Journal of Official Statistics, Vol. 19, No. 4, 2003, pp. 321–331

The Eleventh Morris Hansen Lecture Information and Statistical Data: A Distinction With a Difference

Norman M. Bradburn[1] *and Miron L. Straf*[2]

We distinguish between *information* about individuals and *statistical data*, a representation of information that does not identify any individual. Implications of this distinction are: (1) neither privacy nor confidentiality pertains to statistical data but rather to information; (2) individuals may have the right to control their information, but not the right to control statistical data derived from that information; (3) information should be protected by providers from release, but statistical data need to be safeguarded with sanctions on users; (4) zero tolerance can be a standard for the release of information, but, for statistical data, the standard must be to exercise reasonable care so that the risk of an identification is very small. Finally, we discuss how recent legislation in the U.S. can be used to strengthen confidentiality, increase research access to statistical and other data, and make available more useful statistical data.

Key words: Confidentiality; data access; privacy; research data; statistical purposes.

1. Introduction

The credibility of the information that guides our public policies depends upon the willingness of people and businesses to respond to the censuses and surveys of our federal statistical agencies completely and honestly. That willingness and the quality of the responses depend upon a mutual relationship of respect and trust between the government and its citizens and businesses (Martin, Straf, and Citro 2001). A pillar of that relationship is the principle of protecting the confidentiality of their responses.

Recent landmark legislation in the U.S., the Confidential Information Protection and Statistical Efficiency Act, Title V of the E-Government Act of 2002 (P.L. 107–347), explicitly acknowledges the importance of this trust and extends confidentiality protection to information collected exclusively for statistical purposes by any government agency. The passage of this law is a major accomplishment for which the leadership of the U.S. statistical system deserves praise. Wallman (2003), in this issue, describes the background and particulars of the law.

The law does far more than protect the confidentiality of information. It also provides safeguards under which information can be analyzed for research and other statistical

[1] The National Science Foundation, Social, Behavioral and Economic Sciences, 4201 Wilson Boulevard, Arlington, VA 22230, U.S.A. Email: nbradbur@nsf.gov
[2] The National Academies, Division of Behavioral and Social Sciences and Education, 500 Fifth Street, N.W., Washington, DC 20001, U.S.A. Email: mstraf@nas.edu
Acknowledgments: The authors are grateful for comments from John Fanning, Thomas Jabine, Marilyn Seastrom, Eleanor Singer, Katherine Wallman, and Al Zarate. The opinions expressed herein are those of the authors and not necessarily of their institutions.

purposes. Statistical agencies can use this legislation in creative ways to strengthen the protection of confidentiality while at the same time increasing access to information for these important purposes. In this way, we can respect respondents by both protecting the confidentiality of their information and putting statistics derived from the information they provide to more effective use in benefiting society.

The law is timely. Statistical agencies must find ways to increase access for research and statistical purposes, especially those important to public policy. The demand for data for research to inform public policies and contribute to our well-being is ever-increasing. If statistical agencies do not meet this demand, it will be met by the private sector, which with even today's technology can easily collect and assimilate much more data than our statistical agencies can. An example of research in which data from a marketing firm were analyzed for a public policy issue is given by Putnam (2000).

Private surveys, however, often lack the statistical rigor and quality of those conducted by the statistical agencies and may not cover important topics. Without the leadership of our statistical agencies in providing data for important research and statistical purposes, a type of Gresham's law will prevail. Bad data will drive out good.

If our statistical agencies are unable to provide data for critical research purposes, they could lose their constituencies and, in some cases, the very rationale for collecting their data. It is not a breach of confidentiality that is the biggest threat to our statistical agencies; it is a loss of relevance.

Fortunately, with this new legislation, there is much we can do. Before we describe strategies this law will enable, we first clarify what is meant by confidentiality. That requires drawing an important distinction for what we commonly and ambiguously refer to as ''data.''

2. Information and Statistical Data

We draw a distinction between *information* about individuals and *statistical data*, a representation of information that does not identify any individual. By ''individual'' here, we mean a person, household, establishment, or company.

The representation can be an aggregate, such as a table in the *Statistical Abstract*, or a micro data file for which means have been taken to minimize the probability that any individual can be identified with their information used to produce the file.

Information is about individuals. Statistical data are about quantities that are useful only for describing patterns and relationships among groups of individuals. Statistical data are thus useful only for statistical purposes. In particular, statistical data cannot be used solely by themselves to provide information about any individual.

The problem is that statistical data can be turned into information. But that can only be done through the use of information from another source. There are many ways to minimize the risk of this transformation. We will later describe some of these ways. So, assume, for the moment, this distinction between information and statistical data, and let us examine its implications.

3. Privacy and Confidentiality

Alan Westin developed the concept of information privacy. Its central feature is the right of individuals to control the use of information about themselves. *Information privacy*

pertains to ''the claim of individuals, groups or institutions to determine for themselves when, how, and to what extent information about them is communicated to others'' (Westin 1967, p. 7).

Privacy is a right that pertains to individuals. ''*Confidentiality*,'' on the other hand, pertains to information. Operationally, confidentiality means protecting against the release of information about an individual to others except for specific purposes.

The first implication of the distinction between information and statistical data is that neither privacy nor confidentiality pertains to statistical data. Both privacy and confidentiality apply to the information that is associated with an individual. Statistical data are measurements or some other representation of information that cannot be associated with any particular individual. No element of statistical data is linked to any individual. Therefore, for any individual, statistical data do not describe information about him or her, not even whether the individual is a he or a she.

If statistical data cannot provide information about you personally, then there is no information that you personally need to protect. Confidentiality becomes irrelevant.

By applying the term *confidentiality* to statistical data we do ourselves a great disservice. The term *confidential* evokes an image of a security classification. Confidential information should not be disclosed to others except those with proper clearance and a ''need to know.'' Or the term reflects the image of a confidential communication made to a doctor, priest, attorney, or spouse who cannot be legally compelled to divulge the information.

Neither of these images applies to statistical data. And the more we talk about confidentiality as if these images do apply to statistical data, the more restricted we may become in what types of statistical data we can produce. For example, if we confuse information with statistical data, then so also might the courts, with the result that agencies may be prohibited from issuing even relatively safe public-use data files.

4. Informed Consent

The principle that people should have control over who has access to their information is sometimes taken to mean that statistical data derived from information collected for a purpose to which an individual has given informed consent cannot be used for a research or statistical purpose that is different without seeking the informed consent of the individual for this other purpose. The second implication of the distinction between information and statistical data is that, although individuals may have the right to control their information, they do not have the right to control statistical data derived from that information. Statistical data are not information.

Simply put, statistical data can be used for all research and statistical purposes. It is not a violation of confidentiality to produce statistical data from one's information or to use those statistical data for a purpose different from the one for collecting the information from which the statistical data were derived.

5. Protecting Information and Safeguarding Statistical Data

The third and most important implication of the distinction between information and statistical data is in how we must protect them from misuse that, for example, could cause embarrassment or other harm to an individual.

Information collected by a government agency for statistical purposes must be protected by maintaining confidentiality of the information. The images that confidentiality evokes of security, controlled release, and freedom from being compelled to disclose information are all appropriate here.

For information on individuals, we need stronger locks and fewer keys. Statistical data, however, need to be protected in a different way.

The sine qua non of statistical data is that they are not identified with any individual. Before becoming a statistical data element, an observation must be stripped of immediate identifiers, such as name, address, and social security number. But such an observation could still be identified along with the individual if it includes a unique element that is also in some other file that contains information. In that case, the record of the individual in the information file could be matched to the observation with the unique element. Examples of elements that may be unique include telephone numbers and a history of Medicare payments or social security earnings.

The problem is compounded when a combination of elements may be unique to an individual and that same combination may also be included in some information file. An example of such a combination of elements is detailed geographic location down to the block and a very rare high level of income. Moreover, the identity of an individual with their information might be inferred from special information known to others, such as that one's parent is a respondent to a particular survey.

Agencies take precautions against such identity disclosure through two main strategies. One is to alter the data so as to make identification of any individual with their information nearly impossible, or at least very difficult. This body of techniques is sometimes referred to as ''de-identifying'' or ''masking'' the data. Another common term is ''statistical disclosure control.'' The second is to restrict access to the statistical data files in ways that subject those who have access to severe penalties for disclosing individually identifiable data.

5.1. Altering or ''de-identifying'' data

The simplest and most common form of ''de-identifying'' data files is to remove the link between statistical data and personal information that permits easy identification, such as name, address, date of birth, or any unique identifier, such as social security number, patient record number, or billing number. It is not easy to specify how much information needs to be removed in order for files to be secure. Geographic information, such as zip code, is one of the most problematic types of information, since it is often useful for analytic purposes. Unfortunately, it also reveals data about the individual that, when joined with other publicly available data, may permit identification of individuals either specifically or with a high probability. The increasing availability of large lists of individuals through the World Wide Web and the existence of fast computers have opened up new possibilities for identification through matching, which may cast doubt on the efficacy of our usual practices in ''de-identifying'' data sets (see Sweeney 2001).

A second technique, often used in conjunction with the first, is to alter the raw information by coding it into categories, for example coding age into 5-year categories such

as 20–24 or 65–69, or collapsing categories that have relatively few individuals in them into larger categories, such as combining those with incomes between $100,000 and $199,000 with those who have incomes of at least $200,000 into a category of ''$100,000 or over.'' Again it is difficult to be sure how much collapsing one needs to do to make the file secure from identification. One standard, which may be used for U.S. National Center for Education Statistics data, is that all tabulations that may be produced with a cell size of one or two cases be re-categorized to insure that each cell in the table has at least three cases (NCES 2002).

As data files become larger, particularly with data collected over time or with more detailed geographic information, the difficulty of protecting identities becomes greater. A more controversial technique involves adding statistical ''noise'' to the data, that is, randomly altering some values for measurements on individuals. In this way no one using the data file can be sure that, after an exhaustive cross-classification of characteristics, they have actually obtained the true value. Thus they could not identify an individual with certainty. The deliberate introduction of error into a data file, even to protect confidentiality, is anathema to many researchers. If done properly, however, it does create uncertainty about the relation of any particular measurement to the underlying individual measurements that gave rise to it, so that one can never be sure that a data record that has somehow been identified with an individual through exhaustive analysis of the file is in fact the record of that individual. It could be a randomly altered record and belong to someone else.

The introduction of such random error into the data need not create any bias in the data or affect the validity of any relationships found through statistical analyses. It does, however, reduce the power of the data by making it more ''noisy,'' and, depending upon how much error is introduced, may lead to a failure to detect some valid relationships. The more ''noise'' introduced, the more protection from identification there is, but at the same time, the more difficult it makes it to detect meaningful relationships. There are no good guidelines about how much alteration you need to introduce in order to get a quantifiable increase in protection.

Statistical techniques using multiple imputations are being developed that are an alternative to protecting data files from identification, but at the expense of creating synthetic data. The data reproduce the statistical properties of the actual data file, but no data element is linked to an actual individual. Kennickell has applied this technique in order to release a statistical data file derived from the Survey of Consumer Finance (Mackie and Bradburn 2000).

These techniques are still new, computationally intensive, and controversial. They are an extension of the data altering techniques mentioned above, but done in a more formal and controlled manner. It is too soon to know if they can be widely used to permit public-use statistical data files to be developed from data files with very sensitive information or those that are difficult to ''de-identify'' because of the nature of the data or the characteristics of the individuals in the file that might permit identification.

Even with these techniques, the risk that some individual may be identified or thought to be identified, although small, is never zero. Moreover, in the minds of many researchers, these techniques compromise the utility of the statistical data.

5.2. *Restricting access*

The second general strategy does not do anything to the data, but rather restricts access to the data to researchers who may be subject to the same kinds of sanctions that are imposed on those working directly with the raw data. Within this strategy, there are two forms. The less restrictive is to license researchers to use micro data that do not have the protection afforded in public-use files. In such cases, typically researchers who want to use the micro data have to apply to the holders of the data, describe the data they want to use, make the case for the necessity of having access to the micro data, describe the procedures they will use to protect the data at the site where they will be using them, and sign documents that subject them to penalties if they violate the conditions of use or the confidentiality of the data.

This system is used by some U.S. federal statistical agencies to make data available that would not pass the tests for a publicly available statistical data file. Such micro data are still not raw data and do not contain individual identifiers, but rather contain data at a more disaggregated level that might make it easier to identify a particular individual. We still refer to such data as statistical data, because, with restricted access, oversight, care that nothing leaves the restricted access site that could be used to identify an individual, and penalties for disclosure, measures have been taken to minimize the risk that any individual can be identified.

The stronger version of this strategy is to extend access to micro data to individual researchers after a lengthy application process but only at selected, highly protected sites, sometimes called data enclaves. The users must come to the data enclaves, rather than obtaining the data for their use at their home institutions, and use the data files in the enclave under strict conditions that prevent data from being copied or leaving the enclave. Basically, this strategy extends the confidentiality umbrella given to employees of the data collector to selected other individuals who use the data in research projects consonant with the purposes for which the data were originally collected. The risks of breach of confidentiality are no greater in this case than they are with the original conditions of collection and storage. This strategy provides greater protection, but at the cost of reduced access.

These protective measures – altering data or restricting access to them – are important, but they compromise our ability to use the data and, therefore, reduce their value. As the means become more and more restrictive, fewer statistical data will be available and the utility of what statistical data are available will be reduced. Fortunately, the new legislation broadens the range of options for statistical agencies to make statistical data available in ways that protect the confidentiality of information.

6. The New Confidentiality Law

The Confidential Information Protection and Statistical Efficiency Act of 2002 (CIPSEA, P.L. 107–347, Title V, Subtitle A) places strong sanctions on users who enter an agreement with an agency to use its data for only statistical purposes and to comply with the law:

> Whoever, being an officer, employee, or agent of an agency acquiring information for exclusively statistical purposes, having taken and subscribed the oath of office, or having sworn to observe the limitations imposed by section 512 [limitations on use and disclosure

of data and information], comes into possession of such information by reason of his or her being an officer, employee, or agent and, knowing that the disclosure of the specific information is prohibited under the provisions of this title, willfully discloses the information in any manner to a person or agency not entitled to receive it, shall be guilty of a class E felony and imprisoned for not more that 5 years or fined not more than $250,000 or both. (Sec. 513).

The term ''agent'' includes ''a researcher affiliated with an institution of higher learning . . . and with whom a contract or other agreement is executed, on a temporary basis, by an executive agency to perform exclusively statistical activities under the control and supervision of an officer or employee of that agency'' (Sec. 502, (2)(A)(i)).

This provision can facilitate access to confidential information at restricted sites or through licenses or other agreements. The provision is similar to that in legislation governing data of the National Center for Education Statistics (NCES). With that legislation NCES has been issuing licenses to research institutions that agree to protect the confidentiality of information and subject themselves to unannounced inspections of their procedures and practices.

In the NCES legislation, federal penalties are provided for any person who uses the agency's data ''in conjunction with any other information or technique, to identify any individual student, teacher, administrator, or other individual and who knowingly discloses, publishes, or uses such data for a purpose other than a statistical purpose'' (P.L. 107–279, Sec. 183(d)(6)). The penalties also apply to those who ''use any individually identifiable information . . . for any purpose other than a research, statistics, or evaluation purpose'' or ''make any publication whereby the data furnished by any particular person . . . can be identified'' (Sec. 183(c)(2)).

7. Safeguarding Statistical Data

The NCES legislation appears to go further than CIPSEA by safeguarding statistical data wherever they may be through prohibiting their use to identify any individual. Users who have obtained NCES data, even a public-use data file, are subject to felony penalties if they use those data to identify an individual and knowingly disclose the individual's information.

It is possible that CIPSEA could provide this confidentiality protection for all agencies. The law provides for the Director of the Office of Management and Budget to ''promulgate rules or provide other guidance to ensure consistent interpretation of this Act by the affected agencies'' (Sec. 503(a)). Otherwise, the law could be amended.

Imagine that the law also safeguarded statistical data wherever they may be by prohibiting their use to identify any individual. Statistical data files could then be labeled with a simple, but firm, warning:

Warning. Data in this file are protected by law. It is a felony to identify any individual corresponding to the data in this file.

In fact the National Center for Health Statistics issues a warning to accompany its public-use files (See Box 1). The warning, however, does not have the teeth of a felony penalty. NCES, which has a stronger law, has also adopted an appropriate warning to accompany the statistical data it provides (See Box 2).

Box 1

National Center for Health Statistics Warning for Public-Use Data Files

WARNING – DATA USE RESTRICTIONS!
Read Carefully Before Use.

The Public Health Service Act (Section 308 (d)) provides that the data collected by the National Center for Health Statistics (NCHS), Centers for Disease Control and Prevention (CDC), may be used only for the purpose of health statistical reporting and analysis.

Any effort to determine the identity of any reported case is prohibited by this law.

NCHS does all it can to assure that the identity of data subjects cannot be disclosed. All direct identifiers, as well as any characteristics that might lead to identification, are omitted from the data files. Any intentional identification or disclosure of a person or establishment violates the assurances of confidentiality given to the providers of the information. Therefore, users will:

1. Use the data in these data files for statistical reporting and analysis only.
2. Make no use of the identity of any person or establishment discovered inadvertently and advise the Director, NCHS, of any such discovery (301-458-4500).
3. Not link these data files with individually identifiable data from other NCHS or non-NCHS data files.

By using these data, you signify your agreement to comply with the above-stated statutorily based requirements.

With the onus placed on users to avoid identifying any individual under penalty of law, as in the NCES legislation, statistical agencies could make more statistical data available with greater detail and in more easily accessible and more usable ways without incurring increased risks. The protection would automatically go with the statistical data, from producer to user or from one user to another.

We must not only protect the confidentiality of information collected for statistical purposes, we must also safeguard statistical data derived from that information by such legislative means. For information, we must place strong locks with fewer keys on the producers. For statistical data, we must place restrictions with strong sanctions on any and all users.

The best safeguard for statistical data, however, is to foster a climate to protect them. That can be done through licensing agreements and the rules governing restricted access, in particular, with procedures that continually remind all those working in the area about the importance of protecting confidentiality of information and safeguarding statistical data.

Box 2

National Center for Education Statistics Warning for Public-Use Data Files

WARNING

Under law, public-use data collected and distributed by the National Center for Education Statistics (NCES) may be used only for statistical purposes.

Any effort to determine the identity of any reported case by public-use data users is prohibited by law. Violations are subject to Class E felony charges of a fine up to $250,000 and/or a prison term up to 5 years.

NCES does all it can to assure that the identity of data subjects cannot be disclosed. All direct identifiers, as well as any characteristics that might lead to identification, are omitted or modified in the dataset to protect the true characteristics of individuals. Any intentional identification or disclosure of a person violates the assurances of confidentiality given to the providers of the information. Therefore, users shall:

- Use the data in this dataset for statistical purposes only.
- Make no use of the identity of any person or institution discovered inadvertently, and advise NCES of any such discovery.
- Not link this dataset with individually identifiable data from other NCES or non-NCES datasets.
- To proceed you must signify your agreement to comply with the above-stated statutorily based requirements.

But it must also be done through the education and training we give to future researchers. Taking care to safeguard statistical data must become accepted practice at all research institutions. It must become part of the ethos of science.

8. Zero Tolerance

Many people, especially members of Congress, believe that we can prevent individuals being identified from misuse of statistical data by setting a standard of zero tolerance for such a possibility. The concept of zero tolerance, however, applies to disclosure of information. It does not apply to statistical data.

For statistical data, we seek to minimize the risk that some individual may be identified from the statistical data. The standard, therefore, that we should apply in doing so is to exercise reasonable care.

If we seek legislation to further safeguard statistical data, we can codify this principle of reasonable care in the very definition of statistical data in the legislation:

Statistical data are data derived from information collected for statistical purposes for which an agency has taken reasonable care that these data do not reveal the identification of an individual.

For all or nearly all of our most important statistical data files, there is always a risk that information from another source could be used along with the statistical data to identify an individual. With today's technology, let alone the technology of the future, we may not be able, for a large, complex statistical data file, to reduce the risk to zero that some individual may be identified along with their information, even with means such as creating synthetic statistical data files.

We must explain to those who believe that we could prevent such identifications the necessity of accepting a very small risk. We must accept it because that small risk is balanced against the myriad benefits that statistics and research bring to our lives and our society.

9. Public-use Statistical Data Files

Although a law could safeguard statistical data files prepared for public use, it might be violated with impunity, since it would be too difficult to detect violations. The standard here again must never be one of preventing all possible identity disclosures, but rather to exercise reasonable care. A statistical agency should exercise reasonable care to minimize the risk that some information file generally expected to be publicly available could be matched against the public-use statistical data file in such a way as to identify an individual along with his or her statistical record.

10. Advancing the Culture for Access to Statistical Data

Neither our new confidentiality law nor one to further safeguard statistical data will change the behavior of our statistical agencies overnight. But, with such laws, we can further a culture that seeks to expand access to statistical data.

The Committee on National Statistics, in its report, *Principles and Practices for a Federal Statistical Agency* (Martin, Straf, and Citro 2001), articulated three principles for federal statistical agencies. First is that the agency be in a position to inform public policy. That requires providing relevant statistical data to researchers who can analyze them for public policy purposes.

A second principle is to develop a relationship of mutual respect and trust with respondents who provide the data and with all data subjects whose information it obtains. In particular, we must respect respondents to our censuses and surveys by using the information they provide when it can benefit society.

A third principle is that statistical agencies develop a relationship of mutual respect and trust with those who use their data. What better way could a statistical agency foster that relationship than by making its statistical data available for users to scrutinize as well as analyze. If statistical agencies do not respect these principles, if they use confidentiality as the means to further restrict and diminish the utility of statistical data, they will suffer not a crisis of confidentiality, but, worse, a crisis of confidence.

But, with this new legislation and stronger ways to safeguard statistical data, our statistical agencies can grow to be regarded even more as essential in informing public policies, in guiding businesses and industry, and in providing the information to all members of our society so that they can make better decisions about their future and about their government.

11. References

Mackie, C. and Bradburn, N. (2000). Improving Access to and Confidentiality of Research Data: Report of a Workshop. Committee on National Statistics, National Research Council, Washington, D.C.: National Academy Press.

Martin, M.E., Straf, M. L., and Citro, C. F. (2001). Principles and Practices for a Federal Statistical Agency. Second Edition, Committee on National Statistics, National Research Council, Washington, D.C.: National Academy Press.

National Center for Education Statistics (2002). NCES Statistical Standards, September 2002, available online at ⟨http://nces.ed.gov/statprog/2002/stat_standards.asp⟩, 10 February 2003.

Putnam, R.D. (2000). Bowling Alone: The Collapse and Revival of American Community. New York: Simon & Schuster.

Sweeny, L. S. (2001). Information Explosion. In Confidentiality, Disclosure, and Data Access: Theory and Practical Applications for Statistical Agencies. L. Zayatz, P. Doyle, J. Theeuwes, and J. Lane (eds), Washington, D.C.: The Urban Institute.

Wallman, K. (2003). Privacy and Confidentiality – A New Era. Journal of Official Statistics, 19, 315–319.

Westin, A. F. (1967). Privacy and Freedom. New York: Atheneum.

Received February 2003