THERE OUGHT TO BE A LAW.....

Lois Alexander, Social Security Administration

INTRODUCTION

Many statisticians believe that there ought to be new laws favoring the wider availability of data to statistical users and stronger protections of statistical data from nonstatistical uses. Changes have been proposed for both State and Federal law. 1/

At the Federal level, a draft bill for Confidentiality of Federal Statistical Records has circulated from the Office of Management and Budget (OMB) and has been given serious attention by statistical agencies. 2/ This paper will center on issues of statistical use of information collected by the Federal government, especially the administrative records that Federal agencies maintain in carrying out their programs. It will examine the elements of that draft law with particular reference to administrative records. It will look briefly at existing provisions of Federal confidentiality law and at some of the arguments for legislative change. Finally, it will consider benefits the changes are expected to confer, and new constraints they would impose on the sharing of statistical data.

I. THE DRAFT BILL

A. Threshold Assumptions

An inherent tension exists in Federal confidentiality law, between the belief that Federal records belong to the general public, and the belief that persons have individual rights to the information about them in those records. Privacy law involves an intentional balancing of those views. 3/ Also, the statistician makes a fundamental assumption that may conflict with administrative or law enforcement interests: that it is necessary to give good-faith assurance to persons who contribute information for government statistical uses that the confidentiality of their information will be preserved against administrative uses they have not consented to.

Finally, the U.S. Federal statistical system is neither centralized in its structure and organization, nor uniform in its confidentiality requirements. Its component organizations have their own functions, policies, and independent data needs. In resolving interests that diverge or conflict, there is a complex balance of values and priorities.

The proposed Confidentiality of Statistical Records bill offers an approach to resolving many divergent interests and to establish an element of consistency and stability in confidentiality rules. The rest of this paper will examine that approach, and look briefly at some of the problems that it does not fully resolve.

B. Principles and Objectives

The primary thrust of the draft bill is to differentiate two separate sets of principles for Federal information law; i.e., to develop a distinct and comprehensive set of rules for access and disclosure for statistical data, independent of, but interacting with the rules for records maintained for nonstatistical purposes. 4/ In a nutshell, the logic is that we gain much as a society and lose nothing as individuals from liberal access to information for statistics.

The distinctive feature that characterizes this approach is the one-way flow of identifiers. That is, individually identifiable information from any record source, would be available for certain sanctioned statistical uses, but identifiable statistical information could not be used for any nonstatistical use. Thus, redisclosure could not be made to the administrative source except in a statistical form that does not identify individuals.

Accountability is a common denominator forming a nexus between the privileges and the duties the draft bill would create. The principle of accountability shapes the primary concept of the bill -- specially designated units known as protected Statistical Centers which maintain special files known as Protected Statistical Files. Accountability requirements are imposed through the formality of statute or regulation to designate statistical units that can hold files protected by the law; objective criteria for assigning protected status to statistical files; and strict rules for carrying out procedural and physical safeguards and monitor them.

C. Operation

The three main facets of the bill provide: (a) liberal access to administrative and survey data for statistical use; (b) stringent restrictions on compelled disclosure of statistical data for administrative or compliance use; and (c) protective constraints on the discretionary sharing of identifiable statistical data. These points will be examined in turn.

1. Access - its significance. Access to data for statistical use, particularly access to administrative data, is a primary objective. Liberal access can reduce bias, sample size and the amount of new data to be collected; and can improve sub-sampling methods and the defining of target populations. All these effects reduce the cost of acquiring data.

The statistician wants maximum access to records for sampling and data selection. Though time, money and other resources impose

finite limits on the amount of data that can actually be used, the statistician can control these factors to select data in a scientific way. Missing or unavailable data, however, introduce an element of control by nonstatistical factors that can diminish the value of the data to measure and predict. The structure of the file may affect the mode and efficiency of access.

Census, for instance, has access to its current address listing of (almost) all U.S. households in its Census of Population records, supplemented in its survey records. It uses these as a resource for its continuing efforts to conduct, improve and evaluate its Current Population Survey (CPS) sampling methods and its Census coverage. The Bureau's sampling methods reflect the organization of the files, by household within defined geographical areas. Its usual method is multistage sampling of dwelling units, rather than sampling based on random selection of individuals. 5/

In contrast, the Social Security Administration (SSA) maintains program systems containing individual records of individual workers and beneficiaries, indexed by their Social Security Numbers. To draw a statistically valid national sample of workers covered by the Social Security program, its statisticians have designed digital sampling formulas based on the assignment of SSNs. Because its earnings and benefit records are indexed by SSNs, the agency's statisticians can draw its samples for surveys or for cross-sectional analysis with these formulas. It can also study individuals in its samples over time from the agency's chronological records of earnings and payments. An example is its Continuous Work History Sample (CWHS) which represents one percent of the current work force. At present, this minute sampling fraction produces a sample of about 2-1/2 million individuals, and the file contains a lifetime employment history for each individual. 6/

Thus there are benefits, in terms of both data content and sample design, from unrestricted selection from a file. It is especially important in a file that has been created administratively for a purpose that is different from the study purpose.

Access to identifiable data for epidemiological studies may raise other issues. On the one hand, nothing more than a check of death records may be needed to determine whether particular individuals in a sample have died, and if so, when. At the other extreme, identifying data of deceased persons may be used to contact next-of-kin, former employers, and others. For living individuals it may involve recontacting them for extensive interviews and for medical and psychological testing.

Access to more than one data set containing information about individuals in a given sample presents a different dimension. This aspect is especially important for Federal data users, since the Federal statistical system is not a unified system, but is decentralized among agencies with large systems of individual records. Each system may cover major segments of the U.S. population, with individual units indexed by identifiers (such as the Social Security Number) that are common among the different agencies. Systems have different mixes of information about those individuals, depending on the particular program or mission of the agency. The commonality of individual identifiers makes it possible to combine information about a particular individual from various record sources within an agency, and in principle from the records of other agencies.

Linkage of administrative record data from multiple sources for statistical use is an important and growing area of statistical activity. 7/ The primary purpose of such links is generally to compile more complete data than are available from a single source. Another purpose may be to compare, verify and edit data from different sources. These sources may be surveys or program records, or a combination of both. Individual identifiers are, of course, essential for record linkage.

Under present law, there are substantial impediments to such linkage of data. Some limitations operate to allow only certain persons (e.g., Census employees) to have access to data subject to a particular law. Others, such as the "routine use" and other disclosure rules of the Privacy Act, control the purposes for which identifiable information can be shared. The confidentiality rules of the Internal Revenue Code (enacted in the Tax Reform Act of 1976) are especially stringent, and specify limited classes of authorized recipients, particular classes of data to which those recipients have access, and the particular purposes for which the data can be used. 8/ Part II will present examples of projects in which interagency linkage has been undertaken, and will examine the potential effect of the Confidentiality of Statistical Records bill for this type of project.

2. Compelled disclosure -- the issues.

A recurring nightmare for the statistician is the fear that a court may order the release of data that was obtained under a promise of secrecy. There is substantial belief that the truthfulness and accuracy of statistical information, and sometimes the ability to obtain it at all, are functions of the promise to keep it confidential.

Circumstances that have generated shock waves in the statistical community have usually involved subpoenas or court orders to reveal information collected in confidence, and subsequently considered relevant to some judicial or legislative proceedings. These circumstances are rare but not nonexistent, and situations like the legislative demand for data in the New Jersey negative income tax experiment are widely enough known to raise wide concern. 9/ The law tends to consider

that a judicial order excuses an otherwise valid promise to keep information confidential. Statisticians who are ordered to divulge information they have promised to keep secret have taken little comfort from a judicial excuse. With statisticians, as with newsmen, this issue has produced a collision of values.

When collecting sensitive or potentially embarrassing information the statistician emphasizes confidentiality and the summary form of the study results, to encourage full and open response. The statistician is aware that the respondent is under no obligation to furnish information at all, and rarely sees any direct benefit from participating in a study. The statistician certainly does not want the threat of a court order or the fear that statistical data might be used for enforcement purposes against individual respondents to inhibit their willingness to participate. The reality that the risk is remote will not help the statistician if the respondent's anxiety interferes with good response. When the statistician makes a promise of confidentiality he feels an obligation to honor it, perhaps even honor a promise to resist a court order for disclosure, and risk imprisonment for refusal to comply. However, he understandably resents such a threat, and argues that the law ought to allow him to make and keep a promise of confidentiality when it is required for his work. Without that ability, the consequences of broken promises of confidentiality will impair the future ability to collect data. Whether or not the threat of mandatory disclosure does significant damage generally to the statistical function is a theoretical matter. For the individual data collector, the threat is perceived as an immediate and personal risk.

The government statistician, however, cannot even make a promise of confidentiality unless data collection is done under a law that authorizes such a promise. Unless the Federal agency is authorized by law to withhold particular information, the Freedom of Information Act (FOIA) requires it to disclose it upon request. A requestor can obtain a court order to enforce disclosure. The Census statute and the Internal Revenue Code give protection against such demands, but few other Federal laws do.

The issue is different for the user of previously collected administrative data. The quality or completeness of administrative data is not likely to be affected by its subsequent statistical use. But there is an issue of fairness if individuals are subject to special risk from having their data used for statistics.

Linked data on income from several sources, for example, may indicate that a person selected for study is receiving benefits for which he or she is not eligible, or failing to pay taxes owed. The statistician is concerned to know whether such information is guilty knowledge, and would object to any obligation to disclose it for verification or enforcement.

For example, SSA analysts have encountered difficulty in their efforts to obtain death certificate information from States' Vital Records units. The States fear that information from the records may be used with identifiers by SSA administrators in their fraud detection programs; or may be released by SSA to a commercial or other nonstatistical requestor under the FOIA. SSA's disclosure rules protect information about living persons, but the agency ordinarily releases information about deceased persons unless the disclosure would harm living relatives. When SSA analysts obtain information expressly for statistical purposes, they do not voluntarily disclose identifiable data to anyone not participating in their study. However, the law is not entirely settled on the question of what involuntary disclosure they might be required to make. Because of this uncertainty, it is not clear that the analysts will be able to acquire death certificates to satisfy their data requirements.

The law seems inconsistent when it treats statistical information differently depending on the law of the agency that collects it. The Census Bureau and the National Center for Health Statistics, for instance, operate under statutes that allow them to give the States assurance of confidentiality with certainty. The legal situation of other statistical agencies disregards the important practical fact that the need for confidentiality is not specific to particular agencies but it relates generally to the statistical purpose of obtaining the data, in whatever agency that may be done.

Of course, there are other reasons for attempting to withhold data. An entirely unacceptable reason would be to cover up bad data or dishonest practices on the part of the statistician. The draft confidentiality bill contains provisions intended to protect against such withholding.

3. Discretionary disclosure of statistical data with identifiers.

Voluntary disclosure of data for statistical purposes is sometimes important to the statistician. It may be necessary for the statistician's own projects. In other cases it may be desirable to assist with the statistical work of others. When Federal records are involved, the sharing is restricted by privacy laws.

To support his own work, a statistician may want to obtain information from another source about individuals in the study sample. To do this it may be necessary to identify cases to the other data holder so that the latter can search its records to find and furnish data about the identified individuals or entities. When an agency provides someone else with a

finder list containing individual identifiers, it is actually making a disclosure. On the other hand, a statistical agency may be willing to disclose identifiable data to accommodate another agency's needs, even though the agency making the release has no direct interest in the result. Whether it may do so depends on Federal law.

In an intermediate situation, a statistical unit cooperates in a study in which data are pooled from multiple sources, to be compiled and analyzed in a project of joint interest to several participating statistical units. Such a project assumes disclosure of identifiable data by all, or nearly all, the participants during the processing, and may involve extremely complicated methodological and legal questions of identifiability of the resulting linked microdata. Some examples will be discussed below.

The value of voluntary disclosure may not be apparent to administrative agencies if they see their mission only in terms of particular programs they administer or laws they enforce, and not in terms of their statistical components. The statistical community recognizes a degree of interdependence in the collection and use of data, and in the constant improvement of methodology. Statistical programs benefit indirectly when they can legitimately assist other research even though they do not directly participate in it. Agencies that narrowly restrict their statistical programs and publications lose not only the fruits of their own analysis, but also the seeds of cross-fertilization from the work of others.

The draft bill grows partly out of recognition of long-run benefits from such sharing both to the statistical units and their parent agencies, even though there may be no immediate or exact balancing out between the giver and the receiver of particular knowledge. At the Federal level, it institutionalizes the reasons and rules for sharing data. It also proposes an approach to coordinating policy direction.

The arguments for retaining the decentralized structure of the Federal statistical system still prevail. However, there is wide support for centralizing policy direction to lessen the risks of fragmentation, even though there is no consensus on where those controls should be placed within the political structure.

II. CONSEQUENCES OF THE PROPOSED NEW RULES

Federal confidentiality laws now raise substantial - to some degree impenetrable - barriers to access by statisticians who are not employees of the Federal government, and even by statisticians who are employees of another Federal agency.

The most important change in the proposed bill would give access to virtually all government data, including both administrative and protected statistical data, to a newly-defined class of Federal statistical units. The powerful privilege of access would be based on the statistical nature of the unit's activities (i.e., no administrative or enforcement duties); the confidential character of the data; the requirement of confidentiality as a condition of obtaining that information and above all, to the agency's functional ability to keep its statistical files separate and intact from any nonstatistical use or user. Statistical units not given privileged status by the bill would remain subject to existing rules of access to Federal data for statistical use.

The availability of government program records such as income tax records, Social Security and Medicare records anticipated under the proposed law are meant as inducements to qualify for privileged status of "Protected Statistical Center" under the bill. Another strong inducement is the proposed availability of other organizations' statistical files. Census data are files that are especially attractive to other agencies that would like to be able to acquire them. For others not given special status, the basic form for access would be public use microdata (records about a particular individual) or tables. In addition, a quasi-public use form of release might be available under conditions that would prevent identification of individual entities in a file.

Statistical files in Federal agencies often marry survey and administrative data for individuals in a study. Program records are used as sampling frames for surveys; and the resulting survey data may be linked back to the program records or to other administrative records. When the data sets are all maintained within a single agency, the identifiers (name, SSN, etc.) are likely to be consistent from one data system to another. When interagency linkages are made, technical difficulties may complicate the linkage, and legal restrictions may impede or balk the process.

A simple form of interagency linkage of survey and administrative data is illustrated by the Social Security Number (SSN) validation work performed by SSA for the Bureau of the Census. In certain of its surveys, the Bureau collects SSNs, often reported from memory and sometimes reported by one family member for another. To verify the correctness of the numbers reported to it, Census prepares a finder tape that it furnishes to SSA employees who are also special sworn Census employees. The employees compare the names and SSNs reported to Census with the corresponding information in SSA's file of applications for Social Security Numbers assigned, and use the SSA data to verify or correct data in the file for Census use. 11/

A more complicated process is illustrated by SSA's Retirement History Survey. For that study, the sample of pre-retirement persons aged 58 to 63 in 1969 was selected by Census from its Current Population Survey (CPS) files, and interviewed for SSA. The results of the

survey were linked on an individual basis to earnings histories from SSA's program files. The successful linkage assumes a high degree of coordination between Census and SSA, with safeguards to limit access to identifiable data to SSA employees who were also special sworn employees of the Census Bureau. The linked file was reviewed by both Census and SSA before it was made available in public use form through the National Archives Service.

An example of an internal file based on data from multiple administrative record programs within an agency is SSA's Continuous Disability History Sample (CDHS), a geographically stratified 25 percent sample of persons who have filed for Social Security disabled workers benefits, and for whom a benefit determination has been made. The individual records are compiled from data sets containing personal characteristics, agency decision data, benefit payment data, and earnings history data. The CDHS is valuable as a data base from which custom tabulations are prepared on a reimbursable basis on request. However, its large sampling fraction and risk of individual identification prevent its release in public use microdata form.

It has been possible for SSA to link this file internally with Medicare history information obtained from the Health Care Financing Administration (HCFA), another component of the Department of Health and Human Services. The Privacy Act permits such linkage within the Department. An anomaly of law, however, prevents SSA from releasing the linked file to HCFA, because the file contains earnings information that the Internal Revenue Code makes available to SSA, but not to HCFA. Under the conditions of the proposed statistical bill, SSA could release the file to HCFA if both were Protected Statistical Centers, or if they entered into a special kind of arrangement permitted by the bill.

Another important interagency administrative link is the Continuous Longitudinal Manpower Survey (CLMS) conducted by the Bureau of the Census for the Department of Labor's Employment and Training Administration (ETA). To evaluate selected training programs under the Comprehensive Employment and Training Act (CETA), Census obtained data for CETA participants and for a comparison group of non-participants selected from its own CPS records. To these, it linked data from IRS and SSA records, i.e., earnings records for the pre- and post training periods. The resulting linked files were furnished to the sponsoring agency in public use microdata form. The reimbursable nature of this project, as distinguished from a project performed by Census for its own purposes, raised concern for its status under the disclosure provisions of the Internal Revenue Code. It is not clear that the survey could be replicated, as matters presently are. The proposed Confidentiality of Statistical Records bill would clearly permit future linkage of this type.

Possibly the most ambitious interagency linkages have been undertaken in the Exact Match projects performed jointly by SSA and Census, with cooperation of IRS. The original SSA-Census linkage matched approximately 100,000 records from the 1973 March CPS with corresponding administrative records (approximately 89,000) from SSA's Summary Earnings Record file. This exact match was then expanded to link data from SSA's Master Beneficiary Record file and a limited set of tax items furnished by IRS from 1972 Federal income tax returns. All of the matching processes were performed by Census Bureau employees. These linkages have been thoroughly documented in the series of reports published by SSA as Studies from Interagency Data Linkages.

Subsequent projects linked 1978 and 1980 data, but the later projects involved fewer files and less detail. The intervening enactment of the 1976 Tax Reform Act has complicated the linkage and created serious problems for public use release of the linked files. Current activities are dependent on Census as the catalyst for linkage, and are beset with uncertainty as to the amount of detail and the particular content in files that will be available to the sponsors. The proposed confidentiality bill would remove much of the uncertainty, and create greater parity in access to the resultant linked files by the participating agencies.

Many of the difficulties, uncertainties, and anomalies described above would be resolved in the proposed bill. The Protected Statistical Centers would enjoy expanded access to Census records and tax records, as well as easier transfer of identifiable data among themselves. Beyond that privileged group, however, availability of identifiable statistical data would, if anything, be decreased rather than increased except to non-government statisticians sponsored by Centers.

An earlier recommendation of the Subcommittee on the Use of Administrative Records (a Subcommittee of the Federal Committee on Statistical Methodology) in OFSPS Working Paper 6, was to distinguish natural persons from organizations and other entities when developing standards and practices of record confidentiality. 12/ That recommendation was followed to some extent. However, data about business entities would be subject to the same rules and procedures as data about individuals covered by the proposed law. These rules would not relax the present practices which generally preclude public use release of microdata records about business entities.

III. SUMMARY

Statisticians and other researchers who use government data are restricted in many ways by confidentiality laws, that are largely uncoordinated in their scope and coverage. The "federal statistical establishment" is a

loosely structured federation. Not only is it decentralized, but it encompasses great diversity in the research goals and methodologies of its diffused membership. Perceptions differ as to the problems that confidentiality laws create or solve, and of the solutions that are needed. Diversity increases in the broader context of behavioral and medical research conducted by the federal government, or its social experiments and demonstrations.

There is a growing consensus about certain basic principles for multi-purpose use of information collected by the government from individuals and business entities. These principles focus on safeguards of confidential data; accountability to both the subjects and the providers of statistical data, as well as to the public; and on responsible efforts to avoid improper statistical disclosure. The draft bill attempts to give the decentralized Federal statistical establishment the benefits of a more centralized system: greater parity in agencies' access to one another's administrative and statistical data, and greater uniformity in confidentiality rules.

ACKNOWLEDGEMENT

Affectionate thanks to Weltha Logan and Catherine Murphy-Puryear for doing the impossible.

NOTES AND REFERENCES

1/ See Solutions to Ethical and Legal Problems in Social Research, Academic Press, 1983, edited by Robert Boruch and Joseph Cecil.

2/ The current initiative originated with Dr. James Bonnen in the 1979 Federal Statistical Reorganization Project.

3/ Freedom of Information Act, 5 U.S.C. §552; Privacy Act of 1974, 5 U.S.C. §552a.

4/ Called "functional separation" by the Privacy Protection Study Commission, and defined as "separating the use of information about an individual for a research or statistical purpose from its use in arriving at an administrative or other decision about that individual." Personal Privacy in an Information Society, The Report of the Privacy Protection Study Commission, July 1977.

5/ Technical Paper 40, The Current Population Survey: Design and Methodology, Bureau of the Census, 1978, p. 6.

6/ See Buckler, W. and Smith, C., "The Continous Work History Sample: Description and Contents," Policy Analysis with Social Security Research Files, U.S. Social Security Administration, 1978.

7/ See U.S. Department of Commerce, Office of Federal Statistical Policy and Standards. Statistical Policy Working Paper 6, Report on Statistical Uses of Administrative Records, 1980.

8/ 13 U.S.C. §9; 5 U.S.C. §552a, the Privacy Act of 1974; 26 U.S.C. §6103.

9/ Boruch and Cecil, op. cit.,p.217.

10/ "Such Interesting Tax Returns," The Washington Post, April 27, 1983, p. A22.

11/ This process is not performed for individuals in the sample who refuse to give their SSNs to Census at the time of the Census interview.

12/ U.S. Dept. of Commerce, op. cit., p.2