CONFIDENTIALITY AND THE FUTURE OF THE U.S STATISTICAL SYSTEM

Joseph W. Duncan, Office of Management and Budget

Introduction

The protection of the confidentiality of individual responses to statistical inquiries has long been a paramount consideration in the statistical system. Recently, however, public concern about data record linkages and the privacy of individual records resulted in the passage of the Privacy Act of 1974. This Act, which becomes effective approximately one month from today (September 27, 1975) creates new demands on statistical record systems and presents new challenges to statisticians when using administrative data sources. Therefore, I would like to review with you the statistician's responsibility with respect to confidentiality and to propose some principles concerning where we should go in the future with respect to assuring adequate confidentiality protection for statistical activities. My purpose is to provide a suggested long-range perspective on the confidentiality issues with which we are dealing at the present time.

The Statistician's Responsibility

The statistician has long asserted that the protection of data confidentiality is essential to assure the accuracy of statistical programs. Such protection provides an environment in which a high level of voluntary participation of respondents is assured. Last December, Margaret Martin, in an article entitled "Statistical Legislation and Confidentiality Issues" in the International Statistical Review stated:

> "Even when responses to requests for information are required by law, the success of a statistical program depends in large measure on the willing cooperation of respondents. Respondents who understand the purpose of the inquiry, who sympathize with the intended use of the information, and who believe that providing the Government with the requested information will not harm them are much more likely to answer truthfully and with a minimum of effort on the part of the data collection agency. One element in enlisting such cooperation is the assurance of harmlessness to the respondent, and one of the most common methods for making such assurance in statistical data collection is the provision for keeping the replies confidential."[1/]

The record of statisticians is clear. I do not know of one instance in which there has been a breach of confidentiality pledges by statisticians in the Federal Government. Statisticians have been extremely careful to assure the public and their co-workers that when they have access to administrative records, mailing lists, and individual reports, the identifiable data will be protected and the rights of individuals will not be affected.

In fact, even when statistical aggregates are computed, the statistical agencies have an outstanding record in avoiding leaks of statistical results such as the unemployment rate, Gross National Product estimates, foreign trade balances, population estimates, etc. I am pleased to note that there is not a single case of premature release[2/] that has been brought to my attention since we modified OMB Circular No. A-91 last fall to restrict prerelease access of aggregate results to the President's designated respresentative, the Chairman of the Council of Economic Advisers. That recent revision of A-91 placed clear responsibility upon statistical agencies for keeping the results confidential, and the statisticians have honored their responsibility.[3/]

The statistician's interest in the individual response is primarily related to his examination of the quality of information provided so that he can assess the reliability and utility of statistical aggregates derived from the individual records. The statistician is not interested in information about individuals for action at the individual level. The statistician is interested in information about individuals so that aggregates can be estimated and appropriate analysis can be performed. This distinction concerning the statistician's use of individual records is frequently not fully understood by those who want to restrict the access of statisticians to individual records.[4/]

Finally, before concentrating on the topic of privacy and confidentiality I want to make one important background point that underlies the entire question of seeking statistical information. The imposition of government questionnaires on the public is a clear burden on the respondent's personal time if not on his personal privacy. Therefore, it is essential for statisticians to make certain that the requested information is absolutely necessary and useful. Despite the continuing concerns with reporting burden, it still appears easier to some that it is easier to satisfy a new inquiry rather than maximize the existing data base. Further, the use of the existing base can be considerably enhanced by developing and adhering to more uniform concepts and definitions. Now assuming that needs for data are justified and maximum use of existing data bases are ensured, let us address the privacy and confidentiality issues.

Recent Developments

The purpose of the statistician in developing statistical aggregates is well established. Moreover, the record of the statistician in protecting the confidentiality of sources is unblemished. Nevertheless, it is clear in 1975 that there is growing concern about the character of governmental data collection, both statistical

and nonstatistical, and about the ultimate uses of that information. Particular concern has been expressed about uses of the information held by the Government when the use is not related to the original intent of the data collection. Thus, statisticians using records from tax collection efforts to estimate income distribution, are challenged for using the data in a manner not necessarily intended by the person who was asked to provide the data.

The issue has been posed as an ethical question; is it proper to use the data for a purpose unknown to the provider? The question has significance even if the rights of the particular individual are not directly affected. For example, a participant in a family planning program may not wish to have personal information used in a statistical study of alternative birth control programs because the specific individual may be opposed to consideration of alternative birth control techniques. In another situation, a taxpayer may not want his return to be part of a statistical study of philanthropic support since he is opposed to governmental review of giving patterns to religious organizations.

Both of these examples are given to illustrate the types of concerns which must be explicitly considered. They are not intended to indicate average or widely accepted points of view or to suggest a sound policy position. However, they do represent important social attitudes of persons who want to exercise greater discretion in determining what use will be made of information they provide to the Government. In these cases, harmlessness to the particular individual may not be a sufficient criterion.

In an important way, the current controversy concerning statistical data gathering, began in the mid-1960's with discussions of a National Data Bank. The concepts for a data bank were expressed in the Ruggles Committee Report[5], the Dunn Report[6], the Kaysen Committee Report[7], and the President's Commission on Federal Statistics, which devoted several appendix chapters to topics of confidentiality of statistical data.[8]

The July 1973 HEW report entitled Records, Computers, and the Right of Citizens[9], further addressed these issues and now we have the Privacy Act of 1974 as a specific legislative action designed to assure improved protection of the confidentiality of individual records, and to provide individuals with better information on the types of individual data files which are maintained by the Federal Government. The protection provided is limited by the number of exemptions, provided for in the Act,[10] for disclosure of identifiable data without the individual's prior consent. As social scientists and social statisticians, I do not need to remind members of this audience of the growing disaffection of the general public with institutions of all forms and especially the skepticism and distrust of government which exists today. For the past three years, daily headlines have focused the public on topics such as:

. The abuse of power associated with Watergate,

. The battles of consumers with computer errors affecting their personal credit standing,

. The maintenance of dossiers on radical groups by the Federal Bureau of Investigation,

. Illegal surveillance of mail and individuals by the Central Intelligence Agency, and

. Maintenance of special files on political activists by the Internal Revenue Service.

In this environment the statistician, although he has not been targeted as invading individual rights or privacy, must recognize the growing skepticism of the public with respect to governmental inquiries about the characteristics of individuals.

For the balance of today's discussion, I will not be dealing with these current problems of the interpretation of the Privacy Act directly. These topics will be discussed later in the ASA program in the invited paper session entitled "Privacy, Freedom of Information, and Federal Statistical Programs." Rather, I will discuss some basic principles relating to confidentiality which I believe are important to the Federal Statistical System. I make these points to reflect a personal, not official, OMB perspective. I believe that statisticians need to do much more work on defining and articulating these or related principles so that future legislation and actions will provide a sound balance between the benefits of better knowledge about social processes and the individual's right to fair information policy and practice.

Privacy and Confidentiality

One of the basic distinctions which is important to a discussion of the confidentiality of statistical data systems relates to the difference between individual privacy and individual record confidentiality. As defined in Webster's Third New International Dictionary, privacy is "The quality or state of being apart from the company or observation of others" with a secondary definition relating to "freedom from unauthorized oversight or observation." Clearly the individual who wants to maintain his absolute privacy is unwilling to participate in voluntary statistical inquiries or to provide data about his personal situation other than that which is absolutely required for him to qualify for certain benefits or programs or under penalty of law (e.g., taxes). In extreme cases some individuals seeking absolute privacy will not even apply for participation in various programs because of the personal information required in the application for such programs.[11]

In contrast, confidentiality is defined as "known only to a limited few: not publicly disseminated." Confidentiality is specifically the

quality or state of being confidential (private or secret), i.e., not freely disclosed and a confidential relationship is one which "is indicative of intimacy, mutual trust or willingness to confide." Hence, the confidentiality of information relates to the trust of the provider of the information that the information will not be inappropriately disseminated or used in identifiable form to hurt him. Many of the issues concerning confidentiality relate to definition of the limited few who will have access to the information, and to the specification of the responsibilities of that few.

The general fear of the public is that data provided for an explicit purpose will be misused to affect their rights, benefits, or privileges or to provide an opening to investigation and/or regulation. The computer is seen as the enabling mechanism where information provided in various aspects of one's social contacts will be collated, compared, analyzed, and used to restrict the freedoms of the individual.

With this concern, it is obviously difficult for the statistician to discuss record linkage or access to administrative records without compounding the fear of the individual that the result will be available to the nonstatistician for administrative purposes.

Basic Principles

The basic principles which should be observed by statistical agencies in personal data systems used exclusively for statistical reporting and research are explicitly outlined in the HEW report on Records, Computers and Rights of Citizens and, in large part, are embodied in the Privacy Act of 1974. These principles are:

1. The individual must be informed when "asked to supply personal data for the system whether he is legally required, or may refuse, to supply the data requested, and also of any specific consequences for him, which are known to the organization, of providing or not providing such data."

2. The agency should "assure that no use of individually identifiable data is made that is not within the stated purposes of the system as reasonably understood by the individual, unless the informed consent of the individual has been explicitly obtained."

3. The agency should further "assure that no data about an individual are made available from the system in response to a demand for data made by means of compulsory legal process, unless the individual to whom the data pertain (a) has been notified of the demand and (b) has been afforded full access of the data before they are made available in response to the demand."

These principles are generally acceptable and have been explicitly underscored in OMB's statistical standards since May 1974. The current OMB Circular No. A-46 (as revised May 3, 1974) contains a section on relations with the public (Section 7 in Exhibit A) which states as follows:

> "Finally, maintenance of good relations with the public is essential if Federal statistics are to continue to merit public support. Objectivity and integrity in the compilation and presentation of statistics is the surest means of obtaining such support. Particular attention, however, should be given to relations with respondents and users of the statistics...to the extent possible respondents should be reassured that their interests are being protected. Agencies collecting data for general statistical purposes are usually in a position to assure respondents that the information they supply will be used only for statistical tabulations, and that individual returns will be kept confidential. Respondents to other types of surveys should be informed of the use of the data and extent of the disclosure. Agencies collecting data from business respondents particularly should be aware of the provisions of the Freedom of Information Act (P.L. 89-487) and may need to consult legal counsel on the extent to which confidentiality may be pledged...care must be taken to avoid giving respondents the impression that they must respond to surveys which are voluntary. For this reason, the Office of Management and Budget has prohibited a statement on the form or in the letter of transmittal that this survey is authorized by law in surveys where response is not mandatory. Where response is mandatory, this should be indicated and the applicable statute should be cited.
>
> "If response is voluntary, cooperation can best be obtained by explaining the purposes for which the data are to be used and by stating clearly and persuasively the needs for the data by the Government or the public."[12/]

In addition to the principles outlined in the HEW report quoted earlier, I believe there are further principles which should be pursued in future development of the U.S. statistical system. These are:

1. Statistical agencies should have mandated legislative protection for the confidentiality of information collected solely for statistical purposes. This should apply to both corporate and personal data. The element of trust which is involved in voluntary submission of data should be backed up by clearly mandated protections so there is no uncertainty concerning the confidential nature of the data submission and so that voluntary data collection programs are effective. Even in mandated data collection efforts, it is essential to have cooperation of respondents if the data submission is to be accurate and comprehensive. Protection from disclosures helps assure that the quality of submission is of the highest possible order.

The HEW report suggests the following features for protection against compulsory disclosure:

> "The data to be protected should be limited to those used exclusively for statistical reporting or research. Thus, the protection would apply to statistical-reporting and research data derived from administrative records, and kept apart from them, but not to the administrative records themselves.
>
> "The protection should be limited to data identifiable with, or traceable to, specific individuals. When data are released in statistical form, reasonable precautions to protect against 'statistical disclosure' should be considered to fulfill the obligation to disclose data that can be traced to specific individuals.
>
> "The protection should be specific enough to qualify for non-disclosure under the Freedom of Information Act exemption for matters 'specifically exempted from disclosure by statute.' 5 U.S.C. 552(b)(3).
>
> "The protection should be available for data in the custody of all statistical-reporting and research systems, whether supported by Federal funds or not.
>
> "Either the data custodian or the individual about whom data are sought by legal process should be able to invoke the protection, but only the individual should be able to waive it.
>
> "The Federal law should be controlling; no State statute should be taken to interfere with the protection it provides."[13]

2. The uses of statistical data must be restricted to prevent their use in identifiable form for making determinations which affect a particular respondent. While this is partially covered in the first principle, it should be explicit that the confidentiality of the statistical data means that these data sets are not available for other regulatory, administrative, or judicial purposes within the same agency or department collecting the data. Hence, environmental data collected for statistical purposes should not be used for regulatory purposes. The distinction between regulatory and statistical uses must be made clear at the outset, and there must be no possibility of divergence in these uses. In effect, statistical data in statistical agencies is thus placed in a "protected enclave."

3. Exchange of data among the "protected enclaves" should be feasible under controlled conditions. Comprehensive data systems concerning the interrelationships among various aspects of social and economic patterns requires that various data sets be combined and studied jointly. Once the principle is set forth that the data will only be used for statistical purposes, there should be no concern about the exchange of information among statistical agencies which have "protective enclave" status in law and position to assure confidentiality to provide for data enrichment and correlation analyses.

This principle for statistical data systems is by far the most controversial, especially among those individuals who wish complete knowledge and control of uses of data pertaining to them held by Federal agencies. For the long-range development of sound statistical information or social processes, however, I believe it is essential.

The first step in achieving this situation is the development of a clear legal status for "protected enclaves" for selected statistical agencies in the major departments. The statistical agency must be free of intervention in terms of unauthorized access to data. Employees should be subject to strict ethical standards established with respect to data handling. Once the individual has agreed to provide information for statistical purposes, there should be a mechanism for transferring identifiable data among agencies under controlled conditions. At a minimum this requires:

a. A statement at time of data collection about the character of potential statistical uses;

b. A review agency that has power to authorize transfers;

c. A clear set of criteria that specify when transfer of identifiable data would qualify as being of sufficient public interest to justify the transfer; and

d. A set of procedures to provide for removal of identifiers or destruction of the basic data files after the basic purposes of the transfer have been achieved.

David Hulett has identified some uses which might tend to demonstrate a sufficient public interest to justify a transfer. These are:

> "To avoid an increase in the burden on the public in reporting duplicate information to two different agencies. This principle underlies the Federal Reports Act. In addition, a Federal Paperwork Commission will soon be established to study ways to reduce the burden on the public of Federal requests for information. In its deliberations, the Commission will consider the guarantee of appropriate standards of confidentiality as well as the need of the Government for information. The sharing of data between agencies may well be an important item on the Commission's agenda, since in some cases, the transfer of identifiable information among agencies largely eliminates the need to collect further data.

"To ensure the accuracy, timeliness, and consistency of major statistical or research reports. In some cases, several agencies collect data which are logically related (e.g., production and prices, or income and occupation) and must use consistent samples drawn from the same universe for their data to be related. In most cases, the data which are finally published are collected directly from the respondents.

"To utilize data not obtainable from other sources. In retrospective studies of health or work history, for instance, a given set of data maintained by another agency is simply the only source of information."[14]

4. Administrative data sets should be accessible to statisticians for some statistical uses unrelated to the original data collection. In certain cases statistical agencies need to use administrative records for establishing sample frames for verifying the total universe characteristics. Identifiable data extracted from administrative records for statistical purposes should be held confidential by the statistical agency which receives them in the same manner that data collected directly from the respondent are held confidential. In essence, this suggests the creation of a "protected data set" composed of those items derived from administrative sources for use in the "protected enclave." Thus, subpoena and other access to the original identifiable data would be through the original administrative submission, not through the statistical agency. At the same time, the controlled exchange of data extracted from administrative records among statistical agencies would not be restricted further than the process defined in 3 above, would imply.

The above principles place statistical data in a special class of information. To summarize, it must be made clear from the outset in the laws which would implement these principles that: (a) these data may not be used for determining the benefits, rights, and privileges of individuals or of corporations, and (b) the sole use of these data is for use in determining statistical relationships and preparing statistical aggregates. Such protection of these data would be uniquely strong. Therefore, a controlled exchange of statistical data could appropriately be encouraged to improve the accuracy and comprehensiveness of the various measures employed, as well as to assure reduced costs of data collection and minimum reporting burden.

Further Developments are Needed

To facilitate the development of these principles, it will be useful for statisticians to explore specific techniques such as random rounding of individual data so that sets of microdata can be made more accessible to the public without revealing the characteristics of individual respondents.

I firmly believe that the development of a system of social and demographic accounts, not unlike the National Income Accounting framework for economic statistics, is a necessary future development. This will require statisticians to devise procedures for linking, through statistical matching or direct record linkage, the various data sets which describe important features of socioeconomic groups. Thus, data on education, health care, criminal justice, etc., need to be related in order to develop a comprehensive picture of the social condition. This will undoubtedly require innovative techniques in statistical record kinkage and, insofar as the confidentiality of the individual records is concerned, the pioneering research in this important area must consider ways and means for assuring that confidentiality is not breached.

There are a number of specific areas that require further work by statisticians so that the concept of controlled flow among statistical enclaves[15] can proceed efficiently:

1. The development of optimum grouping techniques, such as those developed by Mosteller, Greenberg, Gastwirth, Kulldorff, et al. These techniques are related to methods based on order statistics which have yielded quite efficient estimates of the parameters of the normal, exponential, and other commonly used distributions in statistics. As the best choice of order statistics, or grouping intervals, depends on the parameters of interest, perhaps methods can be devised which will allow the merging of grouped data which will enable statisticians to estimate the relationships between the basic variables without linking the individual records.

2. The controls on record linkage and the criteria for such exchange need careful conceptual development to assure that the agencies adhere to the basic purposes and principles of confidentiality.

3. Standards for the quantity and quality of data to be linked must be established. Further, specification of time intervals for retention of individual identifiers must be established.

4. Ethical standards and penalties for abuse of these standards should be the subject of wide professional review, perhaps with ASA proposing a set of minimum standards to the agencies.

Finally, of course, the statistics profession has a responsibility for demonstrating to the public the benefits of statistical data gathering, protection, and linkage. The constructive features of the Privacy Act of 1974 must be promoted (letting respondents know how data will be used, what exists, and what files have been developed) and extended (protection from subpoena, etc.).

Conclusions

During this luncheon discussion, I have tried to review some basic principles of statistical treatment of individual data which are important from the viewpoint of protecting the rights of

the individual respondent (person or corporation). The basic proposition of this discussion revolves around the distinction between privacy and confidentiality. In voluntary inquiries the respondent determines what private information he wishes to disclose -- the statistician has a responsibility to protect the confidentiality of that information. For mandatory submissions, the requesting agency must assure that the requested information is necessary, in addition to the clear responsibility to protect the confidentiality of the submission.

Once the data are in the statistical system, I have proposed a set of protected enclaves (statistical units) which are immune from outside access. High ethical standards should be placed upon the members of these units. Given this protection and high standards of performance, it should then be established that controlled flow of data from one enclave to another is appropriate when necessary for improving the accuracy, timeliness, or quality of important statistical information. Finally, I have proposed several areas where further work in statistical technique and professional standards would facilitate the development of this "protected data" system.

Until the development of this more desirable statistical environment, we must work hard to make certain that the individual knows what uses and protections we now give to the data. The dilemmas of balancing the privacy concern and the social need-to-know will continue to challenge all of us.

1/ Margaret Martin, "Statistical Legislative and Confidentiality Issues," International Statistical Review, Volume 42, Number 3, December 1974, page 265.

2/ The author's concern with release policy predates his present role in OMB. See Joseph W. Duncan, et al., "Maintaining the Professional Integrity of Federal Statistics," The American Statistician, Vol. 27, No. 2, April 1973, pages 58-67.

3/ Several instances of prerelease have been noted in the press but all of those known to the author predated the more stringent policy revision by OMB and the President.

4/ This point is discussed by Taeuber, Richard C., in the "Right of Privacy," Bulletin of the American Society for Information Science, Vol. 1, No. 10, May 1975, pages 17 and 18.

5/ Ruggles, Richard, Chairman, Report of the Committee on the Preservation and Use of Economic Data, Social Science Research Council, 1965.

6/ Dunn, Edgar S., Jr., "Review of Proposals for a National Data Center," Statistical Evaluation Report No. 6, Office of Statistical Standards (now Statistical Policy Division of the Office of Management and Budget), November 1965.

7/ Kaysen, Carl, Chairman, "Report of the Task Force on the Storage of and Access to Government Statistics," The American Statistician, Vol. 23, No. 3, June 1969.

8/ U.S. Government Federal Statistics: Report of the President's Commission, Volume II, U.S. Government Printing Office, 1971. (Stock Number 4000-0269).

9/ Report of the Secretary's Advisory Committee on Automated Personal Data Systems, U.S. Department of Health, Education and Welfare, U.S. Government Printing Office, July 1973 (Stock Number 1700-00116).

10/ Duncan, Joseph W., "The Impact of Privacy Legislation on the Federal Statistical System," Review of Public Data Use, Vol. 3, No. 1, January 1975, page 51.

11/ This is discussed in the recently adopted regulations concerning the "Confidentiality of Alcohol and Drug Abuse Patient Records" (Federal Register, July 1, 1975, Vol. 40, No. 127, Part IV) when discussing the topic of privacy and social needs:

> "The other aspect of the right of privacy, which has sometimes been described as the right to be left alone, is the notion that an individual has a right not to be hurt by intrusions into his essentially personal concerns, or to have essentially private information exploited for commercial gain, whether or not the intrusion or exploitation is in connection with any possible governmental action against him. The courts have spoken of a right of privacy in a wide variety of contexts, but they have repeatedly and explicity rejected the notion that anyone has a right to go about his daily affairs encapsulated in an impenetrable bubble of anonymity. The courts have been careful to weigh the competing interests, and the social interest in valid research and evaluation is clearly of sufficient moment to be considered in this process."

12/ OMB Circular No. A-46, pages 10 and 11.

13/ Report of the Secretary's Advisory Committee on Automated Personal Data Systems, U.S. Department of Health, Education and Welfare, U.S. Government Printing Office, July 1973 (Stock Number 1700-00116).

14/ David T. Hulett, "Confidentiality of Statistical and Research Data and the Privacy Act of 1974," Statistical Reporter, June 1975, page 203.

15/ In addition, the confidentiality concerns require specific attention to release of data for statistical use outside the agencies. While this is a broad subject of another article, it does require development of improved techniques for disclosure analysis of merged data sets so that public use samples can be released with assurance that individuals cannot be identified.