ACCESS TO ADMINISTRATIVE RECORDS ON ESTABLISHMENTS AND INDIVIDUALS FOR PUBLIC POLICY ANALYSIS

David A. Hirschberg, Small Business Administration
Vernon Renshaw, Bureau of Economic Analysis

In the course of their administrative activities, Federal, State, and local governments generate an enormous and valuable collection of information that bears on the statistical system. It will be argued that the benefits which might accrue to analysts and policymakers from these collection activities are not being realized. The causes and remedies will be discussed only tangentially. The purpose of this paper is to describe the problem so that remedies and proposals for reform may be more easily understood.

Several administrative data files used in the analysis of 1) regional economics, 2) income distribution and welfare policy, 3) industrial mortality and morbidity, and 4) small business will be discussed. The problems faced by analysts using these data sources will be briefly outlined. It will be of interest to note how restrictions relating to privacy and confidentiality, as well as deficiencies in coordination, timing, comparability, and general quality control significantly reduce the availability and value of administrative data.

The administrative data systems discussed include Social Security Administration (SSA) work history records, Internal Revenue Service (IRS) tax records, civil service records, vital statistics records, medical records, and income transfer records. The IRS and SSA records are central to each of the analytical areas discussed. Although the policy concerns of the analytical areas differ considerably, there are many remarkably similar needs for improving capabilities for associating IRS, SSA, and other administrative records together for statistical applications. Improvements are particularly needed in terms of the capacity for longitudinal association of records, for reliably linking records of individual workers with information relating to the specific business establishments they work for, and for associating together records assembled in separate administrative programs.

THE PROBLEM

"The central problem of data use is one of associating numerical records. No number conveys any information by itself. It acquires meaning and significance only when compared with other numbers. The greatest deficiency of the existing Federal Statistical System is its failure to provide access to data in a way that permits the association of the elements of data sets in order to identify and measure the inter-relationship among interdependent activities."

This statement was made in 1965 by Edgar S. Dunn [6, p.5] in a review of the Ruggles Committee proposal for a national data center. Since Dunn and the Ruggles Committee called attention to problems of data access and association, the problems have been compounded in many areas of data use; problems have become particularly acute in connection with the statistical use of rapidly expanding administrative sources of data. This paper reviews selected data problem areas with special reference to the need for overcoming data deficiencies with improved utilization of administrative records.

The Federal statistical system has been characterized as decentralized. Basically, it is made up of a group of Bureaus engaged in collecting, analyzing, and storing a diverse array of data on demographic, economic, and social variables. The system assigns high priority to timely information on important measures of national economic activity -- for example, the number unemployed, changes in consumer prices, and changes in the gross national product. Every 10 years, the Census Bureau provides detailed demographic and social data for the entire country by detailed geographic areas. The periodic economic censuses provide detailed information on the health of the Nation's business.

In terms of quality and quantity, these social and economic data are the envy of virtually every country in the world. Nevertheless, they have severe limitations, and policymakers are constantly pointing out that they don't have the data they require. In large part, the gaps in policy relevant data stem from deficiencies in statistical access to and utilization of the vast amount of information currently collected as a part of Government administrative programs.

The inadequacy of data for public policy is not a trivial issue. Government, at all levels, represents a major part of the demand for goods and services and, at the same time, utilizes a significant amount of labor and professional talent. Governmental policy, broadly defined, has a direct and indirect impact on virtually every area of our lives. Issues of energy, ecology, education, poverty, health, and urban planning all require improved information not only for the relevant areas involved, but also on their interrelationships with the rest of the economy.

The reasons for data inadequacies are varied and complex, but essentially they reflect the diffused organization of the Federal Government's data collection programs. Basically, there is a lack of an integrative force able to establish compatible survey designs, develop and enforce standards for data collection and archiving, and ensure that data from administrative and regulatory systems are produced in a useful, compatible, timely, and accurate manner.

Although the economic feasibility and technical capacity has existed for several years with the development of the high-speed and large scale com-

puter, the Federal establishment is finding difficulty in structuring statistical data from administrative records which can serve as a guide to agencies in analyzing their missions, establishing standards for performance, evaluating the impacts of regulation, and assessing the costs and benefits of various policy alternatives. In many areas, programs are delayed, and decisions are vulnerable to attack. Groups, particularly at the local level, concerned with roads, schools, hospitals, and urban and regional development are becoming aware of many of the inadequacies and anomalies of available data.

The problem is not that data are not being collected. The growth of both administrative and statistical data programs has been rapid. To reiterate, the problem is that data programs have not been well integrated or coordinated; that is, the information in particular data sets cannot be readily associated with elements of other data sets to measure the impacts of interrelated activities. To some extent, this deficiency has been overcome in the national income accounts and input-output tables by which one can trace and analyze the economy's performance. However, this effort at integration is the exception and often the national accounts must make use of primary data which are substantially inferior to those that might be available with improved coordination in the development of administrative record data sets.

REGIONAL ECONOMIC ANALYSIS

At the regional level, employment and payroll statistics are available from several administrative and closely allied record systems. These include: (1) state unemployment insurance (UI) payroll tax programs, (2) Census County Business Patterns (CBP), (3) economic censuses, and (4) SSA's Continuous Work History Sample (CWHS).

Both the UI and CBP programs provide establishment data at the county-industry level on employment and payrolls. The UI data are reported to the States by firms covered under the unemployment insurance laws. Geographic area reporting is required of multi-establishment firms. The UI system requests these data at the county level. The CBP uses social security data for single establishment firms and uses data for individual establishments of multi-establishment firms obtained from an annual Company Organizational Survey (COS) conducted by the Census Bureau. The Census Bureau, using lists of businesses developed from SSA, IRS, and COS records, conducts periodic censuses on employment, payrolls, sales, business receipts, assets, value added, and other information. The CWHS provides longitudinal demographic information for a sample of workers covered by the Social Security system. For each job held, data are available by sex, race, age, industry, county, and quarterly earnings up to the taxable limit.

These are obviously rich sources of regional data. However, the problem of integrating data sets is formidable. Upon inspection, one can quickly note that CBP, CWHS, and UI data are not consistent as to level and trend. The smaller the region or more detailed the industry, the greater the noncomparability. This difficulty results from a variety of factors, most of which are not readily quantified. For example, multi-establishment firms may be reporting differently to the various agencies. The UI and SSA reporting plans, unlike the Census COS reporting guidelines, permit multi-unit firms to combine establishments into larger (county-level) reporting units. Geographic and industrial breakdowns of data for multi-unit firms are "mandatory" in the UI and Census reporting plans, but voluntary in the SSA plan. The industrial and geographic coding of reporting units, moreover, is not well coordinated among the programs. Coding in the UI system occurs at the State level and record access problems have generally prohibited direct comparisons of UI coding with either Census or SSA coding. There is some coordination and cross checking in the SSA and Census coding procedures, but coordination is limited because of the use of different reporting unit concepts for multi-establishment firms.

There are additional problems that affect the utility of these data aside from uncoordinated procedures for establishment reporting and industrial and geographic coding. The universe covered by the sources differs. For a long time, the UI data excluded small firms for many States; the CWHS covers only about two-thirds of State and local government employment and excludes railroads, virtually all Federal Government workers, and some nonprofit organizations. The Civil Service Commission has maintained a microdata file of a sample of personnel actions noting accessions, separations, and job changes within the Federal establishment, but because of poor data quality and the complexity of using the file, the data have not been merged with the CWHS. Industrial classification problems compound the problems of comparing CBP, CWHS, and UI data. The CWHS classifies some Government workers in private industries providing similar services. This procedure affects, for example, workers in public hospital, transportation, and school systems. The UI data classify all Government establishments in the Government industry. CBP data exclude all Government workers.

Problems of coordinating data programs and accessing basic records have implications beyond reducing the quality and comparability of CBP, CWHS, and UI data. Restricted access to the basic establishment lists used in these programs, for example, has raised the costs and reduced the quality of statistical programs of other Federal agencies needing establishment lists for survey frames or research purposes. Many agencies have incurred substantial expense in purchasing privately-developed proprietary business lists or in developing and maintaining their own lists. The U.S. Postal Service, for example, recently contracted with the Survey Research Center (SRC) of the Institute for Social Research at the University of Michigan, to determine the best way to develop a comprehensive geographic sampling

frame for nonhousehold establishments. The SRC evaluated proprietary establishment lists prepared by National Business Lists and Dun and Bradstreet. The conclusion [18, p.13] was that "The lists themselves are not useful for sampling purposes for several reasons: (1) they do not provide a complete listing of every organization in the areas they purport to cover; (2) they include organizations not in these areas; and (3) possibly, most importantly, they do not provide a relevant measure of size (e.g. total postal expenditures) which is essential if an efficient sample is to be drawn."

Proprietary business establishment lists are generally not developed explicitly for statistical applications, and their suitability for meeting the statistical needs of particular agencies may vary substantially from agency to agency. The SRC recommended that the U.S. Postal Service develop its own nonhousehold sampling frame from postal records. Other agencies, however, have neither needs nor internal resources as elaborate as those of the U.S. Postal Service. In many cases, therefore, reliance on proprietary lists may be the best, or only, available option in the absence of a comprehensive, accessible establishment list prepared specifically for statistical uses. Unfortunately, however, it is often difficult to evaluate how well, or poorly, particular agency needs are being met through the use of proprietary establishment lists.

INCOME DISTRIBUTION AND WELFARE POLICY ANALYSIS

The income transfer and maintenance programs of the Federal Government have a direct impact on millions of people. The effective delivery of services to those experiencing individual and family hardships stemming from work-related problems is a national issue. Attention has been focused on the issues of poverty, and major proposals for welfare reform have been before the Congress for a decade.

For modeling the income transfer, unemployment, social security, and welfare systems, data base concerns have been formidable. In 1976, the GAO prepared a report [9] on the evaluation of the transfer income model. Several problems with the data base were recognized, and suggestions were made to improve the projections.

Basically, the model uses the Census Bureau's Current Population Survey, which includes questions on money income by source. The shortcomings of using such limited information to model the transfer income sector became apparent with the failure of efforts to use the model to understand the sharp increase in the population receiving payments under programs of aid to families with dependent children (AFDC) between 1963 and 1973. Paradoxically, this increase occurred during a period of unprecedented growth in jobs and real income.

It has become clear that the "welfare decision" is highly complex. For various reasons, those receiving welfare payments are poorly counted in censuses and surveys. If counted, such persons generally under-report the income they receive. Moreover, surveys rarely include the noncash benefits that may impinge on the welfare decision--food stamps, rent supplements, Medicaid, subsidized child care, and other transfers.

An analysis of the data base problem suggests that records from a variety of administrative sources must be organized to examine the welfare problem. A major modeling effort by Rydel [17] points out several important shortcomings in the available data. Longitudinal AFDC data on cohorts are virtually nonexistent. In addition, there are limited data on caseload turnover for the various programs. The dynamics of welfare caseload change and the factors affecting the trends are not understood and have not been modeled.

To develop an adequate income transfer model, the interrelationships between the total transfer system and its components must be understood. Data are not available to analyze the impact of food stamps, the demand for Medicaid, rent supplements, unemployment insurance, and child care. Moreover, data on factors that encourage or discourage families from coming onto the welfare rolls--divorce rates, unemployment rates, job opportunities, and illness--must be integrated into the model. This requires a sophisticated data collection system of administrative records, which must be carried out at the local level given the current structure of administrative programs.

The longitudinal information sets that are needed can only be obtained from a data base utilizing administrative and tax records for those participating in the various income transfer programs. Their creation would involve the cooperation of a large number of agencies. For example, the Department of Labor is responsible for the reporting of unemployment compensation; the Department of Agriculture, the food stamp program; Department of Housing and Urban Development reports rent supplements; Social Security Administration reports on regular retirement and supplementary security income programs and, since the dissolution of the Social and Rehabilitation Service, AFDC and Medicaid programs; and the Veterans Administration (VA) maintains veteran and survivor benefits. To compound the problem, each State has an opportunity to set complex eligibility requirements which affect participation and benefit levels.

INDUSTRIAL MORTALITY AND MORBIDITY ANALYSIS

Two recent documents describe the utility of administrative records for monitoring workplace safety and health [12] and for analyzing the health effects of ionizing radiation, and for general epidemiological research. [6]

Statistics on health and safety effects are collected by a series of agencies, similar to the situation in regional employment and income and welfare data. To carry out an effective re-

search program, records containing medical and occupational information must be made available so that the linkage can be made between medical histories and environmental and occupational exposure to disease. Many of the diseases which are occupationally related have long latency periods, and records need to be obtained which go back for extended periods of time. This latency period, sometimes of 25 to 35 years, requires the development of retrospective studies of exposed populations, if timely research findings are to be obtained.

There are several data sets from various sources which, if brought together, would provide additional information on the extent of occupational disease by a significant factor. These include records from the CWHS, SSA disability records and Medicare records maintained by the Health Care Financing Administration, the vital records of health maintained by States, military records maintained by the VA and the Department of Defense, Civil Service records, and IRS records.

A review of the data needs and the limitations of other approaches to epidemiological research require that administrative and vital records must be accessible to researchers conducting epidemiological research. Such research requires the tracing of the health status of large populations of exposed individuals and control groups over long periods of time. Moreover, obtaining the consent of each individual is not feasible, and studies of only those persons available and willing to consent would introduce significant bias.

As mentioned previously, records maintained by Social Security are an important source. These include recent addresses, employer's name and industry classification, and whether a death benefit claim has been filed. Employment histories can be traced by searching the many millions of microfilm and microfiche records obtained from employee reports. This method is expensive and time-consuming, but it is the only source of work history cohorts outside of those included in the statistical samples maintained by Social Security. In addition, the SSA disability program develops information on those applying for, or receiving, disability benefits. Of particular interest is the primary and secondary diagnosis of disability applicants. Medicare records are also available for a 20 percent sample of those eligible, but they have seldom been used in epidemiological research.

Obviously, vital records of death are important. They identify those who died from a particular disease and provide information on occupations, but inadequate access makes these records cumbersome, expensive, and time-consuming to use. The forthcoming National Death Index, now being developed by the National Center for Health Statistics, will provide current information, but those interested in historical records will not be helped. For an extensive discussion of mortality records and related research, the reader is referred to a number of other papers presented at this conference, including [2], [3], [10], [14], [15], and [16].

Internal Revenue Service records provide important data. Information on the taxpayer's address, occupation, last return filed, or estate tax return can significantly reduce the time and expense of locating individuals or determining mortality. The Tax Reform Act of 1976 has, however, severely affected the usefulness of tax return information for epidemiological research, by limiting access to IRS records.

SMALL BUSINESS ANALYSIS

Our knowledge concerning the behavior of small business is limited because of the fragmented nature of surveys, censuses, tax and other governmental administrative reports, and private sources. These data sets are available for small business, but for a variety of reasons they cannot be integrated. In particular, the absence of a uniform numerical identification system becomes a serious obstacle to combining these sources of information. Data on the vital statistics of small business, births, deaths, mergers, and acquisitions are nonexistent. Data on new business starts and failures are extremely important but no Federal agency collects them.

For the most part, financial information is made available for the primary benefit of security market analysts. Little is published for the smaller firms. Unfortunately, data on employment size, production, prices, assets, sales, and plant capacity are not integrated. Most of what we know about employment and production is collected at the establishment level. For the multi-establishment company it is necessary to bring these data together with information on profitability, savings, investments, cash flows, and research and development expenditures, which are generally available only at the company level.

A significant source of administrative information for small business policy analysis is business tax returns. However, a number of shortcomings affect their utility. Access to data is limited because of confidentiality provisions. In addition, IRS data identify business size in terms of asset size. Because there is no measure of employment size on the tax return, one of the most significant inputs to the productive process cannot be obtained from tax records. Another shortcoming is the availability of current information. The latest corporate tax return data are for 1974.

As in all administrative or survey systems, the problem of establishment versus enterprise reporting presents problems to the user, only more so with IRS data. The business unit reporting is the taxpaying unit. It may be either an establishment or an enterprise depending on the convenience of the company. These factors not only affect the analysis of firm performance by size, they preclude any geographic breakdown of firm data.

Dun and Bradstreet provides data from their credit and marketing operations for over 4.3 million selected establishments. This file has significant potential for use in regional and small business analysis. Although a great deal of work has been done with these data by David Birch (MIT) and others, only a limited analysis of the coverage problems has been undertaken, and the biases which might affect the conclusions drawn from these data are not yet fully understood.

CONCLUSION

We have tried to illustrate how administrative records for business establishments and individuals are important for statistical and policy purposes. In meeting the data requirements of the four policy areas considered in the paper, very similar information structures are needed. Similar requirements will also be found in other policy areas one may investigate, including agriculture, education, housing, transportation, and energy.

In each of the areas considered in the paper, longitudinal records for individuals and business establishments are needed to understand the dynamics of change for policy purposes. The CWHS, however, is the only longitudinal statistical file assembled from administrative records. No comparable longitudinal files are available for business establishments, or for other key individual record sets such as IRS individual income tax records. Longitudinal data for business establishments would be particularly valuable for analysis in the areas of small business, regional economics, and industrial mortality and morbidity. And supplementation of the CWHS with longitudinal data from other individual record systems (especially IRS records and income transfer records) is vital if administrative records are to be used for analyzing many issues relating to income distribution and welfare policy as well as regional economics, industrial mortality and morbidity, and even small business.

In addition to longitudinal association of records, policy analysis also often requires association of information about individual workers and the particular establishments they work for. Again, the CWHS is the only major file providing an individual-establishment link; it does so only by transferring establishment information (industry and geographic codes) to individual work records. There is no program to summarize worker characteristics associated with individual establishments for, say, small business or industrial mortality and morbidity studies. Perhaps of even greater concern, however, the establishment reporting procedures underlying the CWHS do not always permit individuals to be associated with a unique establishment (i.e., a business operation at a single location), and therefore CWHS establishment records often cannot be reliably associated with establishment data from other record systems.

Clearly there is a need to coordinate the design and implementation of the statistical and administrative programs of various agencies if the information collected by the Government is to be well suited and cost effective for meeting the needs of policy analysis. This need is particularly strong in the case of statistical use of administrative records, because the responsibility for collecting and maintaining administrative records is dispersed among many agencies with widely varying missions and generally limited concern for potential statistical applications. Moreover, even when administrators are interested in better coordination, the necessary access to administrative records for making statistical applications can easily be thwarted because of concerns for ensuring the confidentiality of individual records and protecting the individual right to privacy.

The right to privacy--the right to be left alone--is a cornerstone of a democratic system. Other papers in this session, particularly [1], will be addressing the complex legal issues relating to privacy and confidentiality. It is sufficient to note here that there are a variety of approaches that show promise of meeting the needs of the statistical system while simultaneously ensuring confidentiality and the right to privacy.

We have worked with various microdata for more than a decade. These record systems included both administrative and survey data. Because identifiers were removed, carefully scrambled, or "top-coded" for high incomes, we have been able to use these records for statistical purposes and were effectively precluded from identifying individuals. These and other measures, such as recommended by Dunn [7], would improve access to data and ensure confidentiality and privacy.

It is necessary to merge selected information from a variety of administrative sources in order to meet many statistical needs. We are not, however, recommending the creation of large data banks. As noted by Hansen [11], large data banks with much detailed information for identifiable individuals not only could pose a threat to privacy and confidentiality, but also would represent an inefficient approach to meeting most statistical needs. Instead of "blindly" assembling records into data banks, therefore, attention should be focused on improving coordination of statistical and administrative programs so that much needed improvements can be introduced into the design, accessing, archiving, and documentation procedures of Federal administrative records.

REFERENCES

[1] Alexander, L., Statistical Progeny of Administrative Records: The Legal Issues, 1979 American Statistical Association Proceedings, Survey Research Methods Section.

[2] Alvey, W. and Aziz, F., Quality of Mortality Reporting in SSA Linked Data, 1979 American Statistical Association Proceedings, Survey Research Methods Section.

[3] Caldwell, S.B. and Diamond, E., Income Differentials in Mortality: Some Preliminary Results Based on IRS-SSA Linked Data, 1979 American Statistical Association Proceedings, Survey Research Methods Section.

[4] Cartwright, D.W. and Armknecht, P.A., Statistical Uses of Administrative Records, 1979 American Statistical Association Proceedings, Survey Research Methods Section.

[5] Cole, S.I., Records and Privacy, Draft, Subgroup of the Interagency Task Force on the Health Effects of Ionizing Radiation, HEW, 1979.

[6] Dunn, E.S., Review of Proposal for a National Data Center, Memorandum Report to the Office of Statistical Standards, U.S. Bureau of the Budget, December 1965.

[7] Dunn, E.S., Social Information Processing and Statistical Systems -- Change and Reform, Part IV, "Information Processing in an Open Society," John Wiley and Sons, New York, 1974.

[8] Garnick, D.H. and Gonzalez, M., Statistical Uses of Administrative Records: Where Do We Go from Here? 1979 American Statistical Association Proceedings, Survey Research Methods Section.

[9] General Accounting Office, An Evaluation of the Use of the Transfer Income Model (TRIM) to Analyze Welfare Programs, 1976.

[10] Gittelsohn, A., The Feasibility of Establishing a National Automated Mortality Surveillance System, 1979 American Statistical Association Proceedings, Social Statistics Section.

[11] Hansen, M.H., The Role and Feasibility of a National Data Bank, Based on Matched Records, and Alternatives, in Federal Statistics: Report of the President's Commission, Vol. II, 1971, pp.1-62.

[12] Interagency Task Force on Workplace Safety and Health, Draft Final Report of the Interagency Task Force on Workplace Safety and Health, 1978.

[13] Knott, J. Major Administrative Record Files: Documentation and Potential Uses, 1979 American Statistical Association Proceedings, Survey Methods Research Section.

[14] Koteen, G. and Grayson, P., Quality of Information on Tax returns, 1979 American Statistical Association Proceedings, Survey Research Methods Section.

[15] Rosen, S. and Taubman, P., Changes in Sociodemographic Determinants of Mortality Rates, 1979 American Statistical Association Proceedings, Survey Research Methods Section.

[16] Rosenberg, H.M., et al., Occupation and Industry Information from Death Certificate Assessment of the Completeness of Reporting, 1979 American Statistical Association Proceedings, Survey Research Methods Section.

[17] Rydel, C.P. et al., Welfare Caseload Dynamics in New York City, Rand Corporation, 1974.

[18] Survey Research Center, Institute for Social Research, University of Michigan, A Quantitative Description of the Current Nonhousehold Mailstream, 1978.