RESEARCH NEEDS FOR RECORD MATCHING

Sam Shapiro, Health Insurance Plan of Greater New York
and
Paul M. Densen, New York City Department of Health

The value for research of bringing together information on two or more unrelated records for an individual or family has long been recognized. Indeed, the present article could just as readily have been prepared 10 years ago as today. In fact a proposal for establishing a permanent record matching system on a national scale was advanced by Dunn, 15-20 years ago.(1) But this proved to be too comprehensive and complex and fell by the wayside.

With the passage of time, record matching has been resorted to with increasing frequency and it is likely that the use of this procedure will continue to expand. It is the purpose of this paper to examine some of the substantive and methodological circumstances that have made "record matching" a vital issue for many research interests. Illustrations are drawn primarily from research that has been completed or is in progress.

Definition of record matching and record systems under consideration

For the current discussion, record matching refers to the process of collating the records on specific individuals from two or more sets of records collected through independent means or for different reasons. The objective of this procedure is to combine the information on these records for research purposes. Usually at least one of the files of records being searched is large, relates to the total population or a defined segment of it, and is the result of a regular on-going collection process. Such a file is ordinarily designed to last for a long time.

There are today several networks of records of this type. Prominent among them are the Decennial Census with its broad range of sociological, economic and housing information regarding a sample of the population; the vital registration system with its basic facts of birth, death, marriage and divorce; the Old Age and Survivors Insurance records which now cover well over 90 per cent of the gainfully employed and will eventually provide a history of the working life of individuals employed in particular industries and geographic areas; the Internal Revenue Service's income tax returns; and the health insurance companies' claims and medical records which contain data on hospital episodes involving at least 75 per cent of the population and information about medical care outside the hospital for an expanding sector of the population. To these could be added the networks of life insurance, welfare, hospital, and educational records and the myriad of specialized sets of records that may exist in a particular community.

Substantive needs for record matching

In a sense, the growth, organization and strengthening of these vast repositories of records is a reflection of the increasing complexity of our society and of the economic, social and health problems it faces. The record systems are of course repeatedly being treated as closed, self-contained sources of information and mined for data relevant to outstanding problems of the day. But for some of the most pressing questions, this approach is inadequate. The reasons are quite apparent. The range of observations collected through a single record system is restricted and frequently the issue under study concerns events for which the system was not designed to collect information in the first place. Many of these problems or limitations are, however, overcome when data in two or more existing record systems can be made to supplement each other or when a routinely collected set of records is used as an adjunct to records generated for specific research purposes. These two types of situations are discussed more fully below in relation to the issues that led to the use of a record-matching procedure.

Studies dependent on multiple sets of established record systems

In the health field extensive use for research purposes has been made of two or more sets of established records. Many health departments have for years been matching live birth and infant death records to enlarge knowledge about risk factors affecting mortality in infancy. This activity is now so well established in many states that it is not always appreciated that it was undertaken in response to a major change in the nature of the infant mortality problem. As long as infectious diseases were the primary causes of death among the newborn and progress was being made in reducing the infant mortality rate, requirements for data could in large measure be satisfied from the items on the death record. These included race, sex, geographic area, cause and age at death.

As postnatal environmental conditions receded in importance, pressures developed for learning more about the influence of prenatal and natal circumstances on infant mortality. Despite the fact that the vital records are far from perfect instruments for comprehensive inquiries into these matters, it was recognized that through the mechanism of matched records a broad extension of relevant information could be quickly and economically accomplished. To the usual set of variables on the death records, it became possible through matching to add items on operative procedures at delivery, characteristics of hospital where the birth occurred, birth weight, ages of

parents, birth order, etc., from the birth record for new types of investigations.(2)

One of the most intriguing exploitations of information in completely unrelated record systems has been carried out by Mancuso and Coulter. Their general interest was in determining "how various cohort industrial population groups, differing primarily in the exposure of the working environment in the same and different geographic areas, vary among each other in morbidity and mortality patterns".(3) This was pursued to good effect in a study of mortality risks among employees in an asbestos plant. The investigators used the BOASI records to establish a cohort of persons working in a particular asbestos manufacturing company during 1938-1939. This cohort was then followed to determine subsequent mortality. Death claims for benefits related to the cohort were used as a basis for locating death certificates on file in vital statistics offices. Death record information was supplemented where possible by reports of microscopic findings and then combined with the BOASI record data for analytical purposes. An association was found between employment in the asbestos industry and mortality due to asbestoses and cancer of the lung and of the peritoneum.

Mancuso and Coulter concluded from this and other experiences that through the merging of data on existing health, insurance, employment and earnings records, cohort studies of great importance for industrial health programs would become feasible. The link that they saw was the Social Security number. It should be noted that many of the same records are pertinent to other areas of interest. Occupational mobility and changes in economic status at various stages of an individual's working life are of considerable significance to behavioral scientists concerned with, for example, parameters of family formation and stability, changes in social class structure, and emotional disorders.

Another study of particular interest is that being conducted by Hauser and Kitagawa using the enumeration schedules filled out in the 1960 Decennial Census and the copies of certificates of deaths that occurred during the 4 month period May-August 1960.(4) The desired end result is the location of the appropriate set of Census data for each of the persons who died and the merging of all of the pertinent facts of death with the sociological and economic data collected through the census. This is unquestionably the most formidable record matching operation ever undertaken in this country, involving as it does records for the 179 million persons enumerated and for a half million deaths.

It is not the purpose here to dwell on the matching problems--that is the prerogative of those attempting to solve them--but rather to indicate that this difficult activity is aimed at relieving some of the basic shortcomings of the death record. It is no longer possible to satisfy the needs for mortality information through the limited range of parameters available on this record. However, the linking of the death record with the Census record should open the door to the exploration of the relationship between income, education, occupation, housing and other environmental factors, and death due to specific chronic and acute conditions.

Additional uses made of established record systems to satisfy research needs come to mind. Lew and associates investigated mortality risks associated with deviations from average weight and with high blood pressure.(5) Medical observations made at the time individuals applied for life insurance were linked to mortality records over a long period of time to arrive at relative risk factors which are of great utility not only for underwriting purposes but also for the current efforts to understand for example the relevance of weight and blood pressure to mortality due to coronary heart disease.

Also, Newcombe and co-workers developed a project to link routinely collected population records from which information could be obtained on differential fertility and mortality in families carrying hereditary defects.(6) The basic step in the program consisted of linking birth registrations (about 400,000 in the years 1946-1958) with the marriage registration of parents (114,000 for the same period) in the Canadian province of British Columbia. The unique feature of this study is that Newcombe applied computer technology to a large scale record matching problem.

Finally, through the linking of marriage, birth and divorce records, Christenson has probed into factors associated with child spacing, premarital pregnancy and marital dissolution.(7) The starting point was the file of records of marriages that occurred during a particular period in three local areas. These were then matched against birth record files covering varying lengths of time after the marriage. In two of the three communities the search also included divorce records. The data derived from these three sets of records proved to be useful for a number of speculations regarding, for example, the relationship of sexual permissiveness in a culture to premarital pregnancy and the effects of the pregnancy on hasty marriage and on subsequent divorce.

Studies dependent on special research records linked to established record systems

Even if it were possible to combine all existing record systems, there would be a large deficit in information required for most studies. This is inevitable. Routine records can not be expected to answer all questions for all time even within the field where they are located. Other methods for obtaining data are of course more appropriate for certain types of research and are being used extensively, e.g., personal interview, mail survey, special medical examinations.

Without flexibility in the approach to problems research would quickly dry up. However, an existing record system frequently serves as the most reliable and convenient source of an essential part of the desired information. For example, many long term population studies which start with biological, physical, behavioral or social observations obtained through special inquiries are concerned with the relationship between mortality and these initial characterizations. Persons who die must be identified and facts about their death retrieved from the appropriate death certificates. Matching with death records may involve only those persons known through other sources to have died or with the entire cohort in the absence of such knowledge.

Dorn, Hammond and others integrated smoking histories obtained through a mail survey or personal interview with information on death certificates for persons in their cohorts who died in succeeding years. Through follow-up procedures in the field Hammond determined periodically who was living and who had died.(8) The search for matching death records was then initiated. Despite the fact that he was dealing with a large cohort (188,000 men), Hammond was able to keep track of the survival status of the overwhelming majority in his cohort and to locate the death record for every man reported as having died (11,870 deaths in a 44 month period). As is well known, the information was then used to examine the relationship between smoking and lung cancer and other causes of death. In this study the investigators had many of the specific details needed to search for death records.

Not all studies concerned with mortality are so favorably disposed and the effort required to identify deaths through manual or punched card matching procedures is so great sometimes that it deters much needed research. However, present experiments in matching records through computer techniques will, hopefully, change the situation and improve the prospects for conducting long-term follow-up studies related to health. Reference has already been made to Newcombe's efforts and the problem that led to his use of computers for record matching purposes. Two other studies are cited below as illustrations of the opportunities for research that will occur when computer matching techniques are refined.

In both, the matching operations are being carried out by the Health Insurance Plan of Greater New York in cooperation with the New York City Department of Health. One of the studies is designed to measure the incidence and prognosis of coronary heart disease in a population of 120,000-130,000 persons aged 25-64 years.(9) Incidence is to be determined through a combination of medical and hospital data and the findings of special medical examinations. New cases are being located during a 5 year period; and the course of the disease is being studied also over a 5 year period. The follow-up of patients admitted to the prognosis cohort presents no special problem. There will be about 2,000 such cases and there will be frequent communication with each of these patients or next of kin in case of death. Mortality can therefore readily be determined and death records obtained.

However, death from coronary disease can occur suddenly among patients not in the prognosis cohort. This circumstance makes it necessary to place the entire population under continuous surveillance for mortality. It is not known to what degree deaths among H.I.P. members are reported as such to the Plan. For this reason, a test is being made which involves matching computer tapes that contain identifying information from the H.I.P. records against similar tapes prepared from records of all deaths among adults that occur in New York City during the study years.

In the other study, two matched samples of 30,000 women each (aged 40-64 years) are being followed for about 10 years to determine differences in mortality from breast cancer.(10) One of the samples is being asked to come in for periodic clinical and x-ray examinations through which it is expected that breast cancer will be diagnosed at a comparatively early stage. The other sample of women will follow their usual practices in seeking medical care. The need for a more efficient method in locating deaths in both cohorts than direct communication with them or their families is critical and the plan is to rely heavily on the same type of computer matching procedures being tested for the Coronary Heart Disease study.

The projects discussed involve large study populations under observation for long periods and require a repetition of the matching process at intervals. There are, however, many other cases in which established record systems are entered only once to amplify the information collected through special studies. Furthermore, the study population is quite small. Anderson, Feldman and Sheatsley supplemented data on hospital and medical utilization and costs obtained through household interviews by obtaining expanded and more specific information on these items from the matched records in the files of health insurance plans.(11) Bright, Lilienfeld and others are following up on persons who were covered in the Hunterdon and Baltimore chronic disease studies about 10 years ago to determine the relationship between morbidity findings on both interview and special medical examinations conducted at the time and subsequent mortality.(12) Death records have an important function in this investigation and must be located.

Methodological reasons for record matching

Record matching is often carried out in response not only to a need for otherwise unavailable data but also because of the opportunity it offers for obtaining improved information. This is true for many of the projects already

discussed. In the case of the Anderson, Feldman and Sheatsley studies much of the information obtained through record matching was already covered in the household surveys. But, through recourse to records they were in a position to correct response errors as well as expand their range of information.

The Hauser-Kitagawa study based on matched census and death records has as one of its major objectives the improvement of mortality rates resulting from the availability of comparable numerators and denominators. Ordinarily rates by age, race, marital status, residence and occupation are computed by using for the numerators information on the death certificate and for the denominators corresponding items on the census schedule. It is well known that for several of these variables major differences exist between the two record sources. Use of the same document (the Census record) does not eliminate all sources of error by any means but it deals effectively with the problem of incomparability.

One additional reason for record matching to be taken up here is related entirely to methodological considerations. Studies have been carried out by means of matched records whose sole purpose has been to measure bias or errors in information obtained through a particular data collection procedure and to find ways of improving the accuracy and completeness of the information.

The National Health Survey has made such methodological studies an integral part of its on-going statistical program. During the past few years, the household survey approach of the National Health Survey has been examined intensively for strengths and weaknesses.(13) Chronic diseases reported in household interviews have been checked against medical record information to determine the completeness of such reports. Similarly indices of accuracy of reporting hospital episodes on interview or in a self enumerated survey were obtained through comparisons with hospital records. In all of these studies, the objective is to learn a great deal more than is now known not only about validity of the information being collected but also what the correlates of poor reporting are. The technique has wide applicability for the testing of other data collecting mechanisms, e.g., mail or telephone surveys, diaries.

Ultimately these methodological studies will clarify the types of data that can be collected by a particular procedure with sufficient accuracy to make them usable. This is a weighty issue for all investigators. From the standpoint of the present paper the important point is that the validation procedure is dependent on the correlation of information in two different sets of records.

The 1940 and 1950 Birth Registration Completeness Tests represent record matching on a "grand" scale to determine the accuracy of a record collection system. In both tests, special infant cards were filled out by Census enumerators for children born in the early part of the census year; in 1940 about 700,000 cards were prepared and in 1950 over 800,000.(14) Birth records were matched against these infant cards to measure the underregistration of live births, and in 1950 the underenumeration of infants. Matching procedures differed in these two tests. For the 1940 test, the approach taken was to match records manually. But in 1950, it was decided to place main reliance on a mechanical procedure based on punched cards. Through the application of a series of alternative matching criteria followed by visual inspection of paired record data, the 1950 test found the new approach efficient, accurate and highly suited to the rapid production of detailed test results.

Validity studies not involving governmental statistics have also been carried out repeatedly. One of the pioneering studies involving matched records was that of Parry and Crossley.(15) Their concern was with reports of voting; ownership of a home, driver's license, or library card; and donations in a community drive for funds. All related records available in the area were systematically scanned to determine the accuracy of the responses to the survey questions.

Conclusion

The need for record matching has been approached through a review of research that has utilized this procedure. Undoubtedly other cases can be cited in which record matching because of its complexity and cost had to be replaced by other, less satisfactory methods of obtaining information. Also, existing records are not always the most complete or precise source for data required and record matching becomes a poor substitute for other methodologies. However, there is often no alternative to record matching and, in fact, many of the studies previously referred to could not have been undertaken without the use of interrelated records. Furthermore, the ever increasing complexity of our society is resulting in the development of voluminous record keeping systems for large segments of the population and these systems represent a reservoir of information regarding many significant health, economic and social problems.

Record matching will however continue to be a laborious, difficult and sometimes impractical process unless it attracts greater attention both from persons responsible for developing record keeping systems and from statisticians concerned with methodology. By its nature, record matching requires a broad perspective of the function of record systems. Often, persons charged with the responsibility of establishing record systems think only in terms of the most immediate administrative needs that have to be satisfied. Even when the outlook is not so narrow, there is almost no attention given to the content or

organization of the records that might facilitate their being matched with other sources of information. This applies to both routinely collected records and special research records.

Today the common thread in almost all records consists of age, sex and address information. This consistency has not occurred because of any interest in record matching and indeed the thread may in some instances be too thin for matching purposes. A prerequisite for a change in the present situation is to lift record matching from its present haphazard state and to consider seriously the basic elements that would facilitate it. This might lead to a convention regarding a minimum set of common items for routine and research records. For example, the addition of social security number which has been suggested by many would provide a potent item of identification for studies involving the adult population.

Not only content needs to be considered but also the organization and storage of record information and efficient techniques for record matching. Some steps have already been taken in this direction. Record systems are increasingly being placed on computer tapes and the application of computer techniques to record matching is being experimented with in several areas, including Canada, California and New York City. These new approaches will hopefully ease the burden of dealing with massive sets of records.

Finally, it must be recognized that we are faced with a set of long term needs for record matching and that the problems will not be resolved by any easy formula. Personnel is needed to concern itself with the elements discussed, i.e., content, organization and technology. The statistician has one of the greatest role in this effort and his training should reflect this responsibility. It is time that the design of record keeping systems for statistical and administrative functions was upgraded to a high priority activity among statisticians and that record matching was accorded the prominent place in the array of methodologies available for research that it merits.

Bibliography

(1) Dunn, H.L., Record Linkage. Amer. Journal of Public Health, 36, 12, Dec. 1946.

(2) National Office of Vital Statistics, Recommendations for Developing Comparable Statistics on Prematurely Born Infants and Neonatal Mortality. Dept. of Health, Education, and Welfare, Dec. 1950.

(3) Mancuso, T.F. and Coulter, E.J., Methodology in Industrial Health Studies. Archives of Environmental Health, 6, Feb. 1963.

(4) Hauser, P.M. and Kitagawa, E., Social and Economic Mortality Differentials in the U.S., 1960: Outline of a Research Project. Social Statistics Section, Amer. Stat. Assn., 1960.

(5) Society of Actuaries, Build and Blood Pressure Study, 1959. Vol. I, Chicago.

(6) Newcombe, H.B. and Rhynas, P.O., Family Linkage of Population Records, Symposium on Medical Electronic Data Processing. Proceedings of the U.N.-W.H.O. Seminar on Use of Vital and Health Statistics for Genetic and Radiation Studies, U.N., 1961.

(7) Christensen, H.T., Cultural Relativism and Premarital Sex Norms. Amer. Sociological Review, 25, Feb. 1960.

(8) Hammond, E.C. and Horn, D., Smoking and Death Rates. Journal of Amer. Med. Assn., 166, Mar. 1958.

(9) Shapiro, S., Balamuth, E., Frank, C.W., Sager, R.V. and Densen, P.M., The H.I.P. Study of Incidence and Prognosis of Coronary Heart Disease: Methodology. In press - Journal of Chronic Diseases.

(10) H.I.P. Annual Statistical Report, 1962. In Press.

(11) Anderson, O.W. and Feldman, J.J., Family Medical Costs & Voluntary Health Insurance: A Nationwide Survey. McGraw-Hill Book Co., 1956.

Anderson, O.W. and Sheatsley, P.B., Comprehensive Medical Insurance. Health Information Foundation Research Series No. 9.

(12) Personal communication. Lilienfeld, A.H.

(13) Health Statistics, U.S. National Health Survey, Health Interview Responses Compared with Medical Records. U.S. Dept. of Health, Education, and Welfare, Series D, No. 5.

Health Statistics, U.S. National Health Survey, Comparison of Hospitalization Reporting. U.S. Dept. of Health, Education, and Welfare, Series D, No. 8.

(14) National Office of Vital Statistics, Vital Statistics of the U.S., 1950, Vol. 1, pp. 108-112, U.S. Dept. of H.E.W.

(15) Parry, H.J. and Crossley H., Validity of Responses to Survey Questions. The Public Opinion Quarterly, 14, 1, Spring 1950.