SOCIAL SECURITY DATA FILES AS A RESOURCE FOR HEALTH RESEARCH

Faye Aziz, Harriet Orcutt and Linda DelBene, Social Security Administration

The intent of this report is to introduce researchers to the general resources on Social Security's administrative and statistical files for health research. We consider this the first step in calling attention to the assets inherent in the files, and in identifying areas with potential for improvement.

CURRENT RESOURCES AT SSA

Social Security touches almost every American. Each individual with a social security number has a record filed with the Social Security Administration (SSA). It is not even unusual for a parent to apply for a number for a child shortly after its birth. As a result, Social Security has names, dates of birth, race, sex, and some geographic information for almost everyone. Information on earnings is accumulated and maintained in Social Security's Summary Earnings Record file. The earnings record contains the demographic information from the application for the social security number, lifetime covered earnings, quarters of coverage, and death information. If a person applies for benefits, a record is created in the Master Beneficiary Record file. This benefit record contains demographic information, historical and current information about entitlement to benefits, benefits paid, dependent beneficiaries, disability, and vital status, as well as the fact of entitlement to hospital and medical insurance.

There are currently about 270 million records on the earnings file--one for every social security number ever issued. As a policy, records are not deleted from Social Security files. When Social Security is informed of a death, the death information is added to the appropriate records and the individual's covered earnings history and benefit history remain as a part of the current file.

In addition to these administrative files, Social Security maintains a system of statistical files called the Continuous Work History Sample (CWHS). As the name implies, the CWHS is a work history. In addition to carrying the individual information from the administrative files, it maintains longitudinal work-related information such as industry and geographic location of the employer. As such, the CWHS is a particularly useful research tool for epidemiology.

RESEARCH USING SSA FILES

The use of Social Security record system information for research makes sense from several perspectives. The information is already being collected for administrative purposes. Using administrative records makes more information available for research without increasing the burden of collecting data on the population. In addition, several types of information collected for administrative purposes, such as earnings data, are likely to be more reliable than the corresponding information collected in surveys.

By linking Social Security record files to external sources of data, the resulting enhanced files can be used for research in many topics. In this report we will focus on their use for health research. First of all, the estimation of mortality and disability, and their interaction with economic variables such as labor supply, is important to Social Security's prediction of the total amount and distribution of benefits. Secondly, Social Security data files have certain unique features which make them particularly attractive for health research. Disability and mortality are rare events. As such, the study of their determinants requires large data bases. Administrative record files, alone and merged with other sources of data, make possible larger, more detailed micro data bases than could ever be collected from a survey. Another unique feature of Social Security records is that they contain information over time on health indicators and variables thought to influence health. This makes possible disentangling the causality responsible for certain differential patterns of health levels observed in the population.

Health research using Social Security data has primarily centered on two types of studies. The first we will call "general policy research"; the second we label the "determination of specific hazards."

General policy research is aimed at determining the impact of various socioeconomic variables (such as education, income, and occupation) on health. Generally, this research seeks to explain the positive relationship between indicators of socioeconomic status and health. The results from such studies have implications for broad policy directions such as whether more emphasis should be put on equalizing health care or improving occupational safety and health in our pursuit of more equalized health levels.[1]

Another type of research focuses on determining and measuring the effect of specific hazards on health. [2] Two categories of hazards can be delineated. First, there are areal pollutants associated with geographic areas. Research assessing the impact of areal pollutants on health has been almost completely limited to cross-sectional comparisons of mortality and other health measures across geographic areas. Using indicators of residence over time could open up a new pathway for measuring the effect of various areal pollutants.

The second type of specific hazards research is the measurement of the effect of industrial work hazards on health. Since the CWHS has information on industry, it can be used for this line of research. Furthermore, the records are longitudinal. Thus, the length of time spent in an industry can be determined, and the exposure to particular hazards measured.

INFORMATION CURRENTLY AVAILABLE ON THE CWHS

In the previous section we described the overall general strength of Social Security data files and the two major types of health research to which these data files have been put to use. We would now like to give a more detailed picture of the data items currently available on the CWHS.

Demographic Data.--The original source of data on age, race and sex for Social Security files is Form SS-5, the Application for a Social Security Number. The applicant is required to provide sex and date of birth information to help assure that each record on file is uniquely identifiable. Information on the applicant's race, however, is provided voluntarily. In recent years, about 2 to 3 percent of applicants for social security numbers have not provided race information on the SS-5.

Industry and Place-of-Work.--The next key items already on the CWHS are industry and place of work. These are obtained from the employer's Application for an Identification Number (Form SS-4) and related forms used periodically to update this information.

The longitudinal nature of the industry information on the CWHS is a unique feature of that file. While there are other sources of industry data such as the Census Bureau's Current Population Survey, they generally only provide that data for a single point in time. On the CWHS there is information on the industry of employment for each job in covered employment over the working career of an individual.

Even though the CWHS does not have specific residence data for workers, it may be possible to use the place-of-work records as a general indication of residence over time. Its usefulness and accuracy for this purpose would, of course, vary among industries and types of establishments. [3]

Earnings Data and Work Experience.--Social Security maintains a record of each person's earnings in covered employment in order to determine the eligibility and amount of benefits an individual worker or dependent is entitled to. The advantage of using administrative record information over survey information for earnings data is that the force of law accompanying the collection of the administrative information is likely to encourage more accurate responses. What is most exciting about this source of information, however, is that it provides a yearly record since 1950 of individual earnings up to the taxable maximum for all jobs in covered employment. And, in fact, starting with the 1978 records, complete earnings information will be recorded for both covered and noncovered employment. This change constitutes a major breakthrough for research in future years.

Disability Data.--Social Security also maintains records on disability status. The determination of disability by Social Security really is the determination of "work disability." That is, people with medically non-severe disabilities may be labeled non-disabled by Social Security if they are considered capable of earning a living. The definition of disability documented in Social Security records then is a narrower definition than is found in other sources such as surveys. It is a nonsubjective assessment of work disability of persons who apply for benefits.

Vital Status.--Social Security's administrative files have long been a source for determining vital status, because deaths are routinely posted to both the beneficiary records and the earnings records. Analysis is underway to examine the extent of death reporting on Social Security's files and to study the nature of under-reporting. Using a sample of individuals known to have died in 1975, we found that about 91 percent of these deaths were recorded on the earnings record. By further examining the beneficiary files, coverage rose even higher. Analysis from an historical perspective shows also that death reporting has risen dramatically in the past few years, particularly since 1973. The last year of deaths studied--1977--shows a reporting rate of about 98.5 percent. Reporting appears to be virtually complete in the recent years for white male decedents aged 65 and over. [4]

INFORMATION THAT WOULD ENHANCE THE CWHS

While Social Security data files are already useful to researchers with health interests, we have begun to consider schemes for enhancing them even further. Below are some of the data items which researchers would like to see added to the CWHS.

Cause of Death.--Cause of death data is of growing interest to epidemiologists and other researchers involved in health and mortality studies. Much thought has gone into developing schemes for routinely adding this item to the CWHS. There are several feasible sources of such information.

First of all, it is possible to purchase death certificates from the states by providing them with certain identifying data which is extracted from the decedent's social security records. However, there is a great deal of time and expense associated with extracting and coding the cause of death directly from the certificate.

A more economical source of cause of death data is the detailed mortality records maintained by the National Center for Health Statistics (NCHS). These files have the specific cause data already coded. However, they must be accessed by providing the death certificate number, which is not present on SSA records.

For deaths occurring in 1979 and beyond, we will be able to access the National Death Index recently established by NCHS. In its present form, the death index does not include coded cause of death; it will, however, furnish the death certificate number and state in which death occurred, thus simplifying the states' search for certificates and improving their results.

The alternative that appears to be the most attractive is to provide the states with an effective incentive to routinely report deaths, or to make it mandatory that they notify SSA of deaths. Another possibility, especially if the lump sum death benefit is abolished, is to establish another incentive for funeral directors to report deaths to Social Security.

Occupation and Education.--Two extremely important variables currently not available on the CWHS are occupation and education. Their absence is often cited by researchers as a major reason for not using the CWHS, despite its other strengths.

One source of occupation information is the death certificate. A project has been underway to add death certificate data to the records of all CWHS decedents for selected years. Obviously, this project will provide occupation data only for deceased individuals. However, this information may be used to compute the relative mortality risks of various groups. Another proposal is to obtain occupation information from Internal Revenue Service (IRS) tax forms. [5] A strong advantage to this approach would be the possibility of obtaining occupation information for the same CWHS individuals over time. Unfortunately, there are also problems. Questions have been raised concerning both the validity and the codeability of the tax return items, and there are severe restrictions placed on access to IRS data stemming from the Tax Reform Act.

There are several surveys which collect both education and occupation information and which could potentially incorporate in their sampling design an overlap with the CWHS. Additionally, some of these surveys follow the same group of households over time and some ask questions about occupational experience over a longer period of time than just the preceding year.

Residence.--In addition to having the geographic location of an employer on the CWHS, worker residence would be an asset for health studies. Generally, Social Security has residence information only for people receiving benefits. Additionally, there have been occasional past matches between Social Security and IRS records which resulted in the addition of IRS address data to the CWHS. However, the Tax Reform Act stopped Social Security's access to these files and the matches were halted.

A new method of acquiring residence data came about as a result of the annual reporting system, which began in 1978. Under annual reporting, addresses are included on the W-2 forms, and some researchers at Social Security have begun an experimental manual pilot study to examine quality and codeability issues for possible inclusion of W-2 addresses in the CWHS. This study is primarily investigating the possibility of obtaining the zip code portion of the addresses from the microfilm copies of the W-2's to enter into the tape records of those workers in the 1-percent CWHS.

Life Style.--Life style variables may be the hardest to capture on Social Security data files. Desirable items would include marital status, length of marriage, number of marriages, family structure, smoking habits, and alcohol consumption. Of these variables, Social Security has only inadequate information on marital status and family structure--and this information is only for certain cases.

The only real opportunity in the past to do analysis dealing with life style has been when Social Security has linked administrative records to survey data. The linkage of SSA data to the 1973 Current Population Survey yielded some information on marital status, marital history, fertility, and family structure for a small sample of individuals. But even in this effort using survey data, variables such as smoking habits and alcohol consumption were not available.

CONCLUSION

In conclusion, the purpose of this report has been to summarize the general resources of Social Security's administrative and statistical files for health research. Further, we wished to call attention to the areas we would like to see enhanced so that we can get feedback from the research community. Finally, our goal is to see the CWHS realize its full potential as a resource for health research.

ACKNOWLEDGEMENTS

The authors wish to thank Warren Buckler and John Hambor for their guidance and helpful comments during the preparation of this report. Creston Smith and Richard Wehrly provided information on related Social Security research and operational activities, and Henry Ezell provided the technical assistance necessary for this project.

NOTES AND REFERENCES

Researchers outside of SSA do not have access to the administrative files. Beginning in 1976, access to the many statistical files has been shackled by the Tax Reform Act. However, efforts are underway to develop methods to release a version of the CWHS for general research.

[1] For an example of general policy research using SSA data, see:

Orcutt, Harriet. "Differential Mortality by Income and Education," 1980 American Statistical Association Proceedings, Social Statistics Section.

[2] For examples of research using SSA data in the determination of specific hazards, see:

Mancuso, Thomas. "Relation of Duration of Employment and Prior Respiratory Illness to Respiratory Cancer Among Beryllium Workers," Environmental Research, Vol. 3, No. 3, March 1972, pp. 1-24.

Goldsmith, John R. and Thresh, Madeline. "Mortality and Industrial Employment (III)," Vol. 106, No. 2, 1977, pp. 109-124.

[3] Much work is being done by the Bureau of Economic Analysis to evaluate the geography and industry data on the CWHS. See:

Levine, Bruce. "Improving Industry and Place-of-Work Coding in the Continuous Work History Samples," 1980 American Statistical Association Proceedings, Section on Survey Research Methods.

_____. "Geographic and Industrial Coding in the CWHS," Bureau of Economic Analysis, unpublished report, April 1981.

U.S. Department of Commerce. Bureau of Economic Analysis. Regional Work Force Characteristics and Migration Data: A Handbook on the Social Security Continuous Work History Sample and Its Application. December 1976.

[4] For more information on death reporting to SSA files, see:

DelBene, Linda and Faye Aziz. "Mortality Coverage in Social Security's Earnings and Benefit Record Systems," 1980 American Statistical Association Proceedings, Section on Survey Research Methods.

Aziz, Faye and Warren Buckler. "Mortality and the Continuous Work History Sample," 1980 American Statistical Association Proceedings, Section on Survey Research Methods.

[5] Results from an initial pilot study on the codeability of occupation information from IRS tax returns are available. See:

Sailer, Peter J., Harriet Orcutt, and Phil Clark. "Coming Soon: Taxpayer Data Classified by Occupation," 1980 American Statistical Association Proceedings, Section on Survey Research Methods.