THE METHODOLOGY OF COMPUTER LINKAGE OF HEALTH AND VITAL RECORDS

David M. Nitzberg, Harvard School of Public Health
Hyman Sardy, Brooklyn College

Introduction

Record linkage is a process whereby records pertaining to the same individual from two or more files are brought together to form a combined record. In order to link records successfully they must be matched on identifying information that is common to the files to be linked and that has (a) high discriminating power, (b) low probability of change during an individual's lifetime, and (c) low likelihood of being recorded erroneously. Such information, together with rules for matching the records, as well as criteria for deciding when a pair of records match sufficiently to be declared a linkage, determine a linkage procedure. Its success is measured by

(1) its speed and cost,
(2) the number of valid linkages produced,
(3) the number of false linkages produced, and
(4) the number of valid linkages missed.

There is, of course, no a priori knowledge of which individuals have records in more than one of the files. Furthermore, there is often only a limited amount of common identifying information available. Also complicating the problem is the presence of noise, that is to say errors, such as surname misspellings and age discrepancies in the information which is available.

This paper briefly discusses our research to develop efficient computer linkage techniques. We are using the IBM 7010 data processing computer at the National Center for Health Statistics to perform actual linkage operations between cohort and death records. However, since our work is still in progress we can only indicate some of the goals we have set for ourselves and present some preliminary results.

Background

Densen and Shapiro (1) have pointed out that the limitations inherent in using existing data (i.e., information collected routinely, not for specific research purposes) can often be overcome by combining the information contained in records from several sources. Moriyama (2) has documented the value of vital records in health research, especially when such records are linked with others.

Computer linkage research has already been carried out (and is continuing) in Canada, England, and the United States (3-11). This research has demonstrated that when there is an abundance of identifying information common to the files to be linked, large-scale computer linkage is possible -- even though noise exists and people do not have unique identity numbers. The research we are conducting, however, is geared to studying the nature of the problem when only limited amounts of identifying information are available. We hope this will lead to the development of computer techniques which will permit linkage prior to the time when widespread use of identity numbers will simplify the methodological problems existing today.

Death Clearance

We chose to do our research in the area of death clearance since it is frequently performed in health research and is methodologically typical of most linkage operations. Death clearance is a process whereby a file of records pertaining to a group of individuals, a cohort file, is linked with a file of death certificates to determine which individuals have died and to extract information from their death certificates. Death clearance by computer would facilitate long-term follow-up studies since large study cohorts could then be placed under automatic surveillance for mortality at frequent intervals, with high precision and at relatively low cost. A contract in 1963 with the New York City Department of Health* by the National Center for Health Statistics (NCHS)** to develop computer death clearance techniques has made the research possible.

The death file we are using is composed of magnetic tapes made from the death index cards routinely keypunched by the New York City Health Department from death certificates. All deaths occurring in the city, as well as deaths of city residents reported to New York City as occurring outside the city, are included. Our file covers the years 1961-1963 (inclusive); Appendix I shows the format of the 1963 cards.

On the tapes used in our computer runs we added the Social Security numbers

*Dr. Paul M. Densen, Deputy Commissioner of Health, New York City, is project director for this contract.

**We are grateful to Mr. S. Binder, Chief of the Data Processing Division of the NCHS for his help.

of the deceased when these were given on the actual death certificates (they are not now keypunched on the index cards) in order to study the value of this unique identity number. One of the several cohorts we are linking has this information recorded. Even though at present many of the deceased do not have Social Security numbers -- 59%* of recent New York City death certificates do not have this item recorded -- an ever increasing proportion of our population will. Furthermore, widespread use of these numbers is being fostered by requirements of the Internal Revenue Service. All of this indicates that even at present the Social Security number is an item of identifying information which may be worth recording and using in linkage operations.

We will limit our discussion here to what we are doing with the largest cohort we are linking, the coronary heart disease (CHD) population of the Health Insurance Plan of Greater New York, better known as HIP. HIP has placed about 120,000 people under observation for specific manifestations of coronary heart disease in order to study its incidence and prognosis. Their study cohort includes all persons 25-64 years of age enrolled in 12 of their medical groups. The prepaid group medical practice setup of HIP with its central record system makes such an undertaking possible (13). Nevertheless, mortality for so large a group over a number of years remains a problem since it is not known what percentage of its members' deaths are reported to them. Obviously, death clearance by manual procedures for such a large group for a number of years is out of the question. With HIP's 1961 enrollment cards (Appendix II shows the format of these records) as our cohort file, we are performing a death clearance operation using computer techniques. This is being done for all members (176,481), not just those between 25 and 64 years old, of the medical groups in the HIP-CHD study for 1961, 1962, and 1963.

Methodology

At present, we have in operational form IBM 7010 programs** with which to explore death clearance techniques. Names are coded phonetically by a Soundex program according to the Russell-Soundex system developed by the Library Bureau of Remington Rand and based on the work, early in this century, of Mr. R. C. Russell (14).

A Matching program brings together records from the HIP and death files having the same Soundex code. The program then compares each of these HIP-death pairs to see whether there is agreement on items of identification which are common to both. For numerical information, it is possible to specify matching rules which permit slight discrepancies (e.g., requiring that $|\text{HIP age} - \text{Death age}| \leq k$ for some value of k)* to be tolerated when deciding whether agreement exists between the fields. For alphabetic information, however, exact agreement is necessary, except in the case of surname prefixes such as Von, Mc, D', etc., which are often keypunched in various ways with respect to the rest of the name. Our program treats as identical such variations as: MC COY and MCCOY, VON MEER and VONMEER, D'AMOS and DAMOS, etc. This class of alphabetic discrepancies became apparent after our initial runs, but were overcome easily by making simple changes to our original program.

Noise suppression techniques such as allowing ± k year differences when matching ages and ignoring blanks and apostrophes appearing within surnames are straightforward. Other sources of noise, as, for example, name misspellings and address spelling variations, are not so easily suppressed. The value of the Soundex name code is that it enables one to bring together records containing similar names on the assumption that many misspellings lead to similar sounding variations of the true spelling. The problem remains, however, of defining how "dissimilar" names can be and still be declared matching -- how can we overcome the loss of potential linkages because of misspellings without thereby creating many false linkages? The results we are beginning to accumulate indicate that, although Soundex is a gross noise filter, e.g., such different and commonly occurring surnames as Jones and James have the same code of J520, it does help overcome many name misspellings. As a result of our experience, we hope to be able to estimate just how many are overcome validly in death clearance operations. We also hope

*Estimated from a sample of 4196 death index cards, and agreeing with a sample we collected of 816 New York City death certificates, for 1961-1963.

**The programs were written by Dr. H. S. Levine (of HIP) and Mr. J. Hayden, especially for this project under the NCHS contract.

*We will get a distribution of age differences for the valid linkages so that an optimum k for death clearance can be estimated.

our results will enable us to modify Soundex so that it can be made a finer, and hence more efficient, noise filter for names.

The Matching program produces a listing of record pairs meeting and surpassing a set of minimum matching criteria specified at the beginning of the computer run. In the case of the HIP-CHD cohort we compared 176,481 people against a total of 281,208 death records in a four-hour computer run. This resulted in 89,306 pairs representing 37,777 different HIP members* meeting our minimum matching criteria, which were exact agreement on Soundex code of first and last names and age agreement within five years. Table 1 gives the number of record pairs matching on the fields specified by asterisks. It should be noted that the numbers given in the table are not cumulative -- that is, the absence of an asterisk means the pairs do not match on that field. Also, the heading "age ± 5 years" means that age agrees within five years but not within one year.

The numbers presented in the table are preliminary. They are based on the initial production run of the programs during which invalid characters due to keypunching errors, blanks and apostrophes in surnames, and several other minor difficulties prevented the matching of a small number of records. A supplementary computer run for these records is planned and our figures will be adjusted accordingly. The changes, however, will not alter the table significantly since the number of records involved is relatively small. Since it is possible to partition the HIP file and do death clearance on each part separately without affecting the accuracy of the operation (although efficiency might be degraded thereby), we have delayed doing this "mopping up" process until we could do it for all of our cohorts at once.

In all, there were 700,738 pairs matching on Soundex, of which 111,252 pairs matched exactly on first and last names as well but whose ages were not within at least five years of one another (and so were not put out by the Matching program). This means there were a total of 132,123 pairs matching exactly on first and last names. This clearly shows that name alone is a very poor identifier of people.

The 7,036 pairs matching exactly on name and age within one year are most interesting since they indicate very clearly the problem of trying to link with insufficient identifying information common to both files. Since only about 3,000 deaths are expected -- of which only about 15% will not be among these 7,036, mainly because of noise -- it is apparent that name and age alone lack the discriminating power upon which to decide which pairs constitute valid linkages.

Looking further at the figures we see that there were 26,075 pairs matching on Soundex of first and last names plus age within one year, and that 7,036 of these matched exactly on first and last names also. This indicates two things: first, that the Soundex code increased the number of possible pairs by a factor of 3.7 (the 26,075 pairs are reduced to 7,036 when exact full name matching is required); and second, that items of identifying information used in conjunction with one another very quickly narrowed down the field of possible pairs (e.g., from 176,481 x 281,208 possible pairs to 132,123 pairs by use of first and last names, and then to 7,036 pairs by use of age to within one year in addition). This seems to indicate that if there were no noise in our data, then one other good variable (e.g., the maiden name of an individual's mother) would probably reduce the 7,036 pairs to the correct 3,000 or so pairs with few false positives. Since noise exists however, schemes such as Soundex must be used to overcome it. This raises the number of possible pairs, which in turn might be reduced if yet another identifying variable were available.

*A given HIP or death record may appear in more than one pair. It is not uncommon for different people with the same common names, like Mary Smith for example, to appear in the HIP file and be within five years of age of several different deceased persons also with the same name. This gives rise to more pairs than the number of different people involved.

Table 1

Number of HIP-Death Pairs Matching
on the Fields and Criteria Specified

Exact Soundex Code of First and Last Names	Exact First Name	Exact Surname	Age ± 5 Years	Age ± 1 Year	Number of Pairs
*			*		14,426
*	*		*		30,818
*		*	*		4,152
*	*	*	*		13,835
*				*	5,615
*	*			*	11,772
*		*		*	1,652
*	*	*		*	7,036
					89,306

Conclusion

By clerically eliminating the pairs that do not seem to constitute valid linkages* and validating the remaining pairs by using information which could not be used by our present Matching program (e.g., address, since only the death file contained this field), we hope to find most of the deaths which did, in fact, occur.

Since this work is exploratory in nature, we wish to find as many of the deaths as possible even though we cannot do it by completely automated means now. HIP has been notified of a large proportion of the deaths of its members, so they will be able to tell us about a number of the deaths our computer-manual death clearance operation missed. Likewise, we will be able to give them a list of deaths about which they had no knowledge and which they will verify through their normal channels of contact with their members. Since the two procedures, HIP's and ours, for finding deaths are independent, we will then be able to estimate the number of deaths we both missed. If this number is small (less than 5% of the total deaths), as we expect it will be, then we will be able to analyze why our death clearance techniques did or did not link the deaths we know about, knowing that the unfound deaths cannot greatly bias these results.

This analysis, together with our other work, should lead to useful estimates of the parameters of death clearance procedures which, in turn, should lead to a better understanding of the methodology of record linkage and to computer techniques for automating the linkage operation. Whether this is actually so, remains to be seen. We feel confident, however, that our attempt to follow a cohort of 176,000 people for deaths for three years with the aid of computer techniques currently available will prove successful.

*This is being done by: (a) Human judgment to decide whether two names having the same Soundex code, but not agreeing exactly, were different names or variants of one another that could reasonably be expected to arise from misspelling or recording error; (b) the use of such other information as date of birth, spouse's name, address, etc., that appears on only one record of the pair and on HIP records other than enrollment cards or the death certificates themselves. Time does not permit fuller discussion of this procedure here.

REFERENCES

1. Densen, P.M. and Shapiro, S. Research needs for record matching. 1963 Proceedings of the Social Statistics Section of the American Statistical Association, pp. 20-24.

2. Moriyama, I.M. Uses of vital records for epidemiological research. J. Chron. Dis. 17:889-897, 1964.

3. Newcombe, H.B., Kennedy, J.M., Axford, S.J. and James, A.P. Automatic linkage of vital records. Science 130(3381):954-959, Oct. 16, 1959.

4. Newcombe, H.B. and Kennedy, J.M. Record linkage. Communications of the Association for Computing Machinery 5(11):563-566, Nov. 1962.

5. Newcombe, H.B. and Rhynas, P.O.W. Family linkage of population records. The use of vital and health statistics for genetic and radiation studies. Proceedings of the Seminar sponsored by the U.N. and the W.H.O. (held in Geneva, 5-9 Sept. 1960), U.N. Publication, Sales No. 61 XVII.8:135-154, 1962.

6. Kennedy, J.M. The use of a digital computer for record linkage. Ibid., pp. 155-159.

7. Binder, S. Present possibilities and future potentialities. Ibid., pp. 161-170.

8. Phillips, W., Bahn, A.K. and Miyasaki, M. Person-matching by electronic methods. Comm. ACM 5(7):404-407, July 1962.

9. Phillips, W. and Bahn, A.K. Experience with computer matching of names. 1963 Proc. Social Statistics Section of the A.S.A., pp. 26-38.

10. Acheson, E.D. Oxford record linkage study: A central file of morbidity and mortality records for a pilot population. Brit. J. Prev. Soc. Med. 18: 8-13, 1964.

11. DuBois, N.S.D. A document linkage program for digital computers. Behavioral Science 10(3):312-319, July 1965.

12. Shapiro, S., Weinblatt, E., Frank, C., Sager, R. and Densen, P. The HIP study of incidence and prognosis of coronary heart disease: Methodology. J. Chron. Dis. 16:1281-1292, 1963.

13. Shapiro, S. Research in prepaid group practice programs. Amer. J. Public Health 54(12):2040-2050, Dec. 1964.

14. Waugh, W.C. Russell-Soundex (An address given by Mr. Waugh). Remington Rand, Inc., Library Bureau, SP-V3012 (undated).

APPENDIX I

Format of the 1963 New York City Death Index Cards

Columns	Field	
1-5	Death Certificate number	
6-9	Month, day, and year of death	
10	Borough where death occurred	
11	Type of institution where death occurred	
12-14	Institution number	
15	Borough of residence	If non-resident of N.Y.C., state and county of residence
16-19	Health area of residence	
20	District of residence	
21	Blank	
22	Sex	
23	Color or race	
24	Marital status	
25-27	Age at death (2 columns for age and one for units -- hours, day, months, years -- of age)	
28	Nativity of deceased (U.S., Puerto Rico, foreign)	
29-32	Cause of death	
33	Operation?	
34	Attendant at autopsy	
35	Religious affiliation of cemetery where buried	
36-40	If death due to accident, type and line number of special accident report, and borough and district of occurrence of accident	
41-44	Medical examiner case number, if any	
45-63	Address of deceased	
64-80	Name of deceased (surname, first name)	

APPENDIX II

Format of the HIP Enrollment Cards

Columns	Field
1-19	Name (surname, first name)
20-22	Blank
23	R = Renewal
24-26	Original Entry Date (month, year) into HIP
27-28	Blank
29-36	Certificate Number (policy number)
37-38	Borough of Residence
39-42	Medical Group
43-46	Blank
47	Associated Hospital Service Class
48	Blank
49	HIP Class
50-51	Number of Persons Covered on Certificate (or Policy)
52-56	Contract Number
57-61	Effective Date of Contract (month, day, year)
62-63	Sex and Family Status
64	Blank
65-67	Month and Year of Birth
68-80	Blank