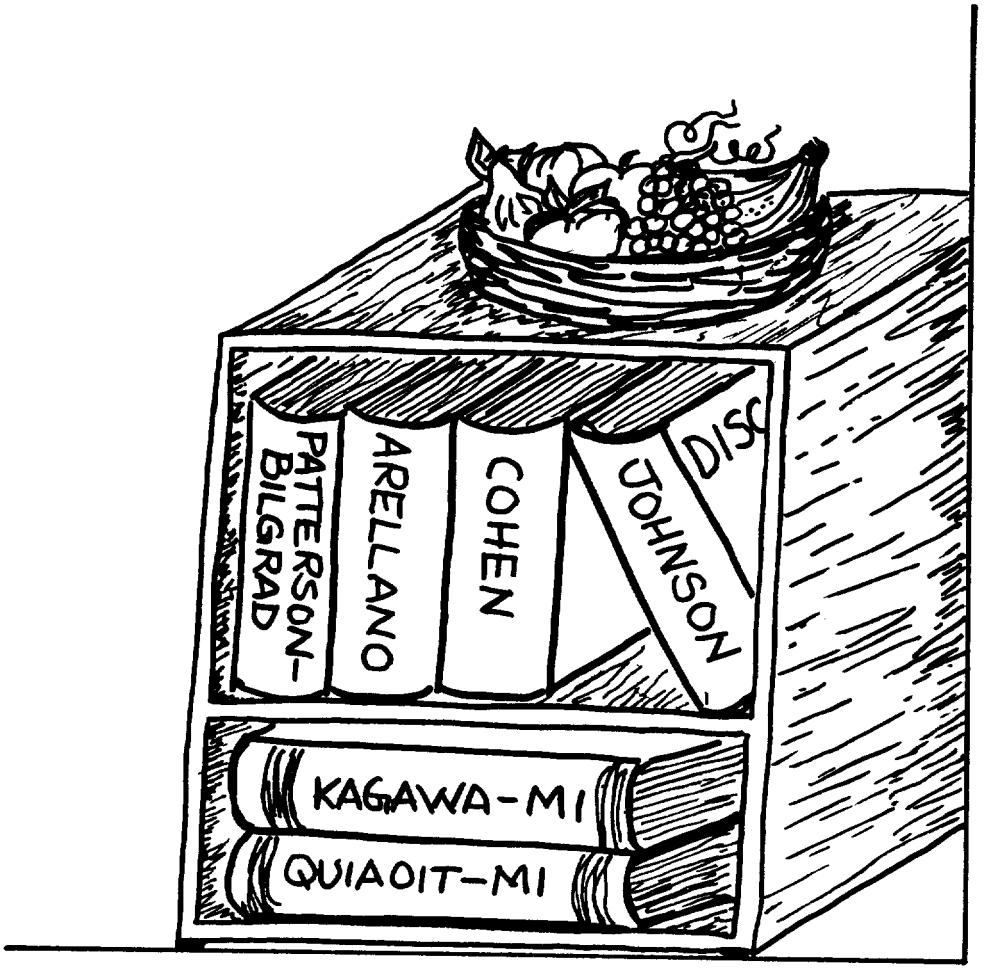
Section IV:
Application
Case Studies I



THE NATIONAL DEATH INDEX EXPERIENCE; 1981-1985

John E. Patterson and Robert Bilgrad, National Center for Health Statistics

The National Death Index (NDI) is a central, computerized index to the death certificates filed in each State vital statistics office. This computer file contains a standard set of identifying information for each person dying in the U.S., beginning with 1979. The NDI was established to assist health and medical investigators in determining whether persons in their studies may have died, and if so, to provide the names of the States in which those deaths occurred, the dates of death, and the corresponding death certificate numbers. The NDI user can then obtain copies of death certificates from the appropriate State offices.

The NDI became operational in November 1981. As of March 31, 1985, the NDI file contained 10.3 million death records for the five-year period 1979-1983. A total of 168 NDI file searches have been performed, involving 2,352,001 records submitted by 99 NDI users. This report gives a brief overview of the NDI users and their research activities, and describes recent evaluations and planned revisions of the NDI matching criteria. Procedures for using the NDI are also presented.

1. OVERVIEW OF NDI USERS

The NDI has been used in a variety of health and medical research projects which rely on the successful ascertainment of the vital status of their study subjects. The research projects of the 99 NDI users have been grouped into five broad research categories in Table 1. These categories are (1) exposure cohorts, involving studies of the effects of being exposed to potential risk factors in the workplace, the environment, or as a result of diagnostic or therapeutic procedures; (2) disease cohorts, involving followup of persons diagnosed as having cancer or other diseases; (3) life style/risk factors, involving studies of the effects of activities such as smoking or drug abuse; (4) clinical trials, primarily involving studies of the potentially beneficial effects of various therapies for specific diseases; and (5) general population cohorts, involving followup of survey participants not selected on the basis of a specific diagnosis or exposure to risk factors.

Forty percent of the NDI users are conducting occupational studies involving followup of rosters of employees to determine whether there have been any harmful effects resulting from their exposures to potentially harmful substances. Most of these studies are being performed by the National Institute for Occupational Safety and Health as well as by oil and chemical companies. Another 28 percent of the NDI users are involved in followup activities on cohorts of persons diagnosed as having cancer or other diseases.

Table 1 also shows the types of organizations using the NDI. It should be noted that while Federal agencies account for only 18 percent of the NDI users, the Federal government is actually providing the funding support for about three-fourths of the studies being performed by universities and consulting firms.

Many of the NDI users are either following cohorts of under 2,500 persons or use the NDI only to check on those study subjects which are considered lost to followup. Almost three-fourths of the users have submitted fewer than 10,000 names. The fewest records submitted for an NDI file search were 7. The largest volume of records was submitted by the Census Bureau for the National Longitudinal Mortality Study being supported by the National Heart, Lung and Blood Institute. Thus far, this study has involved the submission of a test file of 225,875 Census Bureau records and the main study file of 994,195 records. The study's methodology involves a search of the NDI file every two years. The second NDI search for the main study is scheduled for around July 1985 and will involve approximately 1.2 million Census Bureau records.

2. COMPLETENESS AND QUALITY OF NDI AND USER DATA

The effectiveness of the NDI matching process is dependent on the following three factors: (1) the completeness and quality of the death certificate data submitted to the National Center for Health Statistics (NCHS) by the State vital statistics offices for use in creating the NDI file, (2) the completeness and quality of the data provided by the NDI user, and (3) the effectiveness of the NDI matching criteria.

The completeness of the NDI file is probably well in excess of 99 percent. Data on virtually all deaths occurring from 1979 to 1983 have been submitted by the fifty States, the District of Columbia, New York City, Puerto Rico and the Virgin Islands. The NDI file now contains 10.3 million records. Table 2 shows that the completeness of data for most data items exceeds 97 percent except for middle initial (71.7 percent), father's surname (86.2 percent), and social security number (91.0 percent). Although 9.0 percent of the records do not contain social security numbers (as shown in Table 3), only 6.0 percent of the records for persons 22 years and older do not contain such numbers. As might be expected, death records for females have higher percentages of social security numbers not reported than records for males.

It is very difficult to assess the quality of the data on the NDI file, but we have reason to believe that it is probably quite good. The quality of the NDI data is most affected by how the death record information is reported to and recorded by funeral directors. The death certificate is a legal document which must be filed in the State where the death occurs. Most States continually encourage funeral directors to make every effort to obtain accurate information from the person making the funeral arrangements. Funeral directors have a strong incentive for

obtaining and accurately recording good identifying information on each decedent. Their clients would not be pleased if errors appeared on the certificate, since this would very likely delay settlement of claims for life insurance and other survivor benefits. All States perform 100 percent verification of the coding and keying of their records. NCHS also performs various quality control checks as the States' data are received.

The completeness and quality of data submitted by NDI users, on the other hand, vary greatly depending on how the data were collected. The complete and accurate collection of the NDI data items listed in Table 2 will, of course, enhance the effectiveness of any subsequent searches of the NDI file. This table summarizes the overall completeness of the data submitted by NDI users; however, the completeness of each data item varies greatly among the different users, especially for such items as middle initial, social security number, State of residence and State of birth.

Because of the newness of the NDI program, many users did not or could not insure the collection of all of the NDI data items. NCHS strongly encourages investigators who are or will be planning studies to make every possible effort to collect all of the NDI data items, even if the investigators do not have specific plans to conduct a followup of study subjects to ascertain their vital status. Once a study is completed, the same or other health investigators may decide that future followup of the study group may indeed be very useful. Internally, NCHS has instituted a policy requiring each new survey to collect all of the NDI data items, regardless of whether the survey staff or others in NCHS plan to use the NDI to followup on the survey participants in the future.

3. RECENT REVISIONS IN THE NDI MATCHING CRITERIA

When the NDI retrieval program was first designed and implemented, a fairly simple set of seven matching criteria was developed (1) to use most effectively the principal identifiers on the death record; (2) to satisfy the needs of the majority of potential users; (3) to make searches against the NDI very routine, eliminating the need for special programming for individual users; and (4) to take into account the policy concerns of the States. These concerns were very significant and had a major impact on the development of the initial matching criteria. Many States felt that the NDI users should be required to provide a fairly substantial body of identifying information for their subjects. They should not accept matching solely on the basis of social security numbers, for example. A number of States were also concerned about probabilistic matching. They felt that their regulations would prevent them from searching their files on a probabilistic basis, and they did not believe that they could delegate authority to NCHS to do what they would not be permitted to do themselves.

For an NDI record to qualify as a possible match with a given user record, under the initial matching criteria, at least one of the following seven combinations of data items must agree on both records:

1. Social security number, first name.
2. Social security number, last name.
3. Social security number, father's surname.
4. If the subject is female: social security number, last name (user's record) and father's surname (NDI record).
5. Month and exact year of birth, first and last name.
6. Month and exact year of birth, first name, father's surname.
7. If the subject is female: month and exact year of birth, first name, last name (user's record) and father's surname (NDI record).

Nine evaluations of the effectiveness of the above matching criteria have been performed by NCHS and by several NDI users. The results are summarized in Table 4. Each of these evaluations involved study files of known decedents which were searched against the NDI file. In those evaluations where social security numbers were available for a large proportion of decedents, the resulting percentages of true matches (user records which were correctly identified as deceased) ranged from 92.1 percent to 98.4 percent. The differences in these percentages are attributed primarily to differences in the quality of the users' data sets. Three evaluations showed that, without the benefit of any social security numbers true matches amounted to only 79.7 percent [8], 80.0 percent [10], and 81.9 percent [9], primarily because of discrepancies in year of birth and names. However, two other users apparently had much better data on dates of birth and names because they achieved true matches of 91.1 percent [1] and 96.5 percent [3] without the benefit of social security numbers.

Most of our advisers and users have stressed that our first efforts to improve our matching criteria should be to maximize the number of true matches, even if this means a significant increase in the false matches which may be generated as a by-product. Our users have generally found that nearly all false matches can be eliminated easily by simply reviewing the output of the NDI search. This is especially true for small studies. For very large studies computerized processing of the NDI output is necessary to identify true matches and to isolate questionable matches which deserve closer inspection. Several users have developed their own computerized algorithms for this purpose.

As a result of the evaluations mentioned above, NCHS is planning to add five new matching criteria to the initial seven. The five additional matching criteria are listed below and are numbered 8 through 12 to distinguish them from the initial seven. A possible NDI record match would be generated if any of these combinations of data items agree on an NDI and a user record.

8. Month and $\pm$ 1 year of birth, first and last name.
9. Month and $\pm$ 1 year of birth, first and middle initials, last name.

10. Month and exact year of birth, first and middle initials, last name.
11. Month and day of birth, first and last name.
12. Month and day of birth, first and middle initials, last name.

Our evaluations have shown that by also permitting matches on month and day of birth and on month and ± 1 year of birth, the percentage of true matches generated can be increased significantly. One of the NCHS evaluations mentioned previously, involving a cancer registry file containing social security numbers on 85.9 percent of its 2,598 records, showed an increase in true matches from 92.1 percent to 96.2 percent with the addition of the five new matching criteria [8]. The increase in matching effectiveness is greatest, however, for study files having very few or no social security numbers. Another NCHS evaluation involved a file without social security numbers for 607 decedents in the NCHS National Health and Nutrition Examination Study. This evaluation showed an increase in true matches from 81.9 percent to 89.5 percent [9].

The initial retrieval program permitted first names, last names and fathers' surnames to match on the basis of either their exact spelling or Soundex codes. Evaluations showed that the use of Soundex codes often generated agreements on names which were dissimilar, however, causing a number of unnecessary false positives to be generated, while adding very little to the number of true positives. With the planned implementation of the revised matching criteria, the use of Soundex codes will be eliminated. Phonetic matching will be performed only on last names and fathers' surnames and will be based on NYSIIS codes (New York State Identification and Intelligence System). The NYSIIS coding system which will be used was first modified abd tested by the U.S. Department of Agriculture [11] and was subsequently adopted for use in Statistics Canada's Mortality Data Base. The computer program which assigns the modified NYSIIS codes was obtained by NCHS from Statistics Canada.

4. USING THE NDI

As mentioned above, health investigators planning to use the NDI are encouraged to collect as many of the NDI data items as possible and to insure that the data are of good quality. To become an NDI user, health investigators must first complete and submit an NDI application form. Each form is reviewed by the advisers to the NDI program to insure that (1) the proposed use of the NDI is solely for statistical purposes in medical or health research and (2) the applicant provides adequate assurances that the identifying death record information obtained from the NDI and from the State vital statistics offices will be kept confidential and will be used only for the proposed study.

Once the applicant is notified that the application is approved, the NDI user may then submit records for an NDI file search. The user must submit records on a magnetic tape which conforms with the NCHS tape specifications, file format requirements, and coding instructions. Users planning to submit under 300 records have the option of using NCHS coding sheets. The results of an NDI file search are sent to the user (along with the user's data) within three weeks after the user's records are received by the NCHS computer facility.

The user must assess the quality of each possible NDI record match listed and determine which NDI matches are worthy of further investigation. A sample of the planned revision of the NDI Retrieval Report is presented in Table 5. The Retrieval Report lists all user records involved in a match with one or more NDI records. The State of death, death certificate number and date of death are listed for each possible match, along with an indication of which data items are in agreement. Two changes in this report should further assist NDI users in evaluating the quality of possible matches. First, the revised Retrieval Report will show which digits of the social security numbers are in agreement. The current report merely indicates whether or not there was an agreement on the entire social security number. Second, the new report will indicate the extent to which the years of birth disagree; e.g., +1 year, -1 year, -15 years, etc. The current report simply indicates whether or not there is exact agreement on the year of birth.

The user must decide which, if any, of the NDI records are true matches and then obtain copies of the death certificate from the appropriate State vital statistics offices. Most users are interested in obtaining the cause of death from the death certificate. Some users also conduct death record followback activities to the hospitals, physicians, next-of-kin, and/or other persons or establishments indicated on the death certificates. Other users simply obtain copies of certificates to assist in confirming whether a questionable match is actually the person in the study.

Once an application is approved, requests for repeat searches of the NDI file (for additional years of death or for different study subjects) do not need to go through the formal review and approval process again, as long as the information provided in the initial application remains essentially the same. Death records for a particular calendar year are added to the NDI file annually, approximately 12-14 months after the end of that calendar year. Records for deaths occurring in 1984 are scheduled to be added to the NDI file around February 1986.

5. ADDITIONAL REFERENCES CONCERNING THE NDI

In addition to the NDI users' articles and studies cited above, several other articles have been written describing the experience of NDI users [12-15]. There have also been articles written regarding the potential use of the NDI for various studies [16-18]. Finally, papers have been written in which birth certificates from the NCHS 1980 National Natality Survey were searched against the NDI to produce infant mortality rates [19-22]. Copies of these four unpublished papers can be obtained from NCHS [23].

Persons interested in receiving copies of the NDI User's Manual [24] and an NDI Application

Form should write or call:

NATIONAL DEATH INDEX
Division of Vital Statistics
National Center for Health Statistics
3700 East West Highway, Room 1-44
Hyattsville, Maryland 20782
Telephone: (301) 436-8951

NOTES AND REFERENCES

[1] Wentworth, Deborah N., et al., "An Evaluation of the Social Security Administration Master Beneficiary Record File and the National Death Index in the Ascertainment of Vital Status," American Journal of Public Health, Vol. 73, No. 11, November 1983, pages 1270-1274.

[2] Acquavella, J.F., Donaleski, D., and Hanis, N.M., "An Analysis of Mortality Follow-up through the National Death Index for a Cohort of Refinery and Petrochemical Workers," (Accepted for publication in the American Journal of Industrial Medicine, 1985).

[3] Stampher, Meir J., et al., "Test of the National Death Index," American Journal of Epidemiology, Vol. 119, No. 5., 1984, pages 837-839.

[4] Data are based on preliminary, unpublished results of an evaluation of the NDI matching criteria performed by Nancy Fink, Johns Hopkins School of Hygiene and Public Health, 1983.

[5] Lubitz, J. and Pine, P., "Initial Findings: Development and Use of a Linked Medicare-NCHS Mortality File," (Read before the 112th Annual American Public Health Association Meeting, Anaheim, California, November 11, 1984).

[6] Curb, J. David, et al., "Ascertaining Vital Status through the National Death Index and the Social Security Administration," American Journal of Epidemiology, Vol. 121, No. 5, 1985, pages 754-766.

[7] Davis, Kathryn B., et al., "A Test of the National Death Index Using the Coronary Artery Surgery Study (CASS)," (Accepted for publication by Controlled Clinical Trials, 1985).

[8] Patterson, John E., "Evaluation of the Matching Effectiveness of the National Death Index," American Statistical Association Proceedings of the Social Statistics Section, 1983, pages 1-10.

[9] Data are based on unpublished results of an evaluation of both the initial and revised NDI matching criteria performed by Helen Barbano, Division of Analysis, National Center for Health Statistics, 1984.

[10] Rogot, Eugene, et al., "On the Feasibility of Linking Census Samples to the National Death Index for Epidemiologic Studies: A Progress Report," American Journal of Public Health, Vol. 73, No. 11, November 1983, pages 1265-1269.

[11] Lynch, Billy T. and Arends, William L., Selection of a Surname Coding Procedure for the SRS Record Linkage System, Statistical Reporting Service, U.S. Department of Agriculture, February 1977.

[12] Arellano, Max G., et al., "The California Automated Mortality Linkage System (CAMLIS)," American Journal of Public Health, Vol. 74, No. 12, December 1984, pages 1324-1329.

[13] Lubitz, James, "The Cost of Cancer and the Medicare Hospice Benefit," Proceedings of the 19th National Meeting of the Public Health Conference on Records and Statistics, August 1983, pages 145-147.

[14] Nelson, Nancy A. and Van Peenen, P.D.F., "RE: 'Test of the National Death Index'" (letter to the editor), American Journal of Epidemiology, Vol. 121, 1985, page 626, (with reply from the first author, Meir Stampher).

[15] Rogot, Eugene, et al., "The Use of Probabilistic Methods in Matching Census Samples to the National Death Index," Statistics of Income and Related Administrative Record Research: 1983, Statistics of Income Division, Internal Revenue Service, October 1983.

[16] Beebe, Gilbert, "Record Linkage Systems - Canada vs. the United States," American Journal of Public Health, Vol. 70, December 1980, pages 1246-1247.

[17] Edlavitch, S.A., Feinleib, M. and Anello, C., "A Potential Use of the National Death Index for Postmarketing Drug Surveillance," Journal of the American Medical Association, Vol. 253, No. 9, March 1, 1985, pages 1292-1295.

[18] MacMahon, Brian, "The National Death Index," (editorial), American Journal of Public Health, Vol. 73, No. 11, November 1983, pages 1247-1248.

[19] Placek, Paul, et al., "Methodology for the '1980 National Natality Survey/National Death Index Match' Project," 1984 Proceedings of the American Statistical Association, Section on Survey Research Methods.

[20] Keppel, Kenneth G., et al., "Infant Mortality Rates Based on Linked Records from the 1980 National Natality Survey," (Read before the 1985 Annual Meeting of the Population Association of America, Boston, Massachusetts, March 28-30, 1985).

[21] Kessel, Samuel S., et al., "Fetal, Perinatal, and Neonatal Death Rates According to Hospital, Maternal, and Infant Characteristics: United States, 1980," (presented at the 1985 Annual Meeting of the American Statistical Association, Las Vegas, Nevada, August 5-8, 1985).

[22] Placek, Paul J., "Record Linkage Methodologies in the 1980 National Natality Survey (NNS) and 1980 National Fetal Mortality Survey (NFMS)," (presented at the meetings of the International Statistical Institute, Amsterdam, The Netherlands, August 12-22, 1985).

[23] Copies of references [19], [20], [21], and [22] can be obtained by writing to: Natality Statistics Branch, Division of Vital Statistics, National Center for Health Statistics, 3700 East-West Highway, Room 1-44, Hyattsville, Maryland 20782.

[24] User's Manual: The National Death Index, U.S. Department of Health and Human Services, National Center for Health Statistics, DHHS publication number (PHS) 81-1148, September 1981.

Table 1

NATIONAL DEATH INDEX (NDI) USERS AND RECORD VOLUMES

NDI User Characteristics	Users		NDI Searches		User Records	
	Number	Percent	Number	Percent	Number	Percent
Types of Research:						
Total	99	100.0	168	100.0	2,352,001	100.0
Exposure cohorts						
Occupational	40	40.4	57	33.9	636,752	27.1
Environmental	5	5.1	18	10.7	78,824	3.4
Diagnostic/therapeutic	2	2.0	3	1.8	7,566	0.3
Disease cohorts						
Cancer registries	13	13.1	16	9.5	38,002	1.6
Other	15	15.2	18	10.7	42,120	1.8
Life style/risk factors	9	9.1	14	8.3	116,875	5.0
Clinical trials	9	9.1	14	8.3	86,333	3.7
General population cohorts	6	6.1	28	16.7	1,345,529	57.2
Types of NDI Users:						
Total	99	100.0	168	100.0	2,352,001	100.0
Federal Government	18	18.2	62	36.9	1,516,313	64.5
State Government	4	4.0	6	3.6	45,056	1.9
University	28	28.3	37	22.0	327,060	13.9
Private Industry	13	13.1	17	10.1	221,942	9.4
Hospital	19	19.2	22	13.1	63,120	2.7
Consulting firm	17	17.2	24	14.3	178,510	7.6
Record Volume:						
Total	99	100.0	168	100.0	2,352,001	100.0
Under 2,500	42	42.4	45	26.8	29,259	1.2
2,500 - 9,999	29	29.3	38	22.6	165,711	7.1
10,000 - 24,999	12	12.1	31	18.5	225,466	9.6
25,000 - 99,999	13	13.1	33	19.7	513,014	21.8
100,000 - 499,999	2	2.0	7	4.2	424,356	18.0
500,000+	1	1.0	14	8.3	994,195	42.3

Table 2

NUMBER OF RECORDS AND PERCENT COMPLETENESS
OF NATIONAL DEATH INDEX (NDI) AND USER DATA ITEMS

Data Items	NDI File	User Files
No. of Records--------	10,290,730	1,131,931*
Percent Complete:		
Last Name----------	99.9	99.9
First Name----------	99.9	99.7
Middle Initial------	71.7	73.4
Social Security No.-	91.0	84.2
Birth Month---------	98.8	95.7
Birth Day----------	98.7	87.9
Birth Year----------	99.4	97.0
Father's Surname----	86.2	8.9
Sex-----------------	99.9	92.6
Race----------------	97.9	53.1
Marital Status------	99.4	17.9
State of Residence--	99.9	44.2
State of Birth------	99.5	18.6
Age at Death--------	99.9	10.6

* The total number of user records shown excludes 1,220,070 records associated with the National Longitudinal Mortality Study, sponsored by the National Heart, Lung and Blood Institute and involving both the Census Bureau and the National Center for Health Statistics. This large volume of records was eliminated from this table to give a more realistic presentation of the completeness of the data items submitted by the other 98 NDI users.

Table 3

REPORTING OF SOCIAL SECURITY NUMBER ON NATIONAL DEATH INDEX (NDI) RECORDS; BY SEX AND AGE AT DEATH

Age at Death	Number of NDI Records			Percent not Reported WITHIN Age/Sex Group		
	Both Sexes*	Male	Female	Both Sexes*	Male	Female
All Ages--	10,289,958	5,536,778	4,753,180	9.0	7.8	10.3
0-16------	356,704	208,377	148,327	88.6	87.4	90.3
17-21-----	126,475	95,242	31,233	17.8	16.9	20.6
22-59-----	1,965,257	1,279,175	686,082	8.4	7.2	10.6
60+-------	7,841,522	3,953,984	3,887,538	5.3	3.6	7.2

* The record counts and percentages do not include 772 records for which sex was not reported.

Table 4

EVALUATIONS OF THE EFFECTIVENESS OF THE NATIONAL DEATH INDEX (NDI) MATCHING CRITERIA USING RECORDS OF KNOWN DECEDENTS

NDI Users and User Studies*	Known Decedents	True Matches	Percent True Matches
University of Minnesota School of Public Health (Multiple Risk Factor Intervention Trial (MRFIT) for coronary heart disease) [1]....	191	188	98.4
Exxon Corporation Research & Environmental Health Division (Mortality study update of Exxon workers) [2]....	1,449	1,407	97.1
Harvard Medical School (Nurses health study) [3].....	346	334	96.5
Johns Hopkins School of Hygiene and Public Health (Health effects of low-level radiation in shipyard -workers) [4].....	8,947	8,485	94.8
Health Care Financing Administration (Use and costs of Medicare services by cause of death) [5].....	69,631	65,000	93.3
University of Texas at Houston School of Public Health (Hypertension Detection and Follow-up Program post trial survey) [6].....	1,154	1,074	93.1
University of Washington (Coronary Artery Surgery Study) [7].....	370	344	93.0
National Center for Health Statistics Division of Vital Statistics (Evaluation of NDI using cancer registry records) [8]			
INITIAL matching criteria:	2,598	2,394	92.1
Using Social Security Number (SSN).....	2,231	1,874	84.0
Using birth month/year.....	2,596	2,069	79.7
NEW matching criteria	2,598	2,500	96.2
Using SSN	2,231	1,874	84.0
Using birth month/day or birth month/±1 year .	2,596	2,351	90.6
National Center for Health Statistics Division of Analysis (First National Health and Nutrition Examination Survey epidemiologic follow-up) [9]			
INITIAL matching criteria (without SSN)	607	497	81.9
NEW matching criteria (without SSN)	607	543	89.5

* Numbers in brackets refer to studies cited in the NOTES and REFERENCES Section.

Table 5

RETRIEVAL REPORT -- REVISED
(All the information in this example is hypothetical.)

USER REQUEST RECORD (POSSIBLE MATCHES = 4) NDI APPL NO 842899 CONTROL NO 4507

POSSIBLE DECEDENT NAME	FATHERS SURNAME	SOC SEC NO	BIRTH DATE MO DY YR	AGE	SEX	RACE	MS	SOR	SOB	USER DATA
REGINA HANES		000 01 9999	12 10 18	-	F	-	M	PA	LA	011580

POSSIBLE NDI RECORD MATCHES (IN RANKED ORDER)

	STATE OF DEATH	CERT NUMBER	DATE OF DEATH	NAME F	M	L	FATHERS SURNAME	LN/FS	SOC SEC NO	BIRTH DATE MO	DY	YR	AGE	SEX	RACE	MS	SOR	SOB
*	PENNSYLVANIA	861098	02-01-81	X	B	X	-		XXXXXXXXX	X	X	X	-	X	-	X	X	X
	LOUISIANA	421304	07-07-80	X		X	-		--XXXXX-X	X		+01	-	X	-	X		X
A	LOUISIANA	421304	07-07-80	I	B	X	-		--XXXXX-X	X		+01	-	X	-	X		X
	INDIANA	698637	03-21-79	X		N	-	N	---X--X--	X	X	-15	-	X	-	?		

COLUMN HEADING ABBREVIATIONS:

LN/FS = Last name on user record compared to father's surname on National Death Index (NDI) record.

MS = Marital status

SOR = State of residence

SOB = State of birth

SYMBOLS USED WITHIN THE TABLE:

* = All items provided on user record matched exactly with NDI record.

Blank = User and NDI data items did not match.

X = User and NDI data items matched exactly.

\- = Data item not provided by user.
For SSN: specific digits did not match.
For LN/FS: comparison was not attempted.

SYMBOLS (CONTINUED):

? = Insufficient information on NDI record.

A = Alias NDI record.

I = Only first initial of first name matched.

N = Names matched only on NYSIIS codes.

B = Middle initials not provided on either record. This occurrence is treated as a match on middle initial.

+01 = Birth year on the NDI record is one year more than the year on the user record.

-01 = Birth year on the NDI record is one year less than the year on the user record.

-15 = Difference between the two years of birth. (The two-digit birth year on the user record is subtracted from the two-digit birth year on the NDI record. Note: No distinction is made to accomodate birth years in the 1800's versus birth years in the 1900's.)

AN IMPLEMENTATION OF A TWO-POPULATION FELLEGI-SUNTER PROBABILITY LINKAGE MODEL

Max G. Arellano, University of California, San Francisco

I. INTRODUCTION

The California Automated Mortality Linkage System (CAMLIS) has been in operation at the University of California, San Francisco, since the fall of 1981. It was organized under the sponsorship of the Department of Epidemiology and International Health to facilitate the clearance of study population files submitted by qualified investigators against mortality files for the State of California.

The linkage of two independently generated data files has long been thought to be the exclusive province of highly trained clerks because of the need to process the discrepancies which frequently occur between sets of identifying information for the same person on the two files.

A computerized approach to the record linkage problem can adopt either deterministic or probabilistic decision criteria. Deterministic linkage criteria require the formulation of a 'match key' to establish the relationship between records on the two files to be linked. This match key functions on an 'either or' basis, i.e., if an identical value of the match key is found on both files, the records with the identical values are said to be matched. Otherwise, the records are said to be unmatched. In order to perform its required function with minimal error, this match key must possess as many of the characteristics of a unique identifier as possible. Match keys can be constructed from any conceivable combination of last name, first name, sex, social security number, birth date (or portions thereof), or any other identifying items present on the file. Although it is not a true unique identifier, the ready availability of the social security number has led to its widespread use as the match key of choice in deterministic linkage applications.

Probabilistic linkage criteria are based on a linkage weight calculated for each pairwise comparison between records on the two files to be linked; these linkage weights are the sum of component weights calculated for each item of identification contained on the two files. The component weights are functions of occurrence probabilities and of the reliability of the data items. Probabilistic decision criteria provide an attractive alternative to deterministic linkage criteria as a means of computerizing the record linkage activity primarily because: 1) they assign weights in a manner that is consistent with our own human intuition and 2) they can accommodate partial agreements. On the debit side: 1) they require the estimation of many parameters, some of which are inestimable, 2) they are much more difficult to program and 3) they are more costly to use.

Our decision to adopt probabilistic decision criteria was based primarily on our conviction, based on a careful analysis of the available information, that the requirements of investigators in the health and medical care research fields could not be met solely by deterministic linkage criteria. Our experience over the last four years has served to confirm the validity of that decision.

II. THE FELLEGI-SUNTER WEIGHTING ALGORITHM

The Fellegi-Sunter [1] weighting algorithm requires the estimation of two probability distribution functions:

If we let,

p_{jA} = P(Occurrence of the jth configuration in population A)

p_{jB} = P(Occurrence of the jth configuration in population B)

$p_{jA\cap B}$ = P(Occurrence of the jth configuration in $A\cap B$)

$w(\gamma_j)$ = Probability linkage weight for the jth agreement configuration

$m(\gamma_j)$ = P(Occurrence of the jth agreement configuration | the record pairs are associated with members of the matched set)

$= P(\gamma_j | (a,b)\in M)$

$= p_{jA\cap B}(1-e_A)(1-e_B)(1-e_T)$

$u(\gamma_j)$ = P(Occurrence of the jth agreement configuration | the record pairs are associated with members of the unmatched set)

$= P(\gamma_j | (a,b)\in U)$

$= p_{jA}p_{jB}$

Then, $w(\gamma_j) = \log[m(\gamma_j)/u(\gamma_j)]$

Among the obvious difficulties encountered in the implementation of this model are:

(A) It does not address the problem of estimating the e or e_T terms. We generally refer to these as the "component error probabilities."

(B) The $p_{A\cap B}$ term requires information which can only be obtained when the linkage has been completed in a satisfactory manner, if then.

If the populations represented by the files that are being linked can be regarded as samples drawn from the same population, i.e., the "one-population" model, some simplifications can be introduced into the above expressions:

$$m(\gamma_j) = p_j(1-e)^2(1-e_T)$$

$$u(\gamma_j) = p_j^2$$

$$w(\gamma_j) = \log[m(\gamma_j)/u(\gamma_j)]$$

$$= \log[p_j^{-1}(1-e)^2(1-e_T)]$$

Moreover, if the data are being collected continuously, as is generally the case under the circumstances to which the one-population model is

applicable, procedures can readily be developed to iteratively obtain "good" estimates of the component error probabilities. This is, unfortunately, not the case for situations to which the two-population model would generally be applied. For one thing, if the populations being linked do not overlap, the $p_{A\bar{\cap}B}$ term is meaningless. The model also requires estimates of component error probabilities specific to the files that are being linked.

Prior information on the record-pairs that correspond to the intersection of the two populations is obviously desirable, if not absolutely necessary, before probability linkage can be initiated. However, since this is precisely the information we are attempting to obtain by means of probability linkage, if it can be obtained by other means, one may legitimately question the need for probability linkage.

In this paper I will describe the approach that has been adopted by the CAMLIS project to the problem of implementing a two-population Fellegi-Sunter model.

III. THE CAMLIS IMPLEMENTATION OF THE TWO-POPULATION FELLEGI-SUNTER MODEL

Central Concepts

The CAMLIS approach is based on the following central concepts:

(A) A two-stage linkage process, consisting of a deterministic first stage (primarily based on the social security number) followed by a probabilistic second stage, is necessary to achieve the desired performance characteristics. This strategy has several benefits:
  (1) Each stage is capable of detecting valid linkages which will escape detection by the other stage.
  (2) Since deterministic linkage is carried out first, the correctly matched records which it produces can be used to derive estimates of the component error probabilities required by probability linkage.

(B) A phonetic name encoding algorithm with superior operating characteristics must be used to form the basic comparison groups for probability linkage to minimize the number of pairwise record comparisons that must be carried out. We chose to adopt a modified version of the New York State Identification and Intelligence System (NYSIIS) phonetic coding system for this purpose. It is doubtful if CAMLIS could be operated on a cost-effective basis without the use of a phonetic name coding system with the superior performance characteristics of NYSIIS.

(C) A modification of the weighting algorithm for the two-population Fellegi-Sunter model is necessary to compensate for the inestimable parameters.

(D) Component error probabilities can be estimated from the "matched set" produced by first stage or deterministic linkage.

In this presentation, I will focus primarily on points (C) and (D) above, i.e., on our approach to the estimation of the parameters required by the two-population Fellegi-Sunter weighting algorithm.

The Estimation of Relative Frequency Parameters

In CAMLIS applications, a user file, which we denote as file L_A, is linked to a California State mortality file, which we denote as file L_B. Since the characteristics of most user files are significantly different from those of the California mortality file, the two-population model is obviously called for. However, many of the parameters required by the two-population model, e.g., $p_{A\bar{\cap}B}$ and e_A, are inestimable. We therefore carefully scrutinized the expressions for the two probability distribution functions to determine whether a simplification was possible. We first made the observation that the characteristics of the user file are always subsets of the characteristics of the mortality file; we also observed that, for those components that are independent of mortality, $p_A \sim P_{A\bar{\cap}B}$. These observations resulted in the elimination of the p_A term from the weighting algorithm and served to justify the use of relative frequencies derived only from the mortality files. Since these relative frequencies can change over time, files have been developed which contain the necessary relative frequencies at five-year intervals; CAMLIS procedures retrieve them as necessary.

The component for which the assumption is not tenable is birth year; an entirely different approach to weight computation for the birth year component has, therefore, been developed.

The Estimation of Component Error Probabilities

Within the context of a mortality clearance system, it is not possible to derive separate estimates of component error probabilities for files L_A and L_B; there is just not enough information available. We therefore made the simplifying assumption that the corresponding component error probabilities in the two files were identical, i.e., we assume that:

$$e = e_A = e_B$$

Estimates of e and e_T are derived from the matched record-pairs produced by first stage deterministic linkage. To eliminate spurious matches, we require a high concordance among the identifying elements on the two files that are not incorporated into the match key.

The basic algorithm that we utilize to calculate agreement configuration weights is therefore:

$$m(\gamma_j) = p_{jA}(1-e)^2(1-e_T)$$

$$u(\gamma_j) = p_{jA}p_{jB}$$

$$w(\gamma_j) = \log[m(\gamma_j)/u(\gamma_j)]$$

$$= \log[p_{jB}^{-1}(1-e)^2(1-e_T)]$$

IV. CONCLUSION

The Fellegi-Sunter model requires an assumption regarding the independence of the components of the comparison vector; this assumption is frequently a major concern in linkage applications. It is not my intention to minimize the importance of this assumption. The real concern, however, must be the extent to which violations of

this assumption affect the results produced by the model.

(A) The components of the comparison vector should be carefully chosen. Only one of several highly dependent components should be incorporated into the model.

(B) Although it is possible to correct for the effect of dependence, for moderately dependent components, these efforts are hardly ever worth the small gain in precision that can be realized.

(C) We have done a great deal of difference analysis. Our conclusion is that the estimated component error probabilities and relative frequencies must differ considerably from the appropriate values to significantly affect the computed weights.

(D) For matches that achieve a linkage weight significantly greater than the upper threshold value, a bias in the weight is obviously of no consequence. Similarly, for matches that achieve a linkage weight significantly below the lower threshold value, a bias in the weight is also of no consequence. The vast majority of record-pairs achieve either very low or very high linkage weights.

(E) Record-pairs which achieve a linkage weight between the lower and upper threshold values are subject to manual review. Since record-pairs fall into this category because they either contain ambiguous or sparse identifying information, it is extremely doubtful whether they would differ significantly if the weights were computed according to a more precise model. In any case, comparable results could be obtained by redefining the upper and lower threshold values.

The major advantage of probability linkage is that it permits a meaningful ranking of matched record-pairs. The ranking makes it possible to focus review efforts on the comparisons which have been assigned borderline weights. It can readily be shown that the gain achieved by verifying the probability linkage decisions above a certain threshold value and below a certain threshold value is negligible.

Our experience with the Fellegi-Sunter probability linkage criteria has been uniformly favorable. It is our considered opinion, however, that probabilistic linkage and deterministic linkage are best utilized as complimentary procedures and that both are necessary to achieve optimum results.

REFERENCES

[1] Fellegi, I., and Sunter, A. (1969) "A Theory for Record Linkage," Journal of the American Statistical Association, 64, 1183-1210.

DERIVING LABOR TURNOVER RATES FROM ADMINISTRATIVE RECORDS

Malcolm S. Cohen, University of Michigan

U.S. nonagricultural establishments will hire workers new to their firms an estimated 64 million times during 1985. These hiring transactions probably will involve only 12–16 million workers who changed their primary jobs.

An econometric model was constructed using administrative records from Social Security files, and estimates of new hires were made by industry, state, age, race, and sex. When this study was done, Social Security records were available only through the mid–1970s. Wage records used in the administration of the unemployment insurance system were available in sixteen states to verify the accuracy of the econometric estimates. Because wage records were available only for sixteen states, and because of differences in state laws and data processing procedures, wage records could not be used for obtaining national estimates.

Organizationally, this paper is divided into two main sections. In the first, the methodology employed is described. The second presents examples of the various results, as well as some general comments about the usefulness of these administrative records.

METHODOLOGY

Social Security data from a one–percent sample of a continuous work history file for the period 1971–76 were used to construct labor turnover measures. Instructions for using the methodology were given to three government agencies, who then did the matching and provided tabulations for different years. These agencies were the New York Department of Labor, the Social Security Administration, and the Bureau of Economic Analysis. The provisions of the 1976 tax reform act require the Internal Revenue Service to screen the data for possible confidentiality disclosures prior to release. All analyses of Social Security records were from tabulations provided by the government agencies. No Social Security data were released on individual workers or firms.

Employee records were matched with employer records. If a worker's identification number appeared in a firm's file in a given quarter, but did not appear in the file in the previous quarter, the worker was classified as an accession to the firm [1]. If a worker classified as an accession did not work for the firm for the prior four quarters, that worker was classified as a new hire. The decision to use four quarters as a determining factor was somewhat arbitrary. That period of time was chosen because it was long enough to identify workers who return to a firm seasonally, although it would not exclude workers who may have worked for a firm sometime in the more distant past. The higher degree of accuracy that might be attained by matching records several years back, however, was not considered great enough to justify the substantial increase in cost of matching data for more than four quarters [2].

It is also possible to generate other turnover measures using the pattern of employment within the firm. For example, if a worker is present in a given quarter and absent in the next quarter, this is a separation. If a worker is a new hire who continues to work for a period of, say, an additional two quarters, this is a permanent new hire. If a worker is an accession (not employed in previous quarter) who did work for the firm sometime in the previous four quarters, this is a recall. If a worker is an accession and separation in the same quarter, this is a short–term accession. Various turnover measures were developed based on these definitions.

Data were constructed for new hires from quarterly Social Security records from the second quarter of 1972 to the second quarter of 1975. A special tabulation for 1975–76 was used for special analyses but not included in the quarterly analyses used to generate current estimates.

A model was developed to predict new hires. The model's derivation begins with a tautology:

(1) $\Delta E = NH + Recalls - Quits - Layoffs - OS$

where ΔE is change in employment; NH is new hires; and OS is other separations.

From this we obtain:

(2) $NH = \Delta E - Z$

where $Z = Recalls - Quits - Layoffs - OS$

To obtain rates, both series were divided by E. It was assumed that the unemployment rate would be a good proxy for Z. It was assumed that there was a negative correlation between Z and the unemployment rate.

When the equation was estimated, data from the Bureau of Labor Statistics (BLS) 790 series were used for employment, and data from the monthly Current Population Survey were used for unemployment rates and seasonal dummy variables. The final equation was:

$$(3)\quad NHR_t = \alpha_0 + \alpha_1 \%\Delta E_t + \alpha_2 UR_{t-1} + \\ + \alpha_3 S_1 + \alpha_4 S_2 + \alpha_5 S_3 + \alpha_6 D + E_1$$

where NHR is the new hire rate; $\%\Delta E$ is the percentage change in BLS 790 employment; UR is the unemployment rate; S_1,S_2 and S_3 are seasonal dummies for the first three quarters of the year; D is 1 in the first quarter of 1974; and E_1 is a random term.

The dummy variable was used because of a data error in the first quarter of 1974 in the data provided. The coefficient α_1 is expected to be positive, while α_2 is predicted to be negative. The equations were estimated for each state with a total of thirteen observations. The results of the model for fiscal 1975 were simulated to determine goodness of fit.

Figure 1 provides the $\%\Delta E$ and UR_{t-1} parameters, the proportion of variation explained by the model (R^2), actual new hire rate, and percent error in the forecast for all 50 states. All parameters significant at the .05 level are indicated by an asterisk.

One of the difficulties with this model is that data for the dependent variable cannot be obtained from Social Security data beyond 1977 on a quarterly basis. Only annual new hire rates can be computed. These can only be obtained by special arrangements with the Internal Revenue Service and the Social Security Administration. To verify the model in selected states, however, wage records were obtained using similar concepts for workers covered by unemployment insurance. These data can be generated quarterly on a current basis in wage records states. Over 40 states are wage records states. Special arrangements must be made, however, in each state to obtain these data. The arrangements require considerable data processing to match workers and firms over at least four quarters.

Our estimates were compared with the wage records data in sixteen states. The results of the comparisons are shown in Figure 2. The errors are generally relatively small except in Florida. Here, however, the Florida data provided were probably more prone to error than our estimates. The significantly lower reported new hires in Florida probably represents an undercount in the state's processing. The

state used a different processing methodology than the other states.

We simulated our model and obtained new hire estimates for 1975–85 [3].

RESULTS

Figure 3 shows the predicted number of new hires from 1975 through 1985 using our model. Figure 4 illustrates the five states with the largest number of new hires. These states accounted for 40% of all new hires in the United States. Converting the new hires into rates, Figure 5 shows the parts of the United States with the highest and lowest rates. The highest rates are west of the Mississippi. A prominent exception is Florida.

It is also possible to compare new hire rates by industry. Figures 6 and 7 show the industries with the highest and lowest rates, respectively.

In 1985 it is unlikely that social services would be among the high new hire rate industries. This reflects changes in government priorities over the decade. It is probable, however, that the other industries are high and low turnover industries in 1985.

Individuals versus Transactions

One of the difficulties in interpreting our measures is reconciling the incredibly high turnover (e.g., 80% in 1985) with our knowledge of how often workers change jobs. The number of turnover transactions include instances where one worker changed jobs more than once, so the total does not reflect the actual number of workers who changed jobs. Thus, when turnover is expressed as a percentage of employment, the result should not be interpreted as the percentage of workers who changed jobs. To gain some insight into reconciling this apparent dilemma, we developed some special tabulations from 1975–76 Social Security files. First we computed an annualized 84% new hire rate for 1976 by multiplying the rate obtained in the second quarter of 1976 by 4. This is certainly comparable to the rates we had been obtaining for other years. A different analysis was carried out where workers were assigned to their primary jobs, where they earned the most money during 1976. Only 18% of the workers were new hires in their primary jobs, based on the second quarter of 1976. Some of these workers could have accounted for several new hire transactions. Similarly, workers who were not new hires in their primary jobs could be new hires in secondary jobs. Thus, we estimated that of the 64 million new hires, about 14 million workers were new hires in their primary jobs. In another quarter we estimated a ratio which would suggest that slightly under 16 million workers were new hires in their primary jobs. An estimate of 12–16 million seemed appropriate due to the limited number of quarters on which we could base our ratio.

Another comparison we made with our special tabulation was the average number of employers for whom employees worked in different industries. We assigned workers to the employer from whom they received the majority of their earnings and tabulated the number of different employers. Four nonagricultural industries--heavy construction contractors, water transportation, eating and drinking places, and motion pictures--had an average of two or more employers per worker. Water transportation (longshore) averaged 2.5 employers per worker. The industries with an average of 1.25 or fewer employers (with at least 100,000 persons in the industry) included: primary metals, communications, and public utilities.

Areas for Further Research

The information obtained from Social Security records and state unemployment insurance records represent about the only currently comprehensive source of labor turnover data. Our model permits obtaining current estimates from these data. It would be useful to tabulate annual Social Security files to determine labor turnover from more recent Social Security files. It would also be useful to forecast the turnover rates by industry, age, and sex. The 1975–76 special tabulations by person and transaction provide detailed characteristics by state, SMSA, industry, age, wage class, sex, and race. Additional analyses of these data remain to be carried out, as well as additional analyses of separations and short-term new hires. Finally, more efficient forecast estimates can be made by combining cross-section and time-series turnover data.

NOTES AND REFERENCES

[1] A worker's identification number appears in the file if the worker had wages greater than zero in a given quarter.

[2] Using California wage records from the Unemployment Insurance system, the California Employment Development Division did a test of how many fewer new hires there would be if seven quarters were used as a cut-off instead of four, and found only about 2% fewer new hires. (Glen Siebert, Employment Service Potential: Indicators of Labor Market Activity, pp. 48–9. Sacramento, CA: Employment Development Department, 1977.)

[3] For a more complete description of the simulation methodology, see Malcolm S. Cohen and Arthur R. Schwartz, "A New Hires Model for the Private Nonfarm Economy," Economic Outlook for 1984, Department of Economics, University of Michigan, Ann Arbor, 1984.

Figure 1. New Hire Rates by State, Fiscal 1975, % Error, R^2, Selected Coefficients

State	1975 New Hire Rate	1975 % Error	R^2	%E	URLAG
Alabama	19.1	-.3	.943	51.94	-1.59*
Alaska	42.0	5.2	.941	165.85*	2.34
Arizona	24.9	.3	.978	148.44*	-1.65*
Arkansas	22.4	.5	.966	48.90	-2.24*
California	23.4	-.9	.930	87.91	-1.21
Colorado	28.3	1.6	.951	97.29*	-2.75*
Connecticut	15.0	.9	.984	97.25*	-1.03*
Delaware	15.9	-.6	.828	-64.85	-3.25*
D.C.	20.8	-3.5	.822	89.90	-1.49
Florida	26.3	-1.3	.973	178.70*	-2.29*
Georgia	20.2	-1.1	.982	118.40*	-2.24*
Hawaii	20.9	.9	.819	122.97	-.85
Idaho	26.3	.1	.898	68.38	-.52
Illinois	16.8	.1	.988	111.27*	-1.19*
Indiana	15.5	-.9	.992	83.49*	-1.56*
Iowa	18.7	3.0	.951	25.81	-1.61*
Kansas	23.1	3.1	.944	63.74	-1.33
Kentucky	17.7	-1.2	.980	107.40*	-1.07*
Louisiana	26.3	1.7	.890	-15.77	-1.37
Maine	18.2	-3.0	.943	105.95	-.88
Maryland	18.2	-.1	.982	162.01*	-.71
Massachusetts	16.5	-1.9	.976	126.06*	-.81*
Michigan	14.5	-4.1	.935	73.53*	-1.48*
Minnesota	17.3	-.1	.958	62.99	-1.30*
Mississippi	19.5	.2	.938	96.48*	-1.36
Missouri	18.2	.4	.989	99.74*	-1.13*
Montana	23.5	-1.3	.959	191.26*	-.20
Nebraska	20.6	1.7	.971	74.97	-.86
Nevada	33.2	-.5	.975	165.36*	-1.42*
New Hampshire	17.5	-2.0	.917	135.78*	-2.02*
New Jersey	17.1	-.1	.978	121.49*	-1.20*
New Mexico	28.3	-2.3	.916	103.08	-1.61*
New York	15.7	-1.7	.959	109.77*	-1.16*
N. Carolina	16.9	-1.5	.970	112.58*	-2.03*
N. Dakota	22.2	2.0	.902	229.05*	.72
Ohio	15.0	-.3	.996	91.35*	-1.33*
Oklahoma	24.8	-.1	.944	131.44	-1.08
Oregon	23.3	1.1	.925	103.60	-1.22
Pennsylvania	13.9	.2	.980	134.31*	-.96*
Rhode Island	17.8	-1.9	.960	72.75*	-1.84*
S. Carolina	17.6	-1.9	.918	69.73*	-1.77*
S. Dakota	19.9	-2.4	.968	133.96*	-.50
Tennessee	18.1	-.6	.978	93.82*	-1.38*
Texas	27.1	-.3	.977	34.35	-1.57*
Utah	23.9	.0	.967	109.57	-1.20
Vermont	18.0	.9	.821	161.11	-.18
Virginia	18.0	-.2	.970	107.94*	-1.66*
Washington	22.4	.7	.953	141.46*	-.07
W. Virginia	15.7	-2.3	.964	145.31*	-.27
Wisconsin	14.8	-.1	.988	72.78*	-1.39*
Wyoming	33.4	4.4	.899	21.54	-1.22

%E = percentage change in employment
URLAG = unemployment rate in previous quarter
* = coefficient significant at the .05 level
N = 13 for each state

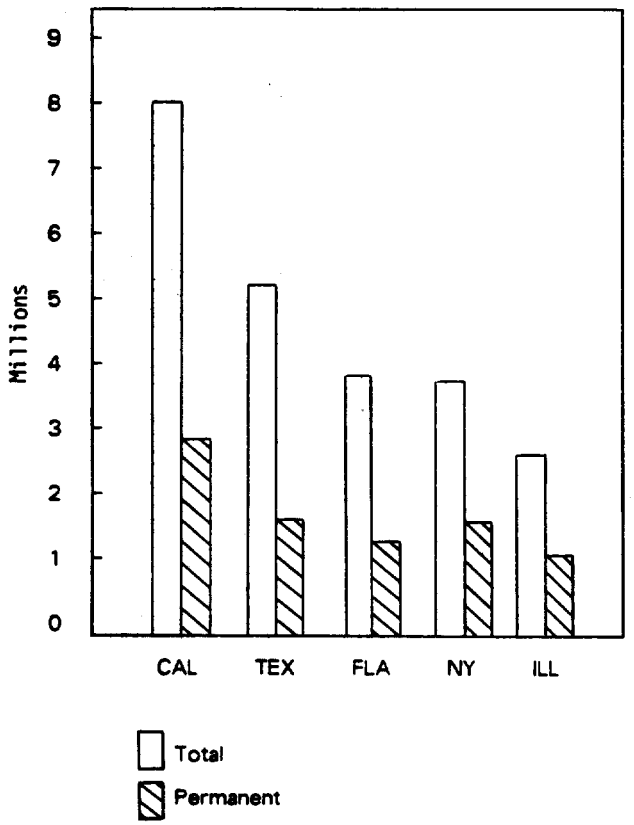
Figure 2. Comparison of New Hire Forecasts with Actual New Hire Data

State	Period	New Hires Reported by State Employment Agencies	Predicted New Hires	% Difference
Arkansas	Fiscal 1979	583,990	603,500	+3.34
Pennsylvania	Fiscal 1976	2,051,553	2,147,100	+4.66
South Dakota	Fiscal 1979	177,433	155,800	-12.19
	Fiscal 1980	142,795	137,500	-3.70
	Fiscal 1981	134,109	142,900	+6.57
Idaho	Fiscal 1976	238,989	241,000	+0.84
California	Fiscal 1976	6,142,625	5,796,000	-5.64
	Fiscal 1977	6,625,804	6,506,800	-1.80
	Fiscal 1978	7,523,644	7,640,400	+1.55
	Fiscal 1979	8,366,534	8,226,400	-1.67
North Dakota	Fiscal 1976	147,081	144,300	-1.88
North Carolina	1979 - 4th Q.	392,663	370,300	-5.71
Nevada	Fiscal 1976	309,100	298,300	-3.48
	Fiscal 1979	452,679	476,800	+5.32
	Fiscal 1980	464,348	466,600	+0.48
	Fiscal 1981	438,880	477,600	+8.95
South Carolina	1979 - 1st-3rd Q.	611,324	627,700	+2.68
	1981 2nd-4th Q.	550,619	522,900	-5.03
Maine	Fiscal 1978	263,175	268,900	+2.17
Illinois	1979 3rd-4th Q.	1,436,475	1,593,500	+10.93
New Mexico	Fiscal 1979	410,927	412,000	+0.26
	Fiscal 1980	378,288	386,200	+2.10
Missouri	1979 -3rd-4th Q.	718,946	670,400	-6.75
	Calendar 1981	1,073,311	1,204,900	+12.26
Iowa	Fiscal 1981	587,016	582,500	-0.77
Mississippi	1981 4th Q.	101,921	107,400	+5.40
Florida	Calendar 1980	2,673,019	3,790,500	+41.81
	Calendar 1981	2,918,487	3,729,700	+27.80

Figure 3. Number of New Hires in the Private Nonfarm Economy by State
(annual totals in thousands)

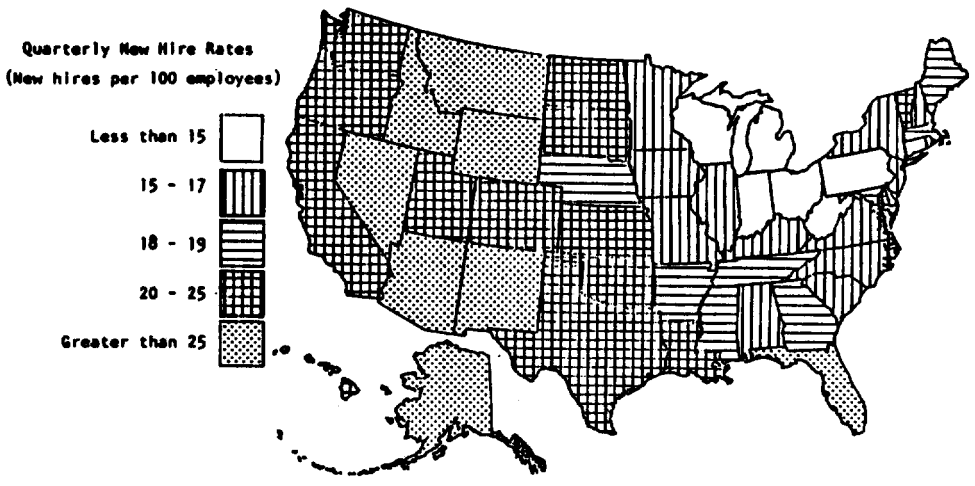
State	1975	1976	1977	1978	1979	1980	1981	1982	1983	1984	1985
Alabama	610.9	693.3	765.6	878.3	895.3	808.1	770.0	620.8	598.7	744.9	829.9
Alaska	178.8	112.3	92.8	102.6	109.2	111.7	126.8	136.2	171.3	162.5	172.3
Arizona	514.5	624.0	735.7	903.1	1005.0	891.5	889.8	730.0	789.2	921.3	1019.3
Arkansas	378.8	443.0	498.7	580.6	606.2	534.0	503.7	392.4	364.2	472.8	541.5
California	5219.2	6059.2	6811.6	7838.4	8294.0	7770.4	7760.8	6700.8	6743.6	8001.6	8532.8
Colorado	656.3	813.2	951.6	1148.8	1242.7	1140.8	1097.3	888.8	913.8	1190.4	1371.7
Connecticut	530.9	653.1	719.2	829.0	852.6	795.4	757.8	642.0	654.6	781.2	862.5
D.C.	183.9	164.2	174.4	196.3	205.8	187.0	180.8	147.8	151.3	174.2	186.5
Delaware	126.4	182.4	203.2	243.4	259.2	220.3	208.1	140.3	138.5	183.7	211.3
Florida	2006.5	2567.0	2968.8	3614.6	3884.4	3790.5	3729.7	3104.6	2983.2	3778.2	4162.8
Georgia	1031.3	1226.7	1388.1	1667.6	1720.3	1547.1	1437.8	1143.7	1223.9	1464.5	1634.2
Hawaii	179.6	205.6	219.9	263.0	268.2	254.5	238.6	213.0	221.4	265.2	278.5
Idaho	207.4	250.0	261.4	289.9	282.7	260.9	260.3	247.2	274.5	290.2	303.1
Illinois	2195.1	2718.2	2813.5	3178.2	3172.0	2826.4	2639.7	2074.3	2241.8	2632.6	2761.8
Indiana	876.1	1090.2	1192.6	1393.0	1351.4	1115.6	1078.8	799.8	844.7	1062.0	1153.4
Iowa	485.4	547.4	602.0	691.4	718.5	633.8	580.4	453.0	405.5	518.4	596.0
Kansas	507.6	564.0	612.7	694.4	726.4	655.0	646.4	532.3	521.9	623.2	680.3
Kentucky	563.8	647.9	733.1	826.1	785.0	689.9	659.4	541.6	633.7	687.9	739.6
Louisiana	868.4	1019.4	1092.7	1239.0	1308.0	1298.7	1295.5	1167.1	1084.5	1240.2	1369.5
Maine	201.8	241.0	247.6	278.0	277.4	261.7	247.4	222.8	237.4	267.8	272.5
Maryland	774.1	859.4	981.1	1111.2	1077.0	1008.1	976.8	872.3	927.6	1007.8	1053.0
Massachusetts	1181.1	1416.4	1535.9	1707.1	1768.3	1697.5	1652.0	1422.0	1525.1	1717.6	1812.5
Michigan	1377.7	1639.4	1862.6	2110.8	2036.3	1676.4	1574.2	1177.4	1292.1	1574.6	1728.0
Minnesota	713.2	823.4	905.2	1065.5	1123.6	993.8	950.0	766.8	766.2	958.9	1074.3
Mississippi	403.6	460.4	517.6	566.2	577.7	507.2	501.5	406.8	429.4	514.0	558.6
Missouri	947.7	1096.5	1201.8	1347.0	1366.0	1184.6	1204.9	1019.6	959.4	1158.4	1251.6
Montana	167.2	213.5	208.6	241.4	219.7	201.2	224.6	183.8	200.5	241.8	249.0
Nebraska	318.7	366.4	381.3	418.9	441.6	399.7	394.3	336.2	317.7	381.7	406.1
Nevada	255.1	319.4	380.8	473.4	488.6	459.4	470.8	400.0	442.6	575.5	637.0
New Hampshire	153.5	208.2	236.8	277.1	292.2	255.4	253.0	183.3	202.5	256.6	278.6
New Jersey	1396.3	1648.9	1793.3	2038.4	2069.0	1917.8	1878.5	1572.4	1628.0	1887.0	2041.0
New Mexico	269.6	312.5	358.1	399.3	414.6	378.9	384.8	334.4	338.0	407.4	454.4
New York	3211.6	3568.6	3809.7	4285.6	4391.2	4072.8	4015.6	3356.4	3296.5	3719.8	4014.8
North Carolina	1035.5	1225.3	1377.6	1622.5	1707.0	1741.9	1378.6	1027.8	1072.5	1366.1	1512.6
North Dakota	134.8	140.5	138.7	160.2	161.8	141.1	159.9	145.9	171.8	185.3	190.0
Ohio	1702.0	2077.8	2324.2	2632.2	2622.1	2204.4	2152.4	1653.7	1595.4	2062.7	2247.7
Oklahoma	632.1	715.4	775.4	904.6	933.1	942.4	972.8	820.1	836.3	975.2	1072.2
Oregon	562.8	661.6	744.8	839.5	884.7	750.3	697.0	592.8	625.4	748.8	813.8
Pennsylvania	1864.6	2214.4	2330.8	2717.4	2651.9	2285.8	2289.0	1628.0	1840.2	2118.2	2271.6
Rhode Island	187.4	223.7	244.6	279.3	286.8	258.4	244.0	188.8	188.5	236.0	265.1
South Carolina	501.3	594.0	649.8	764.0	797.9	719.9	684.2	536.7	508.5	654.4	725.6
South Dakota	119.1	139.9	147.9	159.8	155.1	134.2	141.2	122.7	141.0	155.8	163.8
Tennessee	838.4	975.2	1091.6	1231.3	1223.3	1079.2	1072.2	869.2	850.8	1063.3	1135.4
Texas	3369.7	3886.5	4228.4	4909.2	5327.6	5266.0	5359.2	4772.0	4376.8	5132.4	5760.4
Utah	287.7	337.3	366.9	425.1	434.8	395.6	402.6	350.8	359.1	434.4	475.6
Vermont	96.3	113.2	123.7	136.9	133.7	125.3	126.0	123.0	143.8	149.4	152.7
Virginia	848.0	1014.2	1124.8	1308.3	1378.2	1234.8	1151.6	953.5	912.0	1172.3	1303.1
Washington	828.6	952.7	1034.8	1174.8	1212.2	1085.9	1072.0	1026.9	1156.8	1313.9	1364.2
West Virginia	267.3	288.8	307.0	337.4	353.1	324.6	299.3	259.6	237.7	272.8	294.8
Wisconsin	702.2	830.5	922.1	1075.1	1121.9	932.2	898.1	683.0	667.8	856.0	973.9
Wyoming	123.9	147.2	166.3	193.0	209.8	210.6	212.2	192.7	189.0	222.8	235.9
U.S. Total	42794.9	50296.0	55356.0	63768.0	65824.0	60108.0	58904.0	48876.0	49396.0	58984.0	64196.0

Figure 4. States with the Highest Number of New Hires, 1984



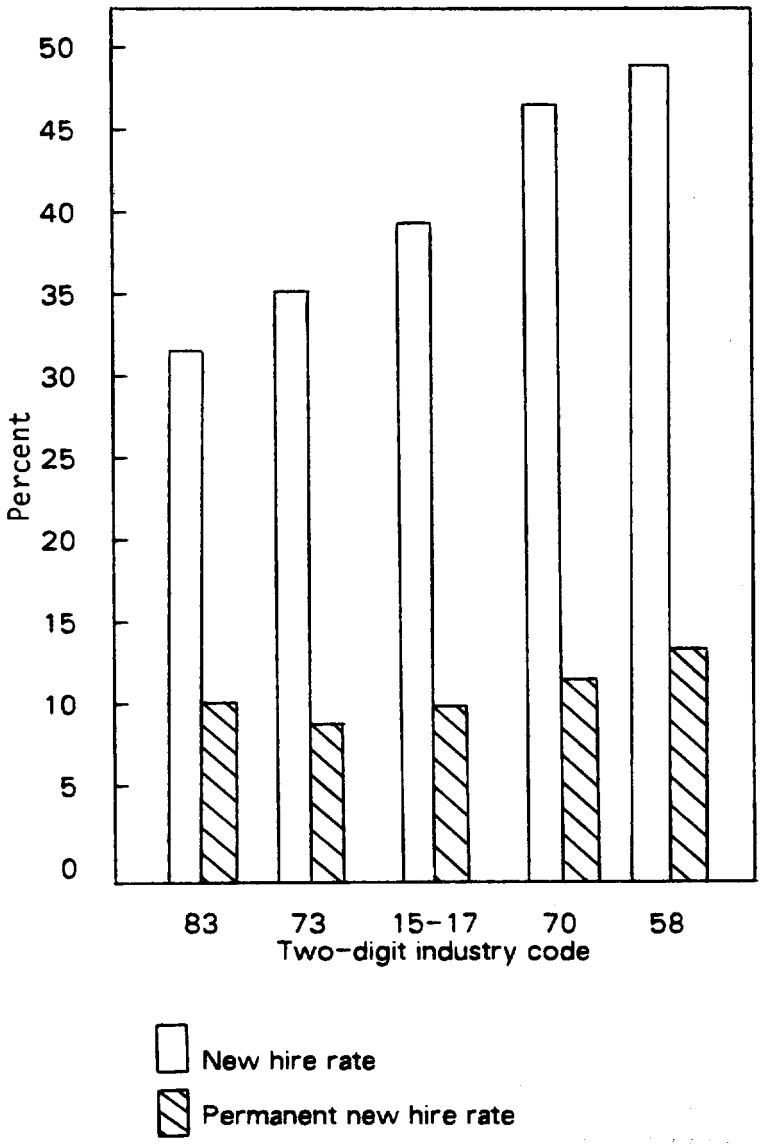
Source: Institute of Labor and Industrial Relations, University of Michigan, November 1983

Figure 5. Projected Quarterly New Hire Rates, 1984.



SOURCE: Institute of Labor and Industrial Relations, University of Michigan. November 1983.

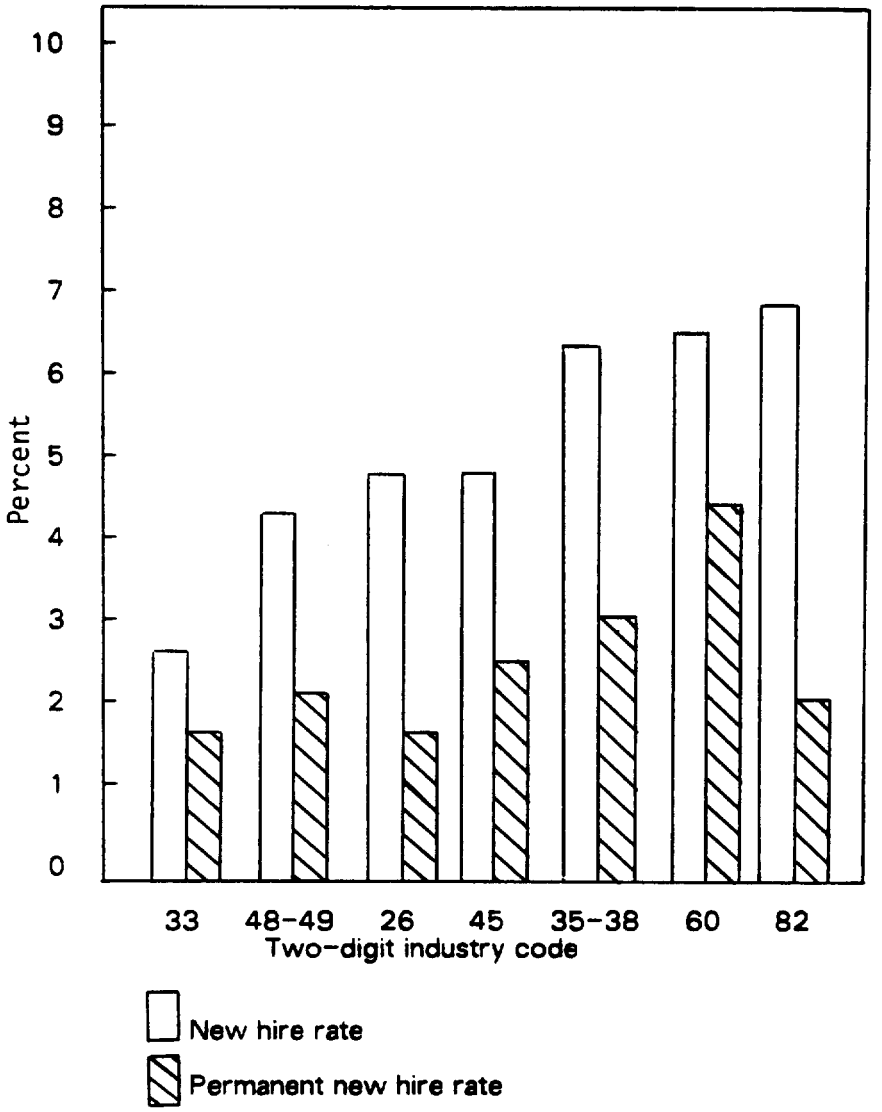
Figure 6. Industries with Highest New Hire Rates, 1975 2nd Quarter



83 - Social services
73 - Business services
15-17 - Construction
70 - Hotel and other lodging
58 - Eating and drinking establishments

Source: Institute of Labor and Industrial Relations, University of Michigan

Figure 7. Industries with Lowest New Hire Rates, 1975 2nd Quarter



33 - Primary metal manufacturing
48–49 - Communications and public utilities
26 - Paper manufacturing
45 - Air transportation
35–38 - Machinery + transportation + instrument manufacturing
60 - Banking
82 - Educational services

Source: Institute of Labor and Industrial Relations, University of Michigan

DISCUSSION

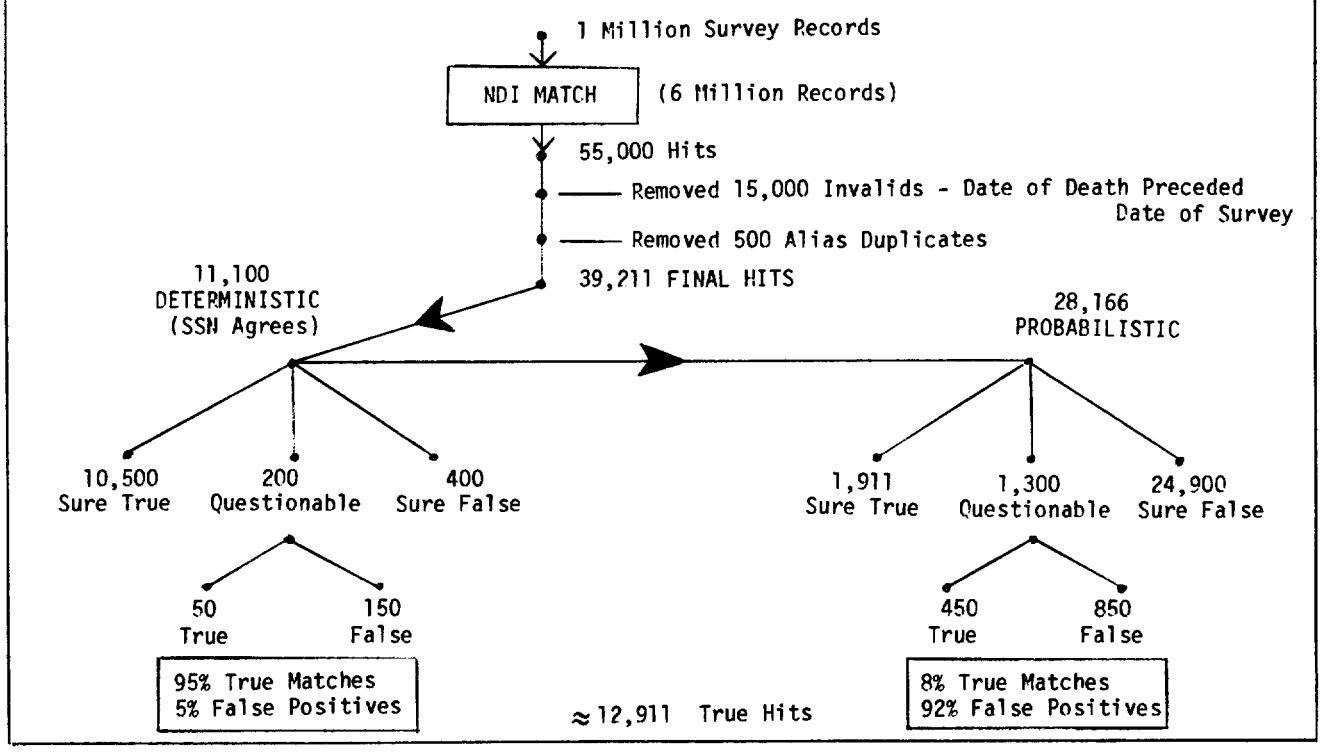
Norman J. Johnson, U.S. Bureau of the Census

I would like to present my discussion of these three papers in terms of points which we have encountered in an application of matching from our project. I have been working on developing the data base for The National Longitudinal Mortality Study (NLMS). This study is being conducted jointly by the National Heart, Lung, and Blood Institute, the National Center for Health Statistics and the U.S. Census Bureau. The primary objectives of the NLMS are to analyze socioeconomic, demographic and occupational differentials in mortality within the United States. A major interest of our analysis will be to compare survival rates of different subsets of the cohorts.

The study population consists of eight cohorts of selected Census samples. Deaths in this population are identified through periodic matching to the National Death Index (NDI), the index discussed in the first paper by Mr. Patterson. As pointed out in that presentation, in terms of number of records submitted for matching, our project is a major user of the National Death Index. The National Longitudinal Mortality Study currently consists of approximately 1 million records from eight cohorts. One match has been made to the NDI, which at the time consisted of approximately 6 million records. We intend to conduct follow-up matches approximately every two years.

The process we used to obtain the final matched records was completed in two steps. First, our files were matched to the NDI using the NCHS criteria. Then, an extensive screening was made of the resulting match using some of the methodologies discussed in presentations given earlier in these sessions to determine the final true match status. This second step involved both computer and manual matching. Our approach in the computer matching phase was similar to that used in the CAMLIS project of Mr. Arellano, the presenter of the second paper of this section. A link was made deterministically for all matches in which there was an exact agreement on social security number. Records not matched deterministically were then matched probabilistically using a modified Newcombe model. Weights for this model were estimated from a subsample of records from the NCHS match which had been reviewed manually to establish correct match status. Three categories of records from the probabilistic match resulted: true, false and questionable matches. Questionable matches were decided on the basis of a manual review. This process and the final results have been schematically diagrammed in Figure 1. From the initial one million records, approximately 12,900 links occurred. The information in the figure also indicates the substantial difference in the true match rate between the deterministic and the probabilistic steps.

Figure 1. -- National Longitudinal Mortality Study



PATTERSON AND BILGRAD

As I mentioned in my introduction, our project is a major user of the National Death Index. Deaths in our cohorts are determined by linking our records to records in this Index. The NDI matching algorithm is, in a sense, deterministic. It uses combinations of five major variables in seven criteria to determine a link. These criteria are soon to be expanded to twelve. A link is made if any one of the seven criteria is satisfied. As other studies continue to match using this index, the NDI may wish to incorporate some probabilistic components into their matching procedure based on the experience of their users. Results from our project may be helpful in this regard.

Five major categories of users were summarized in the presentation. The major users identified are in health-related fields. In many health studies, analysis is done by comparing survival of cohorts, as is the case in our study. Rare events are often of interest and small counts may be greatly affected by match rates. For this reason, in our study, we feel that matching algorithms should put emphasis on detecting true matches, with willingness to manually review more questionable matches, in order to rule out false positives. The additional criteria made available in the new NCHS matching algorithm are a step in the right direction. The expanded criteria will generate more true links as well as more false positives.

The paper presents results of studies to measure the improvements in the match rate to the NDI due to the replacement of the Soundex Code for matching of names by the NYSIIS code. If the NCHS studies of the effects of this change are true, that is, 18 percent fewer true matches and 31 percent fewer false matches could be expected, then, in view of the comments which I made earlier, the Soundex Code would be preferable to us.

ARELLANO

I will focus my discussion on the three points mentioned in the conclusion section of the paper. The paper deals with the use of the Fellegi-Sunter approach in the CAMLIS project to link user files to death certificates from the state of California. The first point discussed concerns the potential for making estimates of error terms in the Fellegi-Sunter model. The estimation of error terms is a major difficulty encountered in application of the theory. In some applications, making simplifying assumptions is the only way to obtain estimates of errors. The similarity of the CAMLIS study and the National Longitudinal Mortality Study may enable us to exchange estimated parameter values once they are obtained.

The conclusion on the robustness of error probability estimates is important and potentially very useful. This quality of the estimates would allow the use of approximate values without great risk of poor matching results and permit a more frequent borrowing of parameter values from other studies. A nice collection of results in the literature demonstrating this robustness would be very useful.

The third point covered in the conclusion deals with the effects of bias. We have observed a positive bias in our scoring algorithm. It would be helpful for us to know if the CAMLIS project has identified any consistent bias in their procedure. If so, what explanation do they have for it?

COHEN

The findings of this particular study are based on the results of a match of two files performed by a Government agency. The match was based on an apparently deterministic match procedure using a certain identification number. The provider of such match results should advise clients of error rates and nonmatch results of similar studies. Error rates of such matches should be required as part of publications and presentations in order to give the reader a chance to determine if any biases have resulted due to the matching procedure. This is similar to documenting which computer and software were used when publishing papers based on computer simulation. In this paper, matching determines the study and data base. What is the error rate in the identification number in both files? Errors in deterministic match variables are more important than in probabilistic match variables. The paper does compare the finding of this study with those of other sources to demonstrate that the match was effective.

The question of what impact effective matching algorithms have on the confidentiality of person records was mentioned in the paper. The law provides specific statements on this subject. Some confidentiality problems were discussed in an earlier session. By linking data from several sources, individual records can be identified more easily. In the case of data collection at the Census Bureau, there is an additional concern. The Bureau is a passive collector of data. Cooperation of the respondent is of crucial importance in obtaining reliable information. As the public becomes aware of our ability to link records from several Governmental agencies, response rates to our questionnaires may decrease, become biased, and possibly inaccurate due to the fear of person-record identification. This is in spite of the potential to provide more beneficial information than would exist without the linked records.

ON MATCHING WITH PERSONAL NAMES

J. T. Kagawa, Cancer Research Center of Hawaii
M.P. Mi, University of Hawaii, Honolulu

In the record linkage process, personal names are important matching criteria for comparing documents to identify information belonging to the same individual or family. The discriminating power of the surname, given name, and middle name for linkage varies depending on the frequencies of various possible configurations in the population. Although the total number of possible configurations of personal names is extremely large, the distribution of these configurations are not uniform.

Due to the many people of different nationalities in Hawaii, the name structure has become very diverse and therefore, offers a good opportunity to study the name configurations that are available in the population. Migratory waves of contract laborers and others seeking new opportunities introduced many new names to Hawaii. Often times, names written in Chinese or Japanese characters had to be phonetically translated and anglicized by immigration officers who had little or no knowledge of these languages. This process created further heterogeneity and inconsistencies within names. It is not uncommon to find two or more different names derived from the same character or to find that one surname was actually derived from two completely different characters. Names were also shortened or modified if they were too difficult to pronounce.

In an attempt to develop an optimal strategic approach for computerized linkage of various documentary sources, studies are being conducted to elucidate the variation in personal names in the population. Some pertinent questions to be answered are: 1) how many possible configurations for surname, given name, and middle initials there are in each racial group? 2) how are these configurations distributed in the population? and 3) is there any evidence of time trends in these distributions or name patterns? Preliminary results from the analysis of the 1942-43 Hawaii Population Registration are presented in this report.

MATERIALS AND METHODS

The Population Registration was conducted in Hawaii during 1942-1943 under martial law. There were a total of 439,601 residents registered and fingerprinted. Eight major racial groups were selected including Caucasian, Hawaiian, Portuguese, Chinese, Filipino, Japanese, Puerto Rican, and Korean. The description of each of these racial groups in Hawaii was given previously by Adams (1937), and Lind (1955).

Recorded configurations for surname, given name and middle intials were tabulated separately by sex and race directly from the 1942-1943 population. For each of the eight racial groups, the name configurations were grouped into four types based on the relative frequency in the registration file. The first type was for unique configurations. The next type was for configurations with a relative frequency less than 0.1 percent. The third type was for configurations of fairly frequent appearance equal to or greater than 0.1 percent but less that 1 percent. Lastly, any configuration with a relative frequency of 1 percent or greater was considered in the fourth group. Since the number of configurations was tabulated directly from the data, which were subject to errors in reporting and recording, possible errors could have been included. Errors could have occurred by insertion, substitution, deletion, and switching of one or more alphabetic letters and such an alteration could or could not be a valid configuration. It was therefore assumed for this analysis that most errors are made accidently, presumably at random, and the altered configuration should be unique.

The relative frequency for each of the configurations for surname, first name, and middle initials was calculated. The relative frequency of the ith configuration is $p_i=m_i/M$, where M is the total number of individuals in the population and m_i the number of individuals having the ith configuration. The probability that two individuals randomly sampled from the population would match on the ith configuration is p_i^2. This also approximates the probability of a chance match for the ith configuration when two documentary sources of vital events from the population are brought together for linkage. The sum of these probabilities over all configurations, that is Σp_i^2, is the probability of a chance match on any configuration for a given criterion. Therefore, the greater the total probability, the less discriminating is the linkage criterion among individuals.

RESULTS AND DISCUSSION

Table 1 gives the number of males and females in each racial group. These groups represented 83 percent of the total population in 1942. The Japanese group was the largest, accounting for 37 percent, and larger than any other two groups combined. The Caucasian group ranked second, followed by the Filipino, Portuguese, Chinese, Hawaiian, Puerto Rican, and Korean. These groups and outcrosses among these groups have contributed to the ethnic diversity of Hawaii's present population.

The surname distributions are shown in Table 2. Data on females were not used because of the possible inclusion of their married surname. The total number of surnames varied greatly from one race to another. There were only 241 configurations in the Korean group,

while the Filipino group had approximately 60 times more configurations. There were no common names in the Filipino group based on the relative frequency of 1 percent or greater. There were a total of only five common names representing only a very small proportion of individuals in the Caucasian, Hawaiian, and Japanese groups. Conversely, a large number of individuals shared more than 12 common names in the Korean and Chinese groups. The total probability of chance match also differed markedly among the eight racial groups. The probability of match between two individuals randomly selected from the population was approximately 6 in 10,000 for the Filipinos as compared to the estimate of 850 in 10,000 for the Koreans. In the Korean group, about one-half of the subpopulation shared four common surnames, namely: Kim (22.4%), Lee (15.2%), Park (6.8%), and Chung (4.5%). A high probability equal to 293 in 10,000 was also found for the Chinese group. There were 25 common surnames shared by 68 percent of the Chinese population. The most common Chinese surnames being Wong (8.1%), Lee (6.3%), Chung (5.2%), Ching (5.1%), and Chang (5.1%).

The distribution of the given name for each racial group is shown in Table 3. The ratio of the number of surname configurations to the number of given names varied from race to race. For the Caucasian, Portuguese, and Hawaiian groups, there were a greater number of surname configurations than given names. This relationship was completely reversed for the Chinese and Koreans. The Japanese and Puerto Rican groups had approximately the same number of surnames and given names. As shown in the table, there were very few common given names. However, these common names accounted collectively for a significant portion of each of the subpopulations. For males, the percentage of the population sharing common names was 65 for the Portuguese, 62 for the Hawaiian, 49 for the Puerto Rican, and 46 for the Caucasian. Among the females, the percentage estimates were lower, varying from 25 to 43. In the Chinese, Japanese, and Korean groups the common given names for males and females were of Western origin. Yoshiko, being a common given name of Japanese origin among the Japanese females was the only exception. As shown with surnames, the probability of chance match for the given name as a matching criterion also varied from race to race. The highest value was 323 in 10,000 for the Portuguese males and the lowest was 33 in 10,000 for the Japanese females. The Portuguese and Hawaiians showed the highest probabilities of chance match for both the male and female given names.

The possibility of time trends of selecting given names was also tested based on the 1942 population file. The recorded given names were tabulated by sex and age for each of the eight racial groups. The age groups were 0-19, 20-49 and 50-99. Except for native Hawaiians, individuals with birth years between 1843-1892 were mainly those who immigrated to the islands. The other two age groups were comprised of a mixture of later arriving immigrants and individuals born in Hawaii. A given name was determined popular if the relative frequency was 1.0 percent or greater of the total number of individuals in each race. The distributions based on age groups also showed variations among the different racial groups.

The majority of the given names of the oldest age groups were the names from their native country. With the influence of Western culture, the given names of the younger age groups showed the trend towards adopting the popular English names of the times. It was also observed that the names in the 20-49 age group of the Japanese continued to be largely Japanese. Although still of Japanese origin, the names were quite distinguishable from those of the older generation. Also the selection of Spanish names for the Filipino group prevailed over the three age groups. The popular English male given names among the racial groups remained unchanged throughout the years. The popular female names showed more distinctive periods of rise and decline, which may be attributed to the influence of literary characters and famous people.

Two middle initials were recorded for individuals registered in the 1942 population file. The middle initials distributions are shown in Table 4. The blank configuration represented 44 percent in the males and 37 percent in the females of the eight racial groups analyzed. The blank response indicated either missing information or a valid configuration. Many immigrants to Hawaii from China, Japan, and Korea did not have middle names. Out of the total possible configurations, the Chinese had the largest number of different combinations for both males and females. Middle initials for the Chinese and Korean groups, mostly comprised of double initials, generated a large number of possible configurations. The frequency of uncommon middle initials was reflected in the lower probability of chance match for both of these groups. The frequencies of common middle initials were high in the remaining racial groups.

The observed variations in name patterns among the different racial groups in Hawaii provides a unique testing ground for the study of record linkage methodology. The analysis of the 1942 Hawaii Population Registration file showed that the distributions of the configurations for surnames, given names, and middle initials were definitely nonuniform. Personal names for the different racial groups maintained varying degrees of discriminating power. A study is being planned to analyze the name structure of the present Hawaii population. There has undoubtedly been many more new names introduced into the population.

REFERENCES

1. Adams, R. 1937. Interracial Marriage in Hawaii. New York: MacMillan. pp. 353.
2. Lind, A.W. 1955. Hawaii's People. Honolulu: University of Hawaii Press. pp 121.

Table 1. Size of Subpopulations

Sex	Racial Groups[1]							
	CAU	PTG	HAW	CHI	FIL	JAP	POR	KOR
No. individuals								
Males	34566	15790	7752	16118	40323	84298	4372	3786
Females	25988	15886	7321	12426	10946	78669	3385	2738

[1]CAU = Caucasian; PTG = Portuguese; HAW = Hawaiian; CHI = Chinese; FIL = Filipino; JAP = Japanese; POR = Puerto Rican; KOR = Korean.

Table 2.--Distribution of Surnames by Racial Groups

Sex / Type[2]	Racial Groups[1]							
	CAU	PTG	HAW	CHI	FIL	JAP	POR	KOR
	Number of Configurations							
Males								
Unique	8548	866	896	240	8960	1111	553	101
Rare	4658	546	943	205	5341	3831	199	48
Fair	79	167	231	76	73	192	157	74
Common	1	16	1	25	0	3	15	18
All	13286	1595	2071	546	14374	5137	924	241
	Σp_i							
Males								
Common	0.01	0.29	0.01	0.69	0.00	0.03	0.32	0.72
Other	0.99	0.71	0.99	0.31	1.00	0.97	0.68	0.28
	$\Sigma p_i^2 \times 10^{-2}$							
Males								
All	0.07	0.83	0.15	2.93	0.06	0.20	1.20	8.50

[1]See Table 1.

[2]Unique = single count in the population; Rare = 0.01% - 0.09%; Fair = 0.10% - 0.99%; Common = 1% or greater.

Table 3.--Distribution of Given Names by Racial Groups

Sex / Type[2]	Racial Groups[1]							
	CAU	PTG	HAW	CHI	FIL	JAP	POR	KOR
	Number of Configurations							
Males								
Unique	1512	432	619	3798	2971	4883	467	1664
Rare	905	239	217	1054	1266	3795	168	253
Fair	113	81	71	99	219	153	98	86
Common	20	23	21	15	7	9	22	14
All	2550	775	928	4966	4463	8840	755	2017
Females								
Unique	1866	723	680	2030	1486	1963	393	730
Rare	869	412	235	570	656	1882	108	99
Fair	165	136	116	137	206	228	138	147
Common	14	15	19	17	5	4	18	13
All	2914	1286	1050	2754	2353	4077	657	989
	Σp_i							
Males								
Common	0.46	0.65	0.62	0.23	0.13	0.13	0.49	0.20
Others	0.54	0.35	0.38	0.77	0.87	0.87	0.51	0.80
Females								
Common	0.25	0.32	0.43	0.24	0.09	0.04	0.36	0.23
Others	0.75	0.68	0.57	0.76	0.91	0.96	0.64	0.77
	$\Sigma p_i^2 \times 10^{-2}$							
Males, all types	1.69	3.23	2.82	0.51	0.49	0.40	1.96	0.43
Females, all types	0.77	1.80	1.59	0.57	0.40	0.33	1.39	0.71

[1]CAU = Caucasian; PTG = Portuguese; HAW = Hawaiian; CHI = Chinese; FIL = Filipino; JAP = Japanese; POR = Puerto Rican; KOR = Korean.

[2]Unique = single count in the population; Rare = 0.01% - 0.09%; Fair = 0.10% - 0.99%; Common = 1% or greater.

Table 4.--Distribution of Middle Initials by Racial Groups

Sex / Type[2]	Racial Groups[1]							
	CAU	PTG	HAW	CHI	FIL	JAP	POR	KOR
	Number of Configurations							
Males								
Unique	122	64	50	72	96	52	15	73
Rare	134	22	22	219	24	8	2	59
Fair	1	4	13	120	7	10	7	92
Commmon	20	17	11	8	17	11	16	5
All	277	107	96	419	144	81	40	229
Females								
Unique	118	84	47	91	96	80	18	73
Rare	107	59	37	179	31	78	2	29
Fair	3	7	16	137	7	11	8	89
Common	20	15	9	18	17	12	14	20
All	248	165	109	425	151	181	42	211
	Σp_i							
Males								
Blanks	0.17	0.39	0.38	0.46	0.34	0.60	0.54	0.61
Common	0.81	0.58	0.55	0.10	0.63	0.36	0.43	0.06
Others	0.02	0.03	0.07	0.44	0.03	0.04	0.03	0.33
Females								
Blanks	0.14	0.30	0.23	0.20	0.39	0.49	0.43	0.31
Commmon	0.83	0.64	0.70	0.32	0.57	0.45	0.52	0.39
Others	0.03	0.06	0.07	0.48	0.04	0.06	0.05	0.30
	$\Sigma p_i^2 \times 10^{-2}$							
Males								
Blanks	2.83	15.35	14.67	21.16	11.57	35.36	28.60	37.13
Common & Others	4.12	2.35	10.46	0.28	2.92	1.60	1.54	0.19
All	6.95	17.70	25.13	21.44	14.49	36.96	30.14	37.32
Females								
Blanks	1.81	9.12	5.25	3.81	15.34	23.79	18.30	9.89
Common & Others	5.25	3.54	14.88	0.96	2.36	2.12	2.69	1.02
All	7.06	12.66	20.13	4.77	17.70	25.91	20.99	10.91

[1]CAU = Caucasian; PTG = Portuguese; HAW = Hawaiian; CHI = Chinese; FIL = Filipino; JAP = Japanese; POR = Puerto Rican; KOR = Korean.

[2]Unique = single count in the population; Rare = 0.01% - 0.09%; Fair = 0.10% - 0.99%; Common = 1% or greater.

SURNAME BLOCKING FOR RECORD LINKAGE

F. Quiaoit, Cancer Research Center of Hawaii, and
M.P. Mi, University of Hawaii, Honolulu

In the linkage between two documentary sources, each record from one source is compared with all the records in the other source. For one-file linkage involving a single source, each record is compared with all other records except itself. In either case, the number of such pair-wise comparisons becomes extremely large even if the size of the documentary source is moderate. The fact that only a small fraction of these comparisons are meaningful emphasizes the need for the grouping of records based on one or more selected items of identifying information. This is known as blocking. Once blocks are formed, the comparison of records is only made between the two corresponding blocks for two-file linkage or within the block for one-file linkage.

In principle, any identifier may be used as a blocking criterion. Surname is often selected for this purpose. Blocking may be made on the whole or part of the surname configuration. The use of a phonetic code on the surname for blocking has become popular in many applications. The objective of the present study was to evaluate the performance of several blocking methods based on prevalent name patterns in various racial groups in a multi-ethnic population, and to test the effects of blocking on linked pairs in which one or both records had known reporting or recording errors in the surname field.

MATERIALS AND METHODS

Data on surnames from the complete 1942-43 Population Registration in Hawaii were used. There were a total of 439,601 individuals registered and fingerprinted under martial law. Eight major racial groups were selected including Caucasian, Portuguese, Hawaiian, Chinese, Filipino, Japanese, Puerto Rican, and Korean. All recorded surname configurations for male subjects were analyzed in the present study. Two methods, namely: the New York State Identification and Intelligence System (NYSIIS) and the Russell's Soundex system were chosen to pre-code surnames phonetically. Under each method, records were blocked with the same code. These two systems were compared specifically to the other five methods of blocking, namely, by the whole surname, first character of surname, first two, three, or four characters of surname, respectively. Criteria such as the total number of blocks formed, distribution of block size, and surname information in matching were used for evaluation.

A set of known linked record pairs was obtained from the linkage project between the 1942 Population Registration file and the death file (1942-79) in Hawaii. It consisted of all male subjects aged 60 and over in the 1942 population who died during the 38-year period from 1942 to 1979. A total of 11,367 linked pairs were established by computer as well as by manual search (Mi et al., 1983). Pairs, in which recorded surname and first name were switched, were excluded. There were 672 pairs with various error conditions in surname. The concordance rate of each method, which is the percentage of record pairs that were properly placed in the same block regardless of these errors, was used for comparison.

RESULTS AND DISCUSSION

The number of male subjects in the 1942 Population Registration is shown for each racial group in Table 1. The total number of recorded configurations for surname varied greatly among racial groups ranging from only 241 in the Korean group to 14,374 among the Filipino. The average number of individuals possessing the same surname varied from 2.6 for the Caucasian group to 29.5 for Chinese men. The value for each racial group was also the average block size when blocking was based on the whole surname of twelve characters. Most of the surname configurations were unique, having only a single representation in the population. These unique configurations included rare spelling variations, and errors in reporting and recording. When a part of the surname was used for blocking, records having the same leading characters in their surname fields were grouped together. As shown in Table 1, the number of blocks increased from an initial maximum of 26, based on the first character of the surname, to several hundreds or thousands using more leading characters for blocking. However, the magnitude of increase was not linear for each additional character used, and varied from one race to another. The distribution of blocks by size also changed. When the whole surname was used for blocking, most blocks were small with 10 or less records. If blocking was based on the first character of surname, the block size increased tremendously. If more leading characters were used, the number of records in each block decreased as expected. The performance of the first four characters of surname for blocking was comparable to the NYSIIS and Soundex method in the percentage distribution of blocks by size in all groups except the Chinese and Koreans. The NYSIIS and Soundex method produced a much higher percentage of large blocks of over 50 records in the Chinese and Korean groups. This was because almost all the Chinese and Korean surnames were five characters or less in length.

It should be emphasized that block size is an important consideration in the choice of a blocking method for linkage. Since the number of pair-wise comparisons is equal to the product of the size of two corresponding blocks in two-file linkage and to the product of the block size and block size minus one in one-file

linkage, a larger block size will greatly affect the cost of a linkage.

The other criterion which deserves attention is the loss of surname information in matching by blocking. Suppose that there is no blocking and the whole documentary source or file is used as a giant block for pair-wise comparison. The amount of information provided by surname in matching is approximately $1 - \Sigma p_i^2$ where p_i is the relative frequency of the i^{th} surname configuration and $\Sigma p_i = 1$. The squared term represents the probability of chance match on the i^{th} configuration. When summed over all configurations, the squared term gives the total probability of chance match in surname. The exact probability of chance match is $1 - \Sigma p_i p_i'$ in the two file linkage where p_i' is the relative frequency of the ith configuration in the second source. If all individuals have the same surname, that is, $p_i = 1$, every record pair must agree on surname and the total probability of chance match reaches the maximum of 1. Under this special condition, surname clearly provides no information. On the other hand, if each individual record has a different surname, the probability of chance match is minimal and the amount of information provided by surname reaches the maximum. When blocking is made based on surname (a part or whole), the newly structured block consists of records of one or more surnames, each with the relative frequency of p_{ij}, the j^{th} surname within the i^{th} block. The relative frequency of the i^{th} block is q_i, and the probability of chance match for records with the i^{th} blocking criterion is q_i^2. The probability of chance match on surname within newly structured blocks is $\Sigma\Sigma p_{ij}^2/\Sigma q_i^2$, and the amount of information of surname in matching is estimated by $1 - \Sigma\Sigma p_{ij}^2/\Sigma q_i^2$. Suppose that the whole surname is used for blocking. Because each block is characterized by a different surname, obviously $\Sigma\Sigma p_{ij}^2/\Sigma q_i^2 = 1$, therefore surname is no longer informative and provides no discrimination among records within any block in which pair-wise comparisons are made.

The average and maximum number of surnames per block and the estimates of surname information in matching under various blocking methods are given in Table 2. When blocking is based on the first character, the amount of surname information was generally high except for the Korean group. The probability of chance match on surname was estimated to be 0.085, the highest among the eight racial groups studied (Kagawa and Mi, 1985). The amount of information decreased rapidly, particularly in the Chinese group, as the number of leading characters for blocking increased. When blocking is based on the NYSIIS and Soundex codes, the amount of information was close to those estimates derived from the blocking based on the first four characters in several racial groups. These phonetic coding methods seemed to be desirable especially for the Chinese and Korean groups, but not for the Japanese. The concordant rate was defined as the percentage of total pairs in which both members were blocked concordantly by a given method. Table 3 gives the estimates of the concordant rate for the four selected methods. The rate over all racial groups was 56.7, 43.9, 56.4, and 64.9 percent, respectively, for blocking based on the first three characters, first four characters, NYSIIS code, and Soundex code of surname. Both NYSIIS and Soundex methods consistently produced a high concordant rate in all racial groups. Because Chinese and Korean surnames are generally short (composed of three to five characters), errors would have to occur in the first few characters. It was anticipated that blocking based on the first three and four characters would not be highly desirable. Among the 672 linked pairs, 176 linked pairs were found to be concordant by all four methods. Erroneous conditions at the end of the surname were not detected even by the modified NYSIIS system. There were 87, 106, 98, 86 and 119 record pairs in which errors occurred in the first, second, third, fourth, and between the fifth and eighth positions, respectively. Therefore, it may be concluded that in a population where spelling variations or errors in reporting and recording usually occur after the fourth position of the surname, these four methods would perform equally well for blocking. Otherwise, NYSIIS and Soundex should be more promising than methods which are based on the use of leading characters.

REFERENCES

Mi, M.P., J.T. Kagawa, and M.E. Earle. 1983. An operational approach to record linkage. Meth. Inform. Med. 22:77-82.

Kagawa, J.T. and M.P. Mi. 1985. On matching with personal names, pp. 269-273 in this volume.

Table 1. Block Characteristics by Methods

Item	Racial Groups[1]							
	CAU	PTG	HAW	CHI	FIL	JAP	PUR	KOR
Number of Male Subjects	34566	15970	7752	16118	40323	84298	4372	3786
Blocking by Complete Surname								
Number of Blocks	13286	1595	2071	546	14374	5137	924	241
Block Size Distribution, %								
1 - 10	96.7	85.1	93.4	77.5	96.6	73.8	92.3	80.1
11 - 50	3.0	10.5	6.4	14.6	3.0	19.9	6.5	13.7
51 - 100	0.2	2.6	0.1	2.0	0.2	3.1	0.8	4.6
101 - 500	0.1	1.6	0.0	5.5	0.1	3.1	0.4	0.8
501 - 1000	0.0	0.2	0.0	1.1	0.0	0.2	0.0	0.8
> 1000	0.0	0.0	0.0	0.4	0.0	0.0	0.0	0.0
Average Size	3	10	4	30	3	16	5	16
Maximum Size	397	550	97	1313	289	1022	288	848
Blocking by First Character of Surname								
Number of Blocks	26	26	23	24	26	25	24	22
Block Size Distribution, %								
1 - 10	3.9	11.5	17.4	12.5	3.9	16.0	8.3	31.8
11 - 50	3.9	19.2	26.1	12.5	3.9	4.0	25.0	27.3
51 - 100	3.9	3.9	21.7	0.0	3.9	8.0	8.3	9.1
101 - 500	15.4	15.4	17.4	45.8	23.1	12.0	50.0	18.2
501 - 1000	15.4	23.1	13.0	16.7	15.4	8.0	8.3	9.1
> 1000	57.7	26.9	4.4	12.5	50.0	52.0	0.0	4.6
Average Size	1329	614	337	672	1551	3372	182	172
Maximum Size	3474	1922	4214	4157	4539	11229	811	1055
Blocking by First 2 Characters of Surname								
Number of Blocks	280	155	142	113	232	178	144	82
Block Size Distribution, %								
1 - 10	34.3	36.1	62.0	39.8	35.8	32.6	58.3	65.9
11 - 50	21.8	26.4	24.7	27.4	17.2	18.0	24.3	15.9
51 - 100	10.0	12.3	4.2	8.0	12.1	10.1	9.7	12.2
101 - 500	28.6	18.7	7.8	18.6	26.3	18.5	7.6	2.4
501 - 1000	5.0	5.8	0.7	3.5	4.7	6.7	0.0	3.7
> 1000	0.4	0.7	0.7	2.7	3.9	14.0	0.0	0.0
Average Size	123	103	54	143	174	474	30	46
Maximum Size	1008	1128	2869	4153	2809	6321	422	872

See note at the end of the table.

Table 1. Block Characteristics by Methods (Continued)

Item	Racial Groups[1]							
	CAU	PTG	HAW	CHI	FIL	JAP	PUR	KOR
Blocking by First 3 Characters of Surname								
Number of Blocks	2212	655	491	354	1880	835	471	179
Block Size Distribution, %								
1 - 10	68.6	68.8	75.6	68.1	66.5	50.1	84.1	77.1
11 - 50	24.5	19.1	18.3	19.5	23.7	24.9	12.3	14.5
51 - 100	3.8	6.6	3.1	3.1	4.9	7.3	2.3	5.6
101 - 500	3.1	4.9	3.1	6.8	4.6	12.7	1.3	1.7
501 - 1000	0.0	0.6	0.0	2.5	0.2	2.9	0.0	1.1
> 1000	0.0	0.0	0.0	0.9	0.0	2.2	0.0	0.0
Average Size	16	24	16	46	21	101	9	21
Maximum Size	471	575	487	1378	740	3879	300	849
Blocking by First 4 Characters of Surname								
Number of Blocks	6941	1112	974	490	5719	1818	709	229
Block Size Distribution, %								
1 - 10	90.6	79.9	82.3	75.9	85.9	61.1	89.0	79.0
11 - 50	8.2	13.1	15.4	13.9	11.9	24.5	9.0	14.9
51 - 100	0.9	4.1	1.4	2.7	1.5	5.9	1.4	4.4
101 - 500	0.3	2.6	0.8	5.9	0.6	6.9	0.6	0.9
501 - 1000	0.0	0.3	0.0	1.0	0.0	0.7	0.0	0.9
> 1000	0.0	0.0	0.0	0.6	0.0	0.8	0.0	0.0
Average Size	5	14	9	33	7	46	6	17
Maximum Size	401	554	255	1322	422	3838	300	848
Blocking by NYSIIS								
Number of Blocks	7293	1025	631	209	6526	1922	649	89
Block Size Distribution, %								
1 - 10	91.7	79.4	80.0	71.8	87.6	55.8	88.4	68.5
11 - 50	7.1	12.5	13.8	12.4	10.7	26.4	9.2	14.6
51 - 100	0.8	4.6	4.3	3.3	1.2	6.8	1.5	10.1
101 - 500	0.4	3.2	1.9	7.7	0.6	10.0	0.8	4.5
501 - 1000	0.0	0.3	0.0	2.9	0.0	0.8	0.0	2.3
> 1000	0.0	0.0	0.0	1.9	0.0	0.2	0.0	0.0
Average Size	5	16	13	77	6	44	7	43
Maximum Size	414	586	406	2311	366	1114	300	965

See note at the end of the table.

Table 1. Block Characteristics by Methods (Continued)

Item	Racial Groups[1]							
	CAU	PTG	HAW	CHI	FIL	JAP	PUR	KOR
	Blocking by Soundex							
Number of Blocks	2864	813	441	161	2779	948	555	86
Block Size Distribution, %								
1 - 10	72.9	73.8	77.1	60.9	66.8	43.1	85.8	62.8
11 - 50	22.1	16.0	15.7	16.2	26.8	26.9	11.5	16.3
51 - 100	3.6	5.8	3.6	4.4	4.8	9.5	1.6	12.8
101 - 500	1.5	4.1	3.0	13.0	1.6	15.5	1.1	5.8
501 - 1000	0.0	0.4	0.7	3.7	0.0	4.3	0.0	2.3
> 1000	0.0	0.0	0.0	1.9	0.0	0.6	0.0	0.0
Average Size	12	20	18	100	15	89	8	44
Maximum Size	449	587	774	2275	352	1395	300	885

[1]CAU = Caucasian; PTG = Portuguese; HAW = Hawaiian; CHI = Chinese; FIL = Filipino; JAP = Japanese; PUR = Puerto Rican; KOR = Korean.

Table 2. Surname Characteristics within Blocks

Blocking Criterion	Racial Groups[1]							
	CAU	PTG	HAW	CHI	FIL	JAP	PUR	KOR
Average Number of Surnames Per Block								
First character	511	61	90	23	553	206	39	11
First 2-characters	48	10	15	5	62	29	6	3
First 3-characters	6	2	4	2	8	6	2	2
First 4-characters	2	1	2	1	3	3	1	1
NYSIIS	2	2	3	3	2	3	1	1
Soundex	5	2	5	3	5	5	2	2
Maximum Number of Surnames Per Block								
First character	1407	184	961	73	1553	834	113	31
First 2-characters	352	100	632	53	962	376	48	22
First 3-characters	178	31	118	12	269	210	23	23
First 4-characters	37	10	60	8	117	89	10	10
NYSIIS	51	13	71	39	52	70	9	
Soundex	68	16	136	24	74	71	15	15
Surname Information in Matching								
First character	0.99	0.89	0.99	0.81	0.99	0.98	0.86	0.47
First 2-characters	0.94	0.70	0.99	0.70	0.97	0.94	0.63	0.29
First 3-characters	0.75	0.32	0.93	0.20	0.85	0.84	0.34	0.08
First 4-characters	0.40	0.14	0.78	0.07	0.57	0.79	0.18	0.02
NYSIIS	0.48	0.17	0.90	0.57	0.46	0.43	0.20	0.25
Soundex	0.64	0.20	0.95	0.54	0.61	0.64	0.27	0.14

[1] CAU = Caucasian; PTG = Portuguese; HAW = Hawaiian; CHI = Chinese; FIL = Filipino; JAP = Japanese; PUR = Puerto Rican; KOR = Korean.

Table 3. Concordant Rate of Blocking

Blocking Method	Racial Groups[1]								
	Total	CAU	HAW	CHI	FIL	JAP	PUR	KOR	OTH
	Number of Linked Pairs with Errors in Surname								
	672	167	77	28	78	222	54	10	36
	Concordant Rate (%)								
First 3-characters	56.7	56.3	62.3	32.1	48.7	54.5	79.6	50.0	63.9
First 4-characters	43.9	50.3	52.0	14.3	32.1	41.4	59.3	20.0	44.4
NYSIIS	56.4	60.5	57.1	57.1	59.0	51.4	70.4	40.0	44.4
Soundex	64.9	66.5	53.3	71.4	71.8	65.3	75.9	50.0	44.4

[1]CAU = Caucasian; PTG = Portuguese; HAW = Hawaiian; CHI = Chinese; FIL = Filipino; JAP = Japanese; PUR = Puerto Rican; KOR = Korean; OTH = All Others.