number out of these that will be linked with matching records in the file being searched (or a reasonably close estimate of Na(L)/Na may be used initially); and
Nb = the size of the file being searched.

To calculate w1, w2...., for reasons of convenience it is desirable to treat separately the data derived from linked pairs and that which applies to unlinked pairs. If w is the net weight for the particular identifier comparison outcome, $\log_2$ (frequency in linked pairs) is the negative component of this net weight, and $\log_2$ (1/ frequency in unlinked pairs) is the positive component of the net weight.

Because the negative components of weight vary with the quality of the file initiating the searches i.e. with the reliability of the identifiers as recorded on that file, these negative components need to be recalculated for each new linkage before the final weighting is done. The data may be obtained initially from preliminary machine linkage, numerical linkages where available, or from manual linkages. Examples of how the weights are obtained are discussed in reference [9].

The positive components tend to be stable where the files being searched are the same on successive occasions (e.g., the death file) and can usually be calculated from the frequencies of the identifier values in that file.

The Art of Record Linkage

The art of designing an efficient computerized linkage operation depends less upon theory than an intuitive perception of how best to carry out the comparisons and what outcomes from these are most likely to be revealing, so that they ought to be recognized by the computer.

Some of the intuitively obvious refinements that have actually been put to use in Statistics Canada's death searches have to do with:

(1) Recognition of partial agreement outcomes, e.g., of
- surnames (three levels of agreement/disagreement);
- given names (eight levels of agreement/disagreement, including agreement truncation where the initials agree);
- birth year (up to 6 levels of agreement/disagreement);
- birth month (3 levels);
- birth day (4 levels).

(2) Recognition of cross-agreement, e.g., of
- initials (where there is no straight agreement);
- month and day of birth - as for initials.

(3) Recognition of degrees of compatibility/incompatibility e.g., in
- last known alive year versus death date (up to 4 levels);
- marital status (up to 4 levels for each status on a search record).

(4) Comparison of place of work versus place of death.

(5) Calculation of age at the time of the matching death to determine the likelihood of death in a particular year using life-table data.
(6) Use of death file size for that same year as influencing the odds for a fortuitous similarity of the identifying particulars.

A potential refinement may be judged worth retaining as a part of the linkage procedure where it is used often enough in doubtful matches, and makes a large enough difference in the final decision to link or not to link, to justify the possible added complexity in the programming. The GIRLS system makes it possible to gather such data after a preliminary linkage and again after a final production run.

The best tactic when designing a linkage procedure for a specific operation is to gather such empirical data after a preliminary linkage so that the procedure can be revised before the final weighting. The information needed earliest has to do with the frequencies among linked pairs of the different comparison outcomes recognized by the preliminary linkage procedures. The tabulations ("info outcomes") should recognize all the comparison outcomes likely to be useful in the decision process.

We often find that what one learns by looking at some manual linkages first can be very helpful in planning a study. This aids in working out the appropriate methods to use and in preparing cost estimates. One may have to decide whether there is enough identifying information available to do the linkage. To get an overall estimate of this, one can first imagine how strongly unfavourable the odds would be if one did not know whether any of the items agreed or disagreed, and were linking to a file of a given size. Then, as one compares each item, in turn, and assumes they agree, this will demonstrate the possible extent of the increase in likelihood favouring correct linkage. One can use a global overall weight for the items employed in this exercise, and hence get a ballpark impression of whether or not there are enough items available to make it work (see Tables 2 and 3 for an example).

After the linkage status decision has been made, the system can identify groups of records which potentially refer to the same entity and it can indicate where conflicts exist. A conflict exists where groups do not fit your requirement e.g., one record relating to more than one death record. In the GIRLS system there are two ways of resolving these conflicts - automatic resolution by the system based on the 'best' linkage, or by manual resolution. A combination of the two often works best.

The retrieval of information operation of the system is designed to quickly and concisely aid the user in making decisions regarding the future direction of the linkage process. The GIRLS system can produce reports at the detailed level on weight sets, linked pairs, group reports, information about the linked pairs, and it can also produce estimates for updating the weights. One may wish to produce reports based only on links for which a given condition is true (e.g., all links above a given weight) or for which a condition using variables on the source records may be true (e.g., all known dead cases as known earlier on the worker's nominal roll file).

VI. FUTURE DIRECTIONS

There are three main directions for our future endeavours:

(1) The improvement and expansion of existing search and linkage facilities which could include further development and enrichment of our current files (e.g., addition of occupation and industry on the death file). The NYSIIS code routine needs to be evaluated more fully taking into account the kinds of names found in Canada. A dictionary of accredited comparison procedures needs to be developed from past linkage studies that could serve as a guide for future studies. Results from earlier studies need to be carefully evaluated,

Table 2. —Example of a Possible Census-to-Death Linkage — Likelihood of Fortuitously Selecting the Correct Death Record, Using no Identifiers Other than Sex (Assumes enumeration in 1971 at age 42, death in 1979 at age 50, and male sex)

COMPARISON ITEMS	ODDS	CUMULATIVE ODDS	WEIGHT	CUMULATIVE WEIGHT	NOTES
			(10 X $\log_2$)		
Random chance of finding death in 1979 male death file, assuming it is there	1/96,532	1/96,532	-166	-166	1
Likelihood of dying in that year, if alive at the beginning of it	1/131	1/12,645,692	- 70	-236	2
Likelihood of being alive at the beginning of 1979 if enumerated in 1971	1/1.04	1/13,151,520	- 1	-237	3

Note: (1) From death file size, for males dying in 1979.
(2) From life tables for likelihood of death in a 12 month period, for a male of age 50.
(3) From life tables, for the likelihood of survival to age 50 among a cohort of males still alive at age 42.

Table 3. --Example of a Possible Census-to-Death Linkage -- Cumulative Effect of Successive Agreements on the Odds in Favour of a Correct Match, when all Identifiers are Present and Agree

IDENTIFIER AGREEING	ODDS	CUMULATIVE ODDS	WEIGHT	CUMULATIVE WEIGHT
			$(10 \times \log_2)$	
(Random chance)	-	1/13,151,520	-237	-237
Surname	2,287/1	1/5,745	+112	-125
First initial	14/1	1/410	+ 38	- 87
Second initial	14/1	1/29	+ 38	- 49
Rest of first name	87/1	3/1	+ 64	+ 15
Marital status	26/1	8/1	+ 14	+ 29
Year of birth	56/1	437/1	+ 58	+ 87
Month of birth	12/1	5,242/1	+ 36	+123
Birth prov/country	8.6/1	45.078/1	+ 31	+154
Ethnicity	3.5/1	157,773/1	+ 18	+172
Parental birthplaces	1.2/1	189,328/1	+ 2	+174
Industry, major	6/1	1,135,968/1	+ 41	+215
Occupation, major	11/1	12,495,648/1	+ 31	+246
Residence province	4.4/1	54,980,851/1	+ 21	+267
Residence city	72/1	3,958,621,272/1	+ 62	+329

particularly with respect to the use of intermediate files and the use of alive follow-up procedures as were used in the Ontario miners study [10]. Further refinements are needed in developing a file of non-links to get weight estimates, particularly where the comparisons are fairly complex (e.g., weighting of forenames).

(2) **The development of new and much needed data bases** which would identify, in a more systematic fashion than heretofore, the occupational and environmental circumstances of people, and which could be used as startinq point files, to initiate the searches for subsequent health histories. Here data collection rules and forms need to be more clearly developed which could be used by industry. Use of new files such as census of agriculture, farm registers, and census of population files can be exploited. The use of existing files for alive and morbidity follow-up need to be explored.

(3) **The exploration with other agencies of any collaborations** that would be productive for generation of the required statistics, and for setting up the necessary communication network and financial support to implement such recommendations.

ACKNOWLEDGMENTS

The author would like to thank Dr. Newcombe for his contribution to this paper. The opinions expressed in this paper are those of the author and do not necessarily represent the views of Statistics Canada.

REFERENCES

[1] Smith, M.E. and Newcombe, H.B., "Automated Follow-up Facilities in Canada." AJPH Vol. 70, No. 12, pp. 1261-1268, 1980.

[2] Smith, M.E., "Long-term Medical Follow-up in Canada." In: Peto, R., Schneidermen, M. eds. Quantification of Occupational Cancer. Banbury Report 9, Cold Spring Harbor Laboratory, pp. 675-688, 1981.

[3] Smith, M.E., "Development of a National Record Linkage Program in Canada." American Statistical Association - 1982 Proceedings of the Section on Survey Research Methods, pp. 303-308, 1982.

[4] World Health Organization, "Guidelines for the Study of Genetic Effects in Human Popu lations." Environmental Health Criteria 46 (in press).

[5] Howe, G.R. and Lindsay, J., "A Generalized Iterative Record Linkage Computer System for Use in Medical Follow-up Studies." Computers and Biomedical Research, Vol. 14, pp. 327-340, 1981.

[6] Smith, M.E. and Silins, J., "Generalized Iterative Record Linkage System." American Statistical Association - 1981 Proceedings of the Social Statistics Section, pp. 128-137, 1981.

[7] Hill, T., "Generalized Iterative Record Linkage System: GIRLS." Research and General Systems, Informatics Services and Development Division, Statistics Canada, Ottawa, 1985.

[8] Lynch, B.T. and Arends, W.L., "Selection of a Surname Coding Scheme for the SRS Record Linkage System." Statistical Reporting Services, U.S. Department of Agriculture, Washington, D.C., 1977.

[9] Newcombe, H.B. and Abbatt, J.D., "Probabilistic Record Linkage in Epidemiology." Red Book Series Report No. 5. Eldorado Resources Limited, Suite 400, 255 Albert Street, Ottawa, Ontario., K1P 6A9, November 1983.

[10] Muller, J., Wheeler, W.C., Gentleman, J.F., Suranyi, G., Kusiak, R., "Study of Mortality of Ontario Miners." International Conference on Occupational Radiation Safety in Mining, Vol. I, pp. 335-343, 1984.

GENERALIZED ITERATIVE RECORD LINKAGE SYSTEM

Ted Hill and Francis Pring-Mill, Statistics Canada

ABSTRACT

The Generalized Iterative Record Linkage System (GIRLS) project was initiated at Statistics Canada in 1978. This paper outlines the concepts behind the system, and summarizes how these have been implemented to provide a powerful tool suitable for a variety of record linkage applications.

1.0 RECORD LINKAGE AND GIRLS

Record linkage is the process of identifying two or more records which refer to the same entity. An entity could be a person, or a business, for example.

In the case where records have unique identifiers (for example, social insurance number), the process of linking is relatively simple as it involves matching on only one field. However in cases where records do not have unique identifiers, information from several fields typically has to be compared to estimate the likelihood that a potential link is a 'true' one. For these cases record linkage is a probabilistic process, and it is for this situation that GIRLS was designed.

GIRLS stands for the "Generalized Iterative Record Linkage System" which has been developed at Statistics Canada, starting in 1978. Since then, the system has been systematically maintained and enhanced on a regular basis.

GIRLS provides a command language in which you can write your own rules for comparing records. Statistically-derived weights are attached to potential links according to the outcomes of these comparisons. Your GIRLS commands are automatically translated into PL/1 (a high-level programming language), compiled, link-edited and executed on the input files to generate an online project database of potential links and the records involved in them. Using other GIRLS commands, you can then query this database to see the results. If these are not satisfactory, you can update the database in various ways, or simply change your comparison rules and try again.

To this end, GIRLS breaks the linkage process into a sequence of distinct phases. Each phase involves deciding on values for system parameters, examining their effect, and adjusting the values as necessary before going on to the next phase. Results from later phases often suggest adjustments to earlier phases. Because phases are distinct, you can easily retrace your steps, run an earlier phase again with new adjustments, run intermediate phases as they are, and quickly catch up to where you were. This is why GIRLS is called an 'iterative' record linkage system.

The principal aims of GIRLS are:

1. To enable you to develop the best comparison rules and statistical weights for the purpose of your linkage project.

2. To provide a convenient framework for this development.

3. To encourage iterative refinement through a sequence of phases and reports.

4. To make the final linkage fast, cheap, and accurate.

Examples of GIRLS applications include:

1. 'Follow-up' studies

   Health Division at Statistics Canada currently runs linkage projects with files provided by employers of individuals exposed to potential health hazards in the course of their work (e.g. uranium miners). These are linked with the Canadian Mortality Database to check that the proportion of matches found is not above normal.

   Such studies can detect risks to health associated with particular occupations, thus pointing the way to causes of disease. They can also aid in testing the long-term effects of curative measures.

2. Building 'case histories'

   Separate records referring to the same person often accumulate in a system. For example, a new record is often made each time an individual is admitted to a hospital. GIRLS can conveniently bring these records together, enabling larger composite records to be made representing 'case histories' for individuals.

2.0 FEATURES OF GIRLS

In the past, record linkage systems have usually been tied to methodologies suited to particular application requirements. GIRLS provides a general solution to developing particular linkage systems.

Its principal features are:

1. A sequence of phases encourages iterative refinement of the linkage process.

2. The full power of database management technology is provided. This includes: automatic maintenance of data integrity across separate files, checkpointing facilities for project recovery, as well as back-up and restore procedures.

3. Both 'one-' and 'two-' file linkages can be performed. (One-file, or internal, linkages can be useful for unduplicating a file or creating composite records.)

4. A variety of samples of records from the input files can be extracted for the purposes of experimenting.

5. A simple but powerful GIRLS command language is provided to write comparison rules, update the project database, and obtain a wide variety of reports at many levels of detail.

6. The commands provided for writing comparison rules can detect full agreement, various levels of partial agreement, disagreement, and missing values. They can also specify cross comparisons of different fields, as well as rules to be executed conditional on the outcomes of previous comparisons.

7. For special purposes you can also write your own PL/1 code and have it included in the Compare program automatically generated from your GIRLS commands.

8. Statistically-derived weights are generated and attached to links to reflect the probability that the records being compared refer to the same entity.

9. Potential links are automatically classified as: rejected, possible, or definite, by comparing link weights against threshold values. You specify these threshold values, and you can easily adjust them. You can also reclassify links manually.

10. Records which refer to the same entity are grouped. Where conflicts exist within groups, these can be resolved either automatically by the system, or manually on a record-by-record basis. (For example, a conflict would exist when records are expected to link to at most one record on the 'other' file, but a group contains some which have linked to several records.)

11. Both batch and online modes are available. Online enables fast iterative adjustment of system parameters by providing quick feedback as to the current state of the project database.

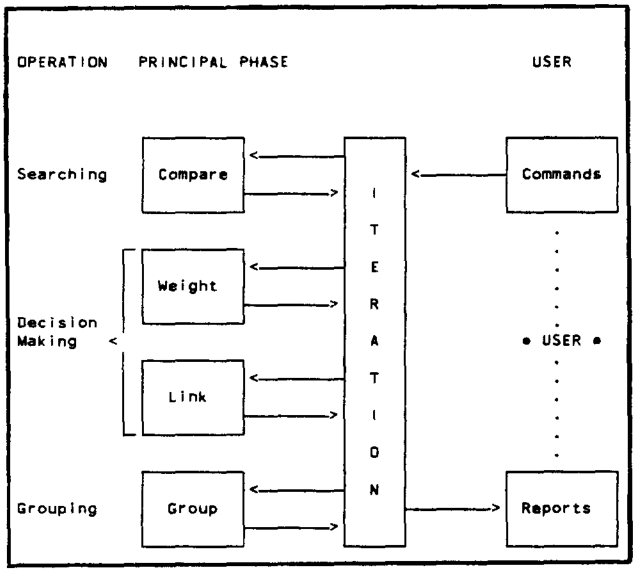
3.0 BASIC OPERATIONS

The phases of the GIRLS system can be grouped into three main operations.

1. Searching.
2. Decision Making.
3. Grouping.

This is shown below:

Figure 1: Basic operations



3.1 Searching

In this operation, pairs of records are compared field by field according to comparison rules you specify. Theoretically, every possible pair of records should be compared. However the number of possible pairs in even a small file is very large. So for practical reasons, records are first blocked into smaller 'pockets' in such a way that it is realistic to look for links only within pockets.

You use GIRLS commands to define your input files, indicate which fields define your pockets, select your sample of records, and specify rules according to which your records are to be compared. Your GIRLS commands are then automatically translated into a PL/1 program, called the Compare program, which is executed on your input files to produce the project database of potential links.

You can write rules to compare fields with values that are: character (e.g. surname), numeric (e.g. birthyear), or coded (e.g. for fields with a small number of discrete values such as birthplace). Your rules can be made conditional on particular outcomes from previous comparisons. You can also specify cross comparisons of different fields (for example, first given name with second given name, in the event that straight comparisons of each field have not already produced an agreement). If your rules do not fit conveniently into the format of the GIRLS command language, you can also write them yourself in PL/1 and have them included in the Compare program.

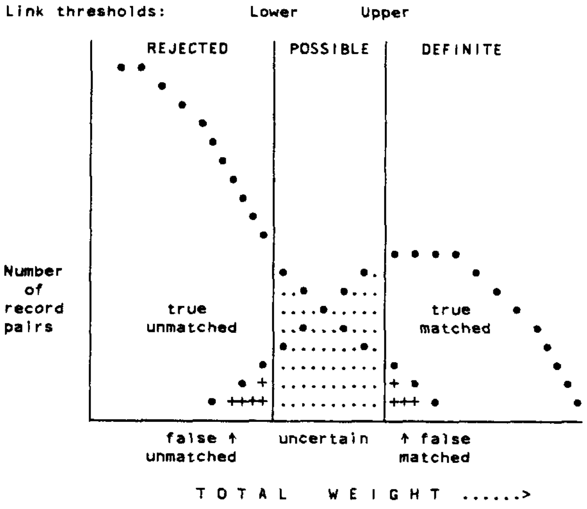
The outcome of having executed a comparison rule can be: agreement, one of various levels of partial agreement, disagreement, or missing. You can specify a 'global' weight to be attached in the event of each one of these possible outcomes.

3.2 Decision Making

In this operation, the potential links generated by the Compare program are evaluated. This involves updating link weights and comparing them against threshold values to decide which to keep and which to reject. Link weights are updated with 'frequency weight sets' which reflect the probability of particular agreements happening by chance. These weights are derived according to formulae developed by Geoff Howe[1], Mike Eagen, and David Binder from methodologies proposed by Howard Newcombe[2], Ivan Fellegi and Alan Sunter[3].

After weight update, the status of links is determined by comparing their total weights against two threshold values. Links with weights above the upper are classified as 'definite', those with weights below the lower threshold are 'rejected', those with weights between the two are 'possible'. This is shown in Figure 2, which is explained as follows:

Figure 2: Link thresholds classify links into three statuses



Let all possible record pairs be divided into two populations: those record pairs which are 'truly matched', and those which are 'truly unmatched'. The goal of the linkage project is then to find the members of the 'truly matched' population. Because it represents all possible record pairs which do not match, the true unmatched population will be far greater than the true matched one. This is shown on the left. The smaller true matched population is shown on the right. The problem is the overlap in the middle, because for these record pairs it is not obvious to which distribution they belong.

The two threshold lines show how GIRLS handles this problem area. Links to the right of the upper threshold are considered 'definite', those to the left of the lower are considered 'rejected', those between the two are considered 'possible'. While permitting flexibility, this approach allows two types of error which any linkage project should aim to minimize.

First is the 'false unmatched' area on the left. These are the record pairs which have been rejected even though they were part of the true matched population. This can happen when information is incomplete or inaccurate on records which 'should' have matched. Second is the 'false matched' area on the right. These are the record pairs which have been accepted even though they were part of the true unmatched population. This can happen when records look very similar even though they refer to different entities, e.g. the different members of the same family. At first glance, these two areas can be minimized simply by setting the thresholds far apart. However this makes for many possible links in between, which will have to be resolved later. By adjusting the thresholds and inspecting various samples of links, however, you can choose the best thresholds for your purposes.

3.3 Grouping

In this operation, the records are grouped according to the status of the links between them. Records may have just one link to another record, or they may have several links to several records. Records joined either by possible or definite links are arranged into 'major' groups - which can be large. Within major groups, records joined by definite links are further arranged into 'minor' groups. A major group may therefore contain several minor groups, and it is the minor groups that contain the best links.

At this stage, 'conflicts' may arise, typically when groups are larger than you want them to be. The system identifies conflicts for you based on your linkage requirement, e.g. one-to-one (i.e. groups are to contain pairs of records only, one from each file). Resolving the conflicts can be done in either of two ways, or both:

1. You can let the system resolve conflicts automatically. This is called 'automatic resolution'. In this case all you specify is your linkage requirement, e.g. one-to-one, many-to-one, or one-to-many.

2. You can resolve the conflicts yourself manually. This is called 'manual resolution'.

You can also use both methods, automatic resolution first followed by an examination of the results and some manual rearrangement where necessary.

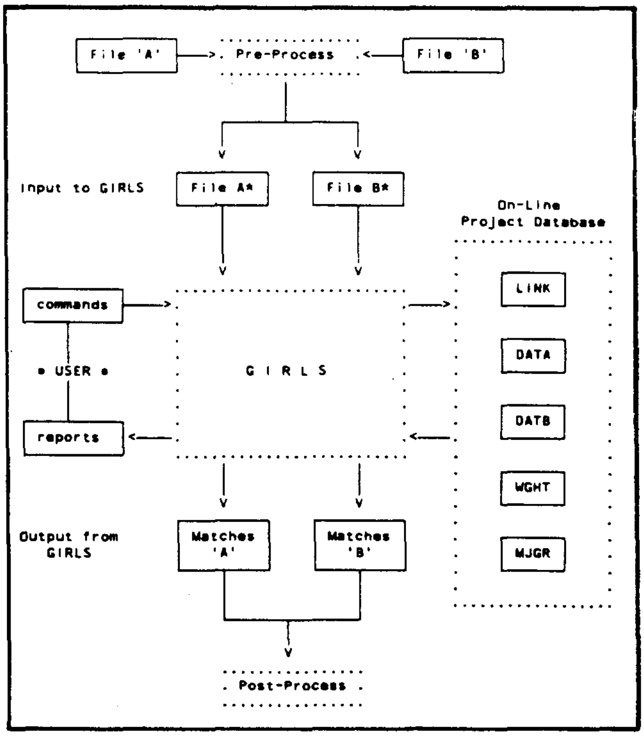
4.0 ENVIRONMENT

4.1 Flowchart

Figure 3 shows a flowchart overview of the system. At the top, two files of records (File 'A' and File 'B') are preprocessed for input to GIRLS. In the middle, records are compared according to your comparison rules, and an online project database is created on the right. This consists of potential links (LINK), the records involved in them (DATA and DATB), together with other files for use later.

On the left, the user is shown interacting with the system via GIRLS commands in the light of the linkage project requirements and feedback from reports as to the current state of the project database. At the bottom, two files of 'matches' are produced. On each output file, each original input record that has been linked is identified by a unique sequence number and has a number identifying the group to which it has been assigned.

Figure 3: Flowchart overview of the system



4.2 Iteration

Iterative refinement of the linkage process can include adjustments to:

1. *COMPARISON RULES*

   From the very many possible links which exist between all possible record pairs, these rules determine which are to be considered the 'potential' links to be written to the project database. These rules can be written, re-written, ordered and re-ordered, so as to produce enough suitable links as efficiently as possible.

2. *WEIGHTS*

   These are attached to links via the comparison rules which applied to the records when the links were formed. It is easy to modify these weights, and thereby select the best ones for your purposes.

3. *THRESHOLD VALUES*

These determine the proportion of definite, possible, and rejected links. The best mixture depends on the aim of a particular linkage project, and is determined by experimenting with the thresholds, and seeing the types of groups which are formed.

For example, for a statistical study it may be satisfactory to find 90% of the links. While for other types of study, it may be necessary not to miss any of them.

4.3 GIRLS Project Files

Making the iterative concept work in practice requires maintaining data integrity across several files when any one of them is being updated. For this reason, an integrated database approach has been taken using the RAPID Database Management System developed at Statistics Canada.[4]The principal RAPID files are:

1. *WEIGHT FILE (WGHT)*

   For each field to be weighted, this contains the values for the field and the frequency weight for each value.

2. *LINK FILE (LINK)*

   For each 'potential' link between a pair of records, this file contains: - the outcomes (agree, disagree) for each comparison rule - the current total weight of the link - the current status of the link (definite, possible, or rejected) - other system control information

3. *DATA FILES (DATA, DATB)*

   These contain the records involved in potential links.

4. *MAJOR GROUP FILE (MJGR)*

   This contains information for each group, enabling reports to be made according to type of group, e.g. "display all groups having more than six records".

4.4 Typical Scenario

A typical (abbreviated) scenario for a GIRLS linkage project might be:

1. Write rules specifying how fields are to be compared.
2. Calculate frequency weight sets (a SAS function is provided to do this job).
3. Use sampling facilities to select a sample of records from the pre-processed input files.
4. Adjust appropriate system parameters, both in batch mode and/or online, until satisfactory results are obtained.
5. Run the full linkage in batch.

Using the system online greatly speeds up the iterative adjustment of linkage parameters. The result can be a linkage process uniquely adapted to the purposes of your linkage project.

Favourable reports from current users include:

- The system is 'comfortable' to use because you remain in control at all stages.
- The command language enables both updates to be made easily, and reports to be obtained to verify intended results.
- Iteration can be continued for as long as it takes for you to be satisfied.

5.0 PHASES

This section briefly outlines the various phases of the GIRLS system. Further details are given in the Strategy Guide and in the User Guide.

5.1 Pre-Process

Purpose: to get files ready for linking

- standardize names and addresses
- validity check
- decide on POCKET
- assign SEQUENCE numbers. (These uniquely identify each record.)
- make duplicate records, when you know records match although they look different. E.g. a record for an individual using her maiden name, and another record for the same individual using her married name.
- recode, e.g. from different codes to common code. (For example, from one hospital coding system to another.)
- encode, e.g. from surname to NYSIIS code
- split files, e.g. by sex, year
- sort files by POCKET

5.2 Weight Creation

Purpose: to create global and frequency weights

- use the provided SAS function to:
  - calculate frequency weights themselves
  - generate GIRLS weight update commands
  - calculate global weights (optional)

"The rarer the value, the higher the FREQ weight."

The frequency weight formula used is:

$$FW_i = 10 \times \log_2 \left(\frac{\text{total number of records}}{\text{No. occurrences of field value i}} \right)$$

where "FWi" is the frequency weight for value "i". For example, the value "SMITH" for the SURNAME field could have a frequency weight of "40".

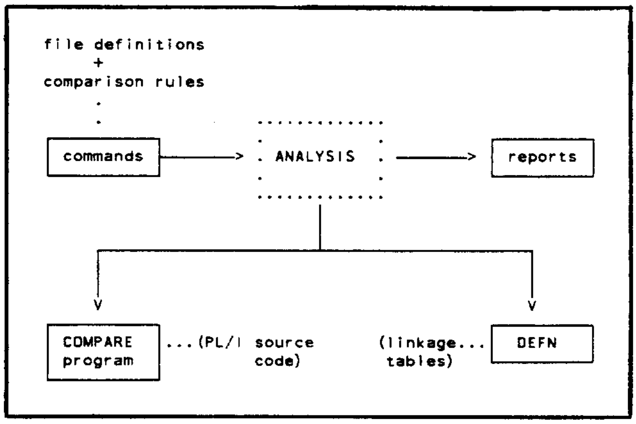
5.3 Analysis

Purpose: to specify comparison rules

- define input files
- choose fields to compare
  - character e.g. surname
  - numeric e.g. birthyear
  - coded e.g. marital status
  - conditional and cross comparisons
  - your own PL/1 code

  choose possible outcomes to weight
  - fully, partially agree
  - disagree
  - missing
- your rules are then translated into a PL/1 program called the 'Compare' program

Figure 4: The Analysis phase



5.4 Compare

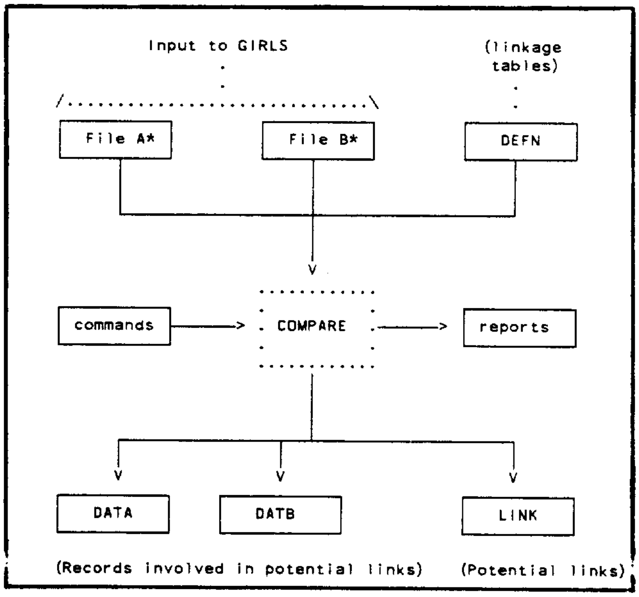
Purpose: to build the linkage database

- set thresholds: upper, lower, and cutoff so as to reject obvious non-matches quickly
- select a sample of pockets with which to experiment
- execute the Compare program

  The comparison rules start assigning global weights to potential links, which are rejected as soon as either current total weight falls below cutoff or if final total weight will be less than the lower threshold.

  The linkage database of potential links and all records involved in them is created.

Figure 5: The Compare phase



5.5 Weight Update

Purpose: to apply and/or modify the weights

- look at link weights 'before'
- apply weights

  You attach frequency weight sets to comparison rules. The system finds all links to which each rule applies and updates the link weights accordingly.
- look at link weights 'after'

5.6 Link

Purpose: to assign statuses to the links

- set a lower and an upper threshold

  The system classifies links by comparing their total weights against these thresholds and assigning a status of definite, possible, or rejected (as explained in Section 3.2).
- inspect results

5.7 Group

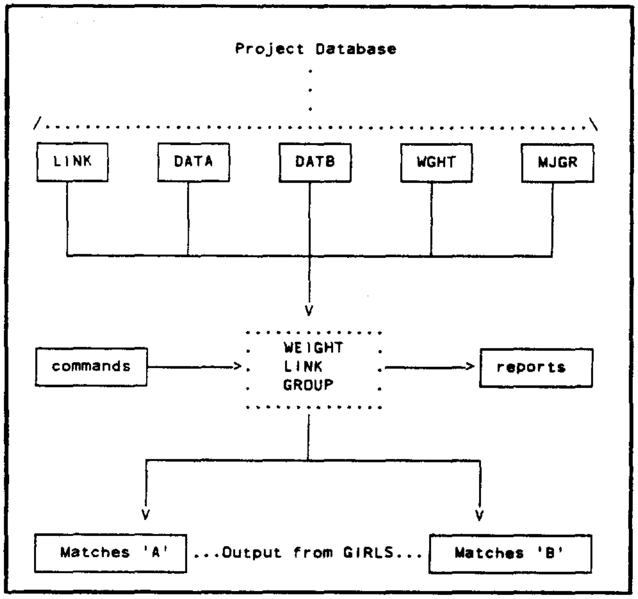
Purpose: to build groups of records

- the system builds 'major' and 'minor' groups of records based on their link status.
  - major groups have both definite and possible links
  - minor groups have definite links only
  - i.e. minor groups contain the best links.
- the system combines groups which share duplicated records. For example, combining a group which contains Mary Smith (maiden) with a group which contains Mary Brown (married).

- resolve group conflicts, either automatically or manually
- output final versions of groups

The Weight, Link, and Group phases are represented below.

Figure 6: The Weight, Link, and Group phases



5.8 Post-Process

Purpose: to use the results of GIRLS

- e.g. for an internal linkage, prepare composite records to represent case histories
- e.g. for a two-file linkage, for each group, generate one record to represent all the members
- create summary files

6.0 EXAMPLE

This is a simple example to show how the GIRLS linkage process works for a two-file linkage.

Part 1 of Figure 7 represents the contents of two files to be linked by GIRLS. File DATA contains 6 records which are to be matched against the 9 records of file DATB. Let the pocket identifier be the SURNAME field (which means that records are compared only if SURNAME agrees on the two records). ROW specifies the row number of the record on the files, and the "..." represents missing data.

Part 2 of Figure 7 shows examples of frequency weights on the WGHT file for the fields SURNAME, MARST and BIRTHYR. (For example, the weight for the surname "Quigley" is "100".) We will be using these weights later to calculate the total weights of links.

Figure 7: Example: two input files and a Weight file

Part 1.— DATA and DATB file

File	ROW	SURNAME	MARST	BIRTHYR
	1	Barnes	01	1950
D	2	Barnes	..	1950
A	3	Jones	03	....
T	4	Jones	02	1960
A	5	Quigley	03	....
	6	Quigley	02	1960
	1	Barnes	01	1950
	2	Barnes	..	1960
D	3	Barnes	02	1960
A	4	Jones	03	....
T	5	Jones	02	1960
B	6	Jones	..	1960
	7	Jones	02	1960
	8	Quigley	02	1970
	9	Quigley	03	1970

Part 2.— WGHT file

SURNAME	MARST	BIRTHYR	WEIGHT
Barnes			40
Jones			10
Quigley			100
	01		10
	02		20
	03		30
		1950	10
		1960	20
		1970	30

The table below shows the links we have on the project database LINK file after executing the Compare phase and applying the frequency weights in the WGHT file. The columns in the table are explained below.

Figure 8: Example: the resulting Link file

LINK ROW	DATA ROW	DATB ROW	SURNAME OUTCOME D(-10)	@SURNAME RESULT	MARST D(-20)	BIRTHYR OUTCOME D(-40)	@BIRTHYR RESULT	TOTWGHT	STATUS
1	1	1	A	Barnes	01	A	1950	60	POS
2	2	1	A	Barnes	M	A	1950	50	POS
3	3	4	A	Jones	03	M	....	40	POS
4	4	5	A	Jones	02	A	1960	50	POS
5	4	7	A	Jones	02	A	1960	50	POS
6	5	8	A	Quigley	D	M	.....	80	DEF
7	5	9	A	Quigley	03	M	.....	130	DEF
8	6	8	A	Quigley	02	D	.....	80	DEF
9	6	9	A	Quigley	D	D	.....	40	POS

• THRESH=(40.75) •

Notes:

1. "LINK ROW" identifies the record number of each link. This identifies the link in subsequent reports.
2. "DATA ROW" and "DATB ROW" indicate the File 'A' and File 'B' records that are involved in a link.
3. "SURNAME" and "BIRTHYR" are fields containing the outcomes of comparison. These are "A" (agree), "D" (disagree), "M" (missing on one or both records).
4. For agreement, the "@SURNAME" and "@BIRTHYR" fields contain the result on which the fields agreed.
5. The "MARST" field contains the outcome of the comparison if it is "D" (disagree) or "M" (missing), or the

result on which the fields agreed if the outcome was agreement.

6. For disagreement, the weights are specified under SURNAME, MARST, and BIRTHYR. E.g. for disagreement on BIRTHYR the weight added is "-40".

7. "TOTWGHT" (total weight) is the sum of the relevant agreement and disagreement weights for each link.

8. "STATUS" shows the link status for each link. This is based on the total weight (TOTWGHT) for the link and the current threshold values (THRESH). In this example, the lower threshold is "40", and the upper "75". "POSS" corresponds to 'possible' and "DEF" to 'definite'. (In this example, comparisons resulting in a total weight less than the lower threshold (40) are excluded from further processing.)

For example, for Link 8 we calculate the total weight (TOTWGHT) from the information on the LINK file, and the weights on the WGHT file, as follows:

Figure 9: Example: calculating the weight for Link 8

Comparison	Value	Weight
SURNAME	QUIGLEY	100
MARST	02	20
BIRTHYR	disagree	-40
	TOTWGHT	= 80

The final table below shows the group numbers assigned to the records after grouping. Records with the same group number refer to the same individual. Records having no group number have no matches on the 'other' file. These groups are based on the DATA ROW, DATB ROW, and STATUS values shown on the LINK file.

For example, Group 1 contains three "Barnes" records: A(1), A(2), and B(1), i.e. two File 'A' records have been grouped with one File 'B' record. If our linkage requirement is one-to-one, then this group contains a 'conflict' which will have to be resolved.

Figure 10: Example: group numbers show the linked records

File	ROW	SURNAME	MARST	BIRTHYR	GROUP
DATA	1	Barnes	01	1950	1
	2	Barnes	...	1950	1
	3	Jones	03	....	2
	4	Jones	02	1960	3
	5	Quigley	03	....	4
	6	Quigley	02	1960	4
DATB	1	Barnes	01	1950	1
	2	Barnes	..	1960	...
	3	Barnes	02	1960	...
	4	Jones	03	....	2
	5	Jones	02	1960	3
	6	Jones	..	1960	...
	7	Jones	02	1960	3
	8	Quigley	02	1970	4
	9	Quigley	03	1970	4

7.0 GIRLS TRAINING

As the GIRLS system is relatively complex, we strongly recommend participating in the introductory Seminar, followed by experimenting with an Example Project that has been set up for training purposes.

7.1 GIRLS Seminar

This is a one-day seminar which covers all aspects of the GIRLS system. It is given by the GIRLS system staff on an ad hoc basis. It requires the use of an overhead projector and can be presented at Statistics Canada or elsewhere. This Seminar is a valuable introduction to the system.

7.2 Example Project

This is a miniature GIRLS linkage project with two small files of test data. It consists of a sequence of batch jobs containing examples of the typical use of GIRLS commands. Submitting these jobs one at a time produces a sequence of listings showing the stages by which the records from the two files become linked. You are also encouraged to make a copy of these jobs, change the commands, and then re-submit the jobs to see the effect of your changes. This Example Project is a valuable learning tool.

8.0 HARDWARE AND SOFTWARE REQUIREMENTS

GIRLS requires the following hardware and software:

- IBM 370 compatible hardware with at least two million bytes of storage (real or virtual).
- The OS MVS or MVT operating system.
- The RAPID database management system.
- The IBM PL/1 compiler.
- Direct access storage devices (3330, 3350, 3380 etc.)
- The following are not mandatory but are highly desirable: SAS (Statistical Analysis System) in order to use the Weight Creation function, TSO or ISPF.

NOTES AND REFERENCES

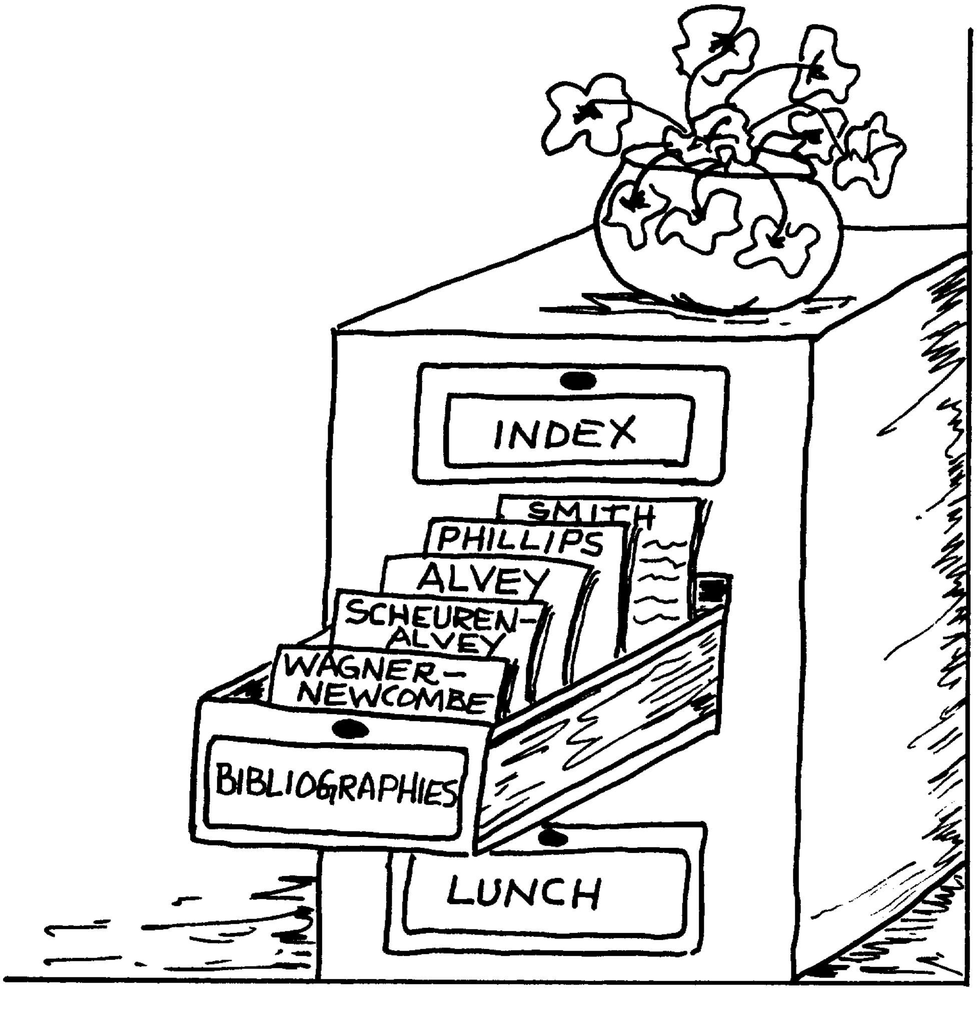
[1] Howe, G.R. and Lindsay, J.(1981). A generalized iterative record linkage system for use in medical follow-up studies. Computers and Biomedical Research, vol 14, 327-340.

[2] Newcombe, H.B. (1967). Record linking: the design of efficient systems for linking records into individual and family histories. American Journal of Human Genetics, vol 19, 335-359.

[3] Fellegi, I.P. and Sunter, A.B. (1969). A theory of record linkage. Journal of the American Statistical Association, vol 64, 1183-1210.

[4] RAPID Database Management System. Informatics Systems Division, Research and General Systems Subdivision, Statistics Canada.

Appendix A: Selected Bibliographies of Exact Matching Methodologies and Applications



UPDATED BIBLIOGRAPHY OF WORK ON EXACT MATCHING

Compiled through 1985 by
Wendy Alvey, Internal Revenue Service

This bibliography is one of the products which grew out of the Workshop on Exact Matching Methodologies, held in Arlington, Virginia, May 9-10, 1985. It draws on references from papers presented and suggested citations provided by participants who attended that conference. The aim was to round out the other bibliographic materials on matching included here, making them more current and filling in some of the historical gaps. The starting place for the effort was an earlier collection, which focused on U.S. linkage techniques during the period 1950-1974:

> Scheuren, Fritz and Alvey, Wendy. (1974) "Selected Bibliography on the Matching of Person Records from Different Sources," Proceedings of the American Statistical Association, Social Statistics Section, pp. 151-154 (pp. 347-356 in this volume).

SCOPE AND LIMITATIONS

The primary emphasis of the present bibliography is on major methodological developments and applications involving exact matching during the past eleven years. Many of the citations document recent linkage efforts involving matches of administrative and survey records for statistical purposes. The references are believed to be less complete in other areas, especially in epidemiological applications. For citations in that area, see:

> Wagner, G. and Newcombe, H.B. (1970) "Record Linkage: Its Methodology and Application in Medical Data Processing," Methods of Information in Medicine, vol. 9, no. 2, pp. 121-138 (pp. 357-374 in this volume).

While this bibliography concentrates on matches of individuals, some establishment studies are referenced, as well. However, time constraints prevented us from covering this area completely. For documentation of some of the earlier literature pertaining to matching of businesses see:

> Phillips, Bruce D. (1985) "The Development of the Small Business Data Base of the U.S. Small Business Administration: A Working Bibliography," Record Linkage Techniques--1985, Internal Revenue Service, pp. 375-379 (in this volume).

It is important to note that the present bibliography deals only tangentially with the confidentiality and disclosure issues which are so vital a part of many matching studies. In particular, just some of the more important recent references are cited. Two excellent bibliographies on privacy and confidentiality are:

> Flaherty, David H.; Hanis, Edward H.; and Mitchell, S. Paula. (1979) Privacy and Access to Government Data for Research: An International Bibliography, Mansell Publications, London, U.K; and

> Flaherty, David H. (1985) Privacy and Data Protection: An International Bibliography, Knowledge Industry Publications, Inc., White Plains, N.Y.

See also:

> U.S. Department of Commerce, Office of Federal Statistical Policy and Standards. (1978) Report on Statistical Disclosure and Disclosure Avoidance Techniques, Statistical Policy Working Paper 2, Government Printing Office, Washington, D.C.

Further, it should also be pointed out that no attempt has been made to cover the literature on synthetic or statistical matching. For a summary of the recent work in this area, see:

> Paass, Gerhard. (1985) "Statistical Record Linkage Methodology," a paper presented at the 45th Meeting of the International Statistical Institute, Amsterdam, August 1985;

> Rodgers, Willard. (1984) "An Evaluation of Statistical Matching," Journal of Business and Economic Statistics, American Statistical Association, vol. 2, no. 1, pp. 91-102; and

> Rubin, Donald. (1986) "Statistical Matching Using File Concatenation with Adjusted Weights and Multiple Imputation," Journal of Business and Economic Statistics, American Statistical Association, vol. 4, no. 1, pp. 87-94.

Finally, for a bibliography which focuses mainly on issues only indirectly related to exact matching, see also:

> Smith, Wray. (1985) "Bibliography of Methodological Techniques Related to Exact Matching," Record Linkage Techniques--1985, Internal Revenue Service, pp. 381-382 (in this volume).

ACKNOWLEDGMENT

This bibliography could not have been compiled without the extensive contributions of the Matching Group, Administrative Records Subcommittee, Federal Committee on Statistical Methodology. Thanks to them and to the participants of the Workshop on Exact Matching Methodologies for their suggestions and recommendations.

CITATIONS

Abels, Dennis. (1982) "File Matching Utilizing Automated Heuristic Techniques (FINDIT)," an unpublished working paper, Social and Scientific Systems, Inc., 7101 Wisconsin Avenue, Bethesda, MD 20814.

Acheson, E. D. (1968) Record Linkage in Medicine, E. and S. Livingstone, Edinburgh.

Alvey, Wendy and Aziz, Faye. (1979) "Quality of Mortality Reporting in SSA Linked Data: Some Preliminary Results," Proceedings of the American Statistical Association, Section on Survey Research Methods, pp. 275-279.

Alvey, Wendy and Scheuren, Fritz. (1982) "Background for an Administrative Record Census," Proceedings of the American Statistical Association, Social Statistics Section, pp. 137-146.

Arcangeli, Sam. (1985) "Pilot Test of a Proto-Type Unemployment Insurance Wage Report and Occupational Information System," an unpublished paper, State Job Training Coordinating Council, Tallahassee, FL 32301.

Arellano, Max G. (1976a) "Application of the Fellegi-Sunter Record Linkage Model to Agricultural List Files," an unpublished working paper, Statistical Reporting Service, U.S. Department of Agriculture, Washington, DC 20250.

Arellano, Max G. (1976b) "Calculation of Weights for Partitioned Variable Comparisons," an unpublished working paper, Statistical Reporting Service, U.S. Department of Agriculture, Washington, DC 20250.

Arellano, Max G. (1976c) "The Development of a Linkage Rule for Unduplicating Agricultural List Files," an unpublished working paper, Statistical Reporting Service, U.S. Department of Agriculture, Washington, DC 20250.

Arellano, Max G. (1976d) "The Estimation of P(M)," an unpublished working paper, Statistical Reporting Service, U.S. Department of Agriculture, Washington, DC 20250.

Arellano, Max G. (1976e) "Optimum Utilization of the Social Security Number for Matching Purposes," an unpublished working paper, Statistical Reporting Service, U.S. Department of Agriculture, Washington, DC 20250.

Arellano, Max G. (1976f) "Weight Calculation for the Place Name Comparison," an unpublished working paper, Statistical Reporting Service, U.S. Department of Agriculture, Washington, DC 20250.

Arellano, Max G. (1985) "An Implementation of a Two-Population Fellegi-Sunter Probability Linkage Model," Record Linkage Techniques--1985, Internal Revenue Service, pp. 255-257 (in this volume).

Arellano, Max G. and Arends, William L. (1976) "The Estimation of Component Error Probabilities for Record Linkage Purposes," an unpublished working paper, Statistical Reporting Service, U.S. Department of Agriculture, Washington, D.C. 20250.

Arellano, Max G. and Coulter, Richard W. (1976a) "Weight Calculation for the Given Name Comparison," an unpublished working paper, Statistical Reporting Service, U.S. Department of Agriculture, Washington, D.C. 20250.

Arellano, Max G. and Coulter, Richard W. (1976b) "Weight Calculation for the Middle Name Comparison," an unpublished working paper, Statistical Reporting Service, U.S. Department of Agriculture, Washington, D.C. 20250.

Arellano, Max G. and Coulter, Richard W. (1976c) "Weight Calculation for the Surname Comparison," an unpublished working paper, Statistical Reporting Service, U.S. Department of Agriculture, Washington, D.C. 20250.

Arellano, Max G. et al. (1984) "The California Automated Mortality Linkage System (CAMLIS)," American Journal of Public Health, vol. 74, no. 12, pp. 1324-1330.

Aziz, Faye and Buckler, Warren. (1980) "Mortality and the Continuous Work History Sample," Proceedings of the American Statistical Association, Section on Survey Research Methods, pp. 461-466.

Aziz, Faye; Kilss, Beth; and Scheuren, Fritz. (1978) "1973 Current Population Survey - Administrative Record Exact Match File Codebook, Part I -- Code Counts and Item Definitions," Studies from Interagency Data Linkages, report no. 8, Social Security Administration.

Baldwin, John A. and Acheson, E. Donald (Eds.). (1984) A Textbook of Medical Record Linkage, Oxford University Press.

Beebe, Gilbert W. (1980) "Record Linkage Systems--Canada vs. the United States," American Journal of Public Health, vol. 70, pp. 1246-1247.

Beebe, Gilbert W. (1981) "Record Linkage and Needed Improvements in Existing Data Resources," Banbury Report, no. 9, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, pp. 661-673.

Beebe, Gilbert W. (1985) "Why Are Epidemiologists Interested in Matching Algorithms?" Record Linkage Techniques--1985, Internal Revenue Service, pp. 139-143 (in this volume).

Bentz, Mary. (1985) "Intergenerational Wealth Study: Prospects for Data Analysis and Methodological Research," a paper presented at the Conference on Tax Modelling, Lac Ste. Marie, Quebec, Canada, September 17-19, 1985. (To appear in the Proceedings of that conference, now in progress.)

Bercini, Deborah; Sirken, Monroe; and Mathiowetz, Nancy. (1985) "Results of a Field Test Linking the National Survey of Family Growth to the National Health Interview Survey," a paper to appear in the Proceedings of the American Statistical Association, Section on Survey Research Methods.

Boruch, Robert and Cecil, J.S. (1979) Assuring the Confidentiality of Data in Social Research, University of Pennsylvania Press, Philadelphia.

Boruch, Robert and Stromsdorfer, Ernst. (1985) "Exact Matching of Microdata Sets in Social Research: Benefits and Problems," Record Linkage Techniques--1985, Internal Revenue Service, pp. 145-153 (in this volume).

Bourne, C.P. and Ford, D.F. (1961) "A Study of Methods for Systematically Abbreviating English Words and Names," Journal of the Association of Computing Machinery, vol. 8, pp. 538-552.

Bristol, Ralph B. (1985) "Age and Tax Filing, 1981," SOI Bulletin, vol. 5, no 2, Internal Revenue Service, Washington, DC.

Brown, Rachel. (1984) "Research on the Uses of Administrative Records for the 1990 Census," Proceedings of the American Statistical Association, Social Statistics Section, pp. 443-448.

Buckler, Warren. (1985) "Employer Reporting Unit Match Study (ERUMS): A Progress Report," a paper to appear in the Proceedings of the American Statistical Association, Section on Survey Research Methods.

Childers, Daniel and Hogan, Howard. (1984) "Matching IRS Records to Census Records: Some Problems and Results," Proceedings of the American Statistical Association, Section on Survey Research Methods, pp. 301-306.

Clark, Phil. (1985) "Estimating After-Tax Income Using Matched IRS-Census Data," a paper presented at the Conference on Tax Modelling, Lac Ste. Marie, Quebec, Canada, September 17-19, 1985. (To appear in the Proceedings of that conference, now in progress.)

Cohen, Malcolm. (1985) "Deriving Labor Turnover Rates From Administrative Records," Record Linkage Techniques--1985, Internal Revenue Service, pp. 259-266 (in this volume).

Coulter, Richard W. (1975) "Sampling Size in Estimating Component Error Probabilities," an unpublished working paper, Statistical Reporting Service, U.S. Department of Agriculture, Washington, DC 20250.

Coulter, Richard W. (1976a) "Processing of Comparison Pairs in Which Place Names Disagree," an unpublished working paper, Statistical Reporting Service, U.S. Department of Agriculture, Washington, DC 20250.

Coulter, Richard W. (1976b) "A Weight for 'Junior' vs. Missing," an unpublished working paper, Statistical Reporting Service, U.S. Department of Agriculture, Washington, DC 20250.

Coulter, Richard. (1985) "An Application of a Theory for Record Linkage," Record Linkage Techniques--1985, Internal Revenue Service, pp. 89-96 (in this volume). (This paper is one selection in a series of papers produced by the Department of Agriculture. It contains annotated references to the rest of the papers.

Coulter, Richard W. and Mergerson, James W. (1977) "An Application of a Record Linkage Theory in Constructing a List Sampling Frame," an unpublished working paper, Statistical Reporting Service, U.S. Department of Agriculture, Washington, DC 20250.

Cox, Lawrence. (1980) "Suppression Methodology and Statistical Disclosure Control," Journal of the American Statistical Association, vol. 75, no. 370, pp. 377-385.

Cox, Lawrence and Boruch, Robert. (1985) "Emerging Policy Issues in Record Linkage and Privacy," a paper presented at the 45th Meeting of the International Statistical Institute in Amsterdam, The Netherlands, August 12-22, 1985.

Cox, Lawrence; Johnson, Bruce; McDonald, Sarah-Kathryn; Nelson, Dawn; and Vazquez, Violeta. (1985) "Confidentiality Issues at the Census Bureau," First Annual Research Conference, Bureau of the Census, Washington, DC, pp. 199-218.

Crane, Jane and Kleweno, Douglas. (1985) "Project LINK-LINK: An Interactive Data Base of Administrative Record Linkage Studies," Record Linkage Techniques--1985, Internal Revenue Service, pp. 311-315 (in this volume).

Damerau, F. J. (1964) "A Technique for Computer Detection and Correction of Spelling Errors," Communications of the Association for Computing Machinery, vol 7., pp. 171-176.

DelBene, Linda. (1979) "1973 Current Population Survey - Administrative Record Exact Match File Codebook, Part II -- Companion Datasets and Other Supplementary Information," Studies from Interagency Data Linkages, report no. 9, Social Security Administration.

DelBene, Linda and Aziz, Faye. (1982) "Further Investigation into Mortality Coverage in Social Security Administration Data," American Statistical Association Proceedings, Section on Survey Research Methods, pp. 292-297.

Duncan, George T. and Lambert, Diane. (1986) "Disclosure-Limited Data Dissemination," Journal of the American Statistical Association, vol. 81, no. 393.

Ericksen, Eugene P. and Kadane, Joseph B. (1983) "Using Administrative Lists to Estimate Census Omissions: An Example," Proceedings of the American Statistical Association, Section on Survey Research Methods, pp. 361-366.

Fellegi, Ivan P. (1985) "Tutorial on the Fellegi-Sunter Model for Record Linkage," Record Linkage Techniques--1985, Internal Revenue Service, pp. 127-138 (in this volume).

Fett, Michael J. (1984) "Matching: The Development of Matching Criteria for Epidemiological Studies Using Record Linkage Techniques," International Journal of Epidemiology, vol. 13, no. 3, pp. 351-355.

Fink, Nancy. (1984) "Estimated Sensitivity and Specificity of the National Death Index," School of Hygiene and Public Health, Johns Hopkins University.

Flaherty, David H. (1978) "The Bellagio Conference on Privacy, Confidentiality and the Use of Government Microdata," New Directions in Program Evaluation, vol. 4, pp. 19-30.

Flaherty, David H. (1979) Privacy and Government Data Banks: An International Perspective, Mansell Publications, London, U.K.

Flaherty, David H. (1985) Privacy and Data Protection: An International Bibliography, Knowledge Industry Publications, Inc., White Plains, NY.

Flaherty, David H; Hanis, Edward H.; and Mitchell, S. Paula. (1979) Privacy and Access to Government Data for Research: An International Bibliography, Mansell Publications, London, U.K.

Greenia, Nick. (1985) "1979 Sole Proprietorship Employment and Payroll: Processing Methodology," Record Linkage Techniques--1985, Internal Revenue Service, pp. 285-289 (in this volume).

Haber, Sheldon E. (1985) "Applications of a Matched File Linking the Bureau of the Census' Survey of Income and Program Participation and Economic Data," SIPP Working Paper series, no. 8502, Bureau of the Census, Washington, DC.

Haber, Sheldon; Ryscavage, Paul; Sater, Doug; and Valdisera, Victor. (1984) "Matching Economic Data to the Survey of Income and Program Participation: A Pilot Study," Proceedings of the American Statistical Association, Social Statistics Section, pp. 529-533.

Hartigan, J.A. (1981) "Consistency of Single Linkage for High-Density Clusters," Journal of the American Statistical Association, vol. 76, pp. 388-394.

Hill, Ted. (1983) "Generalized Iterative Record Linkage System (GIRLS)," Documentation, Special Resources Subdivision, Systems Development Division, Statistics Canada, Tunney's Pasture, Ottawa, Ontario, K1A OT6, Canada.

Hill, Ted and Pring-Mill, Francis. (1985) "Generalized Iterative Record Linkage System," Record Linkage Techniques--1985, Internal Revenue Service, pp. 327-333 (in this volume).

Hirschberg, David. (1985) "The Development of the Master Establishment List," Record Linkage Techniques--1985, Internal Revenue Service, pp. 291-295 (in this volume).

Howe, G. R. (1985) "The Use of Computerized Record Linkage in Follow-up Studies of Cancer Epidemiology in Canada," Journal of the National Cancer Institute (in press).

Howe, G. R.; Fraser, D.; Lindsay, J. (1983) "Cancer Mortality (1965-77) in Relation to Diesel Fume and Coal Exposure in a Cohort of Retired Railway Workers," Journal of the National Cancer Institute, vol. 70, no. 6, pp. 1015-1019.

Howe, G. R. and Lindsay, J. P. (1981) "A Generalized Iterative Record Linkage Computer System for Use in Medical Follow-up Studies," Computers and Biomedical Research, vol. 14, pp. 327-340.

Howe, G. R. and Lindsay, J. P. (1983) "A Follow-up Study of a Ten Percent Sample of the Canadian Labor Force: 1. Cancer Mortality in Males, 1965-73," Journal of the National Cancer Institute, vol. 70, no. 1, pp. 37-44.

Howe, G. R.; Lindsay, J.; Coppock, E.; and Miller, A. B. (1979) "Isoniazid Exposure in Relation to Cancer Incidence and Mortality in a Cohort of Tuberculosis Patients," International Journal of Epidemiology, vol. 8, no. 4, pp. 305-312.

Jabine, Thomas. (1985) "Properties of the Social Security Number Relevant to Its Use in Record Linkage," Record Linkage Techniques--1985, Internal Revenue Service, pp. 213-225 (in this volume).

Jabine, Thomas and Scheuren, Fritz. (1985) "Goals for Statistical Uses of Administrative Records: The Next Ten Years," and accompanying discussion comments by William Butz, John J. Carroll, Janet Norwood and Charles Waite, Journal of Business and Economic Statistics, vol. 3, no. 4, pp. 380-404.

Jaro, Matthew. (1984) "Record Linkage Research and the Calibration of Record Linkage Algorithms," Proceedings of the American Statistical Association, Social Statistics Section, pp. 599-601.

Jaro, Matthew. (1985) "Current Record Linkage Research," Record Linkage Techniques--1985, Internal Revenue Service, pp. 317-320 (in this volume).

Jensen, Poul. (1983) "Towards a Register-Based Statistical System--Some Danish Experience," Statistical Journal of the United Nations Economic Commission for Europe, vol. 1, no. 3, pp. 341-365.

Jensen, Poul and Thygesen, Lars. (1985) "Linkage of Records on Objects of Different Kinds: Methodological Problems and Practical Experience," a paper produced at Denmarks Statistik, Copenhagen, Denmark.

Johnson, David P.; Liss, Teri L.: and Witt, Cecilie A. (1984) "1980 AHA Hospital and National Natality/Fetal Mortality Survey Linkage Methodology," Proceedings of the American Statistical Association, Section on Survey Research Methods, pp. 411-414.

Johnson, Norman; Glover, Claudia; and Rogot, Eugene. (1985) "General Mortality Among Selected Current Population Survey Cohorts for 1979-1981," a paper to appear in the Proceedings of the American Statistical Association, Section on Survey Research Methods.

Kagawa, J.T. and Mi, M.P. (1985) "On Matching With Personal Names," Record Linkage Techniques--1985, Internal Revenue Service, pp. 269-273 (in this volume).

Kasprzyk, Daniel. (1983) "Social Security Number Reporting, the Use of Administrative Records, and the Multiple Frame Design in the Income Survey Development Program," Technical, Conceptual and Administrative Lessons of the ISDP, Martin David (Ed.), Social Science Research Council, Washington, DC, pp. 123-144.

Katz, Arnold; Teuter, Klaus; and Sidel, Paul. (1984) "Comparison of Alternative Ways of Deriving Panel Data from the Annual Demographic File of the Current Population Survey," Review of Public Data Use, vol. 12, pp. 35-44.

Kelley, R. Patrick. (1984) "Blocking Considerations for Record Linkage Under Conditions of Uncertainty," Proceedings of the American Statistical Association, Social Statistics Section, pp. 602-605.

Kelley, R. Patrick. (1985) "Advances in Record Linkage Methodology: A Method for Determining the Best Blocking Strategy," Record Linkage Techniques--1985, Internal Revenue Service, pp. 199-203 (in this volume).

Kestenbaum, Bert. (1985) "The Measurement of Early Retirement," Journal of the American Statistical Association, vol. 90, no. 389, pp. 38-45.

Kilss, Beth and Alvey, Wendy (Eds.). (1985) Record Linkage Techniques--1985, Internal Revenue Service.

Kilss, Beth: Oh, H. Lock; and Scheuren, Frederick. (1977) "1964 Current Population Survey - Administrative Record Pilot Link File Codebook," Studies from Interagency Data Linkages, report no. 7, Social Security Administration.

Kilss, Beth and Scheuren, Frederick. (1978) "The 1973 CPS-IRS-SSA Exact Match Study," Social Security Bulletin, Department of Health, Education, and Welfare, vol. 41, no. 10, pp. 14-22.

Kilss, Beth; Scheuren, Fritz; and Buckler, Warren. (1980) "Goals and Plans for a Linked Administrative Statistical Sample," Proceedings of the American Statistical Association, Section on Survey Research Methods, pp. 450-455.

Kirchhoff, Bruce A. and Hirschberg, David A. (1981) "Small Business Data Base: Progress and Potential," Proceedings of the American Statistical Association, Section on Survey Research Methods, pp. 533-537.

Kirkendall, Nancy. (1985) "Weights in Computer Matching: Applications and an Information Theoretic Point of View," Record Linkage Techniques--1985, Internal Revenue Service, pp. 189-197 (in this volume).

Lubitz, J. and Pine, P. (1984) "Initial Findings: Development and Use of a Linked Medicare/NCHS Mortality File," a paper presented at the annual meeting of the American Public Health Association in Anaheim, CA, November 1984.

Lynch, B. T. and Arends, W.L. (1977) "Selection of a Surname Coding Procedure for the SRS Record Linkage System," Sample Survey Branch, Research Division, Statistical Reporting Service, U.S. Department of Agriculture, Washington, D.C.

MacMahon, Brian. (1983) "The National Death Index," American Journal of Public Health, vol. 73, no. 11, pp. 1247-1248.

Madigan, F.C. and Wells, H.B. (1976) "Report on Matching Procedures of a Dual Record System in the Southern Philippines," Demography, vol. 13, pp. 381-395.

Mi, M. (1967) "Record Linkage and Other Genetic Studies in Hawaii," Proceedings of the III International Congress of Human Genetics, J.A. Crow, and J.V. Neel (Eds.), pp. 489-496.

Mi, M., Kagawa J., Earle, M. (1983) "An Operational Approach to Record Linkage," Methods of Information in Medicine, vol. 22, pp. 77-82.

Miettinen, O.S. (1970) "Matching and Design Efficiency in Retrospective Studies," American Journal of Epidemiology, vol. 91, no. 2, pp. 111-117.

Morgan, H.L. (1970) "Spelling Correction in Systems Programs," Communications of the Association of Computing Machinery, vol. 13, pp. 90-94.

Nelson, D.O. (1976) "On the Solution of a Polynomial Arising During the Computation of Weights for Record Linkage Purposes," an unpublished working paper, Statistical Reporting Service, U.S. Department of Agriculture, Washington, DC 20250.

Newcombe, H. B. (1967) "Record Linking: The Design of Efficient Systems for Linking Records into Individual and Family Histories," American Journal of Human Genetics, vol. 19, pp. 335-359.

Newcombe, H. B. (1969) "The Use of Medical Record Linkage for Population and Genetic Studies," Methods of Information in Medicine, vol. 8, pp. 7-11.

Newcombe, H. B. and Abbatt, J.D. (1983) "Probabilistic Record Linkage in Epidemiology: Principles Employed," Eldorado Resources Ltd.

Newcombe, H. B., Smith, M.E., and Abbatt, J.D. (1982) "Linkage Procedures for the Eldorado Mortality Searches--ENL-Link 2," Eldorado Nuclear Limited.

Newcombe, H. B.; Smith, M.E.; Howe, G.R.; Mingay, J.; Strugnell, A.; and Abbatt, J.D. (1983) "Reliability of Computerized Versus Manual Death Searches in a Study of the Health of Eldorado Uranium Workers," Computers and Biomedical Research, vol. 13, no. 3, pp. 157-169.

Paass, G. (1985a) "Disclosure Risk and Disclosure Avoidance for Microdata," a paper presented at the International Association for Social Service Information and Technology, May 1985.

Paass, Gerhard. (1985b) "Statistical Record Linkage Methodology," a paper presented at the 45th meeting of the International Statistical Institute, Amsterdam, August 1985.

Patterson, John. (1980) "The Establishment of a National Death Index in the United States," Cancer Incidence in Defined Populations, Cairns, Lyon, and Skolnick (Eds.), Cold Spring Harbor Laboratory, Banbury Report 4, pp. 443-447.

Patterson, John. (1983) "Evaluation of the Matching Effectiveness of the National Death Index," Proceedings of the American Statistical Association, Social Statistics Section, pp. 1-10.

Patterson, John and Bilgrad, Robert. (1985) "The National Death Index Experience: 1981-1985," Record Linkage Techniques--1985, Internal Revenue Service, pp. 245-254 (in this volume).

Petska, Thomas. (1985) "Record Linkages Used in SOI Studies of the Business Sector," a paper presented at the Conference on Tax Modelling, Lac Ste. Marie, Quebec, Canada, September 17-19, 1985. (To appear in the Proceedings of that conference, now in progress.)

Phillips, Bruce. (1985) "The Development of the Small Business Data Base of the U.S. Small Business Administration: A Working Bibliography," Record Linkage Techniques--1985, Internal Revenue Service, pp. 375-379 (in this volume).

Plewes, Thomas. (1985) "Confidentiality Principles and Practice," First Annual Research Conference, Bureau of the Census, Washington, DC, pp. 219-226.

Pollock, J. and Zamora, A. (1984) "Automatic Spelling Correction in Scientific and Scholarly Text," Communications of the Association of Computing Machinery, vol. 27, pp. 358-368.

Prochaska, Dean; Dea, Jane; and Gaulden, Tommy. (1979) "Record Linkage for Development of the 1978 Census of Agriculture Mailing List," Proceedings of the American Statistical Association, Business and Economic Statistics Section, pp. 177-182.

Quiaoit, F. and Mi, M.P. (1985) "Surname Blocking for Record Linkage," Record Linkage Techniques--1985, Internal Revenue Service, pp. 275-281 (in this volume).

Redfern, Phillip. (1983) "A Study of the Future of the Census of Population: Alternative Approaches," a report commissioned by the Statistical Office of the European Communities.

Rodgers, Willard. (1984) "An Evaluation of Statistical Matching," Journal of Business and Economic Statistics, American Statistical Association, vol. 2, no. 1, pp 91-102.

Rogot, E.; Feinleib, M.; Ockay, K.A.; Schwartz, S.; Bilgrad, R.; and Patterson, J. (1983) "On the Feasibility of Linking Census Samples to the National Death Index for Epidemiologic Studies: A Progress Report," American Journal of Public Health, vol 73, pp. 1265-1269.

Rogot, Eugene; Schwartz, Sidney; O'Conor, Karen; and Olsen, Christina. (1983) "The Use of Probabilistic Methods in Matching Census Samples to the National Death Index," Proceedings of the American Statistical Association, Section on Survey Research Methods, pp. 319-324.

Rogot, Eugene; Sorlie, Paul; Glover, Claudia; and Johnson, Norman. (1985) "Mortality by Cause of Death Among Selected Current Population Survey Cohorts for 1979-1981," a paper to appear in the Proceedings of the American Statistical Association, Section on Survey Research Methods.

Rubin, Donald. (1986) "Statistical Matching Using File Concatenation with Adjusted Weights and Multiple Imputation," Journal of Business and Economic Statistics, American Statistical Association, vol. 4, no. 1, pp. 87-94.

Sailer, Peter. (1985) "The Longitudinal Sample in the Study of Tax-Related Issues," a paper presented at the Conference on Tax Modelling, Lac Ste. Marie, Quebec, Canada, September 17-19, 1985. (To appear in the Proceedings of that conference, now in progress.)

Sater, Douglas K. (1985) "Enhancing Data from the Survey of Income and Program Participation with Data from Economic Censuses and Surveys--A Brief Discussion of Matching Methodology," Record Linkage Techniques--1985, Internal Revenue Service, pp. 297-301 (in this volume).

Scheuren, Frederick; Bridges, Benjamin; and Kilss, Beth. (1973) "Subsampling the Current Population Survey: 1963 Pilot Link Study," Studies from Interagency Data Linkages, report no. 1, Social Security Administration.

Scheuren, Frederick; Herriot, Roger; Vogel, Linda; Vaughan, Denton; Kilss, Beth; Tyler, Barbara; Cobleigh, Cynthia; and Alvey, Wendy. (1975) "Exact Match Research Using the March 1973 Current Population Survey--Initial Stages," Studies from Interagency Data Linkages, report no. 4, Social Security Administration.

Scheuren, Frederick; Kilss, Beth; and Cobleigh, Cynthia. (1975) "1973 Current Population Survey - Summary Earnings Record Exact Match File Codebook, Part II -- Supplemental Information," Studies from Interagency Data Linkages, report no. 6, Social Security Administration.

Scheuren, Frederick; Kilss, Beth; and Oh, H. Lock. (1973) "Coverage Differences, Noninterview Nonresponse, and the 1960 Census Undercount: 1963 Pilot Link Study," Studies from Interagency Data Linkages, report no. 2, Social Security Administration.

Scheuren, Frederick J. and Oh, H. Lock. (1975) "Fiddling Around with Nonmatches and Mismatches," Proceedings of the American Statistical Association, Social Statistics Section, pp. 627-633.

Scheuren, Frederick; Oh, H. Lock; Vogel, Linda; and Yuskavage, Robert with Kilss, Beth and DelBene, Linda. (1981) "Methods of Estimation for the 1973 Exact Match Study," Studies from Interagency Data Linkages, report no. 10, Social Security Administration.

Scheuren, Frederick; Vaughan, Denton; and Alvey, Wendy. (1975) "1973 Current Population Survey - Summary Earnings Record Exact Match File Codebook, Part I -- Basic Information," Studies from Interagency Data Linkages, report no. 5, Social Security Administration.

Scheuren, Fritz. (1983) "Design and Estimation for Large Federal Surveys Using Administrative Records," Proceedings of the American Statistical Association, Section on Survey Research Methods, pp. 377-381.

Scheuren, Fritz. (1985a) "Evaluating Manpower Training: Some Notes on Data Handling Issues," a paper commissioned by the Job Training Longitudinal Survey Panel.

Scheuren, Fritz. (1985b) "Methodologic Issues in Linkage of Multiple Data Bases," a paper commissioned for the National Academy of Sciences' Panel on Statistics of an Aging Population. Also in Record Linkage Techniques--1985, Internal Revenue Service, pp. 155-178 (in this volume).

Scheuren, Fritz. (1985c) "Role of Government Statistics in a Democratic Society: A Federal Producer's Point of View," a paper presented at the Annual Meeting of the Society of Actuaries and the American Academy of Actuaries, New Orleans, Louisiana, October 14, 1985. (To appear in an abbreviated form in The Record.)

Scheuren, Fritz and Alvey, Wendy. (1974) "Selected Bibliography on the Matching of Person Records from Different Sources," Proceedings of the American Statistical Association, Social Statistics Section, pp. 151-154.

Scheuren, Fritz; Oh, H. Lock; and Alvey, Wendy with Kilss, Beth; and DelBene, Linda. (1980) "Matching Administrative and Survey Information: Procedures and Results of the 1963 Pilot Link Study," Studies from Interagency Data Linkages, report no. 3, Social Security Administration.

Smith, M. E. (1982) "A Development of a National Record Linkage Program in Canada," Proceedings of the American Statistical Association, Section on Survey Research Methods, pp. 303-308.

Smith, M.E. (1985) "Record Keeping and Data Preparation Practices to Facilitate Record Linkage," Record Linkage Techniques--1985, Internal Revenue Service, pp. 321-326 (in this volume).

Smith, M. E. and Newcombe, H. B. (1975) "Methods for Computer Linkage in Hospital Admission - Separation Records into Cumulative Health Histories, " Methods of Information in Medicine, vol. 14, pp. 118-125.

Smith, M. E., and Newcombe, H. B. (1979) "Accuracies of Computer Versus Manual Linkages of Routine Health Records," Methods of Information in Medicine, vol. 18, pp. 89-97.

Smith, M.E., Newcombe, H.B., and Dewar, R.A.D. (1983) "Automated Nationwide Death Clearance of Provincial Cancer Register Files--The Alberta Cancer Registry Study," Proceedings of the American Statistical Association, Section on Survey Research Methods, pp. 300-305.

Smith, M. E. and Silins, J. (1984) "Generalized Iterative Record Linkage System," (an excerpt) Statistical Uses of Administrative Records: Recent Research and Present Prospects, vol. 2, pp. 513-522, Internal Revenue Service, Washington, DC.

Smith, Wray. (1984) "Background Research and Issue Identification in Exact Matching of Microdata," Task 5 Report, Mathematica Policy Research, Inc., Washington, DC.

Smith, Wray. (1985) "Bibliography of Methodological Techniques Related to Exact Matching," Record Linkage Techniques--1985, Internal Revenue Service, pp. 381-382 (in this volume).

Smith, Wray and Scheuren, Fritz. (1985a) "Multiple Linkage and Measures of Inexactness: Methodology Issues," a paper presented at the Workshop on Exact Matching Methodologies, Arlington, VA, May 9, 1985.

Smith, Wray and Scheuren, Fritz. (1985b) "Some New Methods in Statistical Disclosure Avoidance," a paper presented to the Annual Meeting of the American Statistical Association, Section on Survey Research Methods, Las Vegas, August 1985.

Social Security Administration, Office of Research and Statistics. (1979) LASS Working Notes, nos. 1-7, Beth Kilss, Wendy Alvey, Linda DelBene, Faye Aziz, and Fritz Scheuren (Eds.). (A series of working papers on the feasibility of creating Linked Administrative Statistical Samples for epidemiological research.)

Spruill, Nancy. (1983) "The Confidentiality and Analytic Usefulness of Masked Business Microdata," Proceedings of the American Statistical Association, Section on Survey Research Methods, pp. 602-607.

Spruill, Nancy. (1984) Protecting Confidentiality of Business Microdata by Masking, The Public Research Institute, Alexandria, VA.

State of California, Employment Development Department. (1979) Technical Development Issues: Employment Service Potential, vol. II.

Steinberg, J. and Pritzker, L. (1967) "Some Experiences with and Reflections on Data Linkage in the United States," Bulletin of the International Statistical Institute, vol. 42, pp. 786-805.

Taft, R. (1970) "Name Search Techniques," Project Search (System for Electronic Analysis and Retrieval of Criminal Histories) Special Report No. 1, New York State Identification and Intelligence System (NYSIIS), Bureau of Systems Development, Albany.

Tepping, Benjamin J. (1971) "The Application of a Linkage Model to the Chandrasekar-Deming Technique for Estimating Vital Events," Technical Notes 4, Bureau of the Census.

Thygesen, L. (1983) "Methodological Problems Connected with a Socio- Demographic Statistical System Based on Administrative Records," Bulletin of the International Statistical Institute, vol. 50, book 1, Madrid, pp. 227-242.

U.S. Department of Agriculture, Statistical Reporting Service. (1975-1977). Series of working papers (for the most part unpublished) on the development of a record linkage system. (Individual contributions are listed separately.)

U.S. Department of Agriculture, Statistical Reporting Service. (1976) "Partitioned Variable Comparison/Algorithm for Identifying Configurations," an unpublished working paper.

U.S. Department of Commerce, National Bureau of Standards. (1977) "Accessing Individual Records from Personal Data Files Using Non-Unique Identifiers," NBS Special Publication 500-2.

U.S. Department of Commerce, Office of Federal Statistical Policy and Standards. (1978a) A Framework for Planning U.S. Federal Statistics for the 1980's, Washington, DC.

U.S. Department of Commerce, Office of Federal Statistical Policy and Standards. (1978b) Report on Statistical Disclosure and Disclosure Avoidance Techniques, Statistical Policy Working Paper 2, Government Printing Office, Washington, DC.

U.S. Department of Commerce, Office of Federal Statistical Policy and Standards. (1980) Report on Exact and Statistical Matching Techniques, Statistical Policy Working Paper 5, Government Printing Office, Washington, DC.

U.S. Department of Health and Human Services, Social Security Administration. (1980) Economic and Demographic Statistics, Wendy Alvey, Beth Kilss and Fritz Scheuren (Eds.).

U.S. Department of Health, Education, and Welfare, Social Security Administration. (1973-1980) Series on Studies from Interagency Data Linkages, reports numbers 1-11. (Individual contributions are listed separately.)

U.S. Department of Health, Education, and Welfare, Social Security Administration. (1973) "Some Observations on Linkage of Survey and Administrative Record Data," Studies from Interagency Data Linkages, DHEW Pub. No. (SSA) 74-11750.

U.S. Department of Health, Education, and Welfare, Social Security Administration. (1978) Policy Analysis With Social Security Research Files, Research Report No. 52, DHEW Pub. No. (SSA) 79-11808, Washington, DC.

U.S. Department of Health, Education, and Welfare, Social Security Administration. (1979) Statistical Uses of Administrative Records with Emphasis on Mortality and Disability Research, Linda DelBene and Fritz Scheuren (Eds.).

U.S. Department of Health, Education, and Welfare, Social Security Administration. (1980) "Measuring the Impact on Family and Personal Income Statistics of Reporting Differences Between the Current Population Survey and Administrative Sources," Studies from Interagency Data Linkages, report no. 11.

U.S. Department of the Treasury, Internal Revenue Service. (1981-1984) Series on Statistics of Income and Related Administrative Record Research, Wendy Alvey and Beth Kilss (Eds.).

U.S. Department of the Treasury, Internal Revenue Service. (1984) Statistical Uses of Administrative Records: Recent Research and Present Prospects, vols. 1 and 2, Beth Kilss and Wendy Alvey (Eds.).

Wagner, G., and Newcombe, H.B. (1970) "Record Linkage: Its Methodology and Application in Medical Data Processing (A Bibliography)," Methods of Information in Medicine, vol. 9, no. 2, pp. 121-138.

Wentworth, D.; Neaton, J.; and Rassmussen, W. (1983) "An Evaluation of the Social Security Administration Master Beneficiary Record File and the National Death Index in the Ascertainment of Vital Status," American Journal of Public Health, vol. 73, pp. 1270-1274.

Winkler, William. (1984a) "Issues in Developing Frame Matching Procedures," a paper presented to the American Statistical Association's Committee on Energy Statistics, April 1984.

Winkler, William. (1984b) "Exact Matching Using Elementary Techniques," Proceedings of the Amercan Statistical Association, Section on Survey Research Methods, pp. 237-242.

Winkler, William. (1985a) "Exact Matching Lists of Businesses: Blocking, Subfield Identification and Information Theory," Record Linkage Techniques--1985, Internal Revenue Service, pp. 227-241 (in this volume).

Winkler, William. (1985b) "Preprocessing of Lists and String Comparisons," Record Linkage Techniques--1985, Internal Revenue Service, pp. 181-187 (in this volume).

Ziegler, Martin. (1977) "Efforts to Improve Estimates of State and Local Unemployment," Monthly Labor Review, November, pp. 12-18.

SELECTED BIBLIGRAPHY ON THE MATCHING OF PERSON RECORDS FROM DIFFERENT SOURCES

Compiled through 1974 by
Fritz Scheuren and Wendy Alvey, Social Security Administration

The references listed here are restricted essentially to published information on the results and methodology of matching person records. No material relating to establishment matching is included. Several compilations of articles exist on the confidentiality issues raised by record linkages. As a rule, therefore, we have excluded these citations from the bibliography. Two recent such compilations are:

> *Report of the President's Commission on Federal Statistics*, vol. 1, 1971, pp. 246-254.
>
> U. S. Department of Health, Education, and Welfare. *Records, Computers, and the Rights of Citizens: Report of the Secretary's Advisory Committee on Automated Personal Data Systems*, 1973, pp. 298-330.

Some other limitations should also be mentioned:

1. The listing does not contain citations to studies which began with an administrative record and then drew a sample of cases to be interviewed. (Excluded from the bibliography, therefore, are basically all *reverse* record check studies of financial characteristics, as well as *prospective* epidemiological studies.)

2. Only rereferences to "exact" matching are included. Synthetic or "statistical" matching studies are not shown. (For citations to the literature on synthetic matches, see Radner, D. B. *The Statistical Matching of Microdata Sets: the Bureau of Economic Analysis 1964 Current Population Survey--Tax Model Match*, Yale Univ., New Haven, 1974. See also the *Annals of Economic and Social Measurement*, vol. 3, 1974, where several articles on synthetic matching are presented.)

3. Studies of matching for use in Dual System Estimation are also not included. For a recently published source of information in this area, see Marks, E.S., Seltzer, W. and Krotki, K.J. *Population Growth Estimation: A Handbook of Vital Statistics Measurement*, The Population Council, New York, 1974.

4. Studies involving record linkage in medicine are covered only partially. For additional references in this area the reader might consult Acheson, E.D. *Medical Record Linkage*. Oxford University, London, 1967 or *Record Linkage in Medicine, Proc. Int. Sym., Oxford, July 1967*, Williams and Wilkins, Baltimore, 1968.

COMPLETENESS OF COVERAGE

The bibliography's coverage of *major* U. S. studies involving linkages between survey (or census) schedules and administrative records is believed to be reasonably complete for the period 1950-1974. However, only a few references are given to work done outside the United States and to research engaged in before 1950.

SOURCES

Many of the citations shown here were selected from the following three reference works:

> U. S. Bureau of the Census. Indexes to survey methodology literature, *Technical Paper No. 34*, 1974.
>
> Forsythe, J. *List of References on Results and Methodology of Matching Studies*, 1966 (Unpublished Census Bureau Memorandum).
>
> U. S. Public Health Service. Use of vital and health records in epidemiologic research, *Vital and Health Statistics*, series 4, no.7, 1968.

ACKNOWLEDGMENTS

Special thanks are due to Irene Jacobs, Office of Research and Statistics Library, Social Security Administration, and Patricia Fuellhart and her staff, Statistical Research Division, Bureau of the Census, for their assistance in this effort.

[1] Acheson, E. D. The Oxford Record Linkage Study, Report on the Second Year's Operations, Oxford Regional Hospital Board, Oxford, 1963.

[2] Acheson, E. D. Oxford Record Linkage Study, Brit. Jour. Prev. and Soc. Med., vol. 18, 1964, pp. 8-13.

[3] Acheson, E. D. The Oxford Record Linkage Study, a review of the method with some preliminary results, Proc. Roy. Soc. Med., vol. 57, 1964, pp. 11 ff.

[4] Acheson, E. D., and Evans, J. G. The Oxford Record Linkage Study, Biometrics, vol. 2, 1963, pp. 367 ff.

[5] Anderson, O. W., and Feldman, J. J. Family Medical Costs and Voluntary Health Insurance: a Nationwide Survey, McGraw-Hill, New York, 1956.

[6] Anderson, O. W., and Sheatsley, P. B. Comprehensive Medical Insurance, Health Information Foundation Research Series, No. 9.

[7] Anderson, R., and Anderson, O. W. A Decade of Health Services, Univ. of Chicago, Chicago, 1967.

[8] Bachi, R., Baron, R., and Nathan, G. Methods of record-linkage and applications in Israel, Bulletin of the International Statistical Institute, vol. 42, Proc. of 36th Session, Sydney, 1967, pp. 751-765.

[9] Bahn, A. K. Methodological study of population of out-patient psychiatric clinics, Maryland, 1958-59, Public Health Monograph No. 65., PHS Pub. no. 821, 1961.

[10] Balamuth, E. Health interview responses compared with medical records, U. S. Public Health Service, Vital and Health Statistics, series 2, no. 7, 1965.

[11] Bancroft, G. The American Labor Force: Its Growth and Changing Composition, Wiley, New York, 1958, pp. 151-175.

[12] Belloc, N. B. Validation of morbidity survey data by comparison with hospital records, J. Amer. Stat. Assn., vol. 49, 1954, pp. 832-846.

[13] Binder, S. Present possibilities and future potentialities, The Use of Vital and Health Records for Genetic and Radiation Studies, United Nations, New York, 1962, pp. 161-170.

[14] Bixby, L. E. Income ot people aged 65 or older: overview trom the 1968 survey of the aged, Social Security Bulletin, report no. 1, 1970, pp. 28-34.

[15] Brounstein, S. H. Data Record Linkage Under Conditions of Uncertainty, 7th Annual Conference of the Urban and Regional Information Systems Association, Los Angeles, California, 1969.

[16] Chase, H. C. A study of infant mortality from linked records: method of study and registration aspects, U. S. Public Health Service, Vital and Health Statistics, series 20, no. 7, 1970.

[17] Christenson, H. T., Cultural relativism and premarital sex norms, Amer. Soc. Rev., vol. 15, 1960.

[18] David, M., Gates, W., and Miller, R. Linkage and retrieval of microeconomic data. Heath, Lexington, Mass., 1974.

[19] Davidson, L. Retrieval of misspelled names in an airline passenger records system, Communications of the ACM, vol. 5, 1962, pp. 169-171.

[20] Densen, P. M., and Shapiro, S. Research needs for record matching, Proc. Amer. Stat. Assn. Soc. Stat. Sec., 1963, pp. 20-24.

[21] DuBois, Jr., N. S. D. On the problem of matching documents with missing and inaccurately recorded items (preliminary report), Annals of Mathematical Statistics, vol. 35, 1964, p. 1404.

[22] DuBois, Jr., N. S. D. A document linkage program for digital computers, Behavioral Science, vol. 10, 1965, pp. 312-319.

[23] DuBois, Jr., N. S. D. A solution to the problem of linking multivariate documents, Jour. Amer. Stat. Assn., vol. 64, 1969, pp. 163-174.

[24] Dunn, H. L. Record linkage, American Journal of Public Health, vol. 36, 1946.

[25] Dunn, H. L., and Grove, R. D. Completeness of birth registration in the United States, Estadistica, vol. 1, 1943, pp. 3-17.

[26] Elinson, J., and Trussell, R. E. Some factors relating to degree of correspondence of diagnostic information obtained by household interviews and clinical examinations, American Journal of Public Health, vol. 47, 1957, pp. 311-321.

[27] Fellegi, I. P., and Sunter, A. B. A theory for record linkage, Jour. Amer. Stat. Assn., vol. 64, 1969, pp. 1183-1210.

[28] Ferber, R. The Reliability of Consumer Reports of Financial Assets and Debts, University of Illinois, Urbana, 1966.

[29] General Register Office. 1951 Census of England and Wales, General Report, London, 1958, pp. 41-45, 5056.

[30] Goldberg, I. D., Goldstein, H., Quade, D., and Rogot, E. The use of vital records for blindness research, Amer. Jour. Pub. Health, vol. 54, 1964, pp. 278-285.

[31] Goldberg, S. Discussion of data storage and linkage, Bulletin of the International Statistical Institute, Proc. of 36th Session, Sydney, vol. 42, 1967, pp. 806-808.

[32] Grove, R. D. Studies in the completeness of birth registration, U. S. Bureau of the Census, Vital Statistics -- Special Reports, vol. 27, no. 18, 1943.

[33] Guralnick, L., and Nam, C. B. Census-NOVS study of death certificates matched to census records, The Milbank Memorial Fund Quarterly, vol. 37, 1959, pp. 144-153.

[34] Haber, L. D. Evaluating response error in the reporting of the income of the aged: benefit income, Proc. Amer. Stat. Assn. Soc. Stat. Sec., 1966.

[35] Haenszel, W., Loveland, D. B., and Sirken, M. G. Lung-cancer mortality as related to residence and smoking histories. I. white males, Journal of the National Cancer Institute, vol. 28, no. 4, 1962, pp. 947-1001.

[36] Hambright, T. Z. Comparability of age on the death certificate and matching census record, U. S., May-Aug. 1960, U. S. Public Health Service, Vital and Health Statistics, series 2, no. 29, 1968.

[37] Hansen, M. H. The role and feasibility of a national data bank, based on matched records and interviews, Report of the President's Commission on Federal Statistics, vol. 2, pp. 1-63.

[38] Hauser, P. M., and Kitagawa, E. M. Social and economic mortality differentials in the United States, 1960: outline of a research project. Proc. Amer. Assn. Soc. Stat. Sec., 1960, pp. 116-120.

[39] Hauser, P. M., and Lauriat, P. Record matching—theory and practice (abstract), Proc. Amer. Stat. Assn. Soc. Stat. Sec., 1963, p. 25.

[40] Hedrich, A. W., Collison, J., and Rhoads, F. D. Comparison of birth tests by several methods in Georgia and Maryland, U. S. Bureau of the Census, Vital Statistics -- Special Reports, vol. 7, no. 60, 1939, pp. 681-695.

[41] Hoel, P. G., and Peterson, R. P. A solution to the problem of optimum classification, Annals of Mathematical Statistics, vol. 20, 1949, pp. 433-438.

[42] Hogben, L., Johnstone, M. M., and Cross, K. L. Identification of medical documents, British Med. J., 1948, pp. 632-635.

[43] Horn, W. Reliability Survey: a survey on the reliability of responses to an interview survey, Het PTT-bedrijf, vol. 10, 1960, pp. 105-156.

[44] Horn, W. Nonresponse in an interview survey: some specific phenomena (a case-study), Het PTT-bedrijf, vol. 12, 1963, pp. 11-19.

[45] Jaro, M. Unimatch--a computer system for generalized record linkage under conditions of uncertainty, Spring Joint Computer Conference, 1972, AFIPS--Conference Proceedings, vol. 40, 1972, pp. 523-530.

[46] Johnson, Jr., C. E. Consistency of reporting of ethnic origin in the Current Population Survey, Technical Paper No. 31, 1974, pp. 23-27.

[47] Kaplan, D. L., Parkhurst, E., and Whelpton, P. K. The comparability of reports on occupation from vital records and the 1950 census, Vital Statistics -- Special Reports, vol. 53, no. 1, 1961, pp. 1-27.

[48] Kelly, T. F. Factors affecting poverty: a gross flow analysis, The President's Commission on Income Maintenance Programs: Technical Studies, 1970, pp. 1-82.

[49] Kelly, T. F. The creation of longitudinal data from cross-section surveys: an illustration from the Current Population Survey, Annals of Economic and Social Measurement, vol. 2, 1973, pp. 209-214.

[50] Kelly, W. H. Methods and Resources for the Construction and Maintenance of a Navajo Population Register, a report prepared for the National Cancer Institute by the Bureau of Ethnic Research, Department of Anthropology, University of Arizona, Tucson, 1964.

[51] Kennedy, J. M. Linkage of Birth and Marriage Record Using a Digital Computer, Atomic Energy of Canada, Ltd., Chalk River, Ontario, 1961.

[52] Kennedy, J. M. The use of a digital computer for record linkages, The Use of Vital and Health Statistics for Genetic and Radiation Studies, United Nations, New York, 1962, pp. 155-160.

[53] Kitagawa, E. M., and Hauser, P. M. Methods used in a current study of social and economic differences in mortality, Emerging Techniques in Population Research, Milbank Memorial Fund, New York, 1963, pp. 250-266.

[54] Kitagawa, E. M., and Hauser, P. M. Differential Mortality in the United States: a Study in Socioeconomic Epidemiology, Harvard Univ., Cambridge, 1973, pp. 183-227.

[55] Klebba, A. J., Maurer, J. D., and Glass, E. J. Mortality trends: age, color, and sex, U. S. 1950-69, U. S. Public Health Service, Vital and Health Statistics, series 20, no. 15, 1973.

[56] Livingston, R. Evaluation of the reporting of public assistance income in the special census of Dane County, Wisconsin (May 15, 1968), Proc. Ninth Workshop on Public Welfare Research and Statistics, New Orleans, 1969.

[57] Livingston, R. Evaluating the reporting of public assistance income in the 1966 Survey of Economic Opportunity, Proc. Tenth Workshop on Public Welfare Research and Statistics, 1970.

[58] Lloyd, J. W. Discussion of matching of census and vital records in social and health research: problems and results, Proc. Amer. Stat. Assn. Soc. Stat. Sec., 1965, p. 141.

[59] Loeb, J. Weight at birth and survival of the newborn by age of mother and total-birth order, U. S., early 1950, U. S. Public Health Service, Vital and Health Statistics, series 21, no. 5, 1965.

[60] Mandel, B. J., Wolkstein, I., and Delaney, M. M. Coordination of old-age and survivors insurance wage records and the post-enumeration survey, Studies in Income and Wealth: an Appraisal of the 1950 Census Income Data, Princeton, vol. 23, 1958, pp. 169-178.

[61] Marks, E. S., Mauldin, W. P., and Nisselson, H. The post-enumeration survey of the 1950 censuses: a case history in survey design, Jour. Amer. Stat. Assn., 1953, pp. 220-243.

[62] Marks, E. S., and Waksberg, J. Evaluation of coverage in the 1960 Census of Population through case-by-case checking, Proc. Amer. Stat. Assn. Soc. Stat. Sec., 1966, pp. 62-70.

[63] Masi, A. T., Sartwell, P. E., and Shulman, L. E. The use of record linkage to determine familial occurrence of disease from hospital records (Hashimoto's disease), Am. J. Pub. Health, vol. 54, 1964, pp. 1887-1894.

[64] McCarthy, M. A. Comparison of the classification of place of residence on death certificates and matching census records, U. S. Public Health Service, Vital and Health Statistics, series 2, no. 30, 1969.

[65] Meltzer, J. W., and Hochstim, J. R. Reliability and validity of survey data on physical health, Public Health Reports, vol. 85, 1970, pp. 1075-1086.

[66] Miller, H. P., and Paley, L. R. Income reported in the 1950 census and on income tax returns, Studies in Income and Wealth: an Appraisal of the 1950 Census Income Data, Princeton, vol. 23, 1958, pp. 179-201.

[67] Moriyama, I. M. Uses of vital records for epidemiological research, J. Chron. Dis., vol. 17, 1964, pp. 889-897.

[68] Morrison, F. S., and Blen, A. L. Method of testing in the 1950 census the completeness of birth registration, Estadistica, vol. 7, 1949, pp. 185-193.

[69] Murray, J. H., and Haber, L. D. Methodology and validation, Appendix A of the Aged Population of the United States: the 1963 Social Security Survey of the Aged, by L. A. Epstein and J. H. Murray. SSA/ORS research report no. 19, 1967, pp. 193-219.

[70] Nathan, G. On Optimal Matching Processes, Case Institute of Technology, Cleveland, 1964.

[71] Nathan, G. Outcome probabilities for a record matching process with complete invariant information, J. Amer. Stat. Assn., vol. 62, 1967, pp. 454-469.

[72] Neter, J., Maynes, E. S., and Ramanathan, R. The effect by mismatching on the measurement of response errors, Proc. Amer. Stat. Assn. Soc. Stat. Sec., 1964, pp. 2-8.

[73] Neter, J., Maynes, E. S., and Ramanathan, R. The effect of mismatching on the measurement of response errors, J. Amer. Stat. Assn., vol. 60, 1965, pp. 1005-1027.

[74] Newcombe, H. B. Detection of genetic trends in public health, Effect of Radiation on Human Heredity, World Health Organization, Geneva, 1957, pp. 157-168.

[75] Newcombe, H. B. Environmental versus genetic interpretations of birth-order effects, Eugenics Quarterly, vol. 11, 1964, p. 36ff.

[76] Newcombe, H. B. Panel discussion, session on epidemiological studies, Second International Conference on Congenital Malformations, The International Medical Congress, Ltd., New York, 1964, pp. 345-349.

[77] Newcombe, H. B. Pedigrees for population studies, a progress report, Cold Spring Harbor Symposia on Quantitative Biology, vol. 29, 1964, p. 21ff.

[78] Newcombe, H. B. Population genetics, population records, Methodology in Human Genetics, Holden-Day, New York, 1962, pp. 92-113.

[79] Newcombe, H. B. Risk of fetal death to mothers of different ABO and RH blood type, Am. J. Human Genet., vol. 15, 1963, pp. 449-464.

[80] Newcombe, H. B. Screening for efects of maternal age and birth order in a register of handicapped children, Ann. Human Genet., vol. 27, 1964, pp. 367-382.

[81] Newcombe, H. B. Untapped knowledge of human populations, Transaction of the Royal Society of Canada, vol. 56, series 3, section 3, 1962, pp. 173-180.

[82] Newcombe, H. B., Axford, S. J., and James, A. P. A plan for the study of fertility of relatives of children suffering from hereditary and other defects, Atomic Energy of Canada, Ltd., report no. 511, Chalk River, Ontario, 1957, p. 50.

[83] Newcombe, H. B., James, A. P., and Axford, S. J. Family linkage of vital and health records, Atomic Energy of Canada, Ltd., report no. 470, Chalk River, Ontario, 1957.

[84] Newcombe, H. B., and Kennedy, J. M. Record linkage making maximum use of the discrimination power of identifying information, Communications of the ACM, vol. 5, 1962, pp. 563-566.

[85] Newcombe, H. B., Kennedy, J. M., Axford, S. J., and James, A. P. Automatic linkage of vital records, Science, vol. 130, 1959, pp. 954-959.

[86] Newcombe, H. B., and Rhynas, P.O. Child spacing following stillbirth and infant death, Eugenics Quarterly, vol. 9, 1962, pp. 25-35.

[87] Newcombe, H. B., and Rhynas, P. O. Family linkage of population records, The Use of Vital and Health Statistics for Genetic and Radiation Studies, United Nations, New York, 1962, pp. 135-154.

[88] Newcombe, H. B., and Tavendale, O. G. Effects of father's age on the risk of child handicap or death, Am. J. Human Genet., vol. 17, 1965, pp. 163-178.

[89] Newman, S. M. Problem in mechanizing the search in examining patent applications, Patent Office Research and Development Report No. 3, 1956.

[90] Nicholson, W. The income data series in the graduated work incentive experiment: an analysis of their differences, Chapter 6, Part C, The New Jersey Graduated Work Incentive Experiment, 1974.

[91] Nitzberg, D. M., and Sardy, H. The methodology of computer linkage of health and vital records, Proc. Amer. Stat. Assn. Soc. Stat. Sec., 1965, pp. 100-106.

[92] Ohlsson, I. Merging of data for statistical use, Bulletin of the International Statistical Institute, vol. 42, Proc. of 36th Session, Sydney, 1967, pp. 751-765.

[93] Ono, M., Patterson, G. F., and Weitzman, M. S. The quality of reporting social security numbers in two surveys, Proc. Amer. Stat. Assn. Soc. Stat. Sec., 1968, pp. 197-205.

[94] Perkins, W. M., and Jones, O. D. Matching for census coverage checks, Proc. Amer. Stat. Assn. Soc. Stat. Sec., 1965, pp. 122-139.

[95] Phillips, Jr., W., and Bahn, A. K. Experience with computer matching of names, Proc. Amer. Stat. Assn. Soc. Stat. Sec., 1963, pp. 26-29.

[96] Phillips, Jr., W., Bahn, A. K., and Miyasaki, M. Person-matching by electronic methods, Communications of the ACM, vol. 5, 1962, pp. 404-407.

[97] Phillips, Jr., W., Gorwitz, K., and Bahn, A. K. Electronic maintenance of case registers, Public Health Reports, vol. 77, 1962, pp. 503-510.

[98] Pollack, E. S. Use of census matching for study of psychiatric admission rates, Proc. Amer. Stat. Assn. Soc. Stat. Sec., 1965, pp. 107-115.

[99] President's Committee to Appraise Employment and Unemployment Statistics. Measuring Employment and Unemployment, 1962, pp. 386-394.

[100] Rogers, P. B., Council, C. R., and Abernathy, J. R. Testing death registration completeness in a group of premature infants, Public Health Reports, vol. 76, no. 8, 1961, pp. 717-724.

[101] Schachter, J. Matched record comparison of birth certificate and census information: United States, 1950, U. S. Public Health Service, Vital Statistics — Special Reports, vol 47, no. 12, 1962, pp. 365-399.

[102] Scheuren, F. J., Bridges, B., and Kilss, B. Report no. 1: subsampling the Current Population Survey: 1963 Pilot Link Study, Studies from Interagency Data Linkages, Social Security Administration, 1973.

[103] Scheuren, F. J., and West, G. 1966 and 1967 Survey of Economic Opportunity: Computer Consistency Checks, Office of Economic Opportunity, 1971, pp. 113-124 (Processed).

[104] Schneider, P., and Knott, J. Accuracy of census data as measured by the 1970 CPS-Census-IRS Matching Study, Proc. Amer. Stat. Assn. Soc. Stat. Sec., 1973, pp. 152-159.

[105] Shapiro, S. Estimating birth registration completeness, J. Amer. Stat. Assn., vol. 45, 1950, pp. 261-264.

[106] Shapiro, S. Recent testing of birth registration completeness in the United States, Population Studies, vol. 8, 1954, pp. 3-21.

[107] Shapiro, S., and Densen, P. Research needs for record matching, Proc. Amer. Stat. Assn. Soc. Stat. Sec., 1963, pp. 20-24.

[108] Shapiro, S., and Schachter, J. Birth registration completeness, 1950, Public Health Reports, vol. 67, 1952, pp. 513-524.

[109] Shapiro, S., and Schachter, J. Methodology and summary results of the 1950 birth registration test in the United States, Estadistica, vol. 10, no. 37, 1952, pp. 688-699.

[110] Siegel, J. S. 1970 Census of Population and Housing Evaluation and Research Program, Estimates of Coverage of Population by Sex, Race and Age: Demographic Analysis, U. S. Bureau of the Census, PHC(E)-4, 1974.

[111] Siegel, J. S., and Zelnik, M. An evaluation of coverage in the 1960 census of population by techniques of demographic analysis and by composite methods, Proc. Amer. Stat. Assn. Soc. Stat. Sec., 1966, pp. 71-85.

[112] Silver, J. Discussion of matching of census and vital records in social and health research: problems and results, Proc. Amer. Stat. Assn. Soc. Stat. Sec., 1965, p. 140.

[113] Simpson, J. E., and Von Arsdol, Jr., M.D. The matching of census and probation department record systems, Proc. Amer. Stat. Assn. Soc. Stat. Sec., 1965, pp. 116-121.

[114] Sirken, M. G. Hospital utilization in the last year of life, U. S. Public Health Service, Vital and Health Statistics, series 2, no. 10, 1965.

[115] Sirken, M. G. Research uses of vital records in vital statistics surveys, Research Methods in Health Care, ed. by J. B. McKinlay, Prodist, New York, 1973, pp. 39-46.

[116] Sirken, M. G., Maynes, E. S., and Frechtling, J. A. The survey of consumer finances and the census quality check, Studies in Income and Wealth: an Appraisal of the 1950 Census Income Data, vol. 23, Princeton, 1958, pp. 127-168.

[117] Sirken, M. G., Pifer, J. W., and Brown, M. L. Design of Surveys Linked to Death Records, 1962.

[118] Steinberg, J. Some aspects of statistical data linkage for individuals, Data-Bases, Computers, and the Social Sciences, Wiley, New York, 1970, pp. 238-251.

[119] Steinberg, J. Some observations on linkage of survey and administrative record data, Studies from Interagency Data Linkages, Social Security Administration, 1973, pp. 1-14.

[120] Steinberg, J., Hearn, S., and Deutch, J. Social Security Administration's evaluation and measurement system, Proc. Amer. Stat. Assn. Soc. Stat. Sec., 1966, pp. 262-268.

[121] Steinberg, J., and Pritzker, L. Some experiences with and reflections on data linkage in the United States, Bulletin of the International Statistical Institute, vol. 42, 1967, pp. 786-805.

[122] Sunter, A. B. A statistical approach to record linkage, record linkage in medicine, Proc. of the International Symposium in Oxford, E. & S. Livingstone, Ltd., London, 1968.

[123] Tepping, B. J. Study of Matching Techniques for Subscriptions Fulfillment, Philadelphia, 1955.

[124] Tepping, B. J. Progress Report on the 1959 Matching Study, National Analysts Inc., Philadelphia, 1960.

[125] Tepping, B. J. A model for optimum linkage of record, J. Amer. Stat. Assn., vol. 63, 1968, pp. 1321-1332.

[126] Tepping, B. J. The application of a linkage model to the Chandrasekar-Deming technique for estimating vital events, Technical Notes 4, U. S. Bureau of the Census, Washington, D. C., 1971.

[127] Tepping, B. J., and Bailar, B. A. Enumerator variance in the 1970 census, Proc. Amer. Stat. Assn. Soc. Stat. Sec., 1973, pp. 160-169.

[128] Tepping, B. J., and Chu, J. T. A Report on Matching Rules Applied to Reader's Digest Data, National Analysts, Inc., Philadelphia, 1958.

[129] Turner, Jr., M. L. A new technique measuring household change, Demography, vol. 4, 1967, pp. 341-351.

[130] U. S. Bureau of the Census. Infant enumeration study: 1950 completeness of enumeration of infants related to: residence, race, birth month, age and education of mother, occupation of father, Procedural studies of the 1950 Census, no. 1, 1953.

[131] U. S. Bureau of the Census. The post-enumeration survey: 1950, *Technical Paper, No. 4*, 1960.

[132] U. S. Bureau of the Census. *Evaluation and Research Program of the U. S. Censuses of Population and Housing, 1960: Record Check Studies of Population Coverage*, series ER60, no. 2, 1964.

[133] U. S. Bureau of the Census. *Evaluation and Research Program of the U. S. Censuses of Population and Housing, 1960: Accuracy of Data on Population Characteristics as Measured by CPS-Census Match*, series ER60, no. 5, 1965.

[134] U. S. Bureau of the Census. *Evaluation and Research Program of the U. S. Censuses of Population and Housing, 1960: the Employer Record Check*, series ER60, no. 6, 1965.

[135] U. S. Bureau of the Census. *Evaluation and Research Program of the U. S. Censuses on Population and Housing, 1960: Effects of Interviewers and Crew Leaders*, series ER60, no. 7, 1968.

[136] U. S. Bureau of the Census. *Evaluation and Research Program of the U. S. Population and Housing, 1960: Record Check of Accuracy of Income Reporting*, series ER60, no. 8, 1970.

[137] U. S. Bureau of the Census. *1970 Census of Population and Housing Evaluation and Research Program: Test of Birth Registration Completenese 1964 to 1968*, PHC(E)-2, 1973.

[138] U. S. Bureau of the Census. *1970 Census of Population and Housing Evaluation and Research Program: the Coverage of Housing in the 1970 Census*, PHC(E)-5, 1973.

[139] U. S. Bureau of the Census. *1970 Census of Population and Housing Evaluation and Research Program: the Medicare Record Check: an Evaluation of the Coverage of Persons 65 Years of Age and Over in the 1970 Census*, PHC(E)-7, 1973.

[140] U. S. Bureau of the Census. *1970 Census of Population and Housing Evaluation and Research Program: the CPS-Census Match*, PHC(E)-11, 1975.

[141] U. S. Public Health Service. *Statistics of the United States*, 1950, vol. 1, 1954, pp. 108-112.

[142] U. S. Public Health Service. Weight at birth and its effect on survival of the newborn in the United States, early 1950, *Vital Statistics — Special Reports*, vol. 39, no. 1, 1954.

[143] U. S. Public Health Service. Reporting of hospitalization in the Health Interview Survey, Public Health Service Publication no. 584-D4. *Vital and Health Statistics*, series 2, no. 6, 1965.

[144] U. S. Public Health Service. Comparison of hospitalization reporting in three survey procedures, *Vital and Health Statistics*, series 2, No. 8, 1965.

[145] U. S. Public Health Service. Interview response on health insurance compared with insurance records, *Vital and Health Statistics*, series 2, no. 18, 1966.

[146] U. S. Public Health Service. A study of infant mortality from linked records, *Vital and Health Statistics*, series 20, nos. 12, 13, and 14, 1972-1973.

[147] U.S. Social Security Administration. Workers covered under Social Security: 1963 administrative information and pilot link results compared, Studies from Interagency Data Linkages, report no. 3.

[148] U.S. Social Security Administration. Report on Policies and Procedures for Establishing Initial Entitlement to RSDI Benefits, series 1, nos. 21-24, 1974.

[149] U.S. Social Security Administration. The 1% Sample Longitudinal Employee-Employer Data File, 1971 (Processed).

[150] Wells, B. Optimum Matching Rules, Univ. of North Carolina, 1974.

Record Linkage

Its Methodology and Application in Medical Data Processing*

A bibliography compiled by
G. WAGNER and H. B. NEWCOMBE

For the further development of medical data processing the method of record linkage is of utmost importance. It is obvious that two or more items of information about persons or person-groups, recorded at different times and at different places, are of much greater significance when available together than when isolated from each other and inaccessible at the same time. The process of linking together information of medical interest pertaining to the same person is called Medical Record Linkage. Today the linkage procedure can be profitably carried out automatically by computers.

Within the last decade a fast growing literature on this subject has been published, so that it has become more and more difficult for the individual scientist to keep abreast with the wealth of publications.

The references given here have been grouped according to the following topics:

A. General Remarks on Record Linkage
B. Methodology of Record Linkage
C. The Identification Problem
D. The Privacy Problem
E. Application of Medical Record Linkage in

1. Patient Care and Medical Data Processing
2. Epidemiology
3. Vital Statistics, Demography
4. Genetics
5. Public Health Services
6. Other and Non-medical Fields

F. Costs of Record Linkage.

We hope that this bibliography will be of some help for more detailed studies in this important field of medical data processing.

G. WAGNER (Heidelberg, Germany)
H. B. NEWCOMBE (Chalk River, Ontario/Canada)

*Reprinted with permission from Methods of Information in Medicine - Methodik der Information in der Medizin, Vol. IX, No. 2, Copyright ©1970 by F.K. Schattauer, pp. 121-138.

A. General Remarks on Record Linkage

Acheson, E. D.: Medical Record Linkage.
(Oxford University Press, London-New York-Toronto 1967).

Acheson, E. D.: Some remarks on contemporary British medical statistics.
J. roy. statist. Soc., Series A, *131:* 9—12, 1968.

Acheson, E. D.: Linkage of medical records.
Brit. med. Bull. *24:* 206—209, 1968.

Acheson, E. D.: Computers and medical record linkage.
WHO-Seminar on the public health uses of electronic computers, London, 1968.
(WHO Document Euro 3092/6, WHO Regional Office for Europe, Copenhagen 1968).

Acheson, E. D. (Edit.): Record Linkage in Medicine.
(E. & S. Livingstone Ltd., Edinburgh and London 1968).

Acheson, E. D.: Medical record linkage.
Meth. Inform. Med. *8:* 1—6, 1969.

Acheson, E. D.: Personal Record Linkage.
In E. Fritze und G. Wagner (Hrsg.): Dokumentation des Krankheitsverlaufs. Bericht über die 13. Jahrestagung der GMD in Bochum, 30. 9. — 2. 10. 1968.
(F. K. Schattauer Verlag, Stuttgart-New York 1969).

Acheson, E. D. and Forbes, J. A.: Experiment in the retrieval of information in general practice. A preliminary report.
Brit. J. prev. soc. Med. *22:* 105—109, 1968.

Ad Hoc Committee on the Implications of Record Linkage for Health-Related Research: Health Research Uses of Record Linkage in Canada.
(Med. Res. Counc. of Canada, Ottawa 1968).

Bahn, A. K.: Mental health clinic statistics: Needs, sources, methods.
Publ. Hlth. Rep. *69:* 619—625, 1954.

Bahn, A. K.: Experience and philosophy with regard to case registers in health and welfare.
Commun. ment. Hlth J. *3:* 245—250, 1965.

Bejerot, N.: Sjukvärdens journal-och arkiv problem.
Svenska Läk.-Tidn. *62:* 1346—1359, 1965.

Billeter, B. M.: Anforderungen an Konzeption und Aufbau einer Datenbank.
ADL-Nachr. No. 58, 652—656, 1969.

B.M.A. Planning Unit: Computers in Medicine.
Report of the Working Party on Computers in Medicine.
(B.M.A. House, London 1969).

Bothwell, P. W.: Routine, records and research.
Med. Rec. *5:* 359—365, 1961.

Bothwell, P. W.: Conceptual and organizational problems in medical and particularly epidemiologic research.
Publ. Hlth *76:* 360—367, 1962.

Cheeseman, E. A.: Medical record linkage in Northern Ireland — reconnaissance and proposals with particular reference to problems of identification.
In E. D. Acheson (Edit.): Record Linkage in Medicine, pp. 70—76.
(E. & S. Livingstone Ltd., Edinburgh and London 1968).

Cohen, B. M.: Methodology of record follow-up studies on veterans.
Amer. J. publ. Hlth *43:* 1292—1298, 1953.

Davies, M.: Toward a medical data bank for a total population.
Datamation *15:* 257—260, 1969.

Doll, R.: Documentazione ospedaliera nell'epoca dei calcolatori elettronici.
Appl. bio-med. calc. elettr. *4:* 61—72, 1969.

Douglas, J. W. B. and Blomfield, J. M.: The reliability of longitudinal surveys.
Milbank mem. Fd Quart. *34:* 227—252, 1956.

Dubois, N. S. and D'Andrea, J. R.: A document linkage program for digital computers.
Behav. Sci. *10:* 312—319, 1965.

Dunn, H. L.: Record Linkage.
Amer. J. publ. Hlth *36:* 1412—1416, 1946.

Dunn, H. L.: Elements of a coordination system of vital records and statistics.
Publ. Hlth Rep. *68:* 793—801, 1953.

Dunn, H. L.: A national identity registration system to synthesize social statistics.
Estadist. J. Int.-Amer. Statist. Inst. *11:* 605—615, 1953.

Dunn, H. L. and Gilbert, M.: Public health begins in the family.
Publ. Hlth Rep. (Wash.) *71:* 1002—1010, 1956.

Farr, W.: in »Report on Army Medical Statistics« by Lord Herbert, Sir A. Tulloch and Dr. W. Farr.
(British Parliamentary Paper, No. 366, 1861).

Fritze, E. und Wagner, G. (Hrsg.): Dokumentation des Krankheitsverlaufs. Probleme der Erfassung des zeitlichen Krankheitsablaufes und des Medical Record Linkage.
Verh. Bericht 13. Jahrestagung GMD, Bochum 30. 9. — 2. 10. 1968. (F. K. Schattauer Verlag, Stuttgart-New York 1969).

Godber, G.: Opening Remarks.
In E. D. Acheson (Edit.): Record Linkage in Medicine, pp. 1—4.
(E. & S. Livingstone Ltd., Edinburgh and London 1968).

Heasman, M. A.: Recent developments in hospital in-patient statistics in Scotland.
Med. Rec. *8:* No. 3, 2—6, 1967.

Kempf, B.: Le traitement de l'information dans les hôpitaux.
Maroc méd. *45:* 15—17, 1966.

Kempf, B.: Contribution à l'étude de l'automatisation des dossiers médicaux.
Maroc méd. *45:* 45—52, 1966.

Kilpatrick, S. I., Mathers, J. D. and Stevenson, A. C.: The importance of population fertility and consanguinity data being available in medico-social studies.
Ulster med. J. *24:* 113—122, 1955.

Maron, M. E.: Large scale data banks.
Spec. Libr. *60:* 3—9, 1969.

Marshall, J. T.: Canada's national vital statistics index.
Popul. Stud. *1:* 204—211, 1947.

Massé, L. et Reynaud, J.: Les banques d'informations sanitaires et la »symnemonique«.
Rev. franç. aff. soc. *23:* 20—30, 1969.

Mosbech, J.: Medical record linkage. (Dan.)
Ugeskr. Laeg. *129:* 1733—1734, 1967.

Newcombe, H. B.: Untapped knowledge of human populations.
Trans. roy. Soc. Can., Sect. III, *56:* 173—180, 1962.

Newcombe, H. B.: Record linkage: concepts and potentialities. In Medical Research Council: Mathematics and Computer Science in Biology and Medicine, pp. 43—49.
(H.M.S.O., London 1965).

Newcombe, H. B.: Products from the early stages in the development of a system of linked records.
In E. D. Acheson (Edit.): Record Linkage in Medicine, pp. 7—33.
(E. & S. Livingstone Ltd., Edinburgh and London 1968).

Ohkura, K.: Use of family registration in the study of human genetics in Japan.
Jap. J. hum. Genet. *5:* 61—68, 1960.

Oxford Record Linkage Study: Alphabetic list of data abstracted — correct 10th September 1969 (mimeograph).
Oxford Regional Hospital Board, Oxford 1970.

Payne, L. C.: An Introduction to Medical Automation.
(Pitman Med. Publ. Co., Ltd., London 1966).

Randall, H. B.: Strengths and limitations of the cumulative health record.
J. Sch. Hlth *37:* 86—89, 1967.

Royal Commission on Population: Reports and selected papers of the Statistics Committee, Vol. II, pp. 29—48.
(H.M.S.O., London 1950).

Shapiro, S. and Densen, M.: Research needs and record matching.
(Proc. Amer. Statist. Ass., Soc. Statist. Sect. pp. 20—24, 1963).

Smith, A.: Linkage of child health records.
WHO-Paper Euro 0215(7), Copenhagen, 19. Aug. 1969.

Stocks, P.: Measurement of morbidity.
Proc. roy. Soc. Med. *37:* 593—608, 1944.

Study Group on Record Linkage: Progress Report of the Public Health Conference on Records and Statistics.
U.S. Department of Health, Education and Welfare, Document No. 603.5—5/31/66.

Tennison, P.: Passport to immortality: Sir Macfarlane Burnett: Behind and ahead.
The Bulletin (Sydney) *87:* 23—24, 1965.

Thomas, D. S.: Continuous register system of population accounting.
In National Resources Committee (Edit.): The problems of a changing population, pp. 276—297.
(U.S. Govt. Print. Off., Washington, D.C. 1938).

Wagner, G.: Internationales Symposion über Medical Record Linkage, Oxford 17./18. 7. 1967.
Meth. Inform. Med. *4:* 187—190, 1967.

Wagner, G.: Medical Record Linkage — Einführung in das Thema.
In E. Fritze und G. Wagner (Hrsg.): Dokumentation des Krankheitsverlaufs. Bericht über die 13. Jahrestagung der GMD, Bochum, 30. 9. bis 2. 10. 1968.
(F. K. Schattauer Verlag, Stuttgart-New York 1969).

Weir, R. D.: Development of Medical Records in Scotland.
Med. Rec. *8:* 7—10, 1967.

WHO: The public health use of electronic computers.
Report on a Seminar convened by the Regional Office for Europe of the WHO, London, 17—21 June 1968.
WHO-Document EURO 3092, WHO Regional Office for Europe, Copenhagen 1969.

Witts, L. J.: People in confidence; the expanding circle.
In E. D. Acheson (Edit.): Record Linkage in Medicine, pp. 333—338.
(E. & S. Livingstone Ltd., Edinburgh and London 1968).

Yellowlees, H.: Interface problems in medical and administrative patient management.
Proc. Intern. Meet. Automated Data Processing in Hospitals, Elsinore 1966.

Editorial: Order in the medical record.
Lancet *1965*, II, 675—676.

Editorial: Record Linkage.
WHO Chron. *21:* 441—442, 1967.

Editorial: Rethinking medical records.
Lancet *1967*, II, 925—926.

Editorial: Medical record linkage.
Canad. med. Ass. J. *98:* 607—608, 1968.

Editorial: Record linkage.
Brit. med. J. *1968*, III, 116—117.

Editorial: Record linkage.
Brit. J. prev. soc. Med. *23:* 203—204, 1969.

Editorial: Record Linkage Conference.
Note on Meeting held at the Scottish Hospital Centre on Friday, 13th December 1968.
Hlth Bull. (Edinb.) *27:* 31—37, 1969.

B. Methodology of Record Linkage

Acheson, E. D.: The Oxford Record Linkage Study: Report on the second year's operations.
(Oxford Regional Hospital Board, Oxford 1963).

Acheson, E. D.: Oxford Record Linkage Study: A central file of morbidity and mortality records for a pilot population.
Brit. J. prev. soc. Med. *18:* 8—13, 1964.

Acheson, E. D.: The structure, function and cost of a file of linked health data.
In Medical Research Council: Mathematics and Computer Science in Biology and Medicine, pp. 61—69.
(H.M.S.O., London 1965).

Acheson, E. D.: Medical record linkage — the method and its applications.
Roy. Soc. Hlth J. *86:* 12—16, 1966.

Acheson, E. D.: Some potentialities of the computer in national health services.
Proc. Intern. Meet. Automated Data Processing in Hospitals, Elsinore 1966.

Acheson, E. D.: Medical Record Linkage.
(Oxford University Press, London-New York-Toronto 1967).

Acheson, E. D.: Record linkage techniques in studies of the etiology of cancer.
Proc. roy. Soc. Med. *61:* 726—730, 1968.

Acheson, E. D. (Edit.): Record Linkage in Medicine.
(E. & S. Livingstone Ltd., Edinburgh and London 1968).

Acheson, E. D.: The Oxford Record Linkage Study; the first five years.
In E. D. Acheson (Edit.): Record Linkage in Medicine, pp. 40—49.
(E. & S. Livingstone Ltd., Edinburgh and London 1968).

Acheson, E. D. and Evans, J. G.: The Oxford Record Linkage Study: A review of the methods and some preliminary results.
Proc. roy. Soc. Med. *57:* 269—274, 1964.

Acheson, E. D. and Watts, S. P.: Using Computers.
Brit. med. J. *1968*, I, 316.

Ad Hoc Committee on the Implications of Record Linkage for Health-Related Research: Health Research Uses of Record Linkage in Canada.
(Med. Res. Counc. of Canada, Ottawa 1968).

Bahn, A. K.: The development of an effective statistical system in mental illness.
Amer. J. Psychiat. *116:* 798—800, 1960.

Bahn, A. K.: Methodological study of a population of outpatient psychiatric clinics. Maryland 1958—59.
Publ. Hlth Monogr. No. 65. (U.S. Govt. Print. Off., Washington, D.C., 1961).

Bahn, A. K.: Person matching by electronic methods.
Comm. Ass. comput. Mach. *5:* 404—407, 1962.

Binder, S.: Information storage, retrieval and processing: Present possibilities and future potentialities.
In: The Use of Vital and Health Statistics for Genetic and Radiation Studies, pp. 161—167.
(U.N. Publication, Sales No. 61, New York 1962).

Bothwell, P. W.: Routine, records and research.
Med. Rec. *5:* 359—365, 1961.

Bothwell, P. W.: A New Look at Preventive Medicine.
(Pitman Medical Publishing Comp., London 1965).

Christensen, H. T.: The method of record linkage applied to family data.
Marr. Fam. Living *20:* 38—43, 1958.

Cope, C. B.: A centralized nation-wide patient data system.
In E. D. Acheson (Edit.): Record Linkage in Medicine, pp. 34—38.
(E. & S. Livingstone Ltd., Edinburgh and London 1968).

D'Andrea, J. R. and Dubois, N. S.: On the problem of matching documents with missing and inaccurately recorded items (Abstract).
Ann. math. Statist. *35:* 1404—1405, 1964.

Davidson, L.: Retrieval of misspelled names in an airline passenger record system.
Comm. Ass. comput. Mach. *5:* 169—171, 1962.

Davis, L. S., Collen, M. F., and Rubin, L.: Computer-stored medical record.
Comp. biomed. Res. *1:* 452—469, 1968.

Deming, W. E. and Glasser, G. J.: On the problem of matching lists by samples.
J. Amer. statist. Ass. *54:* 403—413, 1959.

Doll, R.: Hospital records in the computer age.
Proc. roy. Soc. Med. *61:* 709—715, 1968.

Doll, R.: Record linkage techniques in studies of the etiology of cancer.
Proc. roy. Soc. Med. *61:* 731—732, 1968.

Dubois, N. S. and D'Andrea, J. R.: A document linkage program for digital computers.
Behav. Sci. *10:* 312—319, 1965.

Fritze, E. und Wagner, G. (Hrsg.): Dokumentation des Krankheitsverlaufs. Probleme der Erfassung des zeitlichen Krankheitsablaufes und des Medical Record Linkage.
Verh. Bericht 13. Jahrestagung GMD,
Bochum 30. 9. — 2. 10. 1968.
(F. K. Schattauer Verlag, Stuttgart-New York 1969).

Galloway, T. McL.: Computers. Their use in local health administration.
Roy. Soc. Hlth J. *86:* 213—216, 1966.

Galloway, T. McL.: The use of an electronic computer in a health department.
Canad. J. publ. Hlth *57:* 331—332, 1966.

Gedda, L. and Milani-Comparetti, M.: Computerization of a permanent twin register: a basic tool in scientific research.
Acta Genet. med. (Roma) *15:* 333—344, 1966.

Guralnick, L. and Nam, C. B.: Census — NOVS study of death certificates matched to census records.
Milbank mem. Fd Quart. *37:* 144—153, 1959.

Hall, P., Mellner, Ch. and Danielsson, T.: J 5 — A data processing system for medical information.
Meth. Inform. Med. *6:* 1—6, 1967.

Hambright, T. Z.: Comparability of age on the death certificate and matching census record.
Vital and Health Statistics, Publ. Hlth Publ. No. 1000 Series 2-No. 29.
(U.S. Dept. H. E. W., Washington, D.C., 1968).

Hauser, P. M. and Lauriat, P.: Record matching — theory and practice (Abstract).
(Proc. Amer. Statist. Ass., Soc. Statist. Sect., p. 25, 1963).

Hubbard, M. R.: Computer Systems for Medical Record Linkage.
(Thesis submitted for the degree of Bachelor of Science in the University of Oxford 1969).

Hubbard, M. R. and Fisher, J. E.: A computer system for medical record linkage.
In E. D. Acheson (Edit.): Record Linkage in Medicine, pp. 157—170.
(E. & S. Livingstone Ltd., Edinburgh and London 1968).

Jahn, E.: Probleme der Zusammenführung zeitlich und örtlich differenter Gesundheitsdaten.
In E. Fritze und G. Wagner (Hrsg.): Dokumentation des Krankheitsverlaufs. Bericht über die 13. Jahrestagung der GMD in Bochum, 30. 9. — 2. 10. 1968.
(F. K. Schattauer Verlag, Stuttgart-New York 1969).

Jainz, M.: Organisation einer Datenbank auf Magnetbändern für die Hautklinik Kiel.
Meth. Inform. Med. *8:* 190—192, 1969.

Jameson, M. J.: A system of recording the family history in general practice.
J. roy. Coll. gen. Pract. *16:* 135—143, 1968.

Kennedy, J. M.: Linkage of birth and marriage records using a digital computer.
Atomic Energy of Canada Ltd., Report No. 1258,
Chalk River, Ontario 1961.

Kennedy J. M.: The use of a digital computer for record linkage.
Proc. Seminar on the Use of Vital and Health Statistics for Genet. and Radiat. Studies, pp. 155—159.
(U. N. Publication, Sales No. 61, New York 1962).

Kennedy, J. M.: File structures for the automatic manipulation of linked records.
In E. D. Acheson (Edit.): Record Linkage in Medicine, pp. 109—118.
(E. & S. Livingstone Ltd., Edinburgh and London 1968).

Kennedy, J. M., Newcombe, H. B., Okazaki, E. A. and Smith, M. E.: List processing methods for organizing files of linked records.
Atomic Energy of Canada Ltd., Report No. 2078,
Chalk River, Ontario 1964.

Kennedy, J. M., Newcombe, H. B., Okazaki, E. A. and Smith, M. E.: Computer methods for family linkage of vital health records.
Atomic Energy of Canada Ltd., Report No. 2222,
Chalk River, Ontario 1965.

Madow, W. G.: Interview data on chronic conditions compared with information derived from medical records.
Vital and Health Statistics, Publ. Hlth Publ. No. 1000-Series 2-No. 23.
(U.S. Dept. H.E.W., Washington, D.C., 1967).

Marksteiner, A. und Wolf, Ch.: Die Normierung inhomogener medizinischer Datenbestände mittels einer EDVA.
In K. Fellinger (Hrsg.): Computer in der Medizin.
Probleme, Erfahrungen, Projekte.
(Verlag Brüder Hollinek, Wien 1968).

McCarthy, M. A.: Comparison of the classification of place of residence on death certificates and matching census records.
Vital and Health Statistics, Publ. Hlth Publ. No. 1000-Series 2-No. 30.
(U.S. Dept. H.E.W., Washington, D.C., 1969).

McKusick, V. A.: Some computer applications to problems in human genetics.
Meth. Inform. Med. *4:* 183—189, 1965.

Moore, F.: Development of a regional health information center.
Proc. 4. IBM Medical Symp., Endicott, N.Y., pp. 225—239.
(Yorktown Heights/N.Y. 1962).

Moore, F.: Health Information Systems.
(Univ. of South. Calif. School of Med., Los Angeles 1963).

Moore, F. J.: Mechanizing a large register of first order patient data.
Meth. Inform. Med. *4:* 1—10, 1965.

Nathan, G.: Outcome probabilities for a matching process with complete invariant information.
J. Amer. statist. Ass. *62:* 454—469, 1967.

Newcombe, H. B.: Population genetics: Population records.
In: Methodology in Human Genetics, pp. 92—113.
(Holden-Day Inc., San Francisco 1962).

Newcombe, H. B.: Screening for effects of maternal age and birth order in a register of handicapped children.
Ann. hum. Genet. *27:* 367—382, 1964.

Newcombe, H. B.: Pedigrees for population studies.
A progress report.
Cold Spr. Harb. Symp. quant. Biol. *29:* 21—30, 1964.

Newcombe, H. B.: Record linkage: concepts and potentialities. In Medical Research Council: Mathematics and Computer Science in Biology and Medicine, pp. 43—49.
(H.M.S.O., London 1965).

Newcombe, H. B.: Use of vital statistics.
In: U.N. World Population Conference, Belgrade, 1965,
Vol. II, pp. 494—497. (United Nations, New York 1965).

Newcombe, H. B.: Record Linking: The design of efficient systems for linking records into individual and family histories.
Amer. J. hum. Genet. *19:* 335—359, 1967.

Newcombe, H. B.: Products from the early stages in the development of a system of linked records.
In E. D. Acheson (Edit.): Record Linkage in Medicine,
pp. 7—33.
(E. & S. Livingstone Ltd., Edinburgh and London 1968).

Newcombe, H. B.: The use of medical record linkage for population and genetic studies.
Meth. Inform. Med. *8:* 7—11, 1969.

Newcombe, H. B.: Die Anwendung des Medical Record Linkage für Bevölkerungs- und genetische Studien.
In E. Fritze und G. Wagner (Hrsg.): Dokumentation des Krankheitsverlaufs. Bericht über die 13. Jahrestagung der GMD in Bochum, 30. 9. — 2. 10. 1968.
(F. K. Schattauer Verlag, Stuttgart-New York 1969).

Newcombe, H. B., James, A. P., Axford, S. J.: Family Linkage of Vital and Health Records:
(1) For population studies of hereditary influences on health,
(2) For verification of status for welfare programmes and other purposes.
Atomic Energy of Canada Ltd., Report No. 470,
Chalk River, Ontario 1957.

Newcombe, H. B. and Kennedy, J. M.: Record linkage.
Making maximum use of the discriminating power of identifying information.
Comm. Ass. comput. Mach. *5:* 563—566, 1962.

Newcombe, H. B. and Kennedy, J. M.: Demographic analysis and computer programs.
In: U.N. World Population Conference, Belgrade 1965,
Vol. III, pp. 251—253.
(United Nations, New York 1965).

Newcombe, H. B., Kennedy, J. M., Axford, S. J. and James, A. P.: Automatic linkage of vital records.
Science *130:* 954—959, 1959.

Newcombe, H. B. and Rhynas, P. O. W.: The cost of individual follow-up studies of large populations.
Atomic Energy of Canada Ltd., Report No. 1255,
Chalk River, Ontario 1961.

Newcombe, H. B. and Rhynas, P. O. W.: Family linkage of population records.
Proc. Seminar on the Use of Vital and Health Statistics for Genet. and Radiat. Studies, pp. 135—153.
(U.N. Publication, Sales No. 61, New York 1962).

Nitzberg, D. M.: Results of research into the methodology of record linkage.
In E. D. Acheson (Edit.): Record Linkage in Medicine,
pp. 187—202.
(E. & S. Livingstone Ltd., Edinburgh and London 1968).

Nitzberg, D. M. and Sardy, H.: The methodology of computer linkage of health and vital records.
(Proc. Amer. statist. Ass., Soc. Statist. Sect. 1965).

Phillips, W., jr.: Record linkage for a chronic disease register.
In E. D. Acheson (Edit.): Record Linkage in Medicine,
pp. 120—151.
(E. & S. Livingstone Ltd., Edinburgh and London 1968).

Phillips, W., jr. and Bahn, A. K.: Experience with computer-matching of names.
(Proc. Amer. statist. Ass., Soc. Statist. Sect.
pp. 26—37, 1963).

Phillips, W., jr., Bahn, A. K. and Miyasaki, M.: Person-matching by electronic methods.
Comm. Ass. comput. Mach. *5:* 404—407, 1962.

Phillips, W., jr., Gorwitz, K. and Bahn, A. K.: Electronic maintenance of case registers.
Publ. Hlth Rep. *77:* 503—510, 1962.

Royal Commission on Population: Reports and selected papers of the Statistics Committee, Vol. II, pp. 29—48.
(H.M.S.O., London 1950).

Shapiro, S. and Densen, M.: Research needs and record matching.
(Proc. Amer. Statist. Ass., Soc. Statist. Sect.
pp. 20—24, 1963).

Shapiro, S. and Schachter, J.: Methodology and summary results of the 1950 birth registration test in the United States.
Estadist. J. Int.-Amer. Statist. Inst. *10:* 688—699, 1952.

Shapiro, S., Weinblatt, E., Frank, C., Sager, R. and Densen, P.: The H.I.P. study of incidence and prognosis of coronary heart disease: Methodology.
J. chron. Dis. *16:* 1281—1292, 1963.

Smith, A. E.: Automatic linkage of medical and vital registration records.
Brit. J. soc. prev. Med. *17:* 185—190, 1963.

Smith, M. E., Schwartz, R. R. and Newcombe, H. B.: Computer methods for extracting sibship data from family groupings of records.
Atomic Energy of Canada Ltd., Report No. 2520,
Chalk River, Ontario 1965.

Smythe, M.: Record Numbering.
In E. D. Acheson (Edit.): Record Linkage in Medicine,
pp. 179—183.
(E. & S. Livingstone Ltd., Edinburgh and London 1968).

Spicer, C. C.: Practical problems of record linkage on a national scale.
In Medical Research Council: Mathematics and Computer Science in Biology and Medicine. (H.M.S.O., London 1965).

Study Group on Record Linkage: Progress Report of the Public Health Conference on Records and Statistics.
U.S. Department of Health, Education and Welfare,
Document No. 603.5—5/31/66.

Sunter, A. B.: A statistical approach to record linkage.
In E. D. Acheson (Edit.): Record Linkage in Medicine,
pp. 89—107.
(E. & S. Livingstone Ltd., Edinburgh and London 1968).

Sunter, A. B. and Felligi, J. P.: An optimal approach to record linkage.
(Proc. 36th Sess. Int. Statist. Inst., Sydney, Australia 1967).

Tepping, B. J.: Study of matching techniques for subscriptions fulfillment.
(National Analysts Inc., Philadelphia 1955).

Tepping, B. J. and Chu, J. T.: A report on matching rules applied to Reader's Digest Data.
(National Analysts Inc., Philadelphia 1958).

Thatcher, R. A.: Medical records in the computer age.
Med. J. Aust. *2:* 234—236, 1968.

Watts, S. P. and Acheson, E. D.: Computer method for deriving hospital in-patient morbidity statistics based on the person as the unit.
Brit. med. J. *1967*, IV, 476—477.

Weir, R. D.: The introduction of record linkage in north-east Scotland.
In E. D. Acheson (Edit.): Record Linkage in Medicine,
pp. 55—60.
(E. & S. Livingstone Ltd., Edinburgh and London 1968).

Yoder, R. D.: Preparing medical record data for computer processing.
Hospitals *40:* 75—76, 1966.

C. The Identification Problem

Acheson, E. D.: Some potentialities of the computer in national health services.
Proc. Intern. Meet. Automated Data Processing in Hospitals, Elsinore 1966.

Acheson, E. D.: Medical Record Linkage.
(Oxford University Press, London-New York-Toronto 1967).

Acheson, E. D.: Linkage of medical records.
Brit. med. Bull. *24:* 206—209, 1968.

Acheson, E. D. (Edit.): Record Linkage in Medicine.
(E. & S. Livingstone Ltd., Edinburgh and London 1968).

Acheson, E. D.: Medical record linkage.
Meth. Inform. Med. *8:* 1—6, 1969.

Acheson, E. D.: Personal Record Linkage.
In E. Fritze und G. Wagner (Hrsg.): Dokumentation des Krankheitsverlaufs. Bericht über die 13. Jahrestagung der GMD in Bochum, 30. 9. — 2. 10. 1968.
(F. K. Schattauer Verlag, Stuttgart-New York 1969).

Acheson, E. D. and Barr, A.: Multiple spells of in-patient treatment in a calendar year.
Brit. J. prev. soc. Med. *19:* 182—191, 1965.

Ad Hoc Committee on the Implications of Record Linkage for Health-Related Research: Health Research Uses of Record Linkage in Canada.
(Med. Res. Counc. of Canada, Ottawa 1968).

Arbeitsausschuß Medizin in der DGD: Ein dokumentationsgerechter Krankenblattkopf für stationäre Patienten aller klinischen Fächer (sog. Allgemeiner Krankenblattkopf).
Med. Dok. *5:* 57—71, 1961.

Bahn, A. K.: Person matching by electronic methods.
Comm. Ass. comput. Mach. *5:* 404—407, 1962.

Bahn, A. K. and Bahn, R.: Considerations in using social security numbers on birth certificates for research purposes.
Publ. Hlth Rep. (Wash.) *79:* 937—938, 1964.

Bennett, A. E. and Holland, W. W.: Towards the development of electronic data-processing systems for medical records.
Lancet 1965, II, 1176—1178.

Billeter, B. M.: Anforderungen an Konzeption und Aufbau einer Datenbank.
ADL-Nachr. No. 58, 652—656, 1969.

Bjarnason, O., Fridriksson, S. and Magnusson, M.: Record linkage in a self-contained community.
In E. D. Acheson (Edit.): Record Linkage in Medicine, pp. 62—68.
(E. & S. Livingstone Ltd., Edinburgh and London 1968).

Brunner, H., Paumgartner, G., Grabner, G., Grabner, H., Marksteiner, A. und Wolf, Ch.: Erfahrungen mit der Dokumentation von Krankengeschichten einer internen Klinik.
In K. Fellinger (Hrsg.): Computer in der Medizin. Probleme, Erfahrungen, Projekte.
(Verlag Brüder Hollinek, Wien 1968).

Cheeseman, E. A.: Medical record linkage in Northern Ireland — reconnaissance and proposals with particular reference to problems of identification.
In E. D. Acheson (Edit.): Record Linkage in Medicine, pp. 70—67.
(E. & S. Livingstone Ltd., Edinburgh and London 1968).

Cope, C. B.: A centralized nation-wide patient data system.
In E. D. Acheson (Edit.): Record Linkage in Medicine, pp. 34—38.
(E. & S. Livingstone Ltd., Edinburgh and London 1968).

Davidson, L.: Retrieval of misspelled names in an airline passenger record system.
Comm. Ass. comput. Mach. *5:* 169—171, 1962.

Fassl, H.: Das Risikopatientenregister der Universitätskliniken Mainz.
Meth. Inform. Med. *7:* 214—218, 1968.

Fritze, E. und Wagner, G. (Hrsg.): Dokumentation des Krankheitsverlaufs. Probleme der Erfassung des zeitlichen Krankheitsablaufes und des Medical Record Linkage.
Verh. Bericht 13. Jahrestagung GMD,
Bochum 30. 9. — 2. 10. 1968.
(F. K. Schattauer Verlag, Stuttgart-New York 1969).

Giersdorf, P., Günther, O., Kreuz, B. und Schneider, W.: Zum Entwurf eines einheitlichen und dokumentationsgerechten Krankenblattkopfes für Patienten in stationären Einrichtungen.
Dtsch. Ges.-wesen *21:* 462—469, 1966.

Heasman, M. A.: Recent developments in hospital in-patient statistics in Scotland.
Med. Rec. *8:* No. 3, 2—6, 1967.

Hogben, L. and Cross, K. W.: The statistical specificity of a code personnel cypher sequence.
Brit. J. soc. Med. *2:* 149—152, 1948.

Hogben, L., Johnstone, M. M. and Cross, K. W.: Identification of medical documents.
Brit. med. J. *1948,* I, 632—635.

Hopkins, R. A. and Gardner, E. A.: Development of a flexible control system in the maintenance of a patient case register.
In Enslein, K. (Edit.): Data Acquisition and Processing, Vol. 3, pp. 191—198.
(Pergamon Press, Ltd., Oxford-London 1964).

Jahn, E.: Probleme der Zusammenführung zeitlich und örtlich differenter Gesundheitsdaten.
In E. Fritze und G. Wagner (Hrsg.): Dokumentation des Krankheitsverlaufs. Bericht über die 13. Jahrestagung der GMD in Bochum, 30. 9. — 2. 10. 1968.
(F. K. Schattauer Verlag, Stuttgart-New York 1969).

Jungner, G.: Health Screening.
Proc. Intern. Meet. Automated Data Processing in Hospitals, Elsinore 1966.

Källen, B. and Winberg, J.: A Swedish register of congenital malformations. Experience with continuous registration during 2 years with special reference to multiple malformations.
Pediatrics *41:* 765—776, 1968.

Kempf, B.: Contribution à l'étude de l'automatisation des dossiers médicaux.
Maroc méd. *45:* 45—52, 1966.

Kennedy, J. M., Newcombe, H. B., Okazaki, E. A. and Smith, M. E.: Computer methods for family linkage of vital health records.
Atomic Energy of Canada Ltd., Report No. 2222,
Chalk River, Ontario 1965.

Ledley, R. S. and Lusted, L. B.: The use of electronic computers in medical data processing. Aids in diagnosis, current information retrieval, and medical record keeping.
IRE Trans. med. Electron. *7:* 31—47, 1960.

Lindberg, D. A. B.: Electronic retrieval of clinical data.
J. med. Educ. *40:* 753—759, 1965.

Maass, W.: Zum Problem der Bildung einer einheitlichen deutschen Bevöikerungsnummer.
ADL-Nachr. H. 25, 254—266, 1963.

Maass, W.: Noch einmal — Zur Frage der Bevölkerungsnummer oder Personalkennziffer.
ADL-Nachr. H. 29, 528—531, 1963.

Marcusson, H.: Zur Erprobung eines neuen Gesundheitsbogens des Jugendgesundheitsschutzes.
Z. ges. Hyg. *11:* 123—140, 1965.

Marthaler, T. M.: A standardized system of recording dental conditions.
Helv. odont. Acta *10:* 1—18, 1966.

Massé, L. et Reynaud, J.: Les banques d'informations sanitaires et la »symnemonique«.
Rev. franç. aff. soc. *23:* 20—30, 1969.

Ministry of Health — Central Health Services Council: The standardisation of hospital medical records.
Report of the Sub-Committee of the Standing Medical Advisory Committee. (H.M.S.O., London 1965).

Moore, F. J.: Mechanizing a large register of first order patient data.
Meth. Inform. Med. *4:* 1—10, 1965.

Newcombe, H. B.: Record linkage: concepts and potentialities.
In Medical Research Council: Mathematics and Computer Science in Biology and Medicine, pp. 43—49.
(H.M.S.O., London 1965).

Newcombe, H. B.: The use of medical record linkage for population and genetic studies.
Meth. Inform. Med. *8:* 7—11, 1969.

Newcombe, H. B.: Record linkage and hospital discharge abstracts.
(Working Paper for the Conference on Hospital Discharge Abstracts, Airlie House,
Warrenton/Va., June 1969; Med. Care, in press).

Newcombe, H. B., James, A. P., Axford, S. J.: Family Linkage of Vital and Health Records:
(1) For population studies of hereditary influences on health,
(2) For verification of status for welfare programmes and other purposes.
Atomic Energy of Canada Ltd., Report No. 470,
Chalk River, Ontario 1957.

Newcombe, H. B., and Kennedy, J. M.: Record linkage. Making maximum use of the discriminating power of identifying information.
Comm. Ass. comput. Mach. *5:* 563—566, 1962.

Nielsen, H.: The central population register and the prospective assignment of a person-number (Dan.).
Ugeskr. Laeg. *129:* 1734—1736, 1967.

Nielsen, H.: The personal numbering system in Denmark.
In E. D. Acheson (Edit.): Record Linkage in Medicine,
pp. 173—177.
(E. & S. Livingstone Ltd., Edinburgh and London 1968).

Nitzberg, D. M.: Results of research into the methodology of record linkage.
In E. D. Acheson (Edit.): Record Linkage in Medicine,
pp. 187—202.
(E. & S. Livingstone Ltd., Edinburgh and London 1968).

Phillips, W., jr.: Record linkage for a chronic disease register.
In E. D. Acheson (Edit.): Record Linkage in Medicine,
pp. 120—151.
(E. & S. Livingstone Ltd., Edinburgh and London 1968).

Postel, H. J.: Die Kölner Phonetik. Ein Verfahren zur Identifizierung von Personennamen auf der Grundlage der Gestaltanalyse.
IBM-Nachr. *19:* 925—931, 1969.

Schental, J. E., Sweeney, J. W., Nettleton, W. J., jr., and Yoder, R. D.: Clinical application of electronic data processing apparatus. III. System for processing of medical records.
J. Amer. med. Ass. *186:* 101—105, 1963.

Scottish Home and Health Department — Scottish Health Services Council: Hospital medical records in Scotland. Development and standardisation.
(H.M.S.O., Edinburgh 1967).

Selmer, E. S.: Registration numbers in Norway: some applied number theory and psychology.
J. roy. statist. Soc., Sect. A. *130:* 225—231, 1967.

Smith, A. E.: Preservation of confidence at the central level.
In E. D. Acheson (Edit.): Record Linkage in Medicine,
pp. 338—345.
(E. & S. Livingstone Ltd., Edinburgh and London 1968).

Study Group on Record Linkage: Progress Report of the Public Health Conference on Records and Statistics.
U.S. Department of Health, Education and Welfare,
Document No. 603.5—5/31/66.

Sunter, A. B.: A statistical approach to record linkage.
In E. D. Acheson (Edit.): Record Linkage in Medicine,
pp. 89—107.
(E. & S. Livingstone Ltd., Edinburgh and London 1968).

Turner, J. G. and Shannon, P. D.: Mechanical identification of hospital records saves time and money.
Canad. Hosp. *33:* 44—50, 1956.

Wagner, G.: Zum Problem der Bildung einer einheitlichen deutschen Bevölkerungsnummer.
ADL-Nachr. H. 27, 411—414, 1963.

Wagner, G.: The development of the standardized medical record in Germany.
Med. Rec. *7:* 183—188, 1965.

Wagner, G.: Medical Record Linkage — Einführung in das Thema.
In E. Fritze und G. Wagner (Hrsg.): Dokumentation des Krankheitsverlaufs. Bericht über die 13. Jahrestagung der GMD in Bochum, 30. 9. — 2. 10. 1968.
(F. K. Schattauer Verlag, Stuttgart-New York 1969).

Wagner, G., Immich, H. und Köhler, C.: Der Krankenblattkopf der Heidelberger Kliniken.
Meth. Inform. Med. *7:* 17—25, 1968.

Wagner, G. und Stutzer, G.: Über die Selektivität der sog. I-Zahl im »Allgemeinen Krankenblattkopf« und die Brauchbarkeit ihrer einzelnen Komponenten.
Meth. Inform. Med. *2:* 148—155, 1963.

WHO: The public health use of electronic computers. Report on a Seminar convened by the Regional Office for Europe of the WHO, London, 17—21 June 1968.
WHO-Document EURO 3092, WHO Regional Office for Europe, Copenhagen 1969.

Wittboldt, S.: MED-Kortet, sjukvårdshandlingen, filmhålkortet och ADB.
Svenska Läk.-Tidn. *62:* 3369—3374, 1965.

Yoder, R. D., Dreyfus, R. H. and Saltzberg, B.: Identification codes for medical records.
Health Serv. Res. *1:* 53—65, 1966.

Yoder, R. D., Swearingen, D. R., Schenthal, J. E., Sweeney, J. W. and Nettleton, W. J., Jr.: An automated clinical information system.
Meth. Inform. Med. *3:* 45—50, 1964.

D. The Privacy Problem

Acheson, E. D.: Medical Record Linkage.
(Oxford University Press, London-New York-Toronto 1967).

Acheson, E. D.: Linkage of medical records.
Brit. med. Bull. *24:* 206—209, 1968.

Acheson, E. D. (Edit.): Record Linkage in Medicine.
(E. & S. Livingstone Ltd., Edinburgh and London 1968).

Acheson, E. D.: Medical record linkage.
Meth. Inform. Med. *8:* 1—6, 1969.

Acheson, E. D.: Personal Record Linkage.
In E. Fritze und G. Wagner (Hrsg.): Dokumentation des Krankheitsverlaufs. Bericht über die 13. Jahrestagung der GMD in Bochum, 30. 9. — 2. 10. 1968.
(F. K. Schattauer Verlag, Stuttgart-New York 1969).

Armer, P.: Social implications of the computer utility.
Doc. P-3642, Rand Corp., Santa Monica/Calif. 1967.

Bahn, A. K., Goldberg, I. D. and Gorwitz, K.: Longitudinal studies using psychiatric case registers.
In E. D. Acheson (Edit.): Record Linkage in Medicine,
pp. 226—249.
(E. & S. Livingstone Ltd., Edinburgh and London 1968).

Baran, P.: Remarks on the question of privacy raised by the automation of mental health records.
Doc. P-3523, Rand Corp., Santa Monica/Calif. 1967.

Bejerot, N.: Sjukvärdens journal-och arkiv problem.
Svenska Läk.-Tidn. *62:* 1346—1359, 1965.

B.M.A. Planning Unit: Computers in Medicine.
Report of the Working Party on Computers in Medicine.
(B.M.A. House, London 1969).

Bruyn, H. D.: Confidentiality in the use of health records.
J. Sch. Hlth *37:* 161—165, 1967.

Butler, R. N.: Privileged communications and research.
Arch. gen. Psychiat. *8:* 139—141, 1963.

Curran, W. J., Stearns, B. and Kaplan, H.: Privacy, confidentiality and other legal considerations in the establishment of a centralized health-data system.
New Engl. J. Med. *281:* 241—248, 1969.

Dunn, E. S., Jr.: The idea of a national data center and the issue of personal privacy.
Amer. Statist. *21:* 21—27, 1967.

Freed, R. N.: A legal structure for a national medical data center.
Proc. AFIPS Fall Joint Computer Conference,
Vol. 33, Part 1, pp. 387—394.
(Thompson Book Comp., Washington, D.C., 1968).

Glazebrook, P. R.: Medical confidences, research and the law.
In E. D. Acheson (Edit.): Record Linkage in Medicine,
pp. 323—333.
(E. & S. Livingstone Ltd., Edinburgh and London 1968).

Hagman, D. G.: The non-litigant patient's right to medical records: medicine vs. law.
J. forens. Sci. *14:* 352—369, 1969.

Hall, P., Engkvist, O., Mellner, C. and Danielsson, T.: The problem of secrecy in mechanical storage of records (Swed.).
Svenska Läk.-Tidn. *62:* 2934—2936, 1965.

Harrison, A.: The problem of privacy in the computer age: an annotated bibliography.
Doc. RM-5495-PR/RC, Rand Corp., Santa Monica/Calif. 1967.

Hartman, J. D.: Medical records and privilege.
Med.-leg. Bull. *178:* 1—3, 1968.

Heasman, M. A.: Manual of hospital morbidity statistics.
WOH-Paper WHO/HS/Nat. Com./147 — 11. June 1963.

Hoffman, L. J.: Computers and Privacy: A survey.
Comput. Surv. *1:* 85—103, 1969.

Kempf, B.: Contribution à l'étude de l'automatisation des dossiers médicaux.
Maroc. méd. *45:* 45—52, 1966.

Keune H. G.: Zu einigen rechtlichen Fragen bei der Benutzung von Krankenunterlagen für wissenschaftliche Zwecke.
Z. ärztl. Fortb. *61:* 887—890, 1967.

Kohlhaas, M.: Zur Weitergabe ärztlicher Befunde an fremde Personen oder Patienten.
Dtsch. med. Wschr. *94:* 1503—1504, 1969.

Ledley, R. S. and Lusted, L. B.: The use of electronic computers in medical data processing. Aids in diagnosis, current information retrieval, and medical record keeping.
IRE Trans. med. Electron. *7:* 31—47, 1960.

Letourneau, C. U.: Erosion of the professional secret.
Hosp. Mgmt *99:* 53—55, 1965.

McCarthy, J.: Information.
Scient. Amer. *215:* 65—72, 1966.

Michael, D. N.: Speculations on the relation of the computer to individual freedom and the right to privacy.
George Washington Law Rev. *33:* 270—286, 1964—65.

Rindani, T. H., Alemeida, V. B. and Kabe, N. G.: Medical record and the law.
J. Ass. Phycns India *13:* 154—159, 1965.

Smith, A. E.: Preservation of confidence at the central level.
In E. D. Acheson (Edit.): Record Linkage in Medicine,
pp. 338—345.
(E. & S. Livingstone Ltd., Edinburgh and London 1968).

Snedeker, L.: On confidentiality and data banks.
New Engl. J. Med. *281:* 269—270, 1969.

Taeuber, C.: Invasion of privacy.
Eugen. Quart. *14:* 243—246, 1967.

U.S. Congress, House Committee on Government Operations: Privacy and the national data bank concept.
(U.S. Govt. Print. Off., Washington, D.C., 1968).

Whittier, J. R.: Research on Huntington's chorea: Problems of privilege and confidentiality.
J. forens. Sci. *8:* 568—575, 1963.

Witts, L. J.: People in confidence; the expanding circle.
In E. D. Acheson (Edit.): Record Linkage in Medicine,
pp. 333—338.
(E. & S. Livingstone Ltd., Edinburgh and London 1968).

Zachary, M. C.: Confidentiality of medical records. Role of the nurse.
Arch. environm. Hlth *19:* 282—285, 1969.

E. Application of Medical Record Linkage in

1.) Patient Care and Medical Data Processing

Acheson, E. D.: Association between ulcerative colitis, regional enteritis and ankylosing spondylitis.
Quart. J. Med. *29:* 489—499, 1960.

Acheson, E. D.: The Oxford Record Linkage Study: Report on the second year's operations.
(Oxford Regional Hospital Board, Oxford 1963).

Acheson, E. D.: Hospital morbidity in early life in relation to certain maternal and foetal characteristics and events at delivery.
Brit. J. prev. soc. Med. *19:* 164—173, 1965.

Acheson, E. D.: Medical record linkage — the method and its applications.
Roy. Soc. Hlth J. *86:* 12—16, 1966.

Acheson, E. D.: Medical Record Linkage.
(Oxford University Press, London-New York-Toronto 1967).

Acheson, E. D.: The Oxford Record Linkage Study; the first five years.
In E. D. Acheson (Edit.): Record Linkage in Medicine,
pp. 40—49.
(E. & S. Livingstone Ltd., Edinburgh and London 1968).

ACHESON, E. D.: Linkage of medical records.
Brit. med. Bull. *24:* 206—209, 1968.

ACHESON, E. D. (Edit.): Record Linkage in Medicine.
(E. & S. Livingstone Ltd., Edinburgh and London 1968).

ACHESON, E. D.: Medical record linkage.
Meth. Inform. Med. *8:* 1—6, 1969.

ACHESON, E. D.: Personal Record Linkage.
In E. Fritze und G. Wagner (Hrsg.): Dokumentation des Krankheitsverlaufs. Bericht über die 13. Jahrestagung der GMD in Bochum, 30. 9. — 2. 10. 1968.
(F. K. Schattauer Verlag, Stuttgart-New York 1969).

ACHESON, E. D. and BARR, A.: Multiple spells of in-patient treatment in a calendar year.
Brit. J. prev. soc. Med. *19:* 182—191, 1965.

ACHESON, E. D. and EVANS, J. G.: The Oxford Record Linkage Study: A review of the methods and some preliminary results.
Proc. roy. Soc. Med. *57:* 269—274, 1964.

ACHESON, E. D. and FELDSTEIN, M.: Duration of stay in hospitals for normal maternity care.
Brit. med. J. *1964,* II, 95—99.

BABIGIAN, H. M., GARDNER, E. A., MILES, H. C. and ROMANO, J.: Diagnostic consistency and change in a follow-up study of 1215 patients.
Amer. J. Psychiat. *121:* 895—901, 1965.

BAHN, A. K.: Methodological study of a population of outpatient psychiatric clinics. Maryland 1958—59.
Publ. Hlth Monogr. No. 65.
(U.S. Govt. Print. Off., Washington, D.C., 1961).

BAHN, A. K., GOLDBERG, I. D. and GORWITZ, K.: Longitudinal studies using psychiatric case registers.
In E. D. Acheson (Edit.): Record Linkage in Medicine, pp. 226—249.
(E. & S. Livingstone Ltd., Edinburgh and London 1968).

BENJAMIN, B.: Assessment of medical care.
Proc. roy. Soc. Med. *60:* 809—813, 1967.

BENNETT A. E. and HOLLAND, W. W.: Towards the development of electronic data-processing systems for medical records.
Lancet 1965, II, 1176—1178.

BROTHERSTON, J.: General discussion and summing up.
In E. D. Acheson (Edit.): Record Linkage in Medicine, pp. 379—381.
(E. & S. Livingstone Ltd., Edinburgh and London 1968).

COPE, C. B.: A centralized nation-wide patient data system.
In E. D. Acheson (Edit.): Record Linkage in Medicine, pp. 34—38.
(E. & S. Livingstone Ltd., Edinburgh and London 1968).

CRONKHITE, L. W.: Patient location control as a first step toward a total information system.
Hospitals *41:* 107—112, 1967.

DAVIES, M.: Toward a medical data bank for a total population.
Datamation *15:* 257—260, 1969.

DOLL, R.: Hospital records in the computer age.
Proc. roy. Soc. Med. *61:* 709—715, 1968.

FAIRBAIRN, A. S.: Comparison of diagnosis on successive hospital admissions.
In E. D. Acheson (Edit.): Record Linkage in Medicine, pp. 215—225.
(E. & S. Livingstone Ltd., Edinburgh and London 1968).

FAIRBAIRN, A. S. and ACHESON, E. D.: The extend of organ removal in the Oxford area.
J. chron. Dis. *22:* 111—122, 1969.

FRITZE, E. und WAGNER, G. (Hrsg.): Dokumentation des Krankheitsverlaufs. Probleme der Erfassung des zeitlichen Krankheitsablaufes und des Medical Record Linkage.
Verh. Bericht 13. Jahrestagung GMD, Bochum
30. 9. — 2. 10. 1968.
(F. K. Schattauer Verlag, Stuttgart-New York 1969).

HALL, P., MELLNER, Ch. and DANIELSSON, T.: J 5 — A data processing system for medical information.
Meth. Inform. Med. *6:* 1—6, 1967.

HEDLEY, A. J., SCOTT, A. M. and DEBENHAM, G.: A computer assisted follow-up register.
Meth. Inform. Med. *8:* 67—77, 1969.

HOBBS, M. S. T. and ACHESON, E. D.: Secondary sex ratio following bleeding in pregnancy.
Lancet 1966, I, 462—463.

KÄLLEN, B. and WINBERG, J.: A Swedish register of congenital malformations. Experience with continuous registration during 2 years with special reference to multiple malformations.
Pediatrics *41:* 765—776, 1968.

MACMAHON, B. and NEWILL, V. A.: Birth characteristics of children dying of malignant neoplasms.
J. nat. Cancer Inst. *28:* 231—244, 1962.

MASI, A. T., SARTWELL, P. E. and SHULMAN, L. E.: The use of record linkage to determine familial occurrence of disease from hospital records. (Hashimoto's disease).
Amer. J. publ. Hlth *54:* 1887—1894, 1964.

MOORE, F.: Health Information Systems.
(Univ. of South. Calif. School of Med., Los Angeles 1963).

MOORE, F. J.: Mechanizing a large register of first order patient data.
Meth. Inform. Med. *4:* 1—10, 1965.

MOSBECH, J.: General discussion and summing up.
In E. D. Acheson (Edit.): Record Linkage in Medicine, pp. 381—383.
(E. & S. Livingstone Ltd., Edinburgh and London 1968).

NEUWIRTH, A. A., REDMOND, M. M. and WILSON, D.:
A »womb to tomb« public health record.
Health (New Haven) *79:* 2—7, 1952.

NEWCOMBE, H. B.: Record linkage and hospital discharge abstracts. (Working Paper for the Conference on Hospital Discharge Abstracts, Airlie House,
Warrenton/Va., June 1969; Med. Care, in press).

PHILLIPS, W., jr.: Record linkage for a chronic disease register.
In E. D. Acheson (Edit.): Record Linkage in Medicine, pp. 120—151.
(E. & S. Livingstone Ltd., Edinburgh and London 1968).

Rang, E. H., Acheson, E. D. and O'Connor, B. T.: Clinical significance of deaths after discharge from hospital unrecorded in the hospital notes.
Lancet *1968*, II, 908—910.

Tötterman, L. E., Forsström, E., Rämö, P. et al.: Completion of the delivery record of the central hospital with data obtained from the maternity card, follow-up examination and auto-anamnesis form (Finn.).
Suom. Lääk. *21:* 2437—2443, 1966.

WHO: The public health use of electronic computers. Report on a Seminar convened by the Regional Office for Europe of the WHO, London, 17—21 June 1968.
WHO-Document *EURO 3092, WHO Regional Office for* Europe, Copenhagen 1969.

Yoder, R. D., Swearingen, D. R., Schental, J. E.: Sweeney, J. W. and Nettleton, W. J., Jr.: An automated clinical information system.
Meth. Inform. Med. *3:* 45—50, 1964.

2.)Epidemiology

Acheson, E. D.: Oxford Record Linkage Study: A central file of morbidity and mortality records for a pilot population.
Brit. J. prev. soc. Med. *18:* 8—13, 1964.

Acheson, E. D.: Record linkage techniques in studies of the etiology of cancer.
Proc. roy. Soc. Med. *61:* 726—730, 1968.

Acheson, E. D. and Bachrach, C. A.: The distribution of multiple sclerosis in U.S. veterans by birthplace.
Amer. J. Hyg. *72:* 88—99, 1960.

Acheson, E. D., Truelove, S. C. and Witts, L. J.: National Epidemiology.
Brit. med. J. *1961*, I, 668.

Ad Hoc Committee on the Implications of Record Linkage for Health-Related Research: Health Research Uses of Record Linkage in Canada.
(Med. Res. Counc. of Canada, Ottawa 1968).

Bahn, A. K.: The development of an effective statistical system in mental illness.
Amer. J. Psychiat. *116:* 798—800, 1960.

Bahn, A. K.: A new psychiatric epidemiology.
Israel Ann. Psychiat. *2:* 11—18, 1964.

Balodimos, M. C. and Hurxthal, L. M.: The remote prediabetic state: effect on infant size, fetal and perinatal mortality.
Geriatrics *21:* 119—127, 1966.

Bidstrup, P. L. and Case, R. A. M.: Carcinoma of lung in workmen in bichromates-producing industry in Great Britain.
Brit. J. industr. Med. *13:* 260—264, 1956.

Bjarnason, O., Fridriksson, S. and Magnusson, M.: Record linkage in a self-contained community.
In E. D. Acheson (Edit.): Record Linkage in Medicine, pp. 62—68.
(E. & S. Livingstone Ltd., Edinburgh and London 1968).

Bothwell, P. W.: A New Look at Preventive Medicine.
(Pitman Medical Publishing Comp., London 1965).

Bragg, R. L.: Risk of admission to mental hospital following hysterectomy or cholecystectomy.
Amer. J. publ. Hlth *55:* 1403—1410, 1965.

Case, R. A. M. and Lea, J.: Mustard gas poisoning, chronic bronchitis and lung cancer.
Brit. J. prev. soc. Med. *9:* 62—72, 1955.

Case, R. A. M.: Cohort studies in assessing environmental hazards.
In E. D. Acheson (Edit.): Record Linkage in Medicine, pp. 207—213.
(E. & S. Livingstone Ltd., Edinburgh and London 1968).

Case, R. A. M. and Hosker, M. E.: Tumours of the urinary bladder as an occupational disease in the rubber industry in England and Wales.
Brit. J. prev. soc. Med. *8:* 39—50, 1954.

Case, R. A. M., Hosker, M. E., McDonald, D. B. and Pearson, J. T.: Tumours of the urinary bladder in workmen engaged in the manufacture and use of certain dyestuff intermediates in the British Chemical Industry.
Brit. J. industr. Med. *11:* 75—104, 1954.

Christopherson, W. M. and Parker, J. E.: Relation of cervical cancer to early marriage and childbearing.
New Engl. J. Med. *273:* 235—239, 1965.

Ciocco, A.: On the mortality in husbands and wives.
Hum. Biol. *12:* 508—531, 1940.

Conterio, F. and Cavalli-Sforza, L. L.: Evolution of the human constitutional phenotype: an analysis of mortality effects.
Ric. scient., Suppl. *29*, 71—78, 1957.

Court-Brown, W. M. and Doll, R.: Mortality from cancer and other causes after radiotherapy for ankylosing spondylitis.
Brit. med. J. *1965*, II, 1327—1332.

Court-Brown, W. M., Doll, R. and Hill, A. B.: Incidence of leukaemia after exposure to diagnostic radiation in utero.
Brit. med. J. *1960*, II, 1539—1545.

Da Silva, H. J., Abbatt, J. D., Da Motta, L. C. and Roriz, M. J.: Malignancy and other late effects following administration of thorotrast.
Lancet *1965*, II, 201—205.

Davies, M.: Toward a medical data bank for a total population.
Datamation *15:* 257—260, 1969.

Davies, M. A., Prywes, R., Tzur, B., Weiskopf, P. and Sterk, U. V.:
The Jerusalem perinatal study. 1. Design and Organization of a continuing, community-based, record-linked survey.
Israel J. Med. Sci. *5:* 1095—1106, 1969.

Doll, R.: Record linkage techniques in studies of the etiology of cancer.
Proc. roy. Soc. Med. *61:* 731—732, 1968.

Doll, R., Jones, F. A., Pygott, F. and Stubbe, J. L.: The risk of gastric cancer after medical treatment for gastric ulcer.
Gastroenterologia *88:* 1—12, 1957.

Dungal, N.: The specific problem of stomach cancer in Iceland.
J. Amer. med. Ass. *178:* 789—798, 1961.

EVANS, J.: Deliberate self-poisoning in the Oxford area.
Brit. J. prev. soc. Med. *21:* 97—107, 1967.

HEASMAN, M. A.: The use of record linkage in long-term prospective studies.
In E. D. Acheson (Edit.): Record Linkage in Medicine, pp. 251—256.
(E. & S. Livingstone Ltd., Edinburgh and London 1968).

HILL, A. B., DOLL, R., GALLOWAY, T. M. and HUGHES, J. P. W.: Virus diseases in pregnancy and congenital defects.
Brit. J. prev. soc. Med. *12:* 1—7, 1958.

HIRAYAMA, T. and WYNDER, E. L.: A study of the epidemiology of cancer of the breast — II. The influence of hysterectomy.
Cancer *15:* 28—38, 1962.

HOBBS, M. S. T. and ACHESON, E. D.: Perinatal mortality and the organization of obstetrics services in the Oxford area in 1962.
Brit. med. J. *1966*, I, 499—505.

HOBBS, M. S. T. and ACHESON, E. D.: Secondary sex ratio following bleeding in pregnancy.
Lancet *1966*, I, 462—463.

KELLER, A. Z.: The epidemiology of lip, oral and pharyngeal cancers, and the association with selected systemic diseases.
Amer. J. publ. Hlth *53:* 1214—1228, 1963.

KILPATRICK, S. I., MATHERS, J. D. and STEVENSON, A. C.: The importance of population fertility and consanguinity data being available in medico-social studies.
Ulster med. J. *24:* 113—122, 1955.

LENZ, W.: Malformations caused by drugs in pregnancy.
Amer. J. Dis. Child. *112:* 99—106, 1966.

MACCACARO, G. A.: Elaborazione elettronica per la medicina preventiva.
Appl. bio-med. calc. elettr. *3:* 259—279, 1968.

MACKENZIE, I.: Breast cancer following multiple fluoroscopies.
Brit. J. Cancer *19:* 1—8, 1965.

MOORE, F.: Health Information Systems.
(Univ. of South. Calif. School of Med., Los Angeles 1963).

MOORE, F.: Metropolitan health information systems.
Proc. 11. Internat. Congr. Med. Rec. Libr., Chicago 1963.

NEWCOMBE, H. B.: Panel discussion. Session on epidemiological studies.
In: 2. Int. Conf. on Congenital Malformations, pp. 345—349.
(International Medical Congress Ltd., New York 1964).

NEWCOMBE, H. B.: Present state and long-term objectives of the British Columbia population study.
In J. F. Crow, J. V. Neel (Edits): Proc. 3. Internat. Congr. Hum. Genet., Chicago 1966, pp. 291—313.
(Johns Hopkins Press, Baltimore 1967).

NEWCOMBE, H. B.: Die Anwendung des Medical Record Linkage für Bevölkerungs- und genetische Studien.
In E. Fritze und G. Wagner (Hrsg.): Dokumentation des Krankheitsverlaufs. Bericht über die 13. Jahrestagung der GMD in Bochum, 30. 9. — 2. 10. 1968.
(F. K. Schattauer Verlag, Stuttgart-New York 1969).

NEWCOMBE, H. B.: The use of medical record linkage for population and genetic studies.
Meth. Inform. Med. *8:* 7—11, 1969.

NEWCOMBE, H. B.: Pooled records from multiple sources for monitoring congenital anomalies.
Brit. J. prev. soc. Med. *23:* 226—232, 1969.

NEWCOMBE, H. B.: The value of Canadian hospital insurance records for detecting increases in congenital anomalies.
Canad. med. Ass. J. *101:* 121—128, 1969.

NEWCOMBE, H. B. and RHYNAS, P. O. W.: The cost of individual follow-up studies of large populations.
Proc. Internat. Population Conf., New York 1961;
Atomic Energy of Canada Ltd., Report No. 1255, Chalk River, Ontario 1961.

PAFFENBARGER, R. S., WOLF, P. A., NOTKIN, J. and THORNE, M. C.: Chronic disease in former college students.
I. Early precursors of fatal coronary heart disease.
Amer. J. Epidem. *83:* 314—328, 1966.

PELL, S.: Epidemiologic studies in a large company based on health and personal records.
Publ. Hlth Rep. *83:* 399—405, 1968.

PHILLIPS, W., jr., GORWITZ, K. and BAHN, A. K.: Electronic maintenance of case registers.
Publ. Hlth Rep. *77:* 503—510, 1962.

RENWICK, D. H.: The combined use of a central registry and vital records for incidence studies of congenital defects.
Brit. J. prev. soc. Med. *22:* 61—67, 1968.

SHAPIRO, S., WEINBLATT, E., FRANK, C., SAGER, R. and DENSEN, P.: The H. I. P. study of incidence and prognosis of coronary heart disease: Methodology.
J. chron. Dis. *16:* 1281—1292, 1963.

SLONE, D., GAETANO, L. F., LIPWORTH, L., SHAPIRO, S., LEWIS, G. P. and JICK, H.: Computer analysis of epidemiologic data of effect of drugs on hospital patients.
Publ. Hlth Rep. *84:* 39—52, 1969.

SMITHELLS, R. W.: Incidence of congenital abnormalities in Liverpool, 1960—64.
Brit. J. prev. soc. Med. *22:* 36—37, 1968.

STEWART, A. and HEWITT, D.: Leukaemia incidence in children in relation to radiation exposure in early life.
In Ebert, M. and Howard, A. (Edits): Current Topics in Radiation Research, Vol. 1, pp. 223—253.
(North-Holland Publishing Co., Amsterdam 1965).

STOCKS, P.: Measurement of morbidity.
Proc. roy. Soc. Med. *37:* 593—608, 1944.

U.S. National Committee on Vital and Health Statistics: Epidemiologic Uses of Vital and Health Statistics. A report of the Sub-Committee on Vital and Health Statistics for Epidemiologic Purposes.
(U.S. Govt. Print. Off., Washington, D.C., 1966).

WATTS, S. P. and ACHESON, E. D.: Computer method for deriving hospital in-patient morbidity statistics based on the person as the unit.
Brit. med. J. *1967*, IV, 476—477.

WELLS, J. and STEER, C. M.: Relationship of leukaemia in children to abdominal irradiation of mother during pregnancy.
Amer. J. Obstet. Gynec. *81:* 1059—1063, 1961.

WHO: The use of electronic computers in health statistics and medical research.
Report on a Symposium convened by the Regional Office for Europe of the WHO, Stockholm, 6—10 June 1966.
WHO Document EURO-341, WHO Regional Office for Europe, Copenhagen *1967*.

WHO: The public health use of electronic computers.
Report on a Seminar convened by the Regional Office for Europe of the WHO, London, 17—21 June 1968.
WHO-Document EURO 3092, WHO Regional Office for Europe, Copenhagen 1969.

Young, M., Benjamin, B. and Wallis, C.: The mortality of widowers.
Lancet *1963*, II, 454—456.

3.) Vital Statistics, Demography

Acheson, E. D.: Medical Record Linkage.
(Oxford University Press, London-New York-Toronto 1967).

Acheson, E. D. (Edit.): Record Linkage in Medicine.
(E. & S. Livingstone Ltd., Edinburgh and London 1968).

Ad Hoc Committee on the Implications of Record Linkage for Health-Related Research: Health Research Uses of Record Linkage in Canada.
(Med. Res. Counc. of Canada, Ottawa 1968).

Alderson, M. and Meade, T.: Accuracy of diagnosis on death certificates compared with that in hospital records.
Brit. J. prev. soc. Med. *21:* 22—29, 1967.

Barrai, I., Moroni, A. and Cavalli-Sforza, L. L.: Further studies on record linkage from parish books.
In E. D. Acheson (Edit.): Record Linkage in Medicine, pp. 270—280.
(E. & S. Livingstone Ltd., Edinburgh and London 1968).

Benjamin, B.: Computers: Their use in medical statistics and health services.
Roy. Soc. Hlth J. *86:* 205—213, 1966.

Boyce, A. J., Küchemann, C. H. and Harrison, G. A.: Neighbourhood knowledge and the distribution of marriage distances.
Ann. hum. Genet. *30:* 335—338, 1967.

Cavalli-Sforza, L. L.: Demographic attacks on genetic problems: Some possibilities and results.
Proc. Seminar on the Use of Vital and Health Statistics for Genet. and Radiol. Studies.
(U.N. Publication, Sales No. 61, New York 1962).

Christensen, H. T.: Studies in child spacing. I. Premarital pregnancy as measured by the spacing of the first birth from marriage.
Amer. soc. Rev. *18:* 53—59, 1953.

Christensen, H. T. and Bowden, O. P.: Studies in child spacing. II. The time-interval between marriage of parents and birth of their first child, Tippecanoe County, Indiana.
Social Forces *31:* 346—351, 1953.

Christensen, H. T. and Meissner, H. H.: Studies in child spacing. III. Premarital pregnancies as a factor in divorce.
Amer. soc. Rev. *18:* 641—644, 1953.

Christensen, H. T. and Rubenstein, B. B.: Premarital pregnancy and divorce.
Marr. Fam. Living *18:* 114—123, 1956.

Douglas, J. W. B. and Blomfield, J. M.: The reliability of longitudinal surveys.
Milbank mem. Fd Quart. *34:* 227—252, 1956.

Dunn, H. L.: Elements of a coordination system of vital records and statistics.
Publ. Hlth Rep. *68:* 793—801, 1953.

Hambright, T. Z.: Comparability of age on the death certificate and matching census record. Vital and Health Statistics, Publ. Hlth Publ. No. 1000-Series 2-No. 29.
(U.S. Dept. H. E. W., Washington, D.C., 1968).

Henry, L.: Problèmes de la recherche démographique moderne.
Population *21:* 1093—1114, 1966.

Hubbard, M. R. and Acheson, E. D.: Notification of death occurring after discharge from hospital.
Brit. med. J. *1967*, III, 612—613.

Keyfitz, N.: Utilisation des machines électroniques dans les calculs démographiques.
Population *19:* 673—682, 1964.

Küchemann, C. H., Boyce, A. J. and Harrison, G. A.:
A demographic and genetic study of a group of Oxfordshire villages.
Hum. Biol. *39:* 251—276, 1967.

Leslie, G. R., Christensen, H. T. and Pearman, G. L.: Studies in child spacing. IV. The time interval separating all children in completed families of Purdue University Graduates.
Social Forces *34:* 77—82, 1955.

Lunde, A. S. (Edit.): The 1970 Census and Vital and Health Statistics. A study group report of the Public Health Conference on Records and Statistics.
Vital and Health Statistics, Publ. Hlth Publ. No. 1000-Series 4-No. 10.
(U.S. Dept. H. E. W., Washington, D. C., 1969).

Madow, W. G.: Interview data on chronic conditions compared with information derived from medical records.
Vital and Health Statistics, Publ. Hlth Publ. No. 1000-Series 2-No. 23.
(U.S. Dept. H. E. W., Washington, D.C., 1967).

Marshall, J. T.: Canada's national vital statistics index.
Popul. Stud. *1:* 204—211, 1947.

Mather, K.: Genetical demography.
Proc. roy. Soc. B *159:* 106—125, 1963.

McCarthy, M. A.: Comparison of the classification of place of residence on death certificates and matching census records. Vital and Health Statistics, Publ. Hlth Publ. No. 1000-Series 2-No. 30.
(U.S. Dept. H. E. W., Washington, D.C., 1969).

Minet, P. L.: Fertilité précoce d'une cohorte de mariages dans une province Canadienne.
Acta genet. (Basel) *14:* 186—196, 1964.

Moore, F.: Development of a regional health information center.
Proc. 4. IBM Medical Symp., Endicott, N.Y., pp. 225—239.
(Yorktown Heights/N.Y. *1962*).

MORONI, A.: Sources, reliability and usefulness of consanguinity data, with special reference to Catholic records. Proc. Seminar on the Use of Vital and Health Statistics for Genetic and Radiation Studies, pp. 109—118.
(U.N. Publication, Sales No. 61, New York 1962).

NEWCOMBE, H. B.: Radiation, genetics and vital statistics: A pilot study in demographic genetics.
(Brochure prepared for the exhibit of Atomic Energy of Canada Ltd. at the 10th Int. Congr. Genet., Montreal 1958).

NEWCOMBE, H. B.: The uniqueness of Canadian vital statistics for studies in population genetics.
Canad. J. Genet. Cytol. *1:* 13—15, 1959.

NEWCOMBE, H. B.: Use of vital statistics.
In: U.N. World Population Conference, Belgrade, 1965, Vol. II, pp. 494—497. (United Nations, New York 1965).

NEWCOMBE, H. B.: Couplage de données pour les études démographiques.
Population *24:* 653—684, 1969.

NEWCOMBE, H. B., JAMES, A. P., AXFORD, S. J.: Family Linkage of Vital and Health Records:
(1) For population studies of hereditary influences on health,
(2) For verification of status for welfare programmes and other purposes.
Atomic Energy of Canada Ltd., Report No. 470,
Chalk River, Ontario 1957.

NEWCOMBE, H. B. and KENNEDY, J. M.: Demographic analysis and computer programs.
In: U.N. World Population Conference, Belgrade, 1965, Vol. III, pp. 251—253. (United Nations, New York 1965).

NEWCOMBE, H. B. and SMITH, M. E.: Changing patterns of family growth: The value of linked records as a source of data.
Popul. Stud. (in press).

RANG, E. H., ACHESON, E. D. and O'CONNOR, B. T.: Clinical significance of deaths after discharge from hospital unrecorded in the hospital notes.
Lancet *1968,* II, 908—910.

Royal Commission on Population: Reports and selected papers of the Statistics Committee, Vol. II, pp. 29—48.
(H.M.S.O., London 1950).

SHAPIRO, S. and SCHACHTER, J.: Methodology and summary results of the 1950 birth registration test in the United States.
Estadist. J. Int.-Amer. Statist. Inst. *10:* 688—699, 1952.

SMITH, A. E.: Automatic linkage of medical and vital registration records.
Brit. J. soc. prev. Med. *17:* 185—190, 1963.

SPICER, C. C.: Practical problems of record linkage on a national scale.
In Medical Research Council: Mathematics and Computer Science in Biology and Medicine.
(H.M.S.O., London 1965).

SUTTER, J., COUX, J.-M. et MUGNIER, M.: Organigrammes pour l'étude mécanographique de la parenté et de la fécondité dans une population.
Population *21:* 75—98, 1965.

THOMAS, D. S.: Continuous register system of population accounting.
In National Resources Committee (Edit.): The problems of a changing Population, pp. 276—297.
(U.S. Govt. Print. Off., Washington, D.C., 1938).

4.) Genetics

ACHESON, E. D.: Hospital morbidity in early life in relation to certain maternal and foetal characteristics and events at delivery.
Brit. J. prev. soc. Med. *19:* 164—173, 1965.

ACHESON, E. D.: Medical Record Linkage.
(Oxford University Press, London-New York-Toronto 1967).

ACHESON, E. D. (Edit.): Record Linkage in Medicine.
(E. & S. Livingstone Ltd., Edinburgh and London 1968).

Ad Hoc Committee on the Implications of Record Linkage for Health-Related Research: Health Research Uses of Record Linkage in Canada.
(Med. Res. Counc. of Canada, Ottawa 1968).

BARR, A. and STEVENSON, A. C.: Stillbirth and infant mortality in twins.
Ann. hum. Genet. *25:* 131—140, 1961.

BARRAI, I., CAVALLI-SFORZA, L. L. and MAINARDI, M.: Studio pilota per la determinazione degli effetti della consanguineita su caratteri esaminati alla visita di leva.
(Atti V. Riun. Scient. A.G.I., pp. 317—331, 1959).

BARRAI, I., CAVALLI-SFORZA, L. L. and MORONI, A.: Record linkage from parish books.
In Medical Research Council: Mathematics and Computer Science in Biology and Medicine, pp. 51—60.
(H.M.S.O., London 1965).

BARRAI, I., MORONI, A. and CAVALLI-SFORZA, L. L.: Further studies on record linkage from parish books.
In E. D. Acheson (Edit.): Record Linkage in Medicine, pp. 270—280.
(E. & S. Livingstone Ltd., Edinburgh and London 1968).

BOYCE, A. J., KÜCHEMANN, C. F. and HARRISON, G. A.: The reconstruction of historical movement patterns.
In E. D. Acheson (Edit.): Record Linkage in Medicine, pp. 303—317
(E. & S. Livingstone Ltd., Edinburgh and London 1968).

CARTER, C. O.: General discussion and summing up.
In E. D. Acheson (Edit.): Record Linkage in Medicine, pp. 383—385.
(E. & S. Livingstone Ltd., Edinburgh and London 1968).

CAVALLI-SFORZA, L. L.: Some notes on the breeding patterns of human populations.
Acta Genet. med. (Roma) *6:* 395—399, 1957.

CAVALLI-SFORZA, L. L.: Some data on the genetic structure of human populations.
Proc. 10. Int. Congr. Genet. Darwin Centenn. Symp. Genet. Evolut., Vol. I, pp. 388—407.
(Univ. of Toronto Press, Toronto 1959).

CAVALLI-SFORZA, L. L.: Demographic attacks on genetic problems: Some possibilities and results.
Proc. Seminar on the Use of Vital and Health Statistics for Genet. and Radiol. Studies.
(U.N. Publication, Sales No. 61, New York 1962).

CHRISTENSEN, H. T.: The method of record linkage applied to family data.
Marr. Fam. Living *20:* 38—43, 1958.

CHUNG, C. S.: Applications of digital computers in human genetics.
Meth. Inform. Med. *3:* 67—72, 1964.

CROW, J. F.: Some possibilities for measuring selection intensities in man.
Hum. Biol. *30:* 1—13, 1958.

DUNN, H. L. and GILBERT, M.: Public health begins in the family.
Publ. Hlth Rep. (Wash.) *71:* 1002—1010, 1956.

EDWARDS, J. H.: The interpretation of pedigree data.
In E. D. Acheson (Edit.): Record Linkage in Medicine, pp. 282—292.
(E. & S. Livingstone Ltd., Edinburgh and London 1968).

EDWARDS, J. H.: Linkage studies of whole populations.
In J. F. Crow and J. V. Neel (Edits): Proc. 3. Int. Congr. Hum. Genet., Chicago 1966, pp. 479—482.
(Johns Hopkins Press, Baltimore 1967).

FALCONER, D. S.: The inheritance of liability to certain diseases, estimated from the incidence among relatives.
Ann. hum. Genet. *29:* 51—76, 1965.

FRITZE, E. und WAGNER, G. (Hrsg.): Dokumentation des Krankheitsverlaufs. Probleme der Erfassung des zeitlichen Krankheitsablaufes und des Medical Record Linkage.
Verh. Bericht 13. Jahrgang GMD, Bochum 30. 9. — 2. 10. 1968.
(F. K. Schattauer Verlag, Stuttgart-New York 1969).

GEDDA, L. and MILANI-COMPARETTI, M.: Computerization of a permanent twin register: a basic tool in scientific research.
Acta Genet. med. (Roma) *15:* 333—344, 1966.

HIGGINS, J. V., REED, E. W. and REED, S. C.: Intelligence and family size: a paradox resolved.
Eugen. Quart. *9:* 84—90, 1962.

HOBBS, M. S. T.: A study of birthweight and other factors in sibships.
In E. D. Acheson (Edit.): Record Linkage in Medicine, pp. 358—367.
(E. & S. Livingstone Ltd., Edinburgh and London 1968).

KENNEDY, J. M., NEWCOMBE, H. B., OKAZAKI, E. A. and SMITH, M. E.: Computer methods for family linkage of vital health records.
Atomic Energy of Canada Ltd., Report No. 2222,
Chalk River, Ontario 1965.

KÜCHEMANN, C. H., BOYCE, A. J. and HARRISON, G. A.: A demographic and genetic study of a group of Oxfordshire villages.
Hum. Biol. *39:* 251—276, 1967.

LABERGE, C.: Prospectus for genetic studies in the French Canadians, with preliminary data on blood groups and consanguinity.
Bull. Johns Hopkins Hosp. *118:* 52—68, 1966.

MANGE, A. P.: Fortran programs for computing Wright's coefficient of inbreeding in human and non-human pedigrees.
Amer. J. hum. Genet. *16:* 484, 1964.

MASI, A. T., SARTWELL, P. E. and SHULMAN, L. E.: The use of record linkage to determine familial occurrence of disease from hospital records. (Hashimoto's disease).
Amer. J. publ. Hlth *54:* 1887—1894, 1964.

MATHER, K.: Genetical demography.
Proc. roy. Soc. B *159:* 106—125, 1963.

MCKUSICK, V. A.: Some computer applications to problems in human genetics.
Meth. Inform. Med. *4:* 183—189, 1965.

MCKUSICK, V. A.: Genealogic and bibliographic applications of computers in human genetics.
In J. F. Crow and J. V. Neel (edits): Proc. 3. Int. Congr. Hum. Genet., Chicago 1966, pp. 483—488.
(Johns Hopkins Press, Baltimore 1967).

MCKUSICK, V. A.: Family-oriented follow-up.
J. chron. Dis. *22:* 1—7, 1969.

MCKUSICK, V. A. and CROSS, H. E.: Generalogical linkage of records in two isolate populations.
In E. D. Acheson (Edit.): Record Linkage in Medicine, pp. 263—269.
(E. & S. Livingstone Ltd., Edinburgh and London 1968).

MCKUSICK, V. A., EGELAND, J. A., ELDRIDGE, R. and KRUSEN, D. E.: Dwarfism in the Amish. I. The Ellis-van Crefeld Syndrome.
Bull. Johns Hopkins Hosp. *115:* 306—336, 1964.

MCKUSICK, V. A., ELDRIDGE, R., HOSTETLER, J. A., RUANGWIT, U. and EGELAND, J. A.: Dwarfism in the Amish. II. Cartilage-hair hypoplasia.
Bull. Johns Hopkins Hosp. *116:* 285—326, 1965.

MCKUSICK, V. A., HOSTETLER, J. A. and EGELAND, J. A.: Genetic studies of the Amish. Background and potentialities.
Bull. Johns Hopkins Hosp. *115:* 203—222, 1964.

MCKUSICK, V. A., HOSTETLER, J. A. et al.: The distribution of certain genes in the Old Order Amish.
Cold Spr. Harb. Symp. quant. Biol. *29:* 99—114, 1964.

MI, M. P.: Record linkage and other genetic studies in Hawaii.
In J. F. Crow and J. V. Neel (Edits): Proc. 3. Int. Congr. Hum. Genet., Chicago 1966, pp. 489—496.
(Johns Hopkins Press, Baltimore 1967).

MILHAM, S.: Increased incidence of anencephalus and spina bifida in siblings of affected cases.
Science *138:* 593—594, 1962.

MILHAM, S.: Random distribution of affected birth ranks in anencephalic and spina bifida sibships with two affected cases.
Nature *200:* 480—481, 1963.

MILLER, J. R.: Human genetics in public health research.
Canad. J. publ. Hlth *7:* 1—8, 1966.

MORONI, A.: Sources, reliability and usefulness of consanguinity data, with special reference to Catholic records.
Proc. Seminar on the Use of Vital and Health Statistics for Genetic and Radiation Studies, pp. 109—118.
(U.N. Publication, Sales No. 61, New York 1962).

MORTON, N. E., CROW, J. F. and MULLER, H. J.: An estimate of the mutation damage in man from data on consanguineous marriages.
Proc. nat. Acad. Sci. *42:* 855—863, 1956.

NEWCOMBE, H. B.: Radiation, genetics and vital statistics: A pilot study in demographic genetics.
(Brochure prepared for the exhibit of Atomic Energy of Canada Ltd. at the 10th Int. Congr. Genet., Montreal 1958).

NEWCOMBE, H. B.: The uniqueness of Canadian vital statistics for studies in population genetics.
Canad. J. Genet. Cytol. *1:* 13—15, 1959.

NEWCOMBE, H. B.: Feasibility of estimating consequences of an increased mutation rate.
In Gardner, L. I. (Edit.): Molecular Genetics and Human Disease, pp. 186—203.
(C. C. Thomas, Springfield/Ill. 1960).

NEWCOMBE, H. B.: Genetics, radiation and people.
Canad. J. Genet. Cytol. *2:* 220—223, 1960.

NEWCOMBE, H. B.: Population genetics: Population records.
In: Methodology in Human Genetics, pp. 92—113.
(Holden-Day Inc., San Francisco 1962).

NEWCOMBE, H. B.: Untapped knowledge of human populations.
Trans. roy. Soc. Can., Sect. III, *56:* 173—180, 1962.

NEWCOMBE, H. B.: Genetic effects in populations, with special reference to studies in man, including ABCC results.
Radiat. Res. *16:* 531—545, 1962.

NEWCOMBE, H. B.: Screening for effects of maternal age and birth order in a register of handicapped children.
Ann. hum. Genet. *27:* 367—382, 1964.

NEWCOMBE, H. B.: Pedigrees for population studies. A progress report.
Cold Spr. Harb. Symp. quant. Biol. *29:* 21—30, 1964.

NEWCOMBE, H. B.: Use of vital statistics.
In: U.N. World Population Conference, Belgrade, 1965,
Vol. II, pp. 494—497.
(United Nations, New York 1965).

NEWCOMBE, H. B.: Environmental versus genetic interpretations of birth order effects.
Eugen. Quart. *12:* 90—101, 1965.

NEWCOMBE, H. B.: The study of mutation and selection in human populations.
Eugen. Rev. *57:* 109—125, 1965.

NEWCOMBE, H. B.: Radiation protection in Canada, Part VI: Problems in the assessment of genetic damage from exposure of individuals and populations to radiation.
Canad. med. Ass. J. *92:* 171—176, 1965.

NEWCOMBE, H. B.: Familial tendencies in diseases of children.
Brit. J. prev. soc. Med. *20:* 49—57, 1966.

NEWCOMBE, H. B.: Present state and long-term objectives of the British Columbia population study.
In J. F. Crow, J. V. Neel (Edits): Proc. 3. Internat. Congr. Hum. Genet., Chicago 1966, pp. 291—313.
(Johns Hopkins Press, Baltimore 1967).

NEWCOMBE, H. B.: Risks to siblings of stillborn children.
Canad. med. Ass. J. *98:* 189—193, 1968.

NEWCOMBE, H. B.: Products from the early stages in the development of a system of linked records.
In E. D. Acheson (Edit.): Record Linkage in Medicine,
pp. 7—33.
(E. & S. Livingstone Ltd., Edinburgh and London 1968).

NEWCOMBE, H. B.: Multigeneration pedigree from linked records.
In E. D. Acheson (Edit.): Record Linkage in Medicine,
pp. 295—301.
(E. & S. Livingstone Ltd., Edinburgh and London 1968).

NEWCOMBE, H. B.: Die Anwendung des Medical Record Linkage für Bevölkerungs- und genetische Studien.
In E. Fritze und G. Wagner (Hrsg.): Dokumentation des Krankheitsverlaufs. Bericht über die 13. Jahrestagung der GMD in Bochum, 30. 9. — 2. 10. 1968.
(F. K. Schattauer Verlag, Stuttgart-New York 1969).

NEWCOMBE, H. B.: The use of medical record linkage for population and genetic studies.
Meth. Inform. Med. *8:* 7—11, 1969.

NEWCOMBE, H. B.: Value of Canadian hospital insurance records in detecting increases in congenital anomalies.
Canad. med. Ass. J. *101:* 121—128, 1969.

NEWCOMBE, H. B.: The use of medical record linkage for monitoring congenital anomalies.
Brit. J. prev. soc. Med. *23:* 226—232, 1969.

NEWCOMBE, H. B., AXFORD, S. J. and JAMES, A. P.: A plan for the study of fertility of relatives of children suffering from hereditary and other defects.
Atomic Energy of Canada Ltd., Report No. 511,
Chalk River, Ontario 1957.

NEWCOMBE, H. B., JAMES, A. P., AXFORD, S. J.: Family Linkage of Vital and Health Records:
(1) For population studies of hereditary influences on health,
(2) For verification of status for welfare programmes and other purposes.
Atomic Energy of Canada Ltd., Report No. 470,
Chalk River, Ontario 1957.

NEWCOMBE, H. B. and RHYNAS, P. O. W.: Child spacing following stillbirth and infant death.
Eugen. Quart. *9:* 25—35, 1962.

NEWCOMBE, H. B. and TAVENDALE, O. G.: Maternal age and birth order correlations: Problems of distinguishing mutational from environmental components.
Mut. Res. *1:* 446—467, 1964.

NEWCOMBE, H. B. and TAVENDALE, O. G.: Effects of father's age on the risk of child handicap and death.
Amer. J. hum. Genet. *17:* 163—178, 1965.

OHKURA, K.: Use of family registration in the study of human genetics in Japan.
Jap. J. hum. Genet. *5:* 61—68, 1960.

REED, E. W. and REED, S. C.: Mental Retardation: A Family Study.
(W. B. Saunders, Philadelphia 1965).

SLATIS, H. M., REIS, R. H. and HOENE, R. E.: Consanguineous marriages in the Chicago region.
Amer. J. hum. Genet. *10:* 446—464, 1958.

SMITH M. E., SCHWARTZ, R. R. and NEWCOMBE, H. B.: Computer methods for extracting sibship data from family groupings of records.
Atomic Energy of Canada Ltd., Report No. 2520,
Chalk River, Ontario 1965.

WHO Expert Committee on Human Genetics: Human Genetics and Public Health.
WHO Technical Report Series No. 282.
(World Health Organization, Geneva 1964).

Woolf, C. M., Stephens, F. E., Mulaik, D. D. and Gilbert, R.E.: An investigation of the frequencies of consanguineous marriages among the Mormons and their relatives in the United States.
Amer. J. hum. Genet. *8:* 236—252, 1956.

Yanase, Y.: The use of the Japanese family register for genetic studies.
Proc. Seminar on the Use of Vital and Health Statistics for Genetic and Radiation Studies, pp. 119—133.
(U.N. Publication, Sales No. 61, New York 1962).

Yerushalmy, J.: Neonatal mortality by order of birth and age of parents.
Amer. J. Hyg. *28:* 244—270, 1938.

Editorial: Genetic studies by record linkage.
Lancet *1965*, I, 1056—1057.

5.) Public Health Services

Abrams, M. E.: An integrated medical records service at Thamesmead.
In E. D. Acheson (Edit.): Record Linkage in Medicine, pp. 370—373.
(E. & S. Livingstone Ltd., Edinburgh and London 1968).

Abrams, M. E., Bowden, K. F., Chamberlain, J. O. P. and Maccallum, I. R.: A computer-based general practice and health centre information system.
J. roy. Coll. gen. Pract. *16:* 415—427, 1968.

Acheson, E. D.: Some potentialities of the computer in national health services.
Proc. Intern. Meet. Automated Data Processing in Hospitals, Elsinore 1966.

Acheson, E. D.: Computers and medical record linkage. WHO-Seminar on the public health uses of electronic computers, London, 1968.
(WHO Document Euro 3092/6, WHO Regional Office for Europe, Copenhagen 1968).

Acheson, E. D.: Record Linkage in Medicine.
(E. & S. Livingstone Ltd., Edinburgh and London 1968).

Acheson, E. D.: Linkage of medical records.
Brit. med. Bull. *24:* 206—209, 1968.

Acheson, E. D.: Medical record linkage.
Meth. Inform. Med. *8:* 1—6, 1969.

Acheson, E. D. and Forbes, J. A.: Experiment in the retrieval of information in general practice. A preliminary report.
Brit. J. prev. soc. Med. *22:* 105—109, 1968.

Bahn, A. K.: Experience and philosophy with regard to case registers in health and welfare.
Commun. ment. Hlth J. *3:* 245—250, 1965.

Bahn, A. K.: Some research tools for community mental health planning and evaluation with particular reference to psychiatric case registers.
(Proc. Internat. Res. Seminar Commun. Mental Hlth Progr., May 1966).

Baldwin, J. A., Innes, G., Millar, W. M., Sharp, G. A. and Dorricott, N.: A psychiatric case register in north-east Scotland.
Brit. J. prev. soc. Med. *19:* 38—42, 1964.

Benjamin, B.: Integrated community health records.
Med. Rec. *10:* 41—47, 1969.

Davies, M.: Toward a medical data bank for a total population.
Datamation *15:* 257—260, 1969.

Fritze, E. und Wagner, G. (Hrsg.): Dokumentation des Krankheitsverlaufs. Probleme der Erfassung des zeitlichen Krankheitsablaufes und des Medical Record Linkage.
Verh. Bericht 13. Jahrestagung GMD, Bochum 30. 9. — 2. 10. 1968.
(F. K. Schattauer Verlag, Stuttgart-New York 1969).

Galloway, T. McL.: Management of vaccination and immunization procedures by electronic computers.
Med. Offr *109:* 232—233, 1963.

Galloway, T. McL.: Computers. Their use in local health administration.
Roy. Soc. Hlth J. *86:* 213—216, 1966.

Galloway, T. McL.: The use of an electronic computer in a health department.
Canad. J. publ. Hlth *57:* 331—332, 1966.

Galloway, T. McL.: Medical records and record linkage in local government.
In E. D. Acheson (Edit.): Record Linkage in Medicine, pp. 80—81.
(E. & S. Livingstone Ltd., Edinburgh and London 1968).

Gardner, E. A.: The use of a psychiatric case register in the planning and evaluation of a mental health program.
Psychiat. Res. Rep., Amer. psychiat. Ass. *22:* 259—281, 1967.

Gorwitz, K., Bahn, A. K., Klee, G. and Solomon, M.: Release and return rates for patients in state mental hospitals of Maryland.
Publ. Hlth Rep. *81:* 1095—1108, 1966.

Hannaford, R.: Automating the administration of public health.
Comput. Dig. *2:* 17, 1967.

Hobbs, M. S. T. and Acheson, E. D.: Perinatal mortality and the organization of obstetrics services in the Oxford area in 1962.
Brit. med. J. *1966*, I, 499—505.

Hobbs, M. S. T. and Acheson, E. D.: Obstetric care in the first pregnancy.
Lancet *1966*, I, 761—764.

Moore, F.: Health Information Systems.
(Univ. of South. Calif. School of Med., Los Angeles 1963).

Moore, F.: Metropolitan health information systems.
Proc. 11. Internat. Congr. Med. Rec. Libr., Chicago 1963.

Neuwirth, A. A., Redmond, M. M. and Wilson, D.:
A »womb to tomb« public health record.
Health (New Haven) *79:* 2—7, 1952.

Ramsay, J. D.: Electronic Scheduling of an Immunization Program — the Weyburn, Saskatchewan, Pilot Project.
Canad. J. publ. Hlth *60:* 459—464, 1969.

RANDALL, H. B.: Strengths and limitations of the cumulative health record.
J. Sch. Hlth 37: 86—89, 1967.

ROSNER, L. J.: Applications of automatic data processing to a public health agency's operations.
Publ. Hlth Rep. *80:* 625—632, 1965.

SAUNDERS, J. and SNAITH, A. H.: Cervical cytology: a computer assisted population screening programme.
Med. Offr *117:* 299—302, 1967.

WEIR, R. D.: The introduction of record linkage in northeast Scotland.
In E. D. Acheson (Edit.): Record Linkage in Medicine, pp. 55—60.
(E. & S. Livingstone Ltd., Edinburgh and London 1968).

6.) Other and non-medical fields

Ad Hoc Committee on the Implications of Record Linkage for Health-Related Research: Health Research Uses of Record Linkage in Canada.
(Med. Res. Counc. of Canada, Ottawa 1968).

BINDER, S.: Information storage, retrieval and processing: Present possibilities and future potentialities.
In: The Use of Vital and Health Statistics for Genetic and Radiation Studies, pp. 161—167.
(U.N. Publication, Sales No. 61, New York 1962).

BOYCE, A. J., KÜCHEMANN, C. F. and HARRISON, G. A.: The reconstruction of historical movement patterns.
In E. D. Acheson (Edit.): Record Linkage in Medicine, pp. 303—317.
(E. & S. Livingstone Ltd., Edinburgh and London 1968).

CHRISTENSEN, H. T.: Cultural relativism and premarital sex norms.
Amer. soc. Rev. *25:* 31—39, 1960.

CROXFORD, A. A.: Record linkage in education.
In E. D. Acheson (Edit.): Record Linkage in Medicine, pp. 351—356.
(E. & S. Livingstone Ltd., Edinburgh and London 1968).

DAVIDSON, L.: Retrieval of misspelled names in an airline passenger record system.
Comm. Ass. comput. Mach. *5:* 169—171, 1962.

DUBOIS, N. S. and D'ANDREA, J. R.: A document linkage program for digital computers.
Behav. Sci. *10:* 312—319, 1965.

TEPPING, B. J.: Study of matching techniques for subscriptions fulfillment.
(National Analysts Inc., Philadelphia 1965).

TEPPING, B. J. and CHU, J. T.: A report on matching rules applied to Reader's Digest Data.
(National Analysts Inc., Philadelphia 1958).

Editorial: Checking the Bouncers.
Time Mag. Jan. 4, 1963.

F. Costs of Record Linkage

ACHESON, E. D.: The structure, function and cost of a file of linked health data.
In Medical Research Council: Mathematics and Computer Science in Biology and Medicine, pp. 61—69.
(H.M.S.O., London 1965).

ACHESON, E. D.: Some potentialities of the computer in national health services.
Proc. Intern. Meet. Automated Data Processing in Hospitals, Elsinore 1966.

ACHESON, E. D.: Medical Record Linkage.
(Oxford University Press, London-New York-Toronto 1967).

ACHESON, E. D. (Edit.): Record Linkage in Medicine.
(E. & S. Livingstone Ltd., Edinburgh and London 1968).

ACHESON, E. D. and FORBES, J. A.: Experiment in the retrieval of information in general practice. A preliminary report.
Brit. J. prev. soc. Med. *22:* 105—109, 1968.

Ad Hoc Committee on the Implications of Record Linkage for Health-Related Research: Health Research Uses of Record Linkage in Canada.
(Med. Res. Counc. of Canada, Ottawa 1968).

NEWCOMBE, H. B. and RHYNAS, P. O. W.: The cost of individual follow-up studies of large populations.
Proc. Internat. Population Conf., New York 1961;
Atomic Energy of Canada Ltd., Report No. 1255, Chalk River, Ontario 1961.

WEIR, R. D.: The introduction of record linkage in northeast Scotland.
In E. D. Acheson (Edit.): Record Linkage in Medicine, pp. 55—60.
(E. & S. Livingstone Ltd., Edinburgh and London 1968).

Addresses of the Authors: Prof. Dr. med. G. Wagner, Institut für Dokumentation, Information und Statistik am Deutschen Krebsforschungszentrum, 6900 Heidelberg, Berliner Straße 27;

Dr. H. B. Newcombe, Head, Biology Branch, Biology and Health Physics Division, Atomic Energy of Canada Limited, Chalk River, Ontario/Canada.

THE DEVELOPMENT OF THE SMALL BUSINESS DATA BASE OF THE U.S. SMALL BUSINESS ADMINISTRATION: A WORKING BIBLIOGRAPHY

Compiled through 1985 by
Bruce D. Phillips, Small Business Administration

The Small Business Data Base (SBDB) is an integrated effort for developing and organizing data on the role of small business in the U.S. economy. It is designed to serve two purposes: (1) to comply with the mandate of the Congress and (2) to meet the needs of the research community for analyzing cause and effect relationships of small business problems and progress. Because of the diversity of small business, no one file or collection of statistics is adequate to meet those needs.

The Congressional mandate as stated in Public Law 96-302 instructs the Office of Advocacy of the Small Business Administration to develop a data base to be used for historical description and policy analysis. The specifics of the law are divided into two parts: the "indicative" data base for creating mailing lists and the "external" data base for descriptive statistics and policy analysis.

The SBDB effort is unique because it stresses maximum use of business microdata, the files of individual firms. Therefore, the file provides information of the "indicative" data base and overcomes the problem of non-comparability of longitudinal data for firms of small and large size classes, as well as other major difficulties of establishing the "external" data base.

The "indicative" data base is the Master Establishment List (MEL) of more than 8.9 million 1984 establishment records. It provides the names, addresses and industry and geographic codes for 3.4 million establishments, and adds employment, sales, age of firm and enterprise linkage for an additional 5.5 million firms and establishments. The 1984 MEL, with information current as of January 1, 1985, will be completed November 30, 1985.

The "external" data base is based upon proprietary files of the Dun and Bradstreet Corporation. The USEEM (United States Establishment and Enterprise Microdata) files are cross-sectional files of about 5 million records each for 1976, 1978, 1980, and 1982. (The 1984 files will be available during the Spring of 1986.) The USELM (United States Establishment Longitudinal Microdata) files are longitudinally linked files covering the same years as the USEEM files. However, USELM files are based on a representative weighted sample of verified and linked establishment records which have been edited for consistency over time. The USELM is an approximate 50 percent of the USEEM files.

There are three methods of accessing data in the Small Business Data Base: by obtaining aggregate data which has been published in summary form, by purchasing detailed data at the U.S. or sub-state levels, and by ordering a customized tabulation on a cost reimbursable basis from the Data Base Branch of the Office of Advocacy (202-634-7550).

The papers in the enclosed bibliography provide a record of the progress to date in the development of the Small Business Data Base including details on increasing the access to it. Both methodological and applications papers are included in the hope that persons interested in small business research will consult both the enclosed studies for reference, as well as design new research applications in areas of importance to the small business community.

INTRODUCTION

The studies below describe the creation, documentation, and applications of the Small Business Data Base of the Office of Advocacy of the U.S. Small Business Administration. The lists below, while comprehensive, examine only the most relevant studies during the years 1980-1985; the bibliography is therefore representative but not necessarily exhaustive. The studies below do, however, provide a recent chronological history on the development of the Small Business Data Base, and examples of some applications using the available data.

Two types of studies have generally been included. First, one collection of papers describes the detailed creation of the three major files of the Small Business Data Base: the USEEM (United States Establishment and Enterprise Microdata) file, the MEL (Master Establishment List), and the USELM (United States Establishment Microdata file). These files contain approximately 5 million, 8 million, and 20 million records, respectively, and their development is described in the papers in this bibliography.

In general, the USEEM file is available on both an enterprise and establishment basis by size class, while the MEL includes USEEM, plus an additional 3 million businesses appearing in yellow page type commercial listings. The MEL is only an establishment file. The USELM file is available for establishments only (by enterprise size class). Both USEEM and USELM files are available for 1976, 1978, 1980, and 1982. (The 1984 files will be available during the Spring of 1986.) The MEL file for 1984 (with data current as of January 1, 1985) will be available on November 30, 1985.

The second group of studies detailed in this bibliography are research applications either using the data files directly, or comparing them with other government data sources, such as from the Bureau of the Census, the Internal Revenue Service, or the Bureau of Labor Statistics. The papers contained in these sections are both by staff members of the Office of Economic Research, as well as by SBA contractors.

It is hoped that the source materials listed below will be periodically expanded and updated as new contracts are completed, and as additional years of data become available. For example, some of the papers in the section describing the Dun and Bradstreet Financial Statistics File are quite preliminary, and are the result of previous attempts to assess the overall quality of the data on an industry specific basis. Ongoing contracts are creating a financial encyclopedia out of this source on an aggregate basis. In still other ongoing research, the development of a longitudinal enterprise file, using the USEEM database, is expected to be completed during the Spring of 1986 for the years 1976-1984. Finally, several papers from ongoing interagency agreements between SBA and other agencies, particularly the Internal Revenue Service, are described which are augmenting the Small Business Data Base.

I. 1984/1985

A. Methodological Papers and File Descriptions

Richard Boden and Bruce D. Phillips, "Uses and Limitations of USEEM/USELM Data." Office of Advocacy, U.S. Small Business Administration, October 1985.

C.D. Day, "1979 Corporation, Partnership and Sole Proprietorship Employment and Payroll Studies: An Initial Look at the Relative Efficiency of Small and Large Business." Prepared under an Interagency Agreement between the Statistics of Income Division, Internal Revenue Service, U.S. Department of the Treasury, and the U.S. Small Business Administration, Office of Advocacy, draft October 1985. (This project entailed a match of SOI files with employment enhancements.)

Nick Greenia, "1979 Sole Proprietorship Employment and Payroll: Processing Methodology," Record Linkage Techniques--1985, Internal Revenue Service, 1985. Prepared under an interagency agreement with the Office of Advocacy, U.S. Small Business Administration.

Lou Jacobsen, "Analysis of the Accuracy of SBA's Small Business Data Base." Prepared by the Hudson Institute of the Center for Naval Analysis under contract to the Office of Advocacy of the U.S. Small Business Administration, August 1985. (This was a matching study between UI and SBDB data.)

Steven Lustgarten and Stavros Thomadakis, "Firm Size and Resource Mobility." (Progress reports available, final report due December 1985). Prepared under contract SBA-7156-0A-83 for the Office of Advocacy of the U.S. Small Business Administration.

Social and Scientific Systems, Inc., "Review of Work Performed During 1984 and Projections for 1985," Washington, D.C., January 1985. Prepared for the Office of Advocacy of the U.S. Small Business Administration under contract number 3-84-6666.

Robert F. Teitel, "The Development Process for the Creation of the SBA Small Business Database Containing the U.S. Establishment and Enterprise Microdata (USEEM). Prepared under contract 9182-0A-83 for the Office of Advocacy of the U.S. Small Business Administration, September 1984.

U.S. Small Business Administration, Office of Advocacy, Office of Economic Research, Data Base Development Division, "Constructing a Business Microdata Base For The Analysis of Small Business Activity," November 1984.

U.S. Small Business Administration, Office of Advocacy, Office of Economic Research, Data Base Development Division, "The Derivation of the U.S. Establishment Longitudinal Microdata (USELM) File: The Weighted Integrated USEEM 1976-1982 Sample," December 1984.

B. Research Applications

Faith Ando, "Distribution of Business Loans, Credit and Investment Capital to Selected Sub-Categories of Small Business." Final report expected November 1985. Prepared by the JACA Corporation for the Office of Advocacy of the U.S. Small Business Administration under contract 6061-0A-82.

Aram Research Associates, "Informal Investor Survey in the Eastern Great Lakes." In preparation for the Office of Advocacy of the U.S. Small Business Administration under contract number SBA-7187-0A-83. Final report expected December 1985.

Jack Faucett, "Procurement Share vs. Industry Share." Final report October 1985. Prepared for the Office of Advocacy of the U.S. Small Business Administration under contract SBA-8566-0A-84.

North River Associates, "Small Business Use of Slack Resources and Service to New and Minor Markets." Final report expected November 1985. Prepared under contract number SBA-7185-0A-83 for the Office of Advocacy of the U.S. Small Business Administration.

Bruce D. Phillips, Hyder Ali Lakhani, and Samuel L. George, "The Economics of Metric Conversion for Small Manufacturing Firms in the United States." Technological Forecasting and Social Change: 25 (2), April 1984, pp. 109-121.

Bruce D. Phillips, "The Effect of Industry Deregulation on the Small Business Sector." Business Economics: 20(1), January 1985, pp. 28-39.

Willard Risdon, "Developing A Key Financial and Income Statements Data Base for Veteran Owned Business." Final report August 1985. Prepared for the Office of Veterans' Affairs of the U.S. Small Business Administration under contract #7215-VA-83.

David Rothenberg, "Differences Between Veteran and Non-Veteran Owned Businesses." Final report August 1985. Prepared for the Office of Veterans' Affairs of the U.S. Small Business Administration under contract #7215-VA-83.

David Rothenberg, "Firm Size and Profitability." Work in progress under contract 86-AER-84-280 for the Office of Advocacy, U.S. Small Business Administration. Final report expected January 1986.

Social and Scientific Systems, Inc., "Review of Work During Last Twelve Months and Work Projection for Next Twelve Months." Washington, D.C., March 1984. Prepared for the Office of Advocacy of the U.S. Small Business Administration under contract no. SBA-4-0-8(a) - C-2136.

II. 1983

A. Methodological Papers and File Descriptions

Candee Harris, "U.S. Establishment and Enterprise Microdata Database Description." Business Microdata Project, The Brookings Institution. Funded under contract to the Small Business Administration, Office of Advocacy, April, 1983.

Candee Harris, "Comparison of County Business Patterns and USEEM Employment Figures," Business Microdata Project, The Brookings Institution. Funded under contract to the Small Business Administration, Office of Advocacy, 1983.

Candee Harris, Handbook of Small Business Data: A Sourcebook for Researchers and Policy-Makers, U.S. Small Business Administration, Office of Advocacy, August, 1983.

David A. Hirschberg, "The Development of a Small Business Data Base: A Progress Report." Appendix B of The State of Small Business: A Report of the President, (GPO, 1983), pp. 271-301.

Hyder Lakhani, "Preliminary Final Report: Validity of the SBA's Master Establishment List, April, 1983," prepared by Social and Scientific Systems, Inc. Funded by the Small Business Administration.

Marjorie Odle and Catherine Armington, "Weighting the 1976-80 and the 1978-80 USEEM Files for Dynamic Analysis of Employment Growth," Business Microdata Project, The Brookings Institution. Funded under contract to the Small Business Administration, Office of Advocacy, Revised April 1983.

B. Research Applications

Catherine Armington, "Further Examination of Recent Sources of Employment Growth: Analysis of the USEEM Data for 1976-80," Business Microdata Project, The Brookings Institution. Funded under contract to the Small Business Administration, Office of Advocacy, March 1983.

Maureen C. Glebes, "An Economic Profile of the State of Indiana," Office of Advocacy, Small Business Administration, Washington, D.C., January 4, 1983.

Thomas A. Gray, with Maureen Glebes and Edward Starr, "Small Business in the U.S. Economy," Chapter 2 in The State of Small Business: A Report of the President, Washington, D.C., GPO, March 1983, pp. 27-58.

Bruce D. Phillips, with William Scheirer, "Small Business Dynamics and Methods for Measuring Job Generation." Chapter 3 in The State of Small Business: A Report of the President," Washington, D.C., GPO, March 1983, pp. 61-88.

Thomas A. Gray and David L. Hirschberg, "Shifts in the Employment Status of Proprietors, 1960-1975," Office of Economic Research, presentation for the Eastern Economics Association, March 10-11, 1983.

Hyder Lakhani, "Econometric Analysis of Profitability of Firms by Size in Retail Trade and Service Industries in 1980," Social and Scientific Systems, Inc. Final report prepared under contract to the Small Business Administration, Office of Advocacy, April 1983.

Social and Scientific Systems, Inc., "Financial Analysis of Firms by Size in Manufacturing, Services and Retail Trade Industries, 1977-1981: Final Report." Prepared under contract to the Small Business Administration, Office of Advocacy, February 1983.

III. 1982

A. Methodological Papers and File Descriptions

Catherine Armington and Marjorie Odle, "Small Businesses -- How Many Jobs?" The Brookings Review, Winter 1982, prepared under contract to the Small Business Administration.

Candee S. Harris, "A Comparison of Employment Data for Several Sources of Business Data: County Business Patterns, Unemployment Insurance and U.S. Establishment and Enterprise Microdata," Working Paper No. 5, Business Microdata Project, The Brookings Institution, Revised March 1982. Prepared under contract to the Office of Advocacy of the Small Business Administration.

Richard Hayes, Kevin Hollenbeck, and Marjorie Odle, "Development of an Enterprise Based Longitudinal Data File." The Policy Research Group, November 1982. Prepared under contract to the Office of Advocacy, Small Business Administration.

Bruce D. Phillips, "The Small Business Data Base and Other Sources of Business Information:

Recent Progress," The State of Small Business: A Report of the President, Washington, D.C., GPO, 1982, pp. 247-281.

Bruce D. Phillips and David A. Hirschberg, "Longitudinal Data for Small Business Analysis" in Development and Use of Longitudinal Establishment Data, U.S. Dept. of Commerce, Bureau of the Census, Economic Research Report ER-4, (GPO, 1982) pp. 93-109.

Paul Rose and Linda B. Taylor, "Size of Employment in SOI: A New Classifier" in U.S. Dept. of the Treasury, Internal Revenue Service, Statistics of Income and Related Administrative Record Research: 1982. U.S. Department of the Treasury, Internal Revenue Service, pp. 35-41.

Social and Scientific Systems, Inc., "Preliminary Report on the Development of the Master Establishment List," November 2, 1982. Funded under contract to the Office of Advocacy, Small Business Administration.

Social and Scientific Systems, Inc., "Technical and Analytical Support Provided to SBA During Fiscal Year 1982," November 12, 1982. Prepared under contract for the Office of Advocacy, Small Business Administration.

Nancy L. Spruill, "Measures of Confidentiality," Statistics of Income and Related Administrative Record Research: 1982, U.S. Department of the Treasury, Internal Revenue Service, 1982, pp. 131-137.

Dun and Bradstreet Financial Statistics File Papers

Alan Unger, "The Finstat Project Phase I: Descriptive Statistics and Quality Assessment of Financial Data on the Services Industries." Prepared by Group Operations, Inc., under contract for the Office of Advocacy, Small Business Administration, March, 1982.

Applied Systems Institute, "Development and Implementation of Automated Finstat Imputing Algorithms Phase I." Prepared under contract for the Office of Advocacy, Small Business Administration, March 1982.

Delta Research Corporation, "Finstat File Retail Sector (SIC Codes 5200-5999): Editing and Analysis Report." Prepared under contract for the Office of Advocacy, Small Business Administration, March 1982.

System Sciences Incorporated, "Phase I Final Report on the Investigation of the Dun and Bradstreet Finstat File Agricultural Sector." Prepared under contract for the Office of Advocacy, Small Business Administration, March 1982.

ESR Associates, "Analysis of Finstat Construction File." Prepared under contract for the Office of Advocacy, Small Business Administration, March 1982.

B. Research Applications

David L. Birch and Susan MacCracken, "The Small Business Share of Job Creation: Lessons Learned from the Use of a Longitudinal File." MIT Program on Neighborhood and Regional Change, Cambridge, Mass. Prepared under contract for the Office of Advocacy, Small Business Administration, November 1982.

Maureen C. Glebes, "Economic Profiles for Selected States: Connecticut, Florida, Hawaii, Illinois, New York, Pennsylvania, Tennessee, Texas, West Virginia, Wisconsin, Wyoming," Office of Advocacy, Small Business Administration, November, 1982.

David Hirschberg and Bruce D. Phillips, "Using Financial Data to Evaluate the Status of Small Business," in Statistics of Income and Related Administrative Record Research: 1982, U.S. Department of the Treasury, Internal Revenue Service, pp. 71-75.

Bruce D. Phillips, with William Whiston, Alice Cullen, and David Hirschberg, "Small Business in the U.S. Economy," Chapter 1 in The State of Small Business: A Report of the President, Washington, D.C., GPO, March 1982, pp. 37-60.

Bruce D. Phillips, with William Whiston, Alice Cullen, and David Hirschberg, "Current and Historical Trends in the Small Business Sector," Chapter 2 in The State of Small Business: A Report of the President, Washington, D.C., GPO, March, 1982, pp. 63-105.

Bruce D. Phillips and William Knight, "The Davis-Bacon Act Reconsidered: A New Small Business Tax," The Restructuring Economy: Implications for Small Firms, Bentley College, Waltham, Mass., August, 1982, pp. 330-352. Proceedings of the 1982 Small Business Research Conference.

Bruce D. Phillips and Hyder Lakhani, "A Study of Profit by Asset Size Class: Two Hypotheses," Small Business Administration, Office of Economic Research, draft, September, 1982.

IV. 1981

A. Methodological Papers and File Descriptions

Catherine Armington and Marjorie Odle, "Associating Establishments into Enterprises for a Microdata File of the U.S. Business Population," Statistics of Income and Related Administrative Record Research, U.S. Department of the Treasury, Internal Revenue Service, 1981, pp. 71-76.

Candee Harris, "Creating a Business Data Base from Dun and Bradstreet Data Files." Working Paper No. 3. Business Microdata Project, The Brookings Institution, March 1981. Prepared under contract for the Office of Advocacy, Small Business Administration.

Bruce A. Kirchhoff and David A. Hirschberg, "Small Business Data Base: Progress and Potential," Statistics of Income and Related Administrative Record Research, U.S. Department of the Treasury, Internal Revenue Service, 1981, pp. 61-67.

Constance Mitchell, Documentation of the Employment Imputation for the IUSBDE Using County Business Patterns Employment Aggregates. Business Microdata Project, The Brookings Institution, January 1981. Prepared under contract for the Office of Advocacy, Small Business Administration.

Constance Mitchell and Matthew Lynde, Documentation of the Imputation of Branch Records for the IUSBDB, Business Microdata Project, The Brookings Institution, July 1981. Prepared under contract for the Office of Advocacy, Small Business Administration.

Bruce D. Phillips, "A Comparison of Three Establishment-Based Data Sources: The Dun and Bradstreet Market Identifier File, County Business Patterns, and Unemployment Insurance (U.I.) Data, 1977-1978." Draft, Office of Economic Research, Small Business Administration, March 1981.

Marjorie Odle, Creating an Interim U.S. Business Data Base (IUSBDB): Documentation of the Match Process Linking the Dun and Bradstreet Data Files, Business Microdata Project, The Brookings Institution, January 1981. Prepared under contract for the Office of Advocacy, Small Business Administration.

B. Research Applications

David L. Birch and Susan MacCracken, "Corporate Evolution - A Micro-Based Analysis." MIT Program on Neighborhood and Regional Change, Cambridge, Mass. Prepared for the Office of Advocacy, Small Business Administration, January 1981.

Bruce D. Phillips, "Recent Trends in the Distribution of Employment by Business Size and Industry," Statistics of Income and Related Administrative Record Research, U.S. Department of the Treasury, Internal Revenue Service, 1981, pp. 77-87.

V. 1980

Methodological Papers and File Descriptions

Catherine Armington, "The Brookings Multi-Establishment Enterprise File," Working Paper No. 1, Business Microdata Project, The Brookings Institution, August 1980. Prepared under contract for the Office of Advocacy, Small Business Administration.

Maureen C. Glebes, "An Economic Profile of the State of Missouri," Office of Advocacy, Small Business Administration, Washington, D.C., November 2, 1980.

Maureen C. Glebes, "An Economic Profile of the State of New Hampshire," Office of Advocacy, Small Business Administration, Washington, D.C., October 26, 1980.

VI. 1979

Methodological Papers and File Descriptions

David L. Hirschberg and Vernon Renshaw, "Access to Administrative Records on Establishments and Individuals for Public Policy Analysis," Bureau of Economic Analysis, draft, 1979, prepared for the 1979 Annual Meetings of the American Statistical Association.

BIBLIOGRAPHY OF METHODOLOGICAL TECHNIQUES RELATED TO EXACT MATCHING

Compiled through 1984 by
Wray Smith, Harris-Smith Research, Inc.

This supplementary bibliography provides references to sources that are not found in most lists of publications in the field of exact matching but which may be useful to the investigator who is looking for additional tools, especially in the areas of contingency table techniques, cluster analysis, pattern recognition, and sequence comparison theory.

AITCHISON, J. and DUNSMORE, I.R. (1975), Statistical Prediction Analysis. Cambridge, UK, and New York, NY: Cambridge University Press.

BARR, A. and FEIGENBAUM, E.A., eds. (1982), The Handbook of Artificial Intelligence: Volume II. Los Altos, CA: William Kaufmann, Inc.

BARR, R.S. and TURNER, J.S. (1982), "Microdata File Merging Through Large-Scale Network Techniques," in D. Klingman and J. Mulvey, eds., Network Models and Associated Applications. Amsterdam: North-Holland Publ. Co.

BELLMAN, R. (1961), Adaptive Control Processes: A Guided Tour. Princeton: Princeton University Press.

BERGER, J.O. and WOLPERT, R.L. (1984), The Likelihood Principle: A Review, Generalizations, and Statistical Implications. IMS Lecture Notes - Monograph Series, Vol. 6. Hayward, CA: Inst. of Math. Statistics.

BISHOP, Y.M.M. (1971), "Effects of Collapsing Multidimensional Contingency Tables," Biometrics, 27, 545-562.

BISHOP, Y.M.M., FEINBERG, S.E., and HOLLAND, P.W. (1975), Discrete Multivariate Analysis: Theory and Practice, Cambridge, MA: MIT Press.

BOURNE, C.P. and FORD, D.F. (1961), "A Study of Methods for Systematically Abbreviating English Words and Names," Journal of the Association for Computing Machinery, 8, 538-552.

BREIMAN, L., FRIEDMAN, J., OLSHER, R. and STONE, C. (1984), Classification and Regression Trees. Belmont, CA: Wadsworth International Group.

CHVATAL, V. and SANKOFF, D. (1975), "Longest Common Subsequences of Two Random Sequences," Journal of Applied Probability, 12, 306-315.

CORMACK, R.M. (1971), "A Review of Classification," Journal of the Royal Statistical Society, Ser. A, 134, 321-367.

COWAN, C.D. and FAY, R.E., III (1984), "Estimates of Undercount in the 1980 Census," 1984 Proceedings of the American Statistical Association, Section on Survey Research Methods.

DEKEN, J. (1979), "Some Limit Results for Longest Common Subsequences," Discrete Mathematics, 26, 17-31.

DEMPSTER, A.P., LAIRD, N.M., and RUBIN, D.B. (1977), "Maximum Likelihood From Incomplete Data via the EM Algorithm," Journal of the Royal Statistical Society, Ser. B, 39, 1-38.

DOLBY, J.L. (1970), "Some Statistical Aspects of Character Recognition," Technometrics, 12, 231-245.

DUBOIS, D. and PRADE, H. (1980), Fuzzy Sets and Systems: Theory and Applications. New York: Academic Press.

DURAN, B.S. and ODELL, P.L. (1974), Cluster Analysis: A Survey. (Lecture Notes in Economics and Mathematical Systems, 100. Operations Research.) New York: Springer-Verlag.

EVERITT, B. (1974), Cluster Analysis. London: Heinemann.

EVERITT, B. and HAND, D. (1981), Finite Mixture Distributions. London: Chapman and Hall.

FATTI, L.P. (1983), "The Random-Effects Model in Discriminant Analysis," Journal of the American Statistical Association, 78, 679-687.

FIENBERG, S.E. (1977), The Analysis of Cross-Classified Categorical Data. Cambridge, MA: MIT Press.

FOWLKES, E.B. and MALLOWS, C.L. (1983), "A Method for Comparing Two Hierarchical Clusterings" (with discussion), Journal of the American Statistical Association, 78, 553-584.

GOLDSTEIN, M. and DILLON, W.R. (1978), Discrete Discriminant Analysis. New York: Wiley.

GORDON, A.D. (1981), Classification. London: Chapman and Hall.

GUPTA, M.M., RAGADE, R.K., and YAGER, R.R., eds. (1979), Advances in Fuzzy Set Theory and Applications. Amsterdam: North-Holland Publ. Co.

HALL, P.A.V. and DOWLING, G.R. (1980), "Approximate String Matching," Computing Surveys, 12, 381-402.

HAND, D.J. (1981), Discrimination and Classification. New York: Wiley.

HARTIGAN, J.A. (1975), Clustering Algorithms. New York: Wiley.

HARTIGAN, J.A. (1978), "Asymptotic Distributions for Clustering Criteria," Annals of Statistics, 6, 117-131.

HAWKINS, D.M. (1981), Identification of Outliers. NY: Chapman and Hall.

JAZWINSKI, A.H. (1970), Stochastic Processes and Filtering Theory. New York: Academic Press.

KANDEL, A. (1982), Fuzzy Sets and Pattern Recognition. New York: Wiley-Interscience.

KRISHNAIAH, P. and KANAL, L., eds. (1982), Handbook of Statistics, Vol. 2: Classification, Pattern Recognition, and Reduction of Dimensionality. New York: Elsevier/North-Holland.

KU, H.H., VARNER, R.N. and KULLBACK, S. (1971), "On the Analysis of Multi-dimensional Contingency Tables," Journal of the American Statistical Association, 66, 55-64.

LANCASTER, H.O. (1969), The Chi-squared Distribution. New York: Wiley.

LITTLE, R.J.A. (1976), "Inference About Means from Incomplete Multivariate Data," Biometrika, 63, 593-604.

MATHAI, A.M. and RATHIE, P.N. (1975), Basic Concepts in Information Theory and Statistics: Axiomatic Foundations and Applications. New York: Halsted Press/Wiley.

NEGOITA, C.V. and RALESCU, D.A. (1975), Applications of Fuzzy Sets to Systems Analysis. New York: Wiley.

OJA, E. (1983), Subspace Methods for Pattern Recognition. New York: Wiley (Research Studies Press).

OKNER, B.A. (1974), "Data Matching and Merging: An Overview," Annals of Economic and Social Measurement, 2, 347-352.

OKUDA, T., TANAKA, E. and KASAI, T. (1976), "A Method for the Correction of Garbled Words Based on the Levenshtein Metric," IEEE Transactions on Computers, C25, 172-177.

RODGERS, W.L. (1984), "An Evaluation of Statistical Matching," Journal of Business & Economic Statistics, 2, 91-102.

RUBIN, D.B. (1976), "Inference and Missing Data" (with discussion), Biometrika, 63, 581-592.

RUKHIN, A.L. (1984), "Adaptive Classification Procedures," Journal of the American Statistical Association, 79, 415-422.

SANKOFF, D. and KRUSKAL, J.B., eds. (1983), Time Warps, String Edits, and Macromolecules: The Theory and Practice of Sequence Comparison. Reading, MA: Addison-Wesley.

SEBER, G.A.F. (1984), Multivariate Observations. New York: Wiley.

SELLERS, P.H. (1980) "The Theory and Computation of Evolutionary Distances: Pattern Recognition," Journal of Algorithms, 1, 359-373.

SHAFFER, J.P. (1981), "Complexity: An Interpretability Criterion for Multiple Comparisons," Journal of the American Statistical Association, 76, 395-401.

ULAM, S.M. (1972), "Some Combinatorial Problems Studied Experimentally on Computing Machines," in Applications of Number Theory to Numerical Analysis, ed. S.K. Zaremba, New York: Academic Press, 1-3.

VAN RYZIN, J., ed. (1977), Classification and Clustering. New York, NY: Academic Press.

WAGNER, R.A. and FISCHER, M.J. (1974), "The String-to-String Correction Problem," Journal of Association for Computing Machinery, 21, 168-173.

WEGMAN, E.J. and SMITH, J.G. (1984), Statistical Signal Processing. New York: Marcel Dekker.

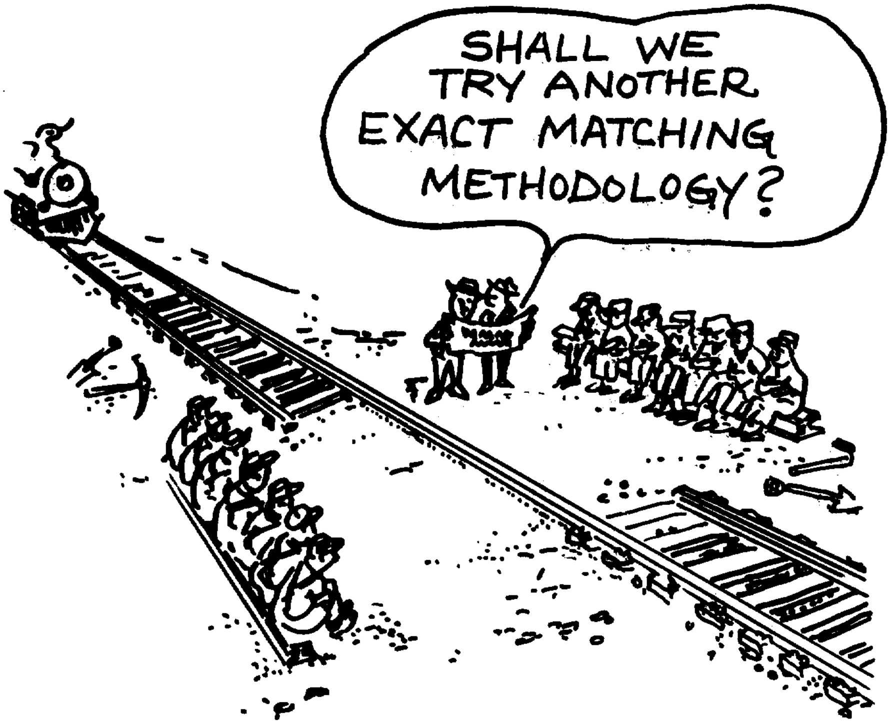
WHITE, D.J. (1969), Dynamic Programming. San Francisco: Holden-Day.

WOLFF, D. and PARSONS, M.L. (1983), Pattern Recognition Approach to Data Interpretation. New York: Plenum Press.

WOODWARD, W.A., PARR, W.C., SCHUCANY, W.R., and LINDSEY, H. (1984), "A Comparison of Minimum Distance and Maximum Likelihood Estimation of Mixture Proportions," Journal of the American Statistical Association, 79, 590-598.

Appendix B: Workshop Particulars —

Workshop Program
List of Attendees
List of Sponsors



THE WASHINGTON STATISTICAL SOCIETY

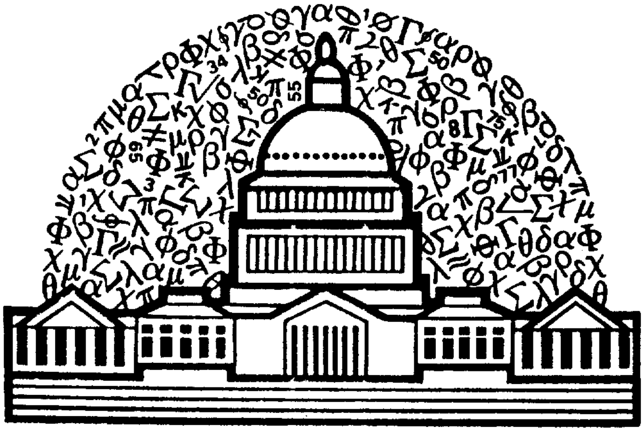
AND THE

FEDERAL COMMITTEE ON STATISTICAL METHODOLOGY

Workshop on Exact Matching Methodologies

MAY 9-10, 1985

ROSSLYN WESTPARK HOTEL
1900 N. FORT MYER DRIVE
ARLINGTON, VIRGINIA 22209



AGENDA

WORKSHOP ON EXACT MATCHING METHODOLOGIES

May 9-10, 1985

Sponsored by
The Washington Statistical Society
and the
Federal Committee on Statistical Methodology

THURSDAY, MAY 9, 1985

8:30 a.m. - 9:30 a.m. Promenade, 2nd Floor
REGISTRATION/COFFEE

9:30 a.m. - 11:30 a.m. Rosslyn A, 2nd Floor
OVERVIEW OF APPLICATIONS AND INTRODUCTION TO THEORY
Chair: FRITZ SCHEUREN, Internal Revenue Service

(9:45) "Tutorial on the Fellegi-Sunter Model for Record Linkage," IVAN FELLEGI, Statistics Canada

(10:15) "Why Are Epidemiologists Interested in Matching Algorithms?" GILBERT W. BEEBE, National Cancer Institute

(10:45) "Exact Matching: The Products of Record Linkage -- Productive, Misleading, and Otherwise," ROBERT F. BORUCH, Northwestern University, and ERNST STROMSDORFER, Washington State University

(11:15) Floor Discussion

11:30 a.m. - 1:00 p.m. Westpark Cafe, 1st Floor
LUNCHEON (Prepaid Event)

1:00 p.m. - 3:00 p.m. Rosslyn A, 2nd Floor
CURRENT THEORY AND PRACTICE
Chair: THOMAS B. JABINE, Consultant, Committee on National Statistics

(1:10) "Multiple Linkage and Measures of Inexactness: Methodology Issues," WRAY SMITH, Mathematica Policy Research, and FRITZ SCHEUREN, Internal Revenue Service

(1:25) "An Information Theoretic Approach to Weights in Computer Matching," NANCY KIRKENDALL, Energy Information Administration

(1:40) "Advances in Record Linkage Methodology: A Method for Determining the Best Blocking Strategy," PATRICK KELLEY, Bureau of the Census

(1:55) "Preprocessing of Lists and Substring Comparison," WILLIAM E. WINKLER, Energy Information Administration

(2:10) Discussant: BENJAMIN TEPPING, Westat, Inc.

(2:25) Discussant: ELI MARKS, Consultant

(2:40) Floor Discussion

3:00 p.m. - 3:30 p.m. Promenade, 2nd Floor
COFFEE BREAK

3:30 p.m. - 5:00 p.m. Rosslyn A, 2nd Floor
APPLICATION CASE STUDIES I
Chair: MARIA ELENA GONZALEZ, Federal Committee on Statistical Methodology

(3:35) "The National Death Index Experience: 1982-1985," JOHN E. PATTERSON and ROBERT BILGRAD, National Center for Health Statistics

(3:50) "An Implementation of a Two-Population Fellegi-Sunter Probability Linkage Model," MAX G. ARELLANO, University of California, San Francisco

(4:05) "Deriving Labor Turnover Rates from Administrative Records," MALCOLM S. COHEN, University of Michigan

(4:20) Discussant: NORMAN JOHNSON, U.S. Bureau of the Census

(4:35) Floor Discussion

5:00 p.m. - 7:00 p.m. Dogwood, 2nd Floor
INFORMAL RECEPTION (With Cash Bar)

Members of the

FEDERAL COMMITTEE ON STATISTICAL METHODOLOGY

(May 1985)

Maria Elena Gonzalez (Chair)
Office of Management and Budget

Barbara Bailar
Bureau of the Census

Norman Beller
National Center for Education Statistics

Yvonne Bishop
Energy Information Administration

Edwin Coleman
Bureau of Economic Analysis

John Cremeans
Bureau of Industrial Economics

Zahava Doering
Defense Manpower Data Center

Daniel Garnick
Bureau of Economic Analysis

C. Terrence Ireland
Department of Defense

Charles Jones
Bureau of the Census

Daniel Kasprzyk
Bureau of the Census

William Kibler
Department of Agriculture

David Pierce
Federal Reserve Bank

Thomas Plewes
Bureau of Labor Statistics

Jane Ross
Social Security Administration

Fritz Scheuren
Internal Revenue Service

Monroe Sirken
National Center for Health Statistics

Thomas Staples
Social Security Administration

Robert Tortora
Department of Agriculture

WORKSHOP ON EXACT MATCHING METHODOLOGIES

FRIDAY, MAY 10, 1985

8:30 a.m. - 9:00 a.m. — Promenade, 2nd Floor
COFFEE

9:00 a.m. - 10:30 a.m. — Rosslyn A, 2nd Floor
APPLICATION CASE STUDIES II
Chair: DANIEL H. GARNICK, Bureau of Economic Analysis

(9:05) "The 1979 Partnership and Sole Proprietorship Employment and Payroll Link Studies," NICK GREENIA, Internal Revenue Service

(9:20) "Creating the Small Business Administration's Master Establishment List," DAVE HIRSCHBERG, Small Business Administration

(9:35) "Enhancing Data from the Survey of Income and Program Participation with Economic Data," DOUGLAS K. SATER, Bureau of the Census

(9:50) Discussant: JOSEPH STEINBERG, Survey Design, Inc.

(10:05) Floor Discussion

10:30 a.m. - 10:45 a.m. — Promenade, 2nd Floor
COFFEE BREAK

10:45 a.m. - 12:15 p.m. — Rosslyn A, 2nd Floor
COMPUTER SOFTWARE AVAILABLE
Chair: HERB MILLER, Social and Scientific Systems, Inc.

(10:50) "Project LINK-LINK: An Interactive Database of Administrative Record Linkage Studies," JANE L. CRANE, National Center for Education Statistics, and DOUGLAS G. KLEWENO, Department of Agriculture

(11:05) "Design and Implementation of a Generalized Record Linkage System," MATTHEW JARO, U.S. Bureau of the Census

(11:20) "Recordkeeping and Data Preparation Practices to Facilitate Record Linkage," MARTHA E. SMITH, Statistics Canada

(11:35) Discussant: TED HILL, Statistics Canada

(11:50) Floor Discussion

WORKSHOP ON EXACT MATCHING METHODOLOGIES

FRIDAY, MAY 10, 1985 -- continued

12:15 p.m. - 1:30 p.m.
LUNCH BREAK

1:30 p.m. - 4:30 p.m.
COMPUTER SOFTWARE WORKSHOP (Three Continuous Concurrent Sessions)

(1:30) (2:30) (3:30)	JANE L. CRANE, National Center for Education Statistics, will demonstrate LINK-LINK.	Rosslyn A, 2nd Floor
(1:30) (2:30) (3:30)	MATTHEW JARO, Bureau of the Census, will demonstrate the Census Bureau's Matching System.	Rosslyn A, 2nd Floor
(1:30) (2:30) (3:30)	TED HILL, Statistics Canada, will present a seminar on Statistics Canada's Generalized Iterative Record Linkage System (GIRLS).	Shenandoah A, 2nd Floor

4:30 p.m.
ADJOURN

LIST OF ATTENDEES
WORKSHOP ON EXACT MATCHING METHODOLOGIES

MAY 9-10, 1985

James Alder
Bureau of the Census

Lois Alexander
Social Security Administration

Bernard Altschuler
Bureau of Labor Statistics

Wendy L. Alvey
Internal Revenue Service

Keith Amick
Bureau of the Census

John C. Angle
Internal Revenue Service

Jonathan G. Ankers
Bureau of the Census

Sam Arcangeli
State of Florida

Max Arellano
University of California

Catherine Armington
Applied Systems Institute

Barbara Bailar
Bureau of the Census

Erma Barron
Social Security Administration

Gilbert Beebe
National Cancer Institute

Richard Bell
Social Security Administration

William Bennett
Bureau of Labor Statistics

Mary Bentz
Internal Revenue Service

Paul P. Biemer
Bureau of the Census

Robert Bilgrad
National Center for Health Statistics

Robert F. Boruch
Northwestern University

Chet Bowie
Bureau of the Census

James E. Bozik
Bureau of the Census

Bertie Brame
Internal Revenue Service

Ralph Bristol
Department of the Treasury

Warren L. Buckler
Social Security Administration

Paul Burke
Department of Housing and Urban Development

William P. Butz
Bureau of the Census

Charles Byce
The Brookings Institution

Helen Choi
Internal Revenue Service

Malcolm S. Cohen
University of Michigan

Sherry Courtland
Bureau of the Census

Charles Cowan
Bureau of the Census

Lawrence H. Cox
Bureau of the Census

Pat Crabbe
Internal Revenue Service

Jane Crane
National Center for Education Statistics

Lester Curtin
National Center for Health Statistics

John Czajka
Mathematica Policy Research, Inc.

Robert Dalrymple
Office of Analysis and Evaluation

Ramesh Dandekar
Energy Information Administration

Charles Day
Internal Revenue Service

Linda DelBene
Social Security Administration

Charles Eastlack
Westat, Inc.

Alan Eck
Bureau of Labor Statistics

Marlene Einstein
Bureau of Labor Statistics

Frank Elsen
Energy Information Administration

Ivan P. Fellegi
Statistics Canada

Wayne Finegar
Social Security Administration

Barbara Garner
Bureau of the Census

Daniel H. Garnick
Bureau of Economic Analysis

Tommy W. Gaulden
Bureau of the Census

William Gerber
Internal Revenue Service

Claudia Glover
Bureau of the Census

Maria E. Gonzalez
Office of Management and Budget

Mildred Gray
University of the District of Columbia

Wayne B. Gray
National Bureau of Economic Research

Brian Greenberg
Bureau of the Census

Nick Greenia
Internal Revenue Service

Richard Gress
University of Utah

Lisa Gross
Internal Revenue Service

Jim Harte
Internal Revenue Service

Beth Hill
Bureau of the Census

Ted Hill
Statistics Canada

David Hirschberg
Small Business Administration

K. P. Ho
Atomic Energy Control Board

Paul Holland
Educational Testing Service

Geoffrey Howe
University of Toronto

Terry Ireland
Department of Defense

Richard Irwin
Bureau of the Census

Thomas Jabine, Consultant
Committee on National Statistics

Kathleen Jablonski
Capital Systems Group

Matthew Jaro
Bureau of the Census

Robert Jewett
Bureau of the Census

Norman Johnson
Bureau of the Census

David Judkins
Bureau of the Census

Kenneth Kaplan
Bureau of the Census

Myron Katzoff
Bureau of the Census

Patrick Kelley
Bureau of the Census

Berdj Kenadjian
Internal Revenue Service

Bertram M. Kestenbaum
Social Security Administration

Beth Kilss
Internal Revenue Service

Nancy Kirkendall
Energy Information Administration

Douglas Kleweno
Department of Agriculture

Lynn Kuo
Department of Agriculture

Enrique Lamas
Bureau of the Census

William LaPlant
Bureau of the Census

Terry Lotz
University of Utah

Dr. Joseph L. Lyon
University of Utah

Eli S. Marks
Consultant

Reggie D. Masano
Bureau of the Census

Carlyle Maw
National Center for Education Statistics

Jim McBride
Response Analysis Corporation

Philip McClain
Centers for Disease Control

Nelson McClung
Department of Treasury

Robert J. McIntire
Bureau of Labor Statistics

Michael L. Mersch
Bureau of the Census

M. P. Mi
University of Hawaii

Herbert J. Miller
Social & Scientific Systems, Inc.

Deborah Moore
Bureau of the Census

Jeffrey Moore
Bureau of the Census

Ann J. Nakamura
Revenue Canada Taxation

Dwaine Nelson
Department of Agriculture

Elizabeth Nelson
Internal Revenue Service

Karen V. O'Conor
Bureau of Labor Statistics

H. Lock Oh
Internal Revenue Service

Bill Page
National Academy of Sciences

David P. Paris
Internal Revenue Service

John Patterson
National Center for Health Statistics

John D. Pearson
Energy Information Administration

Thomas Petska
Internal Revenue Service

Henry Power
Department of Agriculture

D. Dean Prochaska
Bureau of the Census

Judith N. Rayner
Westat, Inc.

Eugene Rogot
National Heart, Lung, and Blood Institute

Steve Ropel
Department of Agriculture

Jeffrey Roth
National Research Council

Peter Sailer
Internal Revenue Service

Jean Salter
Bureau of Economic Analysis

Douglas K. Sater
Bureau of the Census

Fritz Scheuren
Internal Revenue Service

Mary Ellen Schiller
Social Security Administration

Jack Schmulowitz
Social Security Administration

Marvin L. Schwartz
Internal Revenue Service

Sidney Schwartz
Bureau of the Census

Daniel F. Skelly
Internal Revenue Service

Martha Smith
Statistics Canada

Wray Smith
Mathematica Policy Research

Joseph Steinberg
Survey Design, Inc.

B. J. Stone
National Cancer Institute

Tim H. Tang
Pacific Gas & Electric Company

Benjamin Tepping
Westat, Inc.

Kathryn Thomas
Bureau of the Census

John Thompson
Bureau of the Census

Wendel L. Thompson
Energy Information Administration

James Tonascia
Johns Hopkins School of Hygiene & Public Health

Barbara Tyler
Social Security Administration

Carol M. Utter
Bureau of Labor Statistics

Richard Wehrly
Social Security Administration

Oliver Wilson
Internal Revenue Service

Robert Wilson
Internal Revenue Service

Barry R. Windheim
Internal Revenue Service

William E. Winkler
Energy Information Administration

Voting Members of the Board of Directors

WASHINGTON STATISTICAL SOCIETY (1984-1985)

President

C. Terrence Ireland
Department of Defense

President Elect

Mary Grace Kovar
National Center for Health Statistics

Past President

Maria Elena Gonzalez
Office of Management and Budget

Secretary

Rich Allen
Department of Agriculture

Treasurer

Bennie Clemmer
Social Security Administration

Representatives-at-Large

Daniel Kasprzyk
Bureau of the Census

Nancy Kirkendall
Energy Information Administration

Susan Miskura
Bureau of the Census

Richard Taeuber
National Center for Education Statistics

Methodology Section Chair

Michael Cohen
National Academy of Sciences

Representative of the ASA Board

Jerry Coffey
Office of Management and Budget

Representative to Council on Chapters

Don King
King Research, Inc.