DOCUMENT RESUME

ED 317 169 IR 014 243

AUTHOR	Duncan, George T.; Pearson, Robert W.
TITLE	Improving Access to Data While Protecting Confidentiality: Prospects for the Future.
PUB DATE	89
NOTE	17p.; Paper presented at the 1989 Joint Statistical Meetings (Washington, DC, August 1989).
PUB TYPE	Viewpoints (120)
EDRS PRICE	MF01/PC01 Plus Postage.
DESCRIPTORS	*Access to Information; Codes of Ethics; *Confidentiality; *Data Analysis; *Databases; *Government Role; National Surveys; Prediction; Public Policy; Researchers; Social Science Research
IDENTIFIERS	Microdata; Public Private Relationship

ABSTRACT

Providing researchers, especially those in the social sciences, with access to publicly collected microdata furthers research while advancing public policy goals in a democratic society. However, while technological improvements have eased remote access to these databases and enabled computer using researchers to perform sophisticated statistical analyses, they have also increased the likelihood that individual records can be re-identified. Government agencies that sponsor the collection of microdata files containing personal and sensitive data are under increased pressure to form policies that allow continued improvements in data access while protecting subject confidentiality. Based on recent developments, this paper reflects on what the near future holds for the resolution of this issue in five arenas: statistical, computer, legal, administrative, and ethical. In the statistical arena, agencies will develop methods of masking data through disclosure limiting and linear affine masking approaches. Electronic gatekeepers and monitors for remote access will be applied in the computer arena. Legal involvement will include legislation that ensures research access and provides sanctions for improper data use. In the administrative arena, agencies will place more responsibility on researchers as data stewards. Finally, ethical concerns will be addressed by developing a researcher's code of conduct. (32 references) (NRP)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED317169

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
This document has been reproduced as received from the person or organization originating it
Minor changes have been made to improve reproduction quality.
Points of view or opinions stated in this document do not necessarily represent official OERI position or policy

IMPROVING ACCESS TO DATA WHILE PROTECTING CONFIDENTIALITY: PROSPECTS FOR THE FUTURE

George T. Duncan
School of Urban and Public Affairs
Carnegie Mellon University
Pittsburgh, PA 15213

Robert W. Pearson
Social Science Research Council
New York City, NY 10158

Paper to be presented
at the
1989 Joint Statistical Meetings
Washington, DC
August

DRAFT--NOT FOR QUOTATION OR CITATION

COMMENTS INVITED

IR014243

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
George T.Duncan
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

BEST COPY AVAILABLE

IMPROVING ACCESS TO DATA WHILE PROTECTING CONFIDENTIALITY: PROSPECTS FOR THE FUTURE

by George T. Duncan and Robert W. Pearson

ABSTRACT

This paper provides a scenario for the future of research access to officially collected microdata. Many researchers--especially those in the social sciences and public health--find access to government databases increasingly useful. The databases themselves are more comprehensive, of better quality, and--with better database management techniques--better structured. Computer communications improvements ease the technology of remote access to these databases. Substantial gains in the performance/cost ratio of computers permit more sophisticated analyses--including statistical graphics, analysis of extreme values, maximum likelihood and Bayesian methods.

At the same time, individuals and firms that provide the data residing on government databases--and the agencies who sponsor the collection of such information--are becoming increasingly sensitive to privacy concerns. Ironically, some of the same technologies that expand analytical capabilities also provide tools to threaten the confidentiality of data records.

As the broker between the data provider and the data user, government agencies are under increased pressure to have policies that both increase data access and insure confidentiality. In response to these cross-pressures, agencies will pursue statistical, administrative, and legal approaches to responsible data dissemination. Recent developments in these approaches are discussed as they relate to improvements in databases, computer and analytical methodologies, and legal and administrative arrangements for access to and protection of official statistics.

IMPROVING ACCESS TO DATA WHILE PROTECTING CONFIDENTIALITY: PROSPECTS FOR THE FUTURE

by George T. Duncan and Robert W. Pearson

I. CONTEMPORARY CONCERNS

Providing researchers with access to data furthers research in the social sciences while advancing accepted public policy goals in a democratic society. It allows reanalysis by groups with different agendas; stimulates new inquiries on important social, economic, and scientific questions; suggests improved measurement and data collection methods; and provides information to improve forecasts and resource allocation (see, e.g., Flaherty (1979) and Fienberg, Martin, and Straf (1985)).

Widespread access to surveys such as the Panel Study of Income Dynamics and the National Longitudinal Surveys of Labor Market Experience, for example, have furthered our understanding of the dynamics of poverty, replacing longheld beliefs about the permanence of poverty with knowledge about the extent to which poverty is both widespread and temporary for a large proportion of the American public (Duncan 1984). Access to computerized criminal history files maintained by the FBI have permitted, for example, longitudinal studies of criminal careers, which have overturned many inferences drawn from previous cross-sectional studies of crime (Blumstein and Cohen 1987).

In spite of the evident value of microdata dissemination, however, serious concerns about access to publicly collected microdata have been raised. Four factors give rise to this contemporary concern about the (re)identification of individual records:

(1) Sophisticated and more widely available computational and analytical technologies make it easier to breach the anonymity of the individuals and institutions who are the subjects of publicly-sponsored surveys and administrative records.

(2) The creation and accumulation of large and detailed microdata files both in government and in business-- increasingly longitudinal in design--make the unique "signatures" of individual records increasingly difficult to disguise prior to their distribution without also degrading the scientific value of the data. Similarly, the increasing possibility of linking distinct data files make the possible disclosure of the identify of individual records easier in principle, which may also contribute to an increasing suspicion among the public that these records have indeed been linked.

(3) A heightened nervousness on the part of those who collect these data that the technology, the detail of records, and the alleged growth in public concern about privacy and confidentiality will diminish the public's trust and cooperation with these data collection programs. As a result, the quality and usefulness of the data themselves will decline as will the ability of the agencies to fulfill their missions.

(4) An increasingly information-based society in which individuals and organizations can gain competitive advantage through intelligence-gathering activities.

Because of its evident value, the demand side for microdata is well-established. For example, Arber's (1988) survey of British academics revealed a strong demand for the release of individual-level samples of census data for re-analysis which used academics' own hardware and software. A 1984 Census Bureau conference, for a second example, saw more than one hundred economists expressing a desire for a public use Longitudinal Establishment file (Govoni-Waite, 1985).

But because of the four factors of contemporary concern the supply side for microdata is hampered by agency concern about disclosure risk. Recent examples of the unmet need for microdata from one important federal statistical agency--the Bureau of the Census--include the following (Gates 1988):

- Researchers at the National Opinion Research Center requested a special 1980 Census public use file with records linked to tract and SMSA data. The study, part of a three-year study of racial segregation in the United States, would link people to their immediate neighborhoods (tracts) and to larger areas in which they live (SMSA). However, tracts (and some SMSA's) contain populations of fewer than 100,000 persons, the cut-off point for identifiable geographic units on publicly released Census microdata.

- Researchers at Princeton University requested the exact date of birth on a microdata tape of the Survey of Income and Program Participation in order to study the Selective Service draft lotteries held in the United States in the 1970s. Because date of birth is available on many administrative record files and is an excellent match key, its inclusion on the tape would have increased the risk of identifying SIPP respondents.

- The Economic Research Service of the Department of Agriculture requested a file showing non-metropolitan status of SIPP respondents in order to assess their economic well-being in terms of wealth, asset holdings, and participation in government programs.

More generally, a number of social science research and public policy studies could be pursued if the present tension between access and confidentiality were better resolved:

- Contextual data could supplement survey data from various administrative agencies. For example, the National Longitudinal Survey of Labor Market Experience Youth Cohort could link neighborhood and administrative data to its individual records to enable the study of the processes by which persistent and concentrated urban poverty result in problems for family processes and individual development.

- The sponsorship of ongoing longitudinal surveys could be transferred from one agency to another as respondents age. The programmatic interests of several statistical agencies are tied to different stages in the life cycles of people, but concerns about confidentiality have made it difficult for agencies to transfer responsibilities for data collection and analysis. For example, the Longitudinal Retirement History Survey has been of interest to the National Institute of Aging and the growing field of research and public policy

concerning America's elderly population, but the transfer of responsibility for these data has been discouraged because of prohibitions on the release of these data and on their linkage to Social Security data, Medicare records, and data from the National Death Index.

- The latest scientific developments in analyzing very large spatial data bases and modeling complex spatial phenomena would be available. These developments, which could help achieve the goal of identifying and explaining human behavior at both the aggregate and individual levels require the use of refined geographic identifiers which are not now generally available.

The concern by agencies for protecting the confidentiality of records is engendered by legal requirements, ethical issues involving actual and implicit commitments made to data respondents, and practical worries about response rates to statistical surveys. An important part of a data-disseminating program is an adequate set of disclosure-limiting procedures which can be affected through various mixes of ethical, legal, administrative, and statistical controls.

Our purpose in this paper is to reflect on what the near future holds for the mediation of concerns about data access and confidentiality. We draw on recent developments, and we paint a hopeful portrait of the future in part to help provide a target or goal--even if always moving--for better accommodating the increasing tension between data access and confidentiality. Indeed, we share the spirit of the positive remarks of Wolf who wrote (1988):

> By applying reasonable care and conforming to reasonable guidelines, we feel that microaggregate files can be created and released to the public without an unacceptable risk of disclosing confidential information.

In brief, our vision of the future for improving access to data looks like this:

- In the statistical arena, agencies will employ masks that are effective yet faithful to the original data. Statistical methods for the analysis of masked data will be developed, cheaply available, and easy to use.

- In the computer arena, electronic gatekeepers and monitors for remote access to, and utilization of computer databases will be widespread.

- The disclosure implications of record linkage and matching procedures will be better understood.

- In the legal arena, legislation will recognize the need for research access and provide for sanctions for improper use of data.

- In the administrative arena, agencies will place more responsibility on researchers as data stewards. Pledges, bonds, and contracts will become an increasingly explicit part of the conditions under which researchers gain access to microdata.

- In the ethical arena, a researcher's code of conduct concerning disclosure will be further developed, widely discussed, and observed in practice.

In the remainder of this paper we will explore this vision in each of the arenas in more detail. We begin with the statistical arena, and look at masks as a technique for disclosure limitation.

II. STATISTICAL ARENA: MASKING OF MICRODATA

At present, some microdata files are released after steps have been taken to limit the possibility of disclosing or reidentifying a record. In 1963, the U.S. Census Bureau, for example, used the release of sampled data as a disclosure-limiting device when it began providing public use microdata from the 1960 decennial census as a one-in-one-thousand sample file (Gates, 1988).

The data held by the agency is a file represented by an n-by-p matrix X. Each of the n rows gives individual data on each of p attributes. Typically, a file records many attributes of respondents, including some which are sensitive (e.g., income, assets, or medical conditions of target individuals).

The data may be masked through such methods as:

(1) releasing only of a sample of the data (subtracting rows from X);

(2) including simulated data (adding rows to X);

(3) blurring (fuzzing individual values in X by random rounding, grouping, adding random error, etc.);

(4) excluding certain attributes (removing columns of X); and

(5) swapping (exchanging blocks of rows in a certain subset of columns of X).

The purpose of masking data is to make it more difficult for a data user to break the confidentiality of the database X by violating the anonymity of one or more units of observation (i.e., people or organizations). In the evocative language of Roberts (1986), we refer to such a user as a statistical spy. It is now generally accepted--perhaps reluctantly by researchers requiring data--that the simple mask of removing columns of X corresponding to obvious identifiers or near identifiers (such as name, social security number, address, or telephone number) is insufficient in many cases to hamper a serious statistical spy (see Paass (1988), just as locking car doors does not deter a serious thief.

It is also increasingly clear that one of the most widely used techniques for masking the data--suppressing or not collecting detailed information about the place of work or residence--has crippled society's ability to study institutions and contextual or spatial dimensions of social phenomena.

A more careful consideration of the deterrence value of various forms of various masking methods is required if data custodians are to be convinced that microdata can be released under statistical controls. Also, the potential of masked data for valid and informative statistical inference must be assessed and new methods of statistical analysis of masked data developed.

In examining the deterrence value of a particular transformation, the beginning point of the disclosure-limiting (DL) approach of Duncan and Lambert (1986, 1989) is

to model the decision problem of the statistical spy in inferring the value of a target Y from the released X. A measure of disclosure risk is the potential of the information in X for inferring Y.

The basic philosophy underlying the DL approach is to deter the statistical spy from taking actions that identify privacy-protected information. Disclosure-limiting procedures raise the price of using the released information in this way sufficiently high so that the spy will not use it to take such actions. The purpose of this approach is not simply to avoid having the spy make correct identifications. It is equally important that the spy be prohibited from making identifications altogether--whether correct or not--because any purported identification can damage a data-disseminating agency and because luring the spy to incorrect identifications can typically only be achieved by releasing misleading data which undermine legitimate research. From a decision-theoretic point of view the idea is to raise the Bayes risk of identification high enough so that the option of no identification is preferred.

This philosophy yields the threshold rule for the agency: release the data if the Bayes risk to the statistical spy exceeds some threshold. Using the DL framework, this section explores disclosure limitation via displaced linear transformations (linear affine masks) of the data matrix X. The approach involves regression of a sensitive target value Y on the masked data. The heuristic motivation behind a regression approach is evident: the statistical spy wants to use the information in the masked data X to infer something about the sensitive target value Y. All probability distributions have the following interpretation: they are the subjective distributions of the statistical spy as they are perceived by the data disseminating agency. The question is whether any useful insights about disclosure limitation can be obtained from this DL approach. The next simple example suggests that the answer to this question is yes.

In seeking to resolve uncertainty about the value of a scalar quantity Y from a scalar quantity X, a statistical spy is taken to have a squared error loss function. Thus the spy minimizes squared error by choosing a predictor of Y to be the conditional expectation of Y given X. Keeping the Bayes risk above a threshold yields simple policy guidance: an agency faced with a statistical spy who has something to gain should not release the value of X if (in the mind of the spy) it is too highly correlated with the value of the sensitive variable Y.

We focus our attention on the use of linear affine masking of the microdata file X. The data user is provided the masked microdata file $M = AXB + C$, and is not given the original data X. The matrix A, as a matrix of row operators, directly transforms the data records in X; so we call A a record transforming mask. The matrix B, as a matrix of column operators, directly transforms the data attributes in X; so we call B an attribute transforming mask. The matrix C displaces AXB by adding stochastic or systematic noise to the data; so we call C a displacing mask. In general, the mask (A, B, C) may depend on the particular values in X. That is the mask components A, B, and C are not necessarily just fixed matrices with constant elements or random matrices with elements that are independent of the values in X. Generally, for reasons of data utility--the data must be analyzed--the data provider must also give the user either the complete specification or certain characteristics of the mask (A, B, C). It is an open question of disclosure-limitation methodology as to how much information should be given the data user about the mask in a particular context (Wolf, 1988).

Linear affine masking is powerful because it encompasses many commonly proposed disclosure-limitation methods. We illustrate this first with record transforming

masks A, second with attribute transforming masks B, and third with displacing masks C. Some of these procedures are discussed in McGuckin and Nguyen (1988a).

RECORD TRANSFORMING MASKS

By changing the form of the record transforming mask A--even with B an identity matrix and C a zero matrix--we can represent some currently proposed disclosure-limitation techniques, such as:

Aggregation across records. For example, averaging all attributes over three similar records.

Suppression of certain records. For example, suppression of records having extreme values on some attributes or suppression of records from small identifiable geographic units. Here the transforming mask is a function of the data file X.

We can also consider a random record transforming mask in which the matrix A has stochastic elements. Special cases of this that are of interest include the following:

Sampling. In sampling r rows of X, the matrix A has 0-1 random entries with a single 1 in each row.

Multiplication of records by random noise. With the matrix A diagonal, each record is multiplied by a random variable.

ATTRIBUTE TRANSFORMING MASKS

By changing the form of the attribute transforming mask B, we can represent the following disclosure-limiting procedures:

Aggregation across certain attributes. For example, the release of total income, rather than salary income, business income, interest income, etc.

Suppression of certain attributes. For example, some attributes--such as identifiers or medical conditions like mental health or HIV infection indicators--may be suppressed.

Multiplication of attributes by random noise.

DISPLACING MASKS

In the case of displacing masks (the matrices A and B are identities), adding C yields the following disclosure-limitation techniques:

Addition of random noise. Adding a random variable to each entry.

Addition of deterministic noise. Adding a specified quantity to each entry.

Often, implemented procedures involve a combination of disclosure-limitation procedures. See, for example, Kim (1986) for a Census Bureau application to the

Continuous Longitudinal Manpower Survey which was conducted for the Bureau of Labor Statistics to evaluate the effectiveness of the Comprehensive Employment and Training Act (CETA) of 1973. The public use files contain earnings data matched to Social Security Administration administrative records. The masking technique involved both the addition of random noise and data transformation. In these cases, the transforming masks A and B are not identity matrices and the displacing mask C is not the zero matrix.

Given the richness of linear affine masks, it is reasonable to ask, "What commonly used (or proposed) disclosure-limiting procedures are not linear affine masks?" Here are some examples:

Attribute-specific aggregation over records. Release of some attribute values unmasked, but aggregating other attribute values--say releasing only averages of interest income for similar records.

Data swapping. Release of records with some, but not all, attribute fields interchanged.

Multiplication by random noise. Multiplying each element of X by mutually independent random variables is not a matrix multiplication or addition.

Random rounding. Rounding each entry to a certain base.

Grouping. Condensing categories for some attributes.

Truncating. Truncating distributions of certain attributes.

Generally, ad hoc arguments have been used to devise disclosure-limitation procedures and to evaluate them in terms of disclosure risk and data utility. Studies to date suggest that particular implementations can result in significant differences between the information provided by the masked data and that available from the original file (see, for example, Wolf (1988) for an assessment of surrogate microaggregate records). This suggests that a more general analysis based on a systematic approach to masking is desirable.

The basic idea in disclosure limitation is to find a mask that leaves the maximum information about X, while at the same time preserves confidentiality. As a generally useful approach, this suggests choosing a mask (A, B, C) to minimize the conditional variance of X given M while maximizing the conditional variance of Y given M. This notion of constrained optimization can be considered consistent with what is reported to be Census Bureau policy: "In practice the Census Bureau has taken disclosure protection as a binding constraint and provided as much data to the public as is possible within this constraint" (McGuckin and Nguyen, 1988b).

Given that the researcher receives the data in the masked form M rather than the original form X, an important question is how best to analyze the data. In the base of sampled data, standard tools are appropriate. But, for example, addition of noise presents measurement error or errors-in-variables problems for the user analyzing the masked data.

III. COMPUTER ARENA: UTILIZATION CONTROL IN NETWORKED INFORMATION SYSTEMS THROUGH ELECTRONIC GATEKEEPERS

In some cases, access to data by a researcher will be controlled by an intermediary--or "gatekeeper"--as contrasted to or in addition to masking micro data files prior to their release. This will be increasingly true with computer data base systems.

As organizations have increasingly employed distributed database systems, new concerns about data integrity and security in information networks have arisen. Of special concern in developing trusted networks is that authorization policies and implementations accommodate the varying levels of security at the network nodes--including the class of home computers with dial-up potential--so that sensitive information can be processed. The initial focus of network security has been on the problem of controlling access to systems and files at a macro level. While necessary, such access control--say, by passwords--is not sufficient to protect the privacy and integrity of sensitive information.

Network security must also encompass utilization control, which can be thought of as access control at a micro level. By analogy, the guard at the art museum's gate qualifies entrants--thereby controlling access to the museum, but additional security measures are needed in utilization of the museum to prevent theft and vandalism--thereby controlling access to the individual works of art.

Increasingly, organizations are establishing statistical databases that reside on computers and contain confidential data or, implicitly, relationships that are of a sensitive nature. Blue Cross and Blue Shield of Massachusetts, for example, has established the Provider Terminal Network, which allows physicians and hospitals to directly verify a patient's status and eligibility. More generally, the increased amount of confidential data transmitted over networks has prompted the Tele-Communications Association and large network users to appeal to the FCC to determine exactly what network data is considered customer proprietary network information. Further, the Computer Security Act of 1987 requires that civilian agencies identify systems containing sensitive information and develop a security plan for each sensitive system. With their proliferation, the data held in these networked systems will become of increasing interest to researchers.

In a network security system, the utilization protection policy is implemented through a security kernel or reference monitor which processes user queries. Macro-level access control techniques prevent unauthorized access to networks by verifying a user's identity prior to allowing the user access to the host or the network. There are many techniques for making access to a network secure, such as authentication, passwords, and encryption (see, e.g., Government Security News, 1988 October 10). Most access control techniques are not fully relevant when a user has legitimate access to certain information, say, certain statistical aggregates, but does not have legitimate access to certain other information, say, medical, sales, or salary information identifiable to a particular individual. Limiting queries to statistical aggregates is insufficient because a series of such queries can readily identify individual information (see, e.g., Ahituv, Lapid, and Neumann (1988)).

Needed are more sophisticated authorization rules that determine what users can do or see. While some formal theory has been developed (see, e.g., Landwehr (1981) and

Denning (1982)), current techniques for utilization control are fairly rudimentary. For example, audit trails operate only ex post facto in establishing what a user has done. Multilevel passwords for applications and records provide only limited flexibility in controlling utilization.

The decision theoretic methods employed by Duncan and Lambert (1986, 1989) can be applied to the case of a database accessed through a computer network. Users query the database according to certain authorization rules. Access and flow controls are governed by a security kernel. A database has been compromised when a database spy has identified a confidential data record or identified a restricted relationship. In this context, five alternative disclosure limitation techniques were prescribed by Shoshani (1982): (1) limiting the query set, (2) limiting the intersection of query sets, (3) random sample queries, (4) partitioning the database, and (5) perturbing data values. These warrant systematic investigation so that networked data base systems can achieve their full potential for the researcher.

IV. LEGAL ARENA: LEGISLATION FOR RESEARCH ACCESS AND SANCTIONS FOR IMPROPER USE OF DATA.

Research access to data is controlled through a variety of regulations and laws. Legislative restrictions on data access include the Tax Reform Act of 1976 (Public Law 94-455) and the Privacy Act of 1974 (5 U.S.C. 552a). Along with other factors, improvements in computer technology motivate some changes in these legal controls. Unfortunately, the development of legal controls often lags changes in technology. "Courts don't recognize substantive difference between manual records and computer records," said Bob Smith, editor of the Privacy Journal, a newsletter. "They don't really grasp that technology itself has changed." Referring to the 1974 Privacy Act, Professor Arthur Miller of the Harvard Law School is quoted as saying that technological improvements in the years since the law was passed had "rendered it obsolete" (both quotations from an article by Cory Dean in the NEW YORK TIMES, 1986 September 29).

Some regulatory attempts to restrict access would, as in the Reagan Administration National Security directive, limit the use of commercial data bases. These attempts were aborted in 1987 under pressure from the American Civil Liberties Union and the Information Industry Association. "Before these computerized information banks were created, such technical reports were scattered in hundreds of arcane journals and libraries. Now the data-base companies collect millions of documents and let customers comb through them in minutes by computer", writes Bob Davis in the WALL STREET JOURNAL, 1987 February 5. In Great Britain, the Data Protection Act of 1984 regulates the storage and processing by computers of data about living individuals. As the Act applies to data held for statistical or research purposes, the Royal Statistical Society formed an ad hoc study group to monitor its impact.

Legislation governing access to data varies from one agency to the next in the United States, and in some cases varies within an agency (e.g., Titles 13 and 15 prescribe different treatments for data collected by the Bureau of the Census). The future is likely to retain this diversity, but some convergence in laws and practices will occur as issues of confidentiality and access arise with each reauthorization of agencies in the future. Convergence that provides for effective research access to data while maintaining sanctions for improper use of data does not occur spontaneously, however. For example, at the state level, the Model State Vital Statistics Act, has guided a number of states in their legislation regarding research access. Yet this Model Act is sufficiently

ambiguous on what is "legitimate research use" to present operational difficulties to state agencies.

The future holds some promise for input to the legislative process as a result of a recently funded study to be undertaken by the Committee on National Statistics of the National Research Council and the Social Science Research Council. The purpose of this study is in part to bring systematic attention to these regulatory and legislative practices.

V. ADMINISTRATIVE ARENA: AGENCIES AS DATA STEWARDS

As both the value of sensitive data and its potential for compromise rise with improved computer technology, agencies and researchers will increasingly understand their role as that of data stewards. As in the biblical parable, the best steward is one who ensures effective use of the data, not the one who protects it against any risk by hiding it. In the administrative arena, our future holds the following:

(1) Increasing attention will be devoted to the theory and practice of informed consent as it relates to providing access to such data. The considerable attention to these issues in biomedicine will be imported and applied to federal statistics. Agencies will draft informed consent agreements for respondents that assure that their privacy rights are protected, that response rates are not lessened, and that legitimate research use of the data is authorized by respondents who are asked to consent to plans to use the data for research purposes. Assurances of these outcomes will follow from a program of pilot studies that will empirically assess these outcomes.

(2) For longitudinal studies, prior informed consent agreements (which may not have foreseen currently needed research uses and therefore failed to properly inform respondents of such use) must consider the explicit and implicit promises, understandings, and concerns of the respondents at the time at which the data were originally collected. When feasible, agencies will return to respondents (or their guardians) to renegotiate informed consent.

(3) Researchers should be subject to a licensing agreement and bond that clearly state their responsibilities or liabilities for violation that agreement.

(4) Agencies will have an affirmative obligation to exert an active review of the uses made of research records when there is some risk of disclosure.

VI. ETHICAL ARENA: RESEARCHER'S CODE OF CONDUCT

In guiding effective use of data, a clearer code of ethics for researchers will emerge. Its rough shape is beginning to form, as the following principles and examples illustrate:

(1) When data are provided in unidentified form, no attempt will be made to establish personal identities of respondents. "The Treasury Department, in response to The BOSTON GLOBE's request for data pertinent to money laundering," reports Cory Dean in the NEW YORK TIMES, 1986 September 29, "established rules for people seeking access to records. Among other things, they must submit 'a detailed statement' of the information sought and how they intend to use it. They must agree to design computer programs that do not elicit the identities of individuals or businesses.' In the event that a search of the data base results in the inadvertent disclosure of personal identifiers,' the regulations say, the researcher must 'terminate the search until appropriate security measures can be implemented; relinquish all records of personal identifiers to Treasury officials; and make no further disclosure of the information.'"

(2) When data are provided in identified form, researchers will protect the confidentiality of their data against outside threat, and will be provided with legislative protection from subpoena of these records for the purpose of identifying individual subjects. A drug manufacturer, for example, recently sought to subpoena the original records of an epidemiological study. The Centers for Disease Control prevailed in protecting the identity of subjects, both in a lower court and in the U.S. Court of Appeals. The court ruled that such records could not become accessible to lawyers who could use them to call the victims of adverse drug reactions as witnesses, and attempt in court to break down the reports these victims had given to epidemiological investigators (Curran (1986))

VI. CONCLUSIONS

We envision that empowered by exponentially improving computer technology researchers will have access to larger and more relevant databases. This emerging capability is an exciting opportunity to better understand the way our economy and society works. We also envision that researchers will show increasing sensitivity to the need for confidentiality because of computer enhanced potential for disclosure. This emerging capability is a sobering call to better respond to the public's call for responsible use of personal and sensitive data.

REFERENCES

Ahituv, Niv, Lapid, Yeheskel, and Neumann, Seev (1988) Protecting statistical databases against retrieval of private information. Computers & Security, 7, 59-63.

Arber, Sara. (1988) "Anonymized Samples of 1991 Census Data", Presented at "Access to the Census: Anonymized Samples and their Alternatives", a seminar sponsored by the Royal Statistical Society, Social Statistics Section, and the Social Research Association, London, England, November 15.

Blumstein, Alfred and Cohen, Jacqueline (1987) Characterizing criminal careers. Science, 28 August, 985-991.

Boruch, Robert. and Cecil, Joseph. (1979) "Report from the United States: Emerging Data Protection and the Social Sciences' Need for Access to Data." In E. Mochmann and P. Muller, eds. Data Protection and Social Science Research New York: Springer-Verlag, 104-128.

Campbell, D. T., Boruch, R. F., Schwartz, R. D. and Steinberg, J. (1977) Confidentiality-preserving modes of access to files and interfile exchange for useful statistical analysis. Evaluation Quarterly, 1, 269-299.

Cox, Lawrence H. (1980) Suppression methodology and statistical disclosure control. Journal of the American Statistical Association, 75, 377-385.

Curran, W. J. (1986) Protecting confidentiality in epidemiologic investigations by the Centers for Disease Control. New England Journal of Medicine, 314, 1027-1028.

Dalenius, Tore (1985) Privacy and confidentiality in censuses and surveys. Annual Meeting of the American Statistical Association.

Dalenius, Tore and Reiss, S. P. (1982) Data swapping: a technique for disclosure control. Journal of Statistical Planning and Inference, 6, 73-85.

DeMillo, R. A., Dobkin, D. P., and Lipton, R. J. (1977) Even databases that lie can be compromised. IEEE Transactions on Software Engineering, SE-4, 73-75.

Denning, Dorothy E. (1982) Cryptography and Data Security. Reading, MA: Addison-Wesley.

Duncan, George T. and Lambert, Diane (1986) Disclosure- limited data dissemination. Journal of the American Statistical Association, 81, 10-28 (with discussion by L. Cox, O. Frank, J. Gastwirth, and H. Roberts) JASA Applications Section Special Invited Address at the ASA Annual Meeting, Las Vegas, August, 1985.

Duncan, George T. and Lambert, Diane (1989) The risk of disclosure for microdata. Journal of Business and Economic Statistics, 7, 207-217.

Duncan, Greg J. (1984) Years of Poverty, Years of Plenty, Ann Arbor: Institute for Social Research, University of Michigan.

Fellegi, I. P. (1972) On the question of statistical confidentiality. Journal of the American Statistical Association, 67, 7-18.

Fellegi, I. P. and Phellops, J. L. (1974) Statistical confidentiality: some theory and applications to data dissemination. Annals of Economic and Social Measurement, 3, 399-409.

Fienberg, Stephen E., Martin, Margaret E., and Straf, Miron, L. (editors) (1985) Sharing Research Data Washington, DC: National Academy Press.

Flaherty, D. H. (1979) Privacy and Government Data Banks: An International Perspective London: Mansell.

Gates, Gerald W. (1988) "Census Bureau Microdata: Providing Useful Research Data While Protecting the Anonymity of Respondents" Proceedings of the Section on Survey Research Methods, American Statistical Association. Presented at the Annual Meeting of the American Statistical Association, New Orleans, August 22-25. Government Security News, volume 7 number 21, 1988 October 10.

Govoni, J. P. and Waite, P. J. (1985) "Development of a Public Use File for Manufacturing", Proceedings of the Section on Business and Economic Statistics, American Statistical Association, 300-302.

Greenberg, Brian. (1988) "Disclosure Avoidance Research for Economic Data", Presented to the Joint Advisory Committee Meeting, October 13-14, Oxon Hill, MD.

Keller-McNulty, Sallie, Unger, Elizabeth A., and McNulty, Mark S. (1989) The protection of confidential data. Paper presented at the 21st Symposium on the Interface: Computing Science and Statistics, April 9-12, Orlando, Florida.

Kim, J. (1986) "A Method for Limiting Disclosure in Microdata Based on Random Noise and Transformation". Proceedings of the Survey Research Section, American Statistical Association, 370-374.

Landwehr, Carl E. (1981) Formal models for computer security. ACM Computing Surveys, 13, 247-278. Reprinted in Advances in Computer System Security, Volume II. (Rein Turn, editor) Dedham, MA: Artech House, 1984.

Leiss, Ernest L. (1982) Principles of Data Security New York:Plenum Press.

McGuckin, R. and Nguyen, S. (1988a) "Use of 'Surrogate Files' to Conduct Economic Studies with Longitudinal Microdata", Proceedings of the Third Annual Research Conference, Bureau of the Census.

McGuckin, R. and Nguyen, S. (1988b) "Public Use Microdata: Disclosure and Usefulness", U.S. Census Bureau. Center for Economic Studies Discussion Paper CES 88-3, September.

Paass, Gerhard (1985) Disclosure risk and disclosure avoidance for microdata. Journal of Business and Economic Statistics, 6, 487-500

Pearson, Robert W. (1987) "Researchers' access to U.S. federal statistics" Items 41:6-11.

Roberts, Harry V. (1986) "Comment on Duncan and Lambert" Journal of the American Statistical Association 81: 25-27.

Shosani, A. (1982) Statistical databases: characteristics, problems, and some solutions. LBL Perspective on Statistical Database Management Lawrence Berkeley Laboratory, University of California, Berkeley, 3-28.

Wolf, Michael. K. (1988) "Microaggregation and Disclosure Avoidance for Economic Establishment Data", Proceedings of the Section on Survey Research Methods, American Statistical Association. Presented at the Annual Meeting of the American Statistical Association, New Orleans,August 22-25.