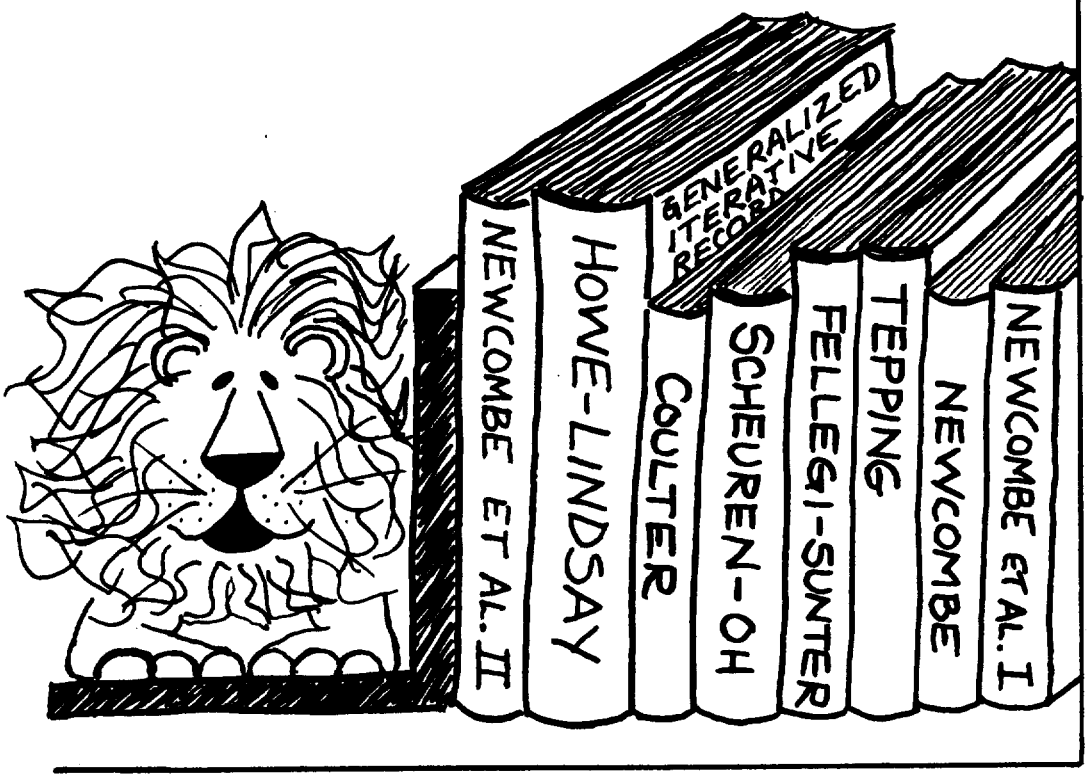
Section I: Selected Background Papers (1959 – 1983)



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. 36*c*).

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.

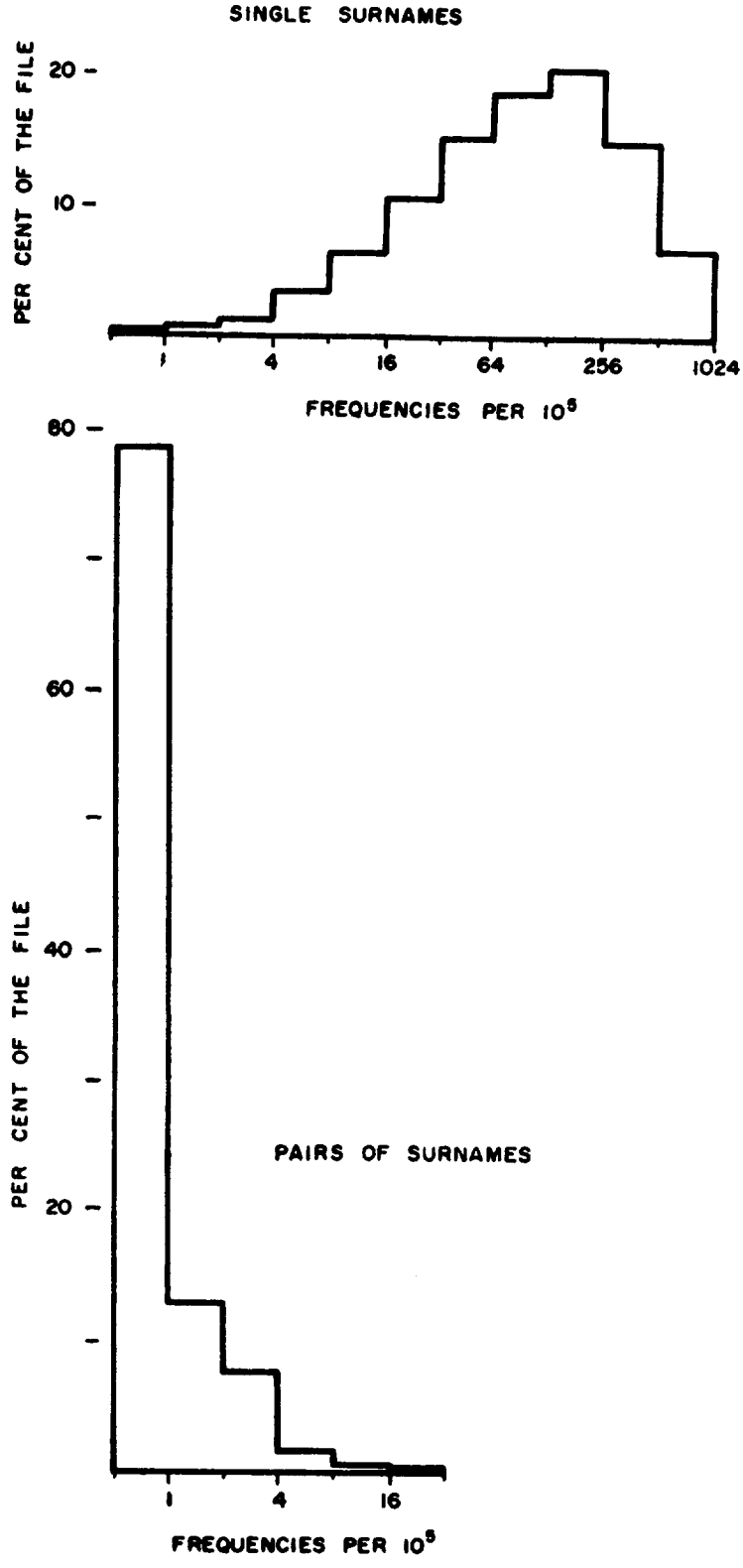


Fig. 1. (Top) Frequency distribution of brides' maiden names, in Soundex coded form, from records of 114,471 marriages in British Columbia for 1946–55. (Bottom) Frequency distribution of family-name pairs for married couples, in Soundex coded form, from the same records. Two East Indian names, of which 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.

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:

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

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

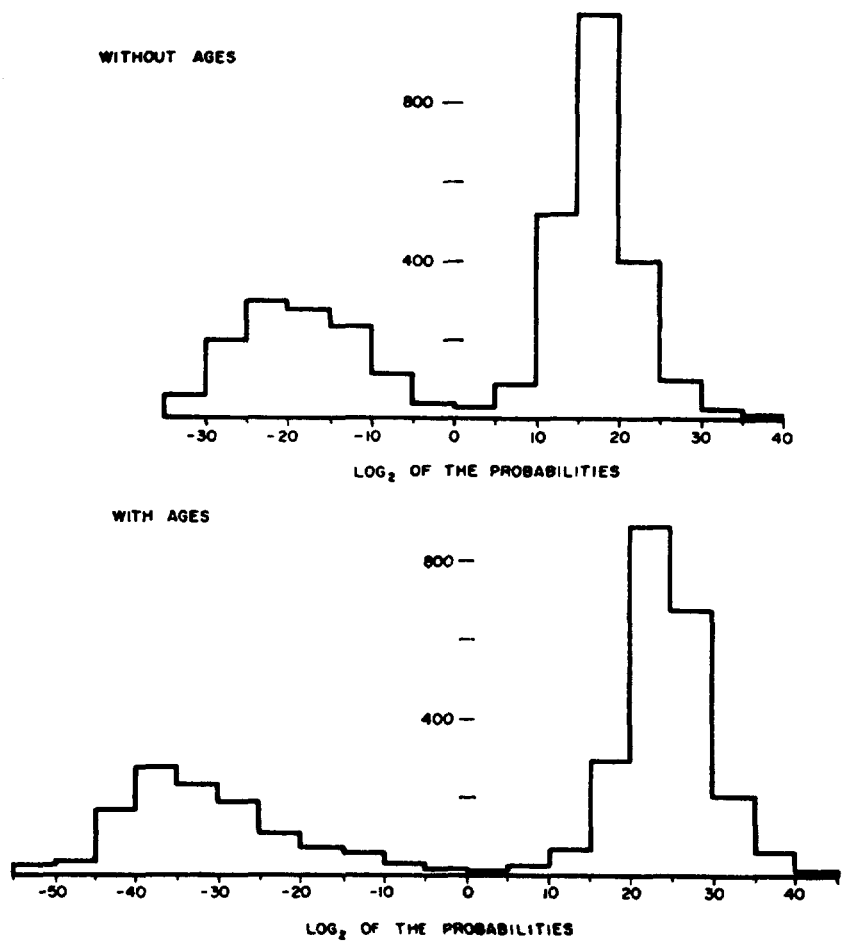


Fig. 2. (Top) Frequency distribution of the probabilities (in binits) obtained on comparing birth and marriage records having identical Soundex code pairs (calculated without using ages), based on records contained in the first fifth of the birth and marriage files (husband's surname beginning with *A*, *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.

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 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.

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 statistics," *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

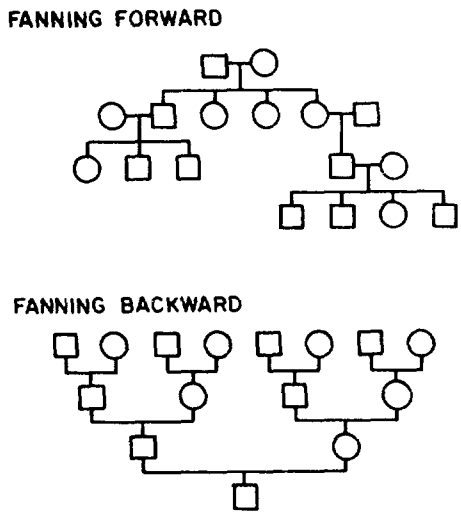


FIG. 1. Conventional pedigree charts. Note the difficulty of representing in a single chart the ancestors, descendants, cousins, and in-laws.

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-

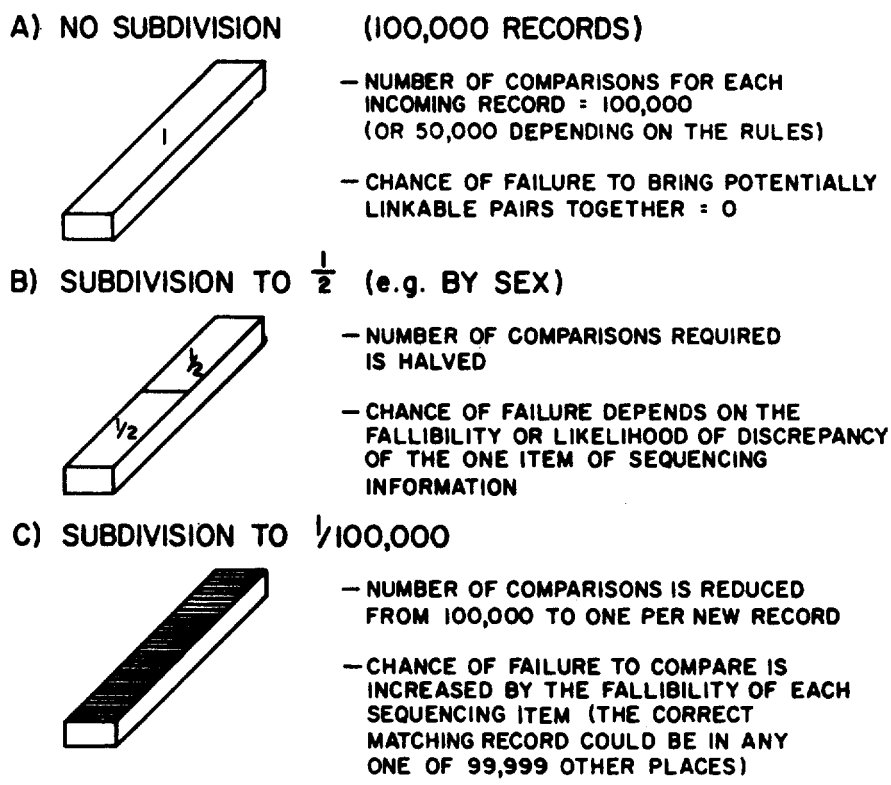


FIG. 2. Optimizing a single sequence search. Subdivision must be based on items of identifying information with the highest efficiency ratios and must be adjusted to an acceptable low level of losses or of wasted comparisons.

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.

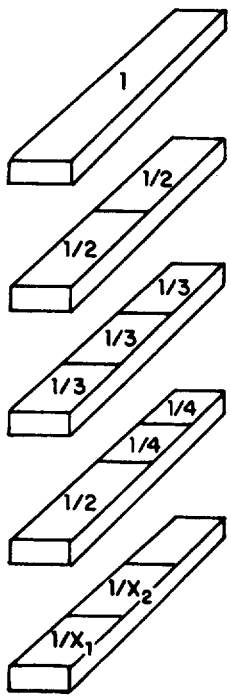
Unlike the coefficient of specificity, which gets smaller as a file becomes more finely divided, the *discriminating power* increases with the extent of the subdivision. Furthermore, it is usually regarded as an "addable" quantity. Thus, the discriminating power may be taken as the logarithm of the inverse 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_s 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

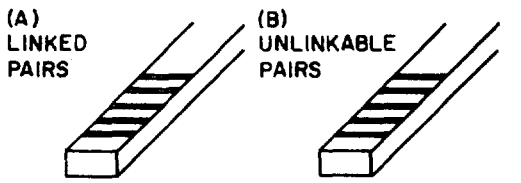
$$W_t = \log_2 (A/B) \qquad (4)$$

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)



("BINIT WEIGHTS" = $\log_2 A/B$)

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 "superfamily," 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 code	Birth order	Family	Child	Disease code
Event records for a sibship (one per event)				
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.			

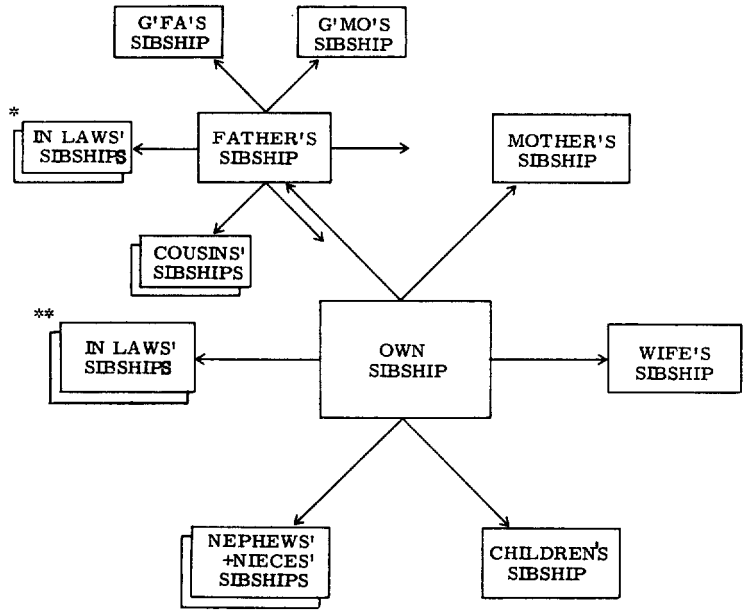
TABLE 12. EXAMPLE OF A TABULATION FROM FAMILY SUMMARY RECORDS

	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 are 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 (C, G, J, K, Q, S, X, Z)	coded 2

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 = A A – A K
01 = A L
02 = A M – A R
03 = A S – A Z
04 = B A A – B A J
05 = B A K – B A Q
06 = B A R
(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 $P[M \mid \gamma]$. (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

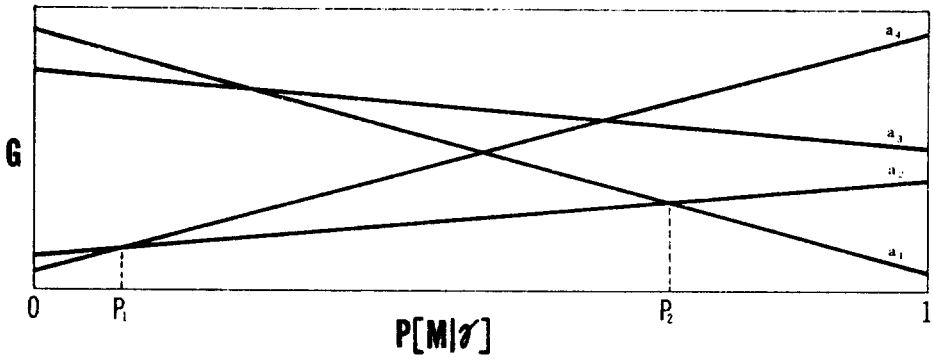


FIG. 1.

$P[M|\gamma]$ for each a_i. In this illustration, the optimum linkage rule specifies:

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

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 $P[M \mid \gamma]$ 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 $P[M \mid \gamma]$ 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 $P[M \mid \gamma]$ 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

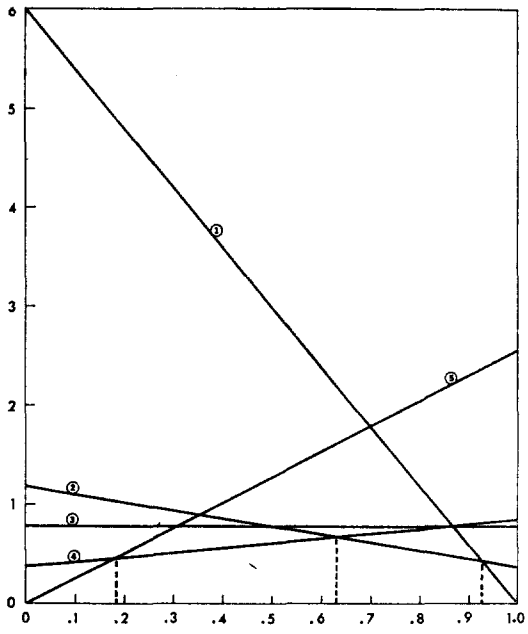


FIG. 2. Cost function for each of five actions, and the optimum action intervals.

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

$$\mu = \sum_{\gamma \in \Gamma} u(\gamma) P(A_1 | \gamma)$$

and

$$\lambda = \sum_{\gamma \in \Gamma} m(\gamma) P(A_3 | \gamma)$$

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 $\gamma \in \Gamma$. 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

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

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

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

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

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

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,

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

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

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

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 completely 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

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

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

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 $\gamma^1, \gamma^2, \cdots, \gamma^K$ 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

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

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*, $\gamma_1^k, \gamma_2^k, \cdots, \gamma_{n_k}^k$. 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 $n_1 \cdot n_2 \cdot \cdots \cdot n_K$. 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

$$m_h = \sum_{\gamma \in S_h} m(\gamma) \tag{35}$$

$$u_h = \sum_{\gamma \in S_h} u(\gamma). \tag{36}$$

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.

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

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

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

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

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

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

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

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

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

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

Note that $P(A_2 \mid M)$ and $P(A_2 \mid U)$ 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,

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

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

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

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

$$P(T_\lambda < w(\gamma) < T_\mu) < P(T'_\lambda < w'(\gamma') < T'_\mu) \tag{47}$$

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)

$$P(M) \cdot \sum_{T_\lambda < w < T_\mu} m(w) + P(U) \cdot \sum_{T_\lambda < w < T_\mu} u(w)$$
$$< P(M) \cdot \sum_{T_\lambda < w < T_\mu} m(w') + P(U) \cdot \sum_{T_\lambda < w' < T_\mu} u(w') \tag{48}$$

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, $n_1 n_2 \cdots n_K$. 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 $|w_j^k|$.
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

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

while

$$\begin{aligned} &P'(w(\gamma) < w(\gamma_h) \mid \gamma \in U) \qquad \text{is estimated by} \\ &\mu_h = \sum_{h'=1}^{h} u(\gamma_{h'})/\pi(\gamma_{h'}). \end{aligned} \tag{53}$$

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

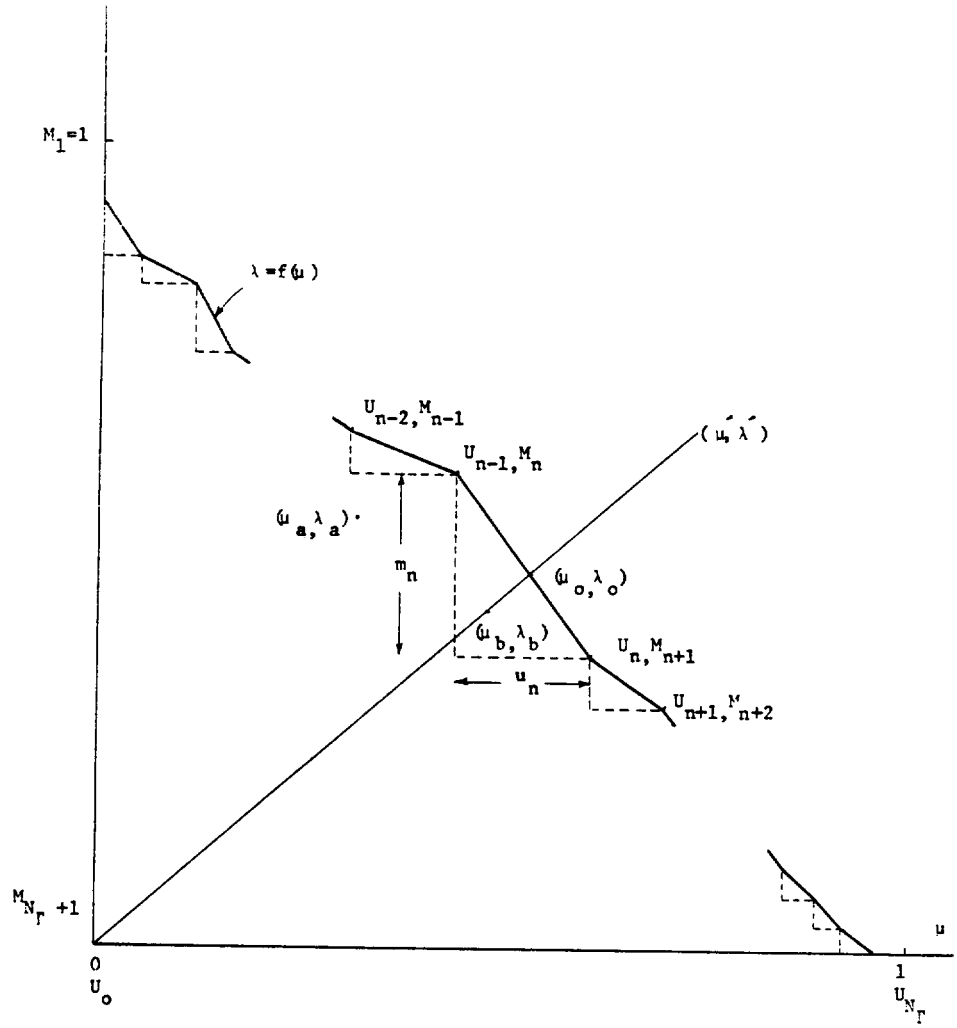


FIG. 1

$$0 < \lambda \leq f(\mu)$$
$$\text{and} \quad 0 < \mu. \tag{5}$$

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

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

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

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

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

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

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

$$P_{i1} = 1 \qquad \text{for } i = 1, 2, \cdots, r-1 \tag{23}$$
$$P_{i2} = 1 \qquad \text{for } i = s+1, s+2, \cdots, N_\Gamma \tag{24}$$

that is, whenever u_i is zero then $P_{i1}=1$ and whenever $m_i=0$ then $P_{i2}=1$.

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

$$\sum_{i=1}^{N\Gamma} m_i P'_{i2} = \sum_{i=1}^{N\Gamma} m_i P_{i2} = \lambda. \tag{30}$$

From (29) and (30) we get

$$\sum_{i=1}^{N\Gamma} m_i(P'_{i1} + P'_{i2}) \leqq \sum_{i=1}^{N\Gamma} (P_{i1} + P_{i2})$$

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

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

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

$$P(A_2 \mid L_0) = 0.$$

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

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

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

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

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

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

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

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

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

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

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

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:

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

The positive root of this equation is

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

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

$$m(\gamma_i^1) = \frac{M_1(\gamma_i^1) - u_3 U_2(\gamma_i^1)}{m_2(m_3 - u_3)}(X^2 + 1) \tag{18}$$

$$m(\gamma_i^2) = \frac{M_2(\gamma_i^2) - u_3 U_1(\gamma_i^2)}{m_1(m_3 - u_3)}(X^2 + 1) \tag{19}$$

$$m(\gamma_i^3) = \frac{M_3(\gamma_i^3) - u_2 U_1(\gamma_i^3)}{m_1(m_2 - u_2)}(X^2 + 1) \tag{20}$$

$$u(\gamma_i^1) = \frac{m_3 U_2(\gamma_i^1) - M_1(\gamma_i^1)}{u_2(m_3 - u_3)} \frac{X^2 + 1}{X^2} \tag{21}$$

$$u(\gamma_i^2) = \frac{m_3 U_1(\gamma_i^2) - M_2(\gamma_i^2)}{u_1(m_3 - u_3)} \frac{X^2 + 1}{X^2} \tag{22}$$

$$u(\gamma_i^3) = \frac{m_2 U_1(\gamma_i^3) - M_2(\gamma_i^3)}{u_1(m_2 - u_2)} \frac{X^2 + 1}{X^2} \tag{23}$$

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

TRUEMATCH -- 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.

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

TRUE NONMATCHES -- 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.

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

Mismatches.--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.)

Truematches.-- 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

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

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/

Alternate Match Rules.--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. "Perfect" Agreement Rule.--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.

Level Comparisons.--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.

Relationship Comparisons.--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.

Reweighting for Nonmatches.--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 $\{P^{MM}_{ij}\}$, is given by

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

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^{TM}_{ij}, times the fraction of the total sample that were truematches, denoted by $(1 - \alpha)$. The second term consists of the mismatch proportion P^{MM}_{ij} 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

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

From (4) this becomes

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

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 $\{P^{TM}_{ij}\}$ by substituting $\hat{\alpha}$, $\hat{P}_{i.}$, and $\hat{P}_{.j}$ in (6). The $\{\hat{P}^{TM}_{ij}\}$ 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 $a_i \in A$, 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 $[\alpha(a_i),\ \alpha(a_j)]$ 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. $m(\gamma^k) = P\{\gamma^k[\alpha(a_i), \alpha(a_j)];\ (a_i, a_j) \in M\}$

2. $u(\gamma^k) = P\{\gamma^k[\alpha(a_i), \alpha(a_j)];\ (a_i, a_j) \in U\}$

A component weight for each γ^k is defined by:

$$w(\gamma^k) = \log_{10}[m(\gamma^k) / u(\gamma^k)].$$

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;

$\sum_{j}^{n} f_{x_j} = N$

$f_{y_1}, f_{y_2}, \ldots, f_{y_n}$ denote the frequencies of the surname realizations;

$\sum_{k}^{n'} f_{y_k} = N, \sum_{d}^{n''} f_{y_d} = N'$

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)

$$m[\gamma_1(x_j)] = (f_{x_j}/N)(1 - e)^2(1 - e_T)$$

$$u[\gamma_1(x_j)] = (f_{x_j}/N)^2$$

$$m[\gamma_2(y_d)] = (f_{y_d}/N')\left[2g_1 e(1 - e)(1 - e_T) + g_1^2 e^2(1 - e_T) + g_2(1 - e)^2 \cdot e_T + 2g_1 g_2 e(1 - e) e_T + g_1^2 g_2 e^2 e_T\right]$$

$u[\gamma_2(y_d)]$ = u(agree on sn code) · u(disagree on sn | agree on sn code)

= u(agree on sn code) · [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

$$m(\gamma_3) = 2(1 - g_1) e(1 - e)(1 - e_T) + (1 - g_1^2) e^2(1 - e_T) + (1 - g_2)(1 - e)^2 e_T + 2(1 - g_1 g_2)e(1 - e)e_T + (1 - g_1^2 g_2) e^2 e_T$$

$$u(\gamma_3) = 1 - \sum_{k=1}^{n'} (f_{y_k}/N)^2$$

$$\text{weight} = w(\gamma_h) = \log_{10}\left[m(\gamma_h)/u(\gamma_h)\right], \quad h = 1,2,3$$

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:

$$m(\gamma_h)' = m(\gamma_h \mid \gamma_3 \text{ does not occur}) = m(\gamma_h) / \left[1 - m(\gamma_3)\right]$$

$$u(\gamma_h)' = u(\gamma_h \mid \gamma_3 \text{ does not occur}) = u(\gamma_h) / \left[1 - u(\gamma_3)\right].$$

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.

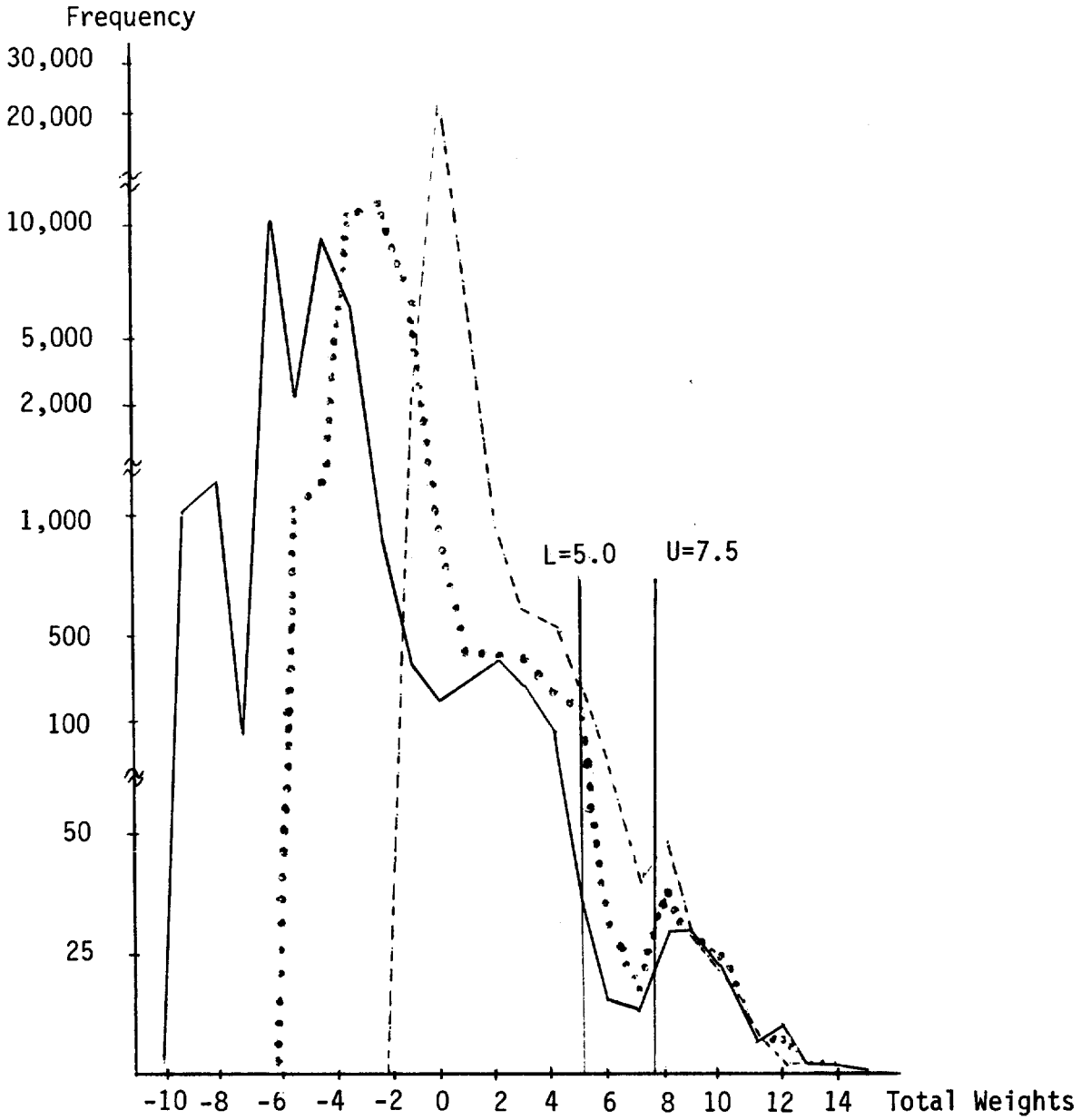
The final parameters to be supplied are the threshold values. It is these two values which ultimately determine the classification of each pair. Fellegi and Sunter suggest a technique of estimating these by sampling from the tails of the m and u probability distributions for the comparison pairs. In 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

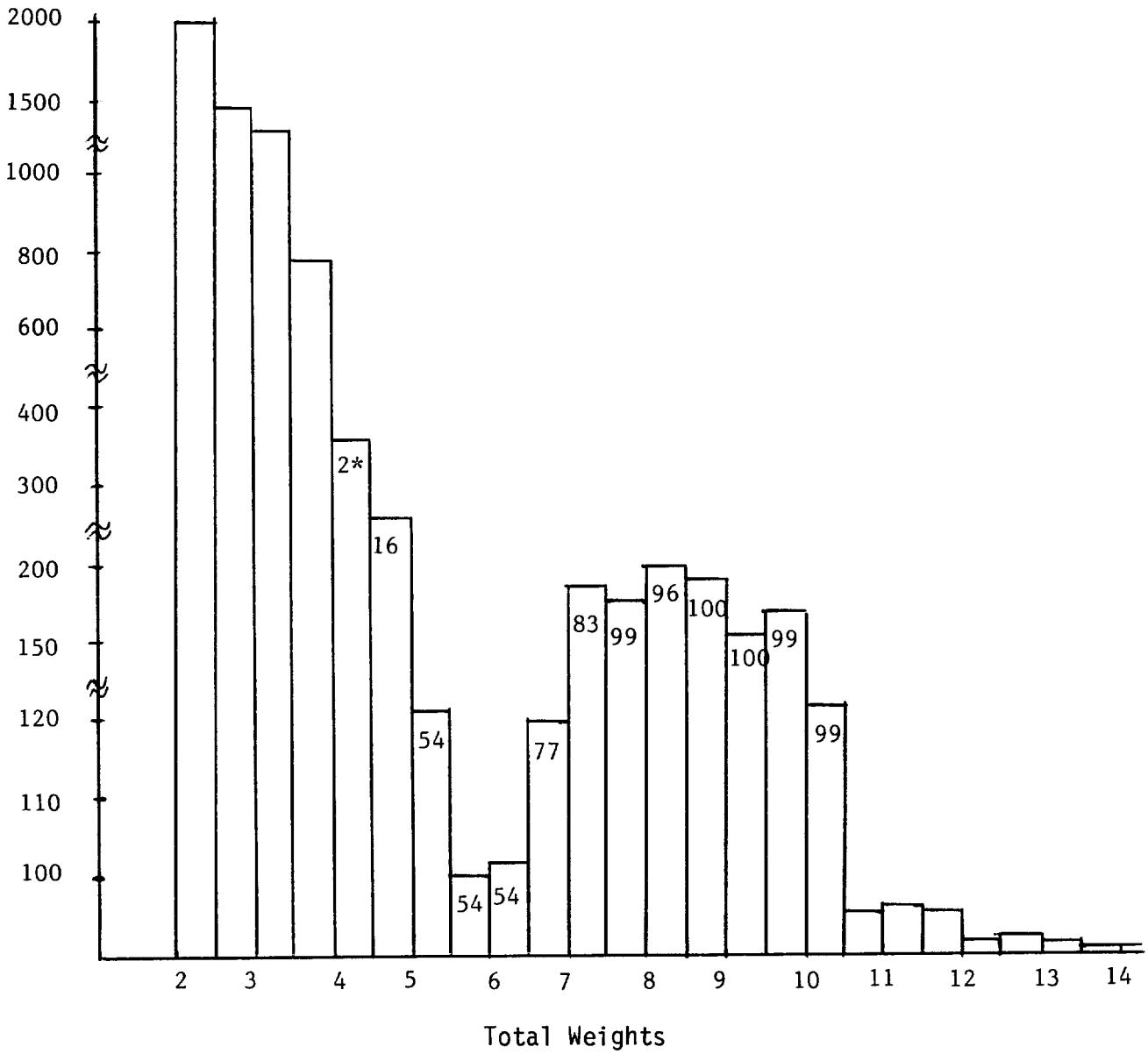
(Approximately 39,000 comparisons)



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 0T6, 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:

$$ w_{1=i} = \log_2 t_1^2 + \log_2 \frac{1}{{}_B f_i}, \qquad [4] $$

$$ w_{2=j} = \log_2 (t_2^2 - t_1^2) + \log_2 \left[\frac{{}_A g_j}{{}_A g_j \; {}_B g_j - \sum_{i \in j} {}_A f_i \; {}_B f_i} \right], \qquad [5] $$

$$ w_{3=k} = \log_2 (t_3^2 - t_2^2) + \log_2 \left[\frac{{}_A h_k}{{}_A h_k \; {}_B h_k - \sum_{j \in k} {}_A g_j \; {}_B g_j} \right], \qquad [6] $$

$$ w_4 = \log_2 (O), \qquad [7] $$

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

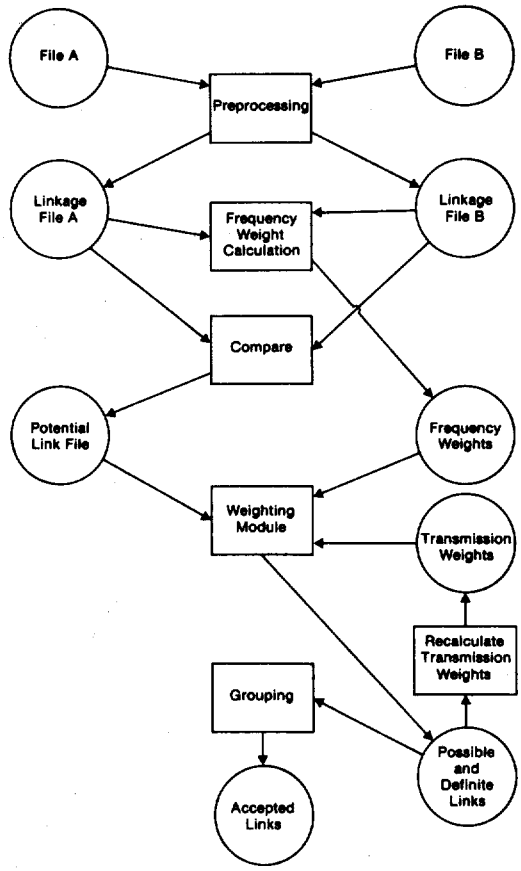


FIG. 1. Generalized iterative record linkage system.

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	
B+	very good possible	35	219
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, one 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 50:50 or 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		1057		–	
+4 to +9	(11:1 to 724:1)	433	2023	424	1845	9	177
+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	1:20	−4.3
10–11	1	10		
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.