Improving Access to and Confidentiality of Research Data: Report of a Workshop

Christopher Mackie and Norman Bradburn, Editors; Committee on National Statistics, National Research Council

ISBN: 0-309-51381-2, 74 pages, 6 x 9, (2000)

Improving Access to and Confidentiality of Research Data

Report of a Workshop

Committee on National Statistics

Christopher Mackie and Norman Bradburn, *Editors*

Commission on Behavioral and Social Sciences and Education

National Research Council

NATIONAL ACADEMY PRESS
Washington, D.C.

NATIONAL ACADEMY PRESS • 2101 Constitution Avenue, N.W. • Washington, D.C. 20418

NOTICE: The project that is the subject of this report was approved by the Governing Board of the National Research Council, whose members are drawn from the councils of the National Academy of Sciences, the National Academy of Engineering, and the Institute of Medicine. The members of the committee responsible for the report were chosen for their special competences and with regard for appropriate balance.

This study was supported by Contract/Grant No. SBR-9709489 between the National Academy of Sciences and the National Science Foundation. Any opinions, findings, conclusions, or recommendations expressed in this publication are those of the author(s) and do not necessarily reflect the views of the organizations or agencies that provided support for the project.

International Standard Book Number 0-309-07180-1

Additional copies of this report are available from National Academy Press, 2101 Constitution Avenue, N.W., Lockbox 285, Washington, D.C. 20055; (800) 624-6242 or (202) 334-3313 (in the Washington metropolitan area); Internet, http://www.nap.edu

Printed in the United States of America.

Suggested citation: National Research Council (2000). *Improving Access to and Confidentiality of Research Data: Report of a Workshop.* Committee on National Statistics, Christopher Mackie and Norman Bradburn, Eds. Commission on Behavioral and Social Sciences and Education. Washington, D.C.: National Academy Press.

THE NATIONAL ACADEMIES

National Academy of Sciences
National Academy of Engineering
Institute of Medicine
National Research Council

The **National Academy of Sciences** is a private, nonprofit, self-perpetuating society of distinguished scholars engaged in scientific and engineering research, dedicated to the furtherance of science and technology and to their use for the general welfare. Upon the authority of the charter granted to it by the Congress in 1863, the Academy has a mandate that requires it to advise the federal government on scientific and technical matters. Dr. Bruce M. Alberts is president of the National Academy of Sciences.

The **National Academy of Engineering** was established in 1964, under the charter of the National Academy of Sciences, as a parallel organization of outstanding engineers. It is autonomous in its administration and in the selection of its members, sharing with the National Academy of Sciences the responsibility for advising the federal government. The National Academy of Engineering also sponsors engineering programs aimed at meeting national needs, encourages education and research, and recognizes the superior achievements of engineers. Dr. William A. Wulf is president of the National Academy of Engineering.

The **Institute of Medicine** was established in 1970 by the National Academy of Sciences to secure the services of eminent members of appropriate professions in the examination of policy matters pertaining to the health of the public. The Institute acts under the responsibility given to the National Academy of Sciences by its congressional charter to be an adviser to the federal government and, upon its own initiative, to identify issues of medical care, research, and education. Dr. Kenneth I. Shine is president of the Institute of Medicine.

The **National Research Council** was organized by the National Academy of Sciences in 1916 to associate the broad community of science and technology with the Academy's purposes of furthering knowledge and advising the federal government. Functioning in accordance with general policies determined by the Academy, the Council has become the principal operating agency of both the National Academy of Sciences and the National Academy of Engineering in providing services to the government, the public, and the scientific and engineering communities. The Council is administered jointly by both Academies and the Institute of Medicine. Dr. Bruce M. Alberts and Dr. William A. Wulf are chairman and vice chairman, respectively, of the National Research Council.

Acknowledgments

The Committee on National Statistics (CNSTAT) appreciates the time, effort, and valuable input of the many people who contributed to the workshop on confidentiality of and access to research data and to the preparation of this report. We would first like to thank those who made presentations, which, along with the background papers prepared for the workshop, helped identify many of the key issues in this area. The comments made by attendees contributed to a broad-ranging exchange of ideas that is captured in this summary report. We are also thankful for the additional input provided by participants on early report drafts. Thanks are due especially to Norman Bradburn, former CNSTAT member, who as workshop chair provided valuable advice during the planning stages and the leadership necessary for conducting a successful workshop. The agenda for the workshop was developed in consultation with Richard Suzman, director of behavioral and social research at the National Institute on Aging, whose input was essential in identifying workshop objectives.

Particular appreciation is due to those who worked to organize the workshop and prepare this report. Christopher Mackie served as study director for the workshop. He led the planning of the workshop, worked to ensure its successful conduct, prepared the report drafts, and revised the report in response to comments from reviewers and workshop participants. Tom Jabine, who served as consultant to the committee, offered valuable guidance on the agenda, the selection of appropriate presenters, and the preparation of this report. He and Heather Koball, research associate in CNSTAT, contributed significantly to our work with a paper on the practices of organizations that

distribute public-use data, which was presented at the workshop and is summarized in this report. The extra time, guidance, and input of the CNSTAT subcommittee for the workshop—Thomas Louis, Roderick Little, and Charles Manski—further enhanced the workshop's outcome and this report. Miron Straf, former CNSTAT director, was responsible for early project development and workshop planning. CNSTAT staff members Terri Scanlan and Jamie Casey were responsible for all of the details involved in organizing the workshop and preparing this report. Rona Briere edited the final draft. Eugenia Grohman, associate director for reports in the Commission on Behavioral and Social Sciences and Education, guided the report through the review process, final editing, and publication.

This report has been reviewed in draft form by individuals chosen for their diverse perspectives and technical expertise, in accordance with procedures approved by the Report Review Committee of the National Research Council. The purpose of this independent review is to provide candid and critical comments that will assist the institution in making the published report as sound as possible and to ensure that the report meets institutional standards for objectivity, evidence, and responsiveness to the study charge. The review comments and draft manuscript remain confidential to protect the integrity of the deliberative process.

We thank the following individuals for their participation in the review of this report: William F. Eddy, Department of Statistics, Carnegie Mellon University; Amitai Etzioni, Department of Sociology, George Washington University; Jack Feldman, Center for Health Affairs/Project HOPE, Bethesda, MD; Olivia Mitchell, Wharton School, University of Pennsylvania; and Richard Rockwell, Inter-university Consortium for Political and Social Research, University of Michigan.

Although the individuals listed above provided constructive comments and suggestions, it must be emphasized that responsibility for the final content of this report rests entirely with the authoring committee and the institution.

Norman Bradburn, *Workshop Chair*

Contents

Preface

The workshop summarized in this report was convened by the Committee on National Statistics (CNSTAT) to promote discussion about methods for advancing the often conflicting goals of exploiting the research potential of microdata and maintaining acceptable levels of confidentiality. The primary sponsor of the workshop was the National Institute on Aging (NIA), but additional funding was received from the Agency for Health Care Policy and Research; the Bureau of Labor Statistics; the National Library of Medicine; the Office of Research and Statistics, Social Security Administration; and the Office of the Assistant Secretary for Planning and Evaluation, U.S. Department of Health and Human Services. Sponsors voiced a common desire to develop research programs aimed at quantitatively assessing the risks of reidentification in surveys linked to administrative data.

Sponsors also stressed the importance of demonstrating and weighing the value of linked data to research and policy. Prior to the CNSTAT workshop, NIA funded a preworkshop conference, organized through the University of Michigan, to illustrate this value—particularly as it applies to research on aging issues. The workshop was designed to advance the dialogue necessary for federal agencies to make sound decisions about how and to whom to release data, and in what cases to allow linkage to administrative records. Sponsors were interested in improving communication among communities with divergent interests, as well as the decision-making frameworks for guiding data release procedures.

This report outlines essential themes of the access versus confidentiality debate that emerged during the workshop. Among these themes are the

tradeoffs and tensions between the needs of researchers and other data users on the one hand and confidentiality requirements on the other; the relative advantages and costs of data perturbation techniques (applied to facilitate public release) versus restricted access as tools for improving security; and the need to quantify disclosure risks—both absolute and relative—created by researchers and research data, as well as by other data users and other types of data.

The workshop was not designed to produce policy recommendations. However, this report does summarize areas of discussion in which common ground among some participants emerged. For example, a subset of participants endorsed the idea that both access and confidentiality can benefit from (1) more coordination among agencies regarding data release procedures and creation of data access outlets, (2) increased communication between data producers and users, (3) improved quantification of the disclosure risks and research benefits associated with different types of data release, and (4) stricter enforcement of laws designed to ensure proper use of restricted access data.

Finally, the report anticipates the direction of future CNSTAT projects. Future work will likely address evolving statistical techniques for manipulating data in ways that preserve important statistical properties and allow for broader general data release; new, less burdensome ways of providing researchers with access to restricted data sets; and the role of licensing coupled with graduated civil and criminal penalties for infringement.

Norman Bradburn, *Workshop Chair*

1

Introduction

In October 1999, the Committee on National Statistics (CNSTAT), in consultation with the Institute of Medicine, convened a 2-day workshop to identify ways of advancing the often conflicting goals of exploiting the research potential of microdata and preserving confidentiality. The emphasis of the workshop was on longitudinal data that are linked to administrative records; such data are essential to a broad range of research efforts, but can also be vulnerable to disclosure. Administrative data are collected to carry out agency missions and constitute the majority of agency data. An additional—much smaller—amount of data is collected specifically for research and other public purposes. It is sometimes feasible and useful to merge the latter data with the more extensive administrative records.

CNSTAT has had an active history working in the area of data confidentiality and access, culminating with the panel study that produced the volume *Private Lives and Public Policies: Confidentiality and Accessibility of Government Statistics* (National Research Council and Social Science Research Council, 1993). That study resulted in a series of recommendations for advancing researchers' access to data without compromising the ability to protect the confidentiality of survey respondents. This workshop brought together several participants from that study and many others representing various communities—data producers from federal agencies and research organizations; data users, including academic researchers; and experts in statistical disclosure limitation techniques, confidentiality policies, and administrative and legal procedures.

KEY ISSUES

The development of longitudinal data sets linked to health, economic, contextual geographic, and employer information has created unique and growing research opportunities. However, the proliferation of linked data has simultaneously produced a complex set of challenges that must be met to preserve the confidentiality of information provided by survey respondents and citizens whose administrative records are entrusted to the government. Unprecedented demand for household- and individual-level data, along with the continuing rapid development of information technology, has drawn increasing attention to these issues. Technological advances have rapidly improved the range and depth of data; opportunities to access, analyze, and protect data have grown as well. However, technology has concurrently created new methods for identifying individuals from available information, of which longitudinal research data are but one of many sources.

Longitudinal files that link survey, administrative, and contextual data provide exceptionally rich sources of information for researchers working in the areas of health care, education, and economic policy. To construct such files, substantial resources must be devoted to data acquisition and to the resolution of technical, legal, and ethical issues. In most cases, requirements designed to protect confidentiality rule out the type of universal, unrestricted data access that custodians—and certainly users—of such databases may prefer.

Several modes of dissemination are currently used to provide access to information contained in linked longitudinal databases. Dissemination is typically restricted either at the source, at the access point, or both. Products such as aggregated, cross-tabulation tables are published regularly and made available to all users, but of course offer no record-level detail. This type of data does not support research into complex individual behavior. Public-use microdata files, on the other hand, offer detail at the individual or household level and are available with minimal use restrictions. However, producers of microdata must suppress direct identifier fields and use data masking techniques to preserve confidentiality. Additional methods, such as licensing agreements, data centers, and remote and limited access, have been developed to limit either the types of users allowed access to the data, the level of data detail accessible by a given user, or both. Restricted access arrangements are generally designed to provide users with more detail than they would get from a public-use file.

It is within this context that the workshop participants debated the key issues, which can loosely be organized at two levels. The first is the tradeoff that exists between increasing data access on the one hand and improving

data security and confidentiality on the other.[1] To examine this tradeoff, it is necessary to quantify, to the extent possible, disclosure risks and costs, as well as the benefits associated with longitudinal microdata and with linking to administrative records. Decisions about what types of data can be made available, to whom, and by what method hinge on the assessment of these relative costs and benefits. Researchers typically appeal for greater access to unaltered data, while stewards of the data are understandably often more focused on assessing and minimizing disclosure risk.

At the second level of discourse, participants discussed alternative approaches to limiting disclosure risk while facilitating data access. Given that all longitudinal microdata require some protections, the compelling question is which approach best serves data users while maintaining acceptable levels of security. The choice reduces essentially to two options: (1) restricting access—physically limiting who gets to see the data, or (2) altering the data sufficiently to allow for safe broader (public) access. Other elements, such as legal deterrents, also come into play. Workshop participants articulated in detail the merits and relative advantages of alternative approaches. Their arguments are summarized in this report.

WORKSHOP GOALS

As noted above, a central objective of the workshop was to review the benefits and risks associated with public-use research data files and to explore alternative procedures for restricting access to sensitive data, especially longitudinal survey data that have been linked to administrative records. Doing so requires considering the impact on each group involved—survey respondents, data producers, and data users—of measures designed to reduce disclosure risk. Presenters from the academic community reviewed the types of research that are enhanced, or only made possible, by the availability of linked longitudinal data. Participants also identified and suggested methods for improving current practices used by agencies and research organizations for releasing public-use data and for establishing restricted access to nonpublic files. The overarching theme was the importance of advancing methods that maximize the social return on investments in research data, while fully complying with legal and ethical requirements.

[1]Early on in the workshop, a participant clarified the distinction between "privacy" and "confidentiality." Privacy typically implies the right to be left alone personally, the right not to have property invaded or misused, freedom to act without outside interference, and freedom from intrusion and observation. In the context of research data, confidentiality is more relevant. The term refers to information that is sensitive and should not be released to unauthorized entities. It was suggested that confidentiality implies the need for technical methods of security.

The workshop, then, was designed with the following goals in mind:

- To review the types of research that are enhanced, or only made possible, using linked longitudinal data.
- To review current practices and concerns of federal agencies and other data producing organizations.
- To provide an overview of administrative arrangements used to preserve confidentiality.
- To identify ways of fostering data accessibility in secondary analysis.
- To assess the utility of statistical methods for limiting disclosure risk.

To date, efforts to address these themes have been hindered by inadequate interaction between researchers who use the data and agencies that produce them and regulate their dissemination. Researchers may not understand and may become frustrated by access-inhibiting rules and procedures; on the other hand, agencies and institutional review boards are not fully aware of how statistical disclosure limitation measures impact data users. The workshop brought the two groups together to help overcome these communication barriers.

REPORT ORGANIZATION

Workshop topics were organized into the following sessions: (I) linked longitudinal databases—achievements to date and research applications, (II) legal and ethical requirements for data dissemination, (III) procedures for releasing public-use microdata files, and (IV) procedures for restricted access to research data files. This report is structured slightly differently to focus on themes as they emerged during the workshop. Chapter 2 outlines the tradeoff between data access and confidentiality. Presentations on the research benefits of linked longitudinal data are summarized, along with discussions of disclosure risk assessment and quantification. Chapter 3 reviews presentations that addressed ethical and legal aspects of data dissemination, as well as discussion on the role of institutional review boards. Chapter 4 summarizes participants' assessments of competing approaches to limiting disclosure risk and facilitating user access; the focus is on two primary competing approaches—data perturbation and access limitation. Agency and organization practices are the subject of Chapter 5. In adition, two appendices are provided: Appendix A is a list of the workshop participants; Appendix B is the workshop agenda.

2

The Data Access, Confidentiality Tradeoff

This chapter presents participants' views on the workshop's overarching theme—the balancing of data protection needs and requirements against the benefit of access to linked longitudinal data. Academic researchers were understandably more vocal about improving data detail and access, while data stewards were more concerned about security and protection of confidentiality. The lines between the two communities are not always clearly drawn, however. For instance, most federal agencies are accountable for both functions, protecting the interests of data subjects through procedures that ensure appropriate standards of privacy and confidentiality, and facilitating responsible dissemination to users (National Research Council and Social Science Research Council, 1993).

It is impossible to specify a universally applicable optimal tradeoff between data access and data protection. The value of data-intensive research is highly variable, as are mission, operation, and data production across agencies. The panel that produced the report *Private Lives and Public Policies* recognized this variability and did not advocate trying to identify such a tradeoff (National Research Council and Social Science Research Council, 1993). However, workshop participants expressed general optimism about the possibilities for developing tools that would enhance, on a case-by-case basis, the ability to increase data access without compromising data protection or, conversely, to increase confidentiality without compromising data access.

ROLE OF LINKED LONGITUDINAL MICRODATA IN RESEARCH AND POLICY

The nation faces a range of increasingly complex policy issues in such areas as social security, health care, population aging, changing savings patterns, advancing medical technology, and changing family structure. Addressing these issues will require increasingly sophisticated data and behavioral modeling. Microdata sets, such as those from the Health and Retirement Study (HRS), offer the most promising means of answering specific questions such as how social security interacts with pensions and savings in household efforts to finance retirement, how social security age eligibility requirements affect retirement rates and timing, and how changes in out-of-pocket medical expenses affect utilization of federal programs.

Survey data sets, particularly those linked to administrative records, facilitate a broad spectrum of research that can shed light on such questions, and that could not otherwise be reliably conducted. Additionally, linking to existing information can streamline the data production process by reducing the need to duplicate survey activities. Linking of survey and administrative data has the potential to both improve data quality and reduce data production costs.

Research Benefits of Linking Survey and Administrative Data

Researchers attending the workshop expressed the view that data linking opens up a wide range of research and modeling options. Richard Burkhauser of Cornell University presented his paper (coauthored by Robert Weathers, also of Cornell), "How Policy Variables Influence the Timing of Social Security Disability Applications," to lead off the opening session. His presentation focused on how the HRS made his study possible. Burkhauser and other participants praised the HRS, calling it a clear "case study" of the potential of linked longitudinal data to advance policy-oriented social science research.

For his application, Burkhauser was able to model the effects of social security on economic behavior, including retirement decisions. This type of analysis follows in the tradition of economic research conducted during the 1970s that used the Retirement History Survey and the Exact Match File, which actually linked Current Population Survey data to Internal Revenue Service (IRS) and social security records.[1] In the opinion of several partici-

[1]These data stimulated significant research in the 1970s and set the agenda for the 1980s. However, during the 1980s, budget cutbacks, combined with the emerging emphasis on confidentiality, affected the ability of the Social Security Administration and other agencies to produce and disseminate new data.

pants, the HRS files, along with new data sets from other sources, such as the Disability Evaluation Study, will be the key inputs for models that will allow researchers to address policy questions about household savings patterns, social security solvency, worker retirement patterns, and other issues that require accurate and detailed financial information.

The HRS longitudinal data set follows about 22,000 persons aged 50 and above and their spouses; data are collected every other year. With the consent of participating individuals, survey data are linked to social security earnings histories and also to some employer pension plan data. Linked records that allow identification at as low as the state level are maintained internally. The HRS is available in a public use-form, and with linkages to social security and pension information under restricted conditions. Access to geocoding is also restricted; to gain such access, a researcher must go through an institutional review board with a research plan and data protection provisions. Firm-level data have been linked as well, and while access to Medicare data has not yet been authorized, work on associated data collection processes and restricted access protocols is in progress. Multiple linkages are also permitted, typically under even more restricted conditions.

The HRS also provides detailed survey information—some of it retrospective—about the health conditions and work patterns of respondents. However, it is the linkage to administrative records that is exceptionally valuable for policy research. Linking to social security data allows researchers to construct earnings histories and to determine each person's potential disability benefits. Past earnings are available to predict potential future earnings for cases in which individuals have not yet applied for benefits. Linking HRS data with administrative records and geographic information allowed Burkhauser and Weathers to answer questions that could not otherwise have been answered. For instance, the authors were able to estimate how disability allowance rates affected the timing and rate of application for benefits. Likewise, they were able to estimate the behavioral impact of expected benefit level, expected earnings, gender effects, and policy variables.

Linking to administrative records can improve data accuracy as well as data scope by giving researchers access to information that individuals may not be able to recall or estimate accurately in a survey context. Survey data can be biased as a result of flaws in respondent memory or understanding of measurement concepts. For instance, the HRS is linked to social security files containing lifetime earnings data that would be virtually impossible for respondents to recall or even find in records. Similar situations arise with regard to medical records. Also, survey data are devalued if they fail to capture variability of parameters accurately, since this variability is central to modeling efforts and, in turn, the ability to answer policy questions. HRS linkages introduce accuracy and detail to the data that are particularly constructive for modeling savings incentives, retirement decisions, and other dy-

namic economic behavior. The authors concluded that the HRS was a clear example in which the potential research benefits of linking were sufficiently large that the Social Security Administration could justify approving the linking proposal with minimal controversy.

Also in the first session, Rachel Gordon of the University of Illinois at Chicago presented a second example illustrating the benefits of data linking. Her research investigates the impact of community context on child development and socialization patterns, as well as the impact of the availability of child care on parents' work decisions. By having access to a National Longitudinal Survey of Youth (NLS-Y) file with detailed geographic codes for survey respondents, Gordon was able to add contextual data, such as the availability of child care, for the neighborhoods in which respondents lived. Gordon's application highlights the tradeoff between data precision and disclosure risks. Access to census tract-level geocoding permits more sensitive construction of community and child care variables central to the study but could, under certain conditions, greatly increase the indentifiability of individual NLS-Y records.

During the general discussion, participants cited other aspects of linking that increase data utility. Linking makes it possible to get more out of isolated data sets that would otherwise have limited application. The process can increase the value of data sets, reduce data collection redundancies, and improve data accuracy in a cost-effective manner, and provides added flexibility to meet unforeseen future research needs with existing data sets. For example, if the Census Bureau's Survey of Income and Program Participation could be linked to administrative tax return records, income data would likely be more accurate and the cost of the survey decreased. Robert Willis of the University of Michigan also made the point that this type of linking reduces response burden. If survey designers know that links can be made to administrative data, they can limit the length of questionnaires.

The benefits of linked data extend beyond social science research, which was the focus of this workshop. Robert Boruch from the University of Pennsylvania pointed out, for example, that linking records is essential for randomized field trials in the social sector (crime, welfare, education). J. Michael Dean from the University of Utah Medical Center reiterated these points for medical research, observing that there is no way to conduct high-quality clinical trials in many different areas without the capacity for linking; the same can be said of work in criminal justice.

Linking also facilitates research on infrequent events, such as rare diseases, that affect only a small percentage of the population. In such cases, working from general sample data does not provide adequate sample sizes for target groups. Population-based data, which are very expensive to collect, are often required. Linking can, in some instances, provide a much less costly substitute.

Assessing and Articulating the Social Value of Research

Robert Boruch noted in his presentation that 20 years ago, despite interagency agreements to link data sources at the IRS, the Census Bureau, and other agencies, examples of the benefits of linking were few and far between. In contrast, current research, as represented by workshop presentations in Session I, demonstrates the extent to which this landscape has changed. Yet several researchers expressed concern that it seems easier to discuss the risks associated with data provision than to communicate the benefits convincingly. They suggested that researchers have done a poor job of publicizing the value of their work. Boruch summarized the view, noting, "We do not sell social research well in this country, and that is part of the reason why at least a half dozen large private foundations are trying to understand how to do that job better."

Of course, it can be difficult for agencies and organizations to assess the value of their data without carefully tracking the numbers and types of users, as well as information on those being denied access. The panel that produced *Private Lives and Public Policies* explicitly recommended establishing procedures for keeping records of data requests denied or partially fulfilled. This recommendation could be expanded to encourage full documentation of data usage. Developing archives and registries of data performance may serve as an effective first step toward fostering understanding by both the public and policy makers of the extent to which data-intensive research provides key information.

Research Impact of Data Alteration Versus Access Restriction

Alternative methods for reducing statistical disclosure risks were discussed at length. At the most general level, the options fall into two categories—data alteration and access restriction. The discussion of these approaches is detailed in Chapters 4 and 5. On balance, researchers participating in the workshop expressed a preference for monitoring the behavior of scientists over altering the content of data sets to permit broader distribution. They favored licensing agreements, data enclaves, and the like over perturbation methods. They also advocated the use of legal remedies (discussed in Chapters 3 and 4) whenever possible as a nonintrusive (to rule-abiding researchers) alternative that rewards responsible data use.

Researchers expressed serious concern about the impact of statistical disclosure limitation techniques that distort variable relationships and that may have an unanticipated (or even anticipated) impact on modeling results. The perception among leading researchers appears to be that altered or, more specifically, synthetic data can solve some problems, but are inadequate for the majority of cutting-edge work. Regardless of its accuracy, this perception

implies that such data may be less likely to be used, which damages the research enterprise. Sophisticated data perturbation also increases the methodological knowledge researchers must have to understand data, utilize them legitimately, and produce interpretable results.

Several researchers cautioned that, although they would rather live with the burden of limited access than deal with synthetic data, data centers and other limited access arrangements do impose costs on research. To the extent that data centers favor large-budget research projects, less well-funded disciplines can be prevented from accessing important data resources. Also, an important reason for accessing a data set is to replicate findings of other researchers. Typically, programs used to create data files must be designed in such a way that they can be shared by others wishing to replicate work. When researchers must repeat the entire data acquisition process, documentation and replication become more cumbersome. Additionally, one participant noted that when projects must be approved by agencies that host data, the potential is created for censorship, as well as milder forms of restriction arising from a lack of familiarity with the scientific literature.

Participants also raised the issue of research timing as it relates to policy relevance. The benefits of research can be dampened when data acquisition is arduous and results are delayed. For instance, it took Burkhauser and Weathers 2 years to gain access to restricted data required for their analysis. Accessing previously linked data will become less time-consuming if data centers are able to streamline their procedures effectively.

At data enclaves, researchers must typically submit programs to center staff who then perform the necessary data processing steps. For Burkhauser and Weathers, this meant enlisting HRS staff to merge the data sets, run the programs, and produce output that could then be rechecked. The length of this iterative process depends on the turnaround time from the data center and the number of adjustments and resubmissions required by the researcher. This process can appear burdensome and inefficient to researchers accustomed to having access to detailed data, doing the coding, and creating extract files themselves. These researchers argue that one must understand the process of research to assess the effectiveness of a remote access program.

Another indirect consequence, noted by Rachel Gordon, of limiting access to small-area geographic codes for survey respondents is that doing so may conceal demand by researchers for better data of one type or another. If researchers know, for example, that they can link contextual information, this type of work will be done, and in the process, the type of information of interest to the research community will be revealed. This feedback can in turn be used by producing agencies and organizations to make funding and allocation decisions.

THE CONFIDENTIALITY PROTECTION CONSTRAINT: ASSESSING THE RISKS

In addition to estimating the value of data access, efficient and balanced policy requires accurately assessing the disclosure risks (and associated social cost) posed by microdata and linking.[2] Risk of disclosure is affected by numerous data set characteristics. Level of geographic detail is often the factor cited first. Small-area geocodes can make reidentification possible using basic statistical techniques. More advanced computer science methods, outlined by Latanya Sweeney of Carnegie-Mellon University, can substantially increase the power of reidentification techniques. As the sample approaches its underlying population, the geographic unit must increase in size if a constant level of protection is to be maintained. Certain types of surveys, such as those for people with rare medical conditions, maintain high sample rates from the subpopulation from which they are drawn. Likewise, geographic units must increase in size with the number of variables that can be cross-referenced if disclosure risk is to be held constant.

In addition to detrimental effects on exposed citizens, disclosure events can negatively impact data-intensive research enterprises. Several workshop participants argued that more work is needed on assessing the impact of disclosure (or perceived high risk) on survey participation. If potential survey participants observe instances of disclosure, or even perceive that confidentiality is becoming less secure, it may become more difficult for data producing organizations and agencies to enlist their cooperation. Arthur Kennickell of the Federal Reserve Board suggested that disclosure of individuals in the Fed's Survey of Consumer Finances might endanger the whole study, even if it were just an annoyance for those involved. Similarly, if potential survey participants believe that linking increases risks, or that all data about them are available through linking, they may be less forthcoming with information and their time. A theme that emerged from the workshop was that advancing access and confidentiality objectives requires cognizance of the relationship between the perceptions of respondents and the ability to collect data.[3] Developing better methods for eliciting consent and educating the public about

[2]The risk-cost relationship question was raised but not answered at the workshop. Risk is a function of both the probability of and the potential damage from disclosure. Participants acknowledged the need to assess disclosure risks; they were less certain about how best to quantify harm—the true cost—that results from disclosure. This question requires additional attention.

[3]A report produced by the Panel on Privacy and Confidentiality as Factors in Survey Response (National Research Council, 1979) provides some evidence about respondent attitudes, indicating that promises about confidentiality and data security are often questioned by the public.

the real risks associated with survey involvement may be a cost-effective use of resources.

Absolute Versus Relative Risks Associated with Data Access and Linking

Discussions of disclosure risk often emphasize the isolated probability that identifiable information will be revealed about an individual from a survey or administrative data set. CNSTAT member Thomas Louis of the University of Minnesota suggested recasting the debate: instead of comparing risks with probability zero, one might consider how the probability of disclosure changes as a result of a specific data release or linkage, or from adding (or masking) fields in a data set. In this context, the question becomes what marginal risk is associated with an action.

For cases in which the same data are available elsewhere, even if not in the same form or variable combination, the added risk of releasing a research data file may be comparatively small. Given the current trend whereby more and more data are becoming available, it may be reasonable to assume that the marginal risk of releasing research data has actually diminished. The validity of this assumption is, at present, unknown since no one has estimated the risk of disclosure as a function of survey inclusion conditional on the existence of data available from other sources. If security risks are rising rapidly in general, the relative risk of scientific survey data may be decreasing.

Robert Gellman, consultant, described the sensitive data that are available from a wide range of sources. Records such as driver's licenses, voter registration information, vehicle licenses, property tax records, arrest records, and political contributions, to name a few, are readily available to the public in many jurisdictions. Additionally, marketers compile this information into lists that are also available, often at low cost. Companies have software designed to locate, retrieve, and cross-reference information to construct detailed consumer profiles and mailing lists. For a given individual, these lists may include name, address, age, date of birth, marital status, income, household size, number of credit cards, occupation, phone number, and even social security number. Much of this information is constructed from linkages across sources. However, these market-oriented data collectors are not typically building their products from longitudinal research data sources.

Some participants argued that policy makers need to consider how rules should differ and how each should apply to different types of data. There appear to be far more recorded instances of breached confidentiality with nonresearch data, or at least data used for nonresearch purposes.[4] If this is

[4]In fact, though there have been numerous cases reported anecdotally in which procedural rules governing data use were violated, there are no known cases in which a respondent was harmed as a result of disclosures from a research data set.

true, legislation designed to protect against data misuse, if not carefully constructed, could extend unnecessarily to research data or, on the other hand, be too lax to safeguard the public from commercial data abuses.

Confusion over confidentiality risks associated with different types of data can inhibit the creation and productive exploitation of legitimate research data. The risks associated with participating in a well-designed and well-conducted scientific survey may be very different from those posed by marketing data; when these distinctions become blurred and all data production activities are viewed homogeneously, the public is more likely to believe that data endanger privacy. Although survey data may not add much to the danger, suspicions about data in general make collecting data more difficult, which in turn constrains data-intensive research.

J. Michael Dean pointed out the importance of comparing the risks posed by linked data against those posed by native data sets—that is, the original data from which the links were made. In many cases, he argued, the native databases are already sensitive and as such require adequate protection. The linking may create a combined data set that increases disclosure risks substantially, but this is not always the case. A disclosure incident occurring from a linked data source is not necessarily caused by the linking; it might have occurred from the stand-alone data as well. Again, the marginal risk associated with the link needs to be evaluated, preferably against the research benefit. This evaluation requires assessing the extent to which stewards and users of data sets are likely to be responsible for leaks. It is necessary to assess how likely people handling medical records or working at agencies are to be sources of leaks relative to users of linked data. Effort expended to protect the security of data at the source versus at the linking phase should be proportional to the relative disclosure risks posed at each point. Unfortunately, such assessments of relative risk have as yet typically not been made.

It is also important to distinguish between organizationally linked data (such as the HRS) and individual user-linked data. The enormous growth of computing power and data availability is constantly changing the cost parameters of "snooping." To begin with, the prospective linker must have software that can be expensive. Some probabilistic matching programs can cost thousands of dollars, while other types of data linking can be performed with tools as simple as Microsoft Access. The cost of linking is a deterrent, but this cost varies significantly with the type of linking performed. Latanya Sweeney reported that she is able to identify individuals from a range of publicly available data sources using simple software in combination with creative methods. However, her applications have thus far focused on localized geographic data sources, and it remains to be seen how far computer science-based methods can go toward identifying individuals from national longitudinal surveys linked to administrative records.

The probability of being able to identify a record within a longitudinal

research database also depends on the snooper's objective. Knowing whether an individual is in a sample makes an enormous difference in terms of identifiability (Juster, 1991). For instance, a snooper can guess with much greater certainty whether an individual is in a localized database of voter registration records than whether an individual is in the HRS. Of course, if the snooper is simply trying to identify a random record, that is another matter altogether.[5] It is important to note that, for sample-based research data sets, a given level of confidentiality can generally be obtained with less protection than is required for population-based data sources, such as voter registration records, municipal registries, local censuses, and hospital administrative records.

A related issue is the need to distinguish among different kinds of data users. At one end of the spectrum are individuals whose data use is motivated by objectives other than disclosing identities intentionally; at the other end are those with vested interests in doing just that. The latter group may be involved in marketing, conflict resolution, law enforcement, or a range of other activities. It may be presumed that researchers typically fall at the "safe" end of the spectrum, and that those seeking information on specific individuals are less likely to rely on research data than on other sources.[6] However, the extent to which this generalization holds is not known.

Nonetheless, several participants expressed the view that access rules must be tailored to reflect risk levels posed by specific types of data users. When access rules are set universally, they typically tend to protect against the most dangerous users, limiting the ability to maximize the social return on data. It is highly unlikely that the same legal framework designed to protect individuals from marketers, employers, or the media is appropriate to apply to researchers and research data. The risks and benefits involved are not comparable.

On the other hand, several participants indicated that it is no simple task to regulate data access by class of user since traditional categories overlap, and data users may work in multiple areas. To establish clear rules and procedures for researchers, these participants suggested first gaining a clearer idea of who needs to be covered and in what way. No one has adequately sorted

[5]This point also relates to the debate between Sweeney and Dean about the ease with which data sources can be linked. The distinction between Sweeney's position—that linking is inexpensive and requires only basic software—and Dean's—that linking is difficult and expensive—is at least partially tied to a snooper's objective. Sweeney's works suggests that it is technically possible to quickly link records of *some* individuals from two large files; at the same time, as Dean argued, building an accurate *comprehensive* linked data set from the two sources may require many identifiers and a high degree of sophistication.

[6]Similar arguments were advanced to support the view that regulating the behavior of data users is more efficient than altering the data to allow broader access. This discussion is reviewed in Chapter 4.

out how identification of the various players, even those in the research community—e.g., top researchers, all researchers, graduate students—should work.

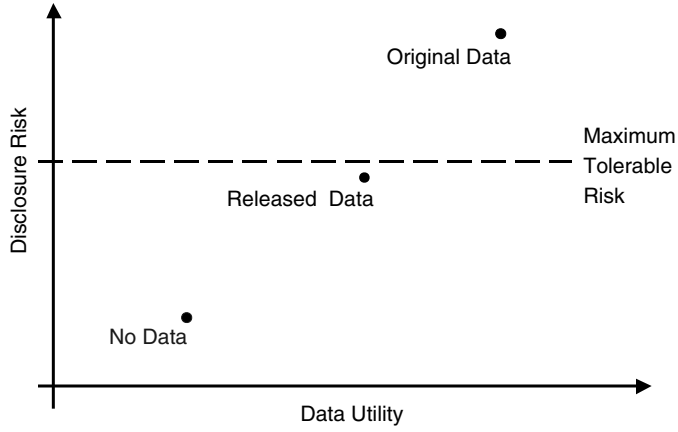
Quantifying Disclosure Risks and Costs, and Research and Policy Benefits

As suggested earlier, a fully informed data release policy requires quantitative estimates of the potential social costs and benefits associated with the release. Unfortunately, given case-by-case variability and uncertainty in the process of scientific discovery, there is no obvious mechanism available that can facilitate evaluation of the balance between disclosure risk and data utility. During Session III of the workshop, Sallie Keller-McNulty of Los Alamos National Laboratory and George Duncan of Carnegie-Mellon University presented a paper that begins to define such an operational construct.

Statistical disclosure risks are defined by the ability of snoopers to draw inferences from de-identified data; linking is generally part of this process. The authors' framework, represented in the diagram below, essentially attempts to maximize data utility by minimizing necessary distortion, subject to a maximum tolerable risk constraint.

Data utility is a function of usefulness for statistical inference, and may be reflected in bias, variance, or mean square error characteristics inherent in the data. Options for maintaining tolerable risk include restricting data detail, restricting access to the data, or a combination of the two. Identifying an optimal data access strategy involves a difficult maximization problem since the parameters are not constant.

In their presentation, Duncan and Keller-McNulty emphasized restric-



tive data release as opposed to restricted data access—specifically (1) data masking and (2) synthetic data. A goal of their work is to quantify disclosure risk and data utility, distinguishing between identity disclosure (e.g., associating the name of a respondent with a piece or pieces of released information) and attribute disclosure (e.g., estimating an attribute value associated with an individual). Keller-McNulty discussed how proponents of the framework might identify important parameters of the process intended to measure risk and data utility generically. Full utilization of their framework would entail comparison of statistical properties of raw versus perturbed data. The eventual goal would be to allow choices in setting parameters that would maximize the tradeoff between data utility and risk and thereby avoid distorting data more than is necessary.

The authors reviewed the algebra of some simple univariate examples that illustrate the framework for assessing risks associated with data dissemination strategies. Their conclusions pointed broadly to the need to (1) extend their univariate examples to more realistic settings, and (2) develop risk and utility measures that are a function of the masking parameters and other matrix factors used to create synthetic records. Finally, the authors noted that a multitude of possible dissemination strategies exist; level of masking and extent of synthetic iteration are specific to data set and application. In no case will the optimal tradeoff imply 100 percent disclosure prevention.

Latanya Sweeney was discussant for the Keller-McNulty and Duncan presentation. She has analyzed technical protections and risks from a computational science perspective, and is concerned with identifying optimal modes of releasing information so that inferences can be controlled—again in an environment where information about individuals is increasing at an exponential rate. It should be noted that Sweeney's comments were directed toward a broad class of information and not specifically toward longitudinal research databases. This broadened focus provided essential context within which the relative risks of different types of data could be discussed.

Sweeney's computational approaches to assessing risk involve constructing functions that relate hierarchical data aggregation to data security. Working at the cell level, she has developed algorithms designed to balance the tension between the usefulness of data, as measured by the degree of aggregation (i.e., number of variables collapsed), and protection of anonymity. The matrix mask described by Keller-McNulty and Duncan could be used as a general representation for the hierarchical aggregation. Sweeney's method structures disclosure limitation into a generalized hierarchy in which variable suppression would typically be at the top. The algorithms compute how much a field is distorted as it is moved up the hierarchy toward suppression. Sweeney then computes a "precision metric" for a particular combination of data based on how far the variables must move up the table to obtain a given level of security.

In principle, one can obtain a given level of risk while minimizing required data distortion for each field and without resorting to uniform cell suppression. In computer programs such as DataFly, which generalizes values based on a profile of data recipients, and u-ARGUS, a similar program used by Statistics Netherlands, statistical disclosure limitation methods are used that do make data safer but less specific and the overall statistical relationships within the data set less accurate. The key is finding optimal levels of distortion across data set fields and exploiting tradeoffs between specificity and anonymity protection. Applying her techniques to u-ARGUS and DataFly, Sweeney concluded that the former tends to underprotect the data, while the latter tends to overdistort.

A central message of Sweeney's presentation was that an incredible range of resources can be directed toward cracking data security, yet technology can also offer solutions to those charged with producing and protecting data. In the future, sophisticated cryptographic techniques may be implemented. For example, researchers may be able to use cryptographic keys to access and link data. In this way, agencies could monitor access to the data and make the results of program runs anonymous. While such techniques are visions of the future, technological solutions, in Sweeney's view, already offer greater protection than legislative approaches, which tend to lag behind real-time needs and reflect outdated technological capabilities.

3

Ethical and Legal Requirements Associated with Data Dissemination

Session II of the workshop was structured to provide an overview of the ethical and legal issues related to data dissemination. Mary Ann Baily, Institute for Ethics, American Medical Association, presented a paper, commissioned for the workshop, titled "Regulating Access to Research Data Files: Ethical Issues." Donna Eden, Office of the General Counsel, U.S. Department of Health and Human Services, outlined her assessment of recent and prospective legislative developments. Finally, Thomas Puglisi of the Office for Protection from Research Risks (OPRR), National Institutes of Health, provided an overview of the role of institutional research boards (IRBs).

ETHICAL ISSUES

Mary Ann Baily described underlying ethical issues raised by the use of microdata, especially longitudinal data that are linked to administrative records. In so doing, she articulated the conflicting rights and obligations of data subjects, producers, and users, and the role of government in providing a structure within which these conflicts can be resolved.

Baily outlined positions at both extremes of the policy debate over data access, then made the case for pursuing a middle ground—striking a balance between the right to be left alone and the obligation to cooperate in the pursuit of communal goals. She discussed activities that are essential to setting appropriate limits on data use and the organizational framework required to carry out these activities. She concluded with observations on the

problem of translating societal recognition of a moral right to privacy into enforceable public policy.

Cases For and Against Unrestricted Access to Microdata for Research

In fields such as health care, education, and economic policy, research based on microdata can illuminate the nature of social problems and the effects of public and private actions taken to ameliorate those problems. Members of society, including data subjects, can benefit from more efficient use of their pooled funds (such as tax dollars and health insurance premiums). Researchers themselves benefit as well, in terms of advancing research programs and fulfilling career goals.

Given these benefits, and given that the cost—often publicly funded—of developing databases is high, it may be asked why databases should not be made widely available to all. Such a policy would maximize benefits, and any costs associated with the additional access could be charged to users when appropriate.

The primary objection to a policy of unrestricted access arises from the potential effect on data subjects, since disclosure of personal information can be harmful. Disclosure of such information may result in being arrested for a crime, being denied eligibility for welfare or Medicaid, being charged with tax evasion, losing a job or an election, failing to qualify for a mortgage, or having trouble getting into college. Disclosure of a history of alcoholism, mental illness, venereal disease, or illegitimacy can result in embarrassment and loss of reputation. Less directly, research results based on personal data can cause harm by affecting perceptions about a group to which a person belongs.

Even in the absence of such concrete effects, disclosure can be seen as a harm in itself, a violation of the fundamental right to privacy, derived from the ethical principle of respect for individual autonomy. Informational privacy has been defined as "the claim of individuals, and the societal value representing that claim, to control the use and disclosure of information about them" (Fanning, 1998:1). Gostin (1995:514) highlights the importance of respect for privacy to the development of a sense of self and personhood: "It is difficult to imagine how, in the absence of some level of privacy, individuals can formulate autonomous preferences, or more basically, develop the capacity to be self-governing." A lack of respect for privacy makes people reluctant to trust others with personal information; for example, they may conceal sensitive information needed by their physicians to provide effective treatment.

Those who argue for recognition of a strong right to informational privacy claim that access to data about individuals should require their explicit consent. Advocates of this position may acknowledge the research potential of personal information databases developed by public and private entities;

many accept the use of such data for socially useful research if the data are aggregated or otherwise processed to prevent identification with a specific person. They maintain, however, that if the data are personally identifiable, researchers must persuade the data subjects to agree voluntarily to each use. To holders of this view, then, the appropriate policy is no access to personally identifiable data without explicit, informed consent.[1]

Restricted Access: What Limits Are Appropriate?

Since unrestricted access can cause harm to individuals and also conflicts directly with respect for individual autonomy, it is not an appropriate policy. On the other hand, requiring explicit, informed consent for any access to personally identifiable data is also problematic. On a practical level, having to obtain meaningful informed consent for every use of data would make much valuable research prohibitively expensive. Meeting this requirement would be costly not only to the research enterprise, but also to data subjects, who would have to spend time submitting to the process.

On a philosophical level, such a policy is focused solely on an individual right and ignores individual responsibilities. The right to informational privacy has never been considered absolute. Governments must collect personal information to function, and members of society have a civic duty to cooperate. For instance, the U.S. Constitution requires that there be a decennial census. Governments require research results to determine areas in which policy action is needed and what form it should take; additional research is needed to determine whether policies have been effective. Research is also an essential element in the support of individual civil rights and the right to a fair trial. Moreover, private organizations must be able to produce research results to carry out their roles effectively. Individuals cannot refuse to provide personal information to private entities such as educational institutions, health care delivery organizations, and employers unless they are willing to forego education, health care, and employment.

In some cases, research can be done on data collected from volunteers, but in others, unacceptable bias would result. Moreover, if data collected from voluntary subjects for one purpose should later become essential for an

[1]Informed consent is defined in *Private Lives and Public Policies* as "a person's agreement to allow personal data to be provided for research and statistical purposes. Agreement is based on full exposure of the facts the person needs to make the decision intelligently, including any risks involved and alternatives to providing the data" (National Research Council and Social Science Research Council, 1993:23). See Chapter 3 of the same report for a full description of the terminology, as well as of the historical development and use of informed consent and notification procedures.

unforeseen but even more beneficial purpose, obtaining individual consent to the new use may be impossible, or at least prohibitively expensive.

For the above reasons, it is unreasonable to allow individuals complete authority to control whether and on what terms they participate in socially important research. Therefore, a balance must be struck between the right to be left alone and the obligation to cooperate in the pursuit of societal goals. According to Baily, the appropriate policy is somewhere in between unlimited access to personally identifiable data and access only with explicit informed consent, with the chosen policy being supported by sufficient security to maintain confidentiality. The difficulty is in reaching agreement on just where the point of balance exists in a particular context. For instance, the appropriate policy is likely to be different for research databases than for other types of data, such as hospital records and marketing information.

Establishing and Enforcing Appropriate Limits

Baily suggested that three kinds of activities are inherently part of imposing appropriate limits on access to microdata:

- *Weighing of relative benefits and costs*—First, data managers must develop information on the benefits associated with research use of personal data and the harm that could result from granting access. This activity includes investigating the attitudes of data subjects. Both benefits and costs may vary significantly in different contexts. The benefits must then be weighed against the costs to determine access policy, with guidance from and accountability to the community as a whole through democratic institutions.
- *Maintenance of confidentiality*—Data managers must be able to enforce whatever limits are established. It is impossible to eliminate the possibility of improper use entirely, but security measures must be adequate to protect legitimate privacy interests. This requires developing information on the risk of misuse in each case, given the nature of the data and their potential uses, and tailoring security measures accordingly. There must also be effective sanctions for violations of confidentiality policies, aimed at preventing improper use, not merely punishing those responsible after misuse occurs.
- *Public education/notification/consent*—Data managers must inform data subjects about information policy and obtain their consent to use of the data when appropriate. Decisions about what people must know, how they should be told, and when consent rather than simple notification is morally necessary are complex, however. It is generally agreed that respect for privacy requires openness about the existence of databases containing personal information and the uses made of the data, regardless of whether explicit consent is required for every use. In practice, however, principles of fair information practice—such as those that form the basis for the federal Privacy Act of

1974—are surprisingly ambiguous.[2] There appears to be acceptance, if not explicit acknowledgment, of the fact that personal data will be collected, some of it without subjects' explicit, freely given consent.

In clarifying the obligations of data managers to inform and seek consent from data subjects, it is useful to think in terms of three levels. There should be a base level of education about the role of data and research in making society run well. The goal is to make sure people understand that information about them is collected and used, with confidentiality safeguards, as a matter of routine, and that it is their civic responsibility to accede to this in exchange for medical progress, an effective educational system, protection of their civil rights, and so on. This level implies a category of "ordinary" research uses producing substantial social benefits with a low risk of harmful disclosure. For research in this category, there is no obligation to notify data subjects about each use or to seek explicit consent, although there should be a way for subjects to learn what research is being done with the data if they wish to do so.

The second level pertains to research uses that differ substantially from the routine, making it reasonable to notify data subjects and provide justification for the use. For example, a new hypothesis about the cause of an illness might lead to new analysis of old data that promises significant benefits with little risk. Alternatively, a significant change in the underlying benefit/cost picture might lead to new kinds of research or new interest in variables not previously examined.

The third level pertains to uses for which explicit, informed consent is required. A research use might fall into this category because the potential for harm is significantly greater relative to societal benefits, or because the degree of actual or perceived harm varies substantially across individuals. Uses for private rather than social benefit also fall into this category.

The above categories suggest a way to think about informing potential participants of current and future uses of the data. The immediate research goals could be explained, and the participants could be informed that the data

[2]These principles—as set forth by Alan Westin and cited by George Duncan in Chapman (1997:336)—are as follows: (1) there must be no secret personal data record-keeping system; (2) there must be a way for individuals to discover what personal information is recorded and how it is used; (3) there must be a way for individuals to prevent information about them that was obtained for one purpose from being used or made available for other purposes without their consent; (4) there must be a way for individuals to correct or amend a record of information about themselves; and (5) an organization creating, maintaining, using, or disseminating records of identifiable personal data must ensure the reliability of the data for their intended use and must take reasonable precautions to prevent misuses of the data.

might also be used for routine socially beneficial research in the future, with confidentiality safeguards and without further notification. An important issue to be addressed is where the linkage of survey data (especially longitudinal data) to administrative records falls along this spectrum. Can confidentiality safeguards and accountability be effective enough, and the risk of harm to individuals low enough, to allow most such linkages to be considered "routine" or "ordinary" research uses in the sense discussed above?

Concluding Remarks

Baily concluded her presentation by offering the following assessment:

- There is consensus on the existence of a right to informational privacy, but not on the extent to which policies should go to protect that right or on how to implement such policies in practice. In a pluralistic society, translating a moral right into enforceable policy is a political problem; inevitably, no one is entirely satisfied with the result.
- It is easier to reach an agreement most people can live with if people understand that the goal is a practical compromise among competing moral visions, not the triumph of their own point of view. Also, the process of achieving compromise must both be and be perceived to be one in which there is ongoing democratic accountability for what happens as a result.
- Finally, it is easier to agree on change when the existing situation is unsatisfactory to nearly everyone, so that there is much to gain from an improved system. At present, personal privacy, even in highly sensitive areas such as medical information, is far less protected than most people realize. The opportunity exists to both improve safeguards on the use of data and increase access to data for socially useful research if the right policies are instituted.

RECENT AND PROSPECTIVE LEGISLATION

During her presentation, Mary Baily asserted that implementation of effective access and confidentiality policies requires carefully constructed social mechanisms. Paramount among these is a consistent national legal framework, designed to indicate to both data producers and users what standards are appropriate and to aid in imposing sanctions for misuse. The need for such a framework, which does not now exist, is a recurrent theme in the literature on privacy and data use. Confirming Baily's assessment, Donna Eden provided an overview of recent and prospective legislative developments. Speaking primarily about health data, Eden observed that coordinated federal privacy legislation to protect health records does not yet exist. This is somewhat of a surprise, given that Congress has been working to enact

comprehensive privacy legislation in health and other areas for many years. In fact, in the Health Insurance Portability and Accountability Act (HIPAA) of 1996, Congress set an August 1999 deadline for the enactment of privacy legislation to protect health records, and provided that if that deadline were not met, the Secretary of Health and Human Services would be required to issue regulations to protect the privacy of certain health data.

Consensus on privacy legislation has not been forthcoming; instead, protections are based on a piecemeal system that originates from both federal and state levels. The primary laws governing data access are the Freedom of Information Act, the Privacy Act, privacy and confidentiality laws specific to individual agencies or purposes, and state laws. The wide variability in statutes governing access to administrative records nationwide makes it difficult for researchers and others to understand applicable rules. Laws dictate what data can be collected, the linkages that can be made, and the protections that are required for data use.

Bills are pending that would regulate federal statutes applying to large databases, including proposals to mandate copyright protection for some. The first major component of legislation pertaining to health information is directed toward encouraging adoption of standards, particularly in the electronic context, to replace an array of currently existing formats. HIPAA requires the Secretary of Health and Human Services to adopt standards developed by organizations accredited by the American National Standards Institute whenever possible. The legislation requires that all health care information transmitted by health plans, clearinghouses, and those providers who conduct business using electronic transactions comply with these standards within 24 months (or, for small health plans, 36 months) of their formal adoption by the Secretary.

Standardization is slated to include the creation of unique national identifiers for health care providers, health plans, employers, and individual patients that will technically facilitate linkages across sources. This practice will ensure that different variables—for example, for diagnosis and procedure data—are coded identically across data sets. The standard for individual patient identifiers is on hold until comprehensive privacy protections are in place. One of the major gaps in current HIPAA requirements is that the standards will not apply to exactly the same data when those data are held by employers, some insurers, and government agencies.

HIPAA also requires the adoption of standards for the security of information transmitted or maintained electronically, and for electronic signatures used in standard health care transactions. The Department of Health and Human Services will issue compliance and enforcement requirements to provide assurance that, if information is misused, there will be redress. This move toward standardization will clearly impact researchers' data collection efforts, particularly with regard to the types of data that can be linked. Stan-

dardization will make reading of medical records much easier, and should significantly simplify the mechanics of data matching and analysis.

Because Congress failed to enact comprehensive health privacy legislation, the Secretary of Health and Human Services is now required to issue privacy regulations. These regulations will be based on recommendations for privacy legislation prepared by the Secretary for Congress in 1997. Statutory language requires that the privacy regulations address (1) recognition of the rights of individual subjects, (2) procedures designed to enforce those rights, and (3) the uses and disclosures permitted. There is as yet no clear administrative mechanism or funding to activate the provisions, which may create additional delays. Moreover, since the new privacy standards will be issued as regulations, not as a statute, state privacy laws will not be preempted; all existing federal statutes, such as the Privacy Act and the Public Health Act, also remain in place.

At this point, it is unclear how effective these efforts to control the uses of data and protect individual confidentiality will be. Uncertainty about effects on access and the potential for disclosure will persist until legislation is formulated. Eden reviewed several current circuit court cases that indicate possible directions the privacy legislation may take.

Under HIPAA, individual data subjects have some limited rights and protections. The statute provides both criminal and civil penalties for disclosure of data in violation of the various standards, along with very limited monetary penalties. Many of the workshop participants expressed the view that these rules need to be strengthened and criminal penalties stiffened. The Freedom of Information Act gives the public certain rights to data held by the federal government. Federal, state, and local governments authorize themselves to use data for basic operations and particular social purposes. Researchers and data collectors have very few explicit legal rights to data.

In conclusion, Eden offered her assessment of practical resolutions for current issues. Given that the prospects for passage of comprehensive privacy laws appear to be remote, she envisions a continued piecemeal approach to legislation. The potential for practical solutions may be greatest in the areas of copyright law and contracts. The Internet offers a wide range of possibilities for creating instant contracts and user agreements. Eden predicted a broad expansion in the use of click-on and other technology-facilitated agreements, most of which offer the promise of enforceability through existing contract law. Additionally, these mechanisms do not require special recognition by Congress of a separate private right of action. Agreement violators can be taken directly to state or, in certain cases, federal court. In some states, data subjects have the right to take legal action against a secondary user or licensee if terms and conditions are violated.

As noted earlier, a number of participants expressed skepticism about the ability of privacy laws to keep pace with technology and to effectively target

individuals and groups that pose the greatest risk to data security. The protection offered by HIPAA, as well as by regulations that will be adopted under this legislation, is limited by its restriction to health plans, clearinghouses, and those providers who conduct business electronically. Eden noted that the Secretary of Health and Human Services is on record as supporting the need for comprehensive federal legislation in this area.

ROLE OF INSTITUTIONAL REVIEW BOARDS

The IRB is the most direct regulatory link between researcher and research data. Thomas Puglisi presented a paper coauthored by Jeffery Cohen, also of OPRR, titled "Human Subject Protections in Research Utilizing Data Files," providing an overview of IRB procedures, limitations, and prospects.

OPRR is in charge of enforcing federal policy for the protection of human research subjects. Nearly every executive branch agency of the federal government that supports human-subject research is a signatory to this policy, which makes researchers subject to a common set of regulations. The IRB process and the requirement for informed consent are the two core protections provided to individuals by these regulations.

Federally funded researchers become subject to the IRB process as soon as they access potentially identifiable private information about living individuals. Information is considered private if an individual could reasonably expect that it would not be used for purposes other than that for which it was provided. The IRB is charged with judging whether risks to data subjects are "reasonable" in relation to the anticipated benefits of data release and whether the risks are minimized by a sound research design, as well as with ensuring that informed consent is acquired and that confidentiality protections underlying data dissemination are adequate.

With regard to federally mandated informed consent, regulations require that subjects be notified about the degree to which the confidentiality of their records will be maintained. Meeting this requirement is not problematic for certain specific-use data sets, but it is typically not possible to provide accurate notification about information that will be linked or otherwise be incorporated in larger data sets. It can be difficult to predict the level of security for data that are extended beyond their original use, as is often the case with clinical or administrative data not initially collected as part of a defined research project.

Similarly, it is frequently impossible to acquire informed consent for research of the type discussed in the workshop session on case studies. In survey-based social science research, IRBs may waive the consent requirement if they find that three conditions are met: (1) the risk to subjects is minimal, (2) use of the information for research will not adversely impact the rights and welfare of the subjects, and (3) it would not be practicable to

obtain informed consent. The second and third conditions usually pose little difficulty; the first, however, requires more judgment. As noted above, there is no set of standards an IRB can apply in deciding whether a given application should be considered as involving minimal risk and whether the confidentiality protections in place are adequate. In fact, it probably is not possible to establish a universal standard, given the case-by-case variation in key parameters.

With regard to data access policy, Puglisi noted that the inherent challenge of the IRB mandate is in interpreting a standard, since the regulations do not establish one. Judgments must be made about what type of data can be released, and in what form, for each research project. The IRB must weigh the risk of information disclosure and the potential ramifications associated with it against the anticipated benefits to research and policy, and then decide on an appropriate level of protection.

During this process, IRB staff must first evaluate the nature and sensitivity of the data. Could disclosure put data subjects at risk of criminal or civil liability? Could it be damaging to financial standing, employability, insurability, or the reputation of an individual or group? Obviously, the more sensitive the information is, the more stringent the protections must be. Risk is a function of the level of identifiability (and data protection) and the sensitivity of the data.

The probability of disclosure and subsequent damage often appears trivial from the perspective of the researcher, but disclosure of private information does occur. Puglisi provided several real-world examples of inadvertent disclosure—one involving a case study and one involving exposure of notes at a professional meeting. He also noted instances in which data were released by investigators to news reporters and to congressional committees. The risks are real, but difficult to quantify.

Frequently, IRBs are advocated in proposed legislation as the vehicle for resolving data access and confidentiality tensions. There are hurdles to overcome, however, if IRBs are to serve this purpose optimally. IRB personnel are often no more competent than other groups to make unstructured decisions. Federal regulations do not require that IRB members have training in statistical disclosure limitation techniques and other methods of protecting confidentiality. Moreover, the federal requirement that IRBs have the scientific expertise to judge research they review is not being met uniformly. These factors make it impossible to carry out cost/benefit assessment, which is, strictly speaking, a violation of federal regulation.

Another problem pointed out by Puglisi is that IRB standards tend to become increasingly restrictive as more procedural constraints are adopted; staff who are inadequately trained may have incentives to err on the safe side. With no increase in the expertise of IRB personnel, this trend is likely to continue. IRB administrators have an obligation to acquire (either internally

or through outside consultants) enough professional expertise to combat fears that an institution will be held liable for a mistake. Communication between researchers and IRBs is also key to improving knowledge levels and enhancing the performance of IRBs.

Another important aspect of efficient IRB utilization involves matching a research project with the appropriate IRB. In most instances, a researcher's local IRB will not be the one best suited to evaluate research value and data security risks. The model Puglisi recommends would require that data access proposals be reviewed at the location where data are maintained. The host IRB is in the best position to balance research potential against confidentiality risks. This approach would allow the host IRB to play an educational role as well, which is appropriate since its staff should be most knowledgeable about specific data characteristics, research applications, and conditions under which data should be shared. In acting as the responsible gatekeeper, the host IRB could provide information to local IRBs that would help streamline the approval process. Researchers could submit judgments from the overseeing IRB, demonstrating to the local IRB that a knowledgeable, respected body has approved confidentiality protections. Robert Willis noted that this is essentially the model that has been implemented successfully at HRS. Given the expanding role of data enclaves, it is likely that, along with auditing procedures developed by the National Center for Education Statistics, the Bureau of Labor Statistics, and the National Science Foundation, IRBs will for the foreseeable future continue to be the central mechanism for monitoring researchers' access to data.

4
Alternative Approaches for Limiting Disclosure Risks and Facilitating Data Access

Presentations throughout the workshop confronted aspects of the restricted data, restricted access debate. Data alteration allows for broader dissemination, but may affect researchers' confidence in their modeling output and even the types of models that can be constructed. Restricting access may create inconveniences and limit the pool of researchers that can use the data, but generally permits access to greater data detail. This chapter reviews the presentations and discussions addressing these two approaches, highlighting their advantages and disadvantages. It also summarizes the discussion of potential technical solutions and the use of legal sanctions to modify the behavior of individuals with access to the data.

DATA ALTERATION

There are both technical and statistical solutions for protecting data security. Several workshop participants argued that these solutions need to be blended with the substantive knowledge of researchers to solve disclosure problems in a way that satisfies all interested communities. This section reviews the discussion of statistical approaches, for which a presentation by Arthur Kennickell, titled "Multiple Imputation in the Survey of Consumer Finances," was the centerpiece. Technical approaches are discussed later in this chapter.

Participants representing the research community expressed frustration with some of the standard perturbation methods employed by the large longitudinal surveys. Finis Welch of Texas A&M University articulated several

concerns. He argued that the introduction of noise to variables can, in certain instances, create major headaches for the modeler. For instance, adding noise to a field that will be used as a dependent variable in models may be acceptable as long as the expected value of the disturbance is zero, and hence efficiency of estimates is preserved. When dispersion is added to fields that will be used as explanatory variables, however, expected errors tend to be correlated with variable values. Welch believes priority should be given to developing perturbation methods that, if invoked, will preserve the key statistical properties of a data set. He made some specific recommendations as well. First, he advocated top-coding at fixed quantile levels, over time, rather than at absolute values; when the percentage of records top-coded changes, serious problems arise in longitudinal or panel modeling contexts. Welch also would like to see scrambling—as opposed to truncation—of "sensitive" data above the top-coded cutoff points to maintain full distribution, but eliminate knowledge of where an individual record fits into the distribution.

Kennickell's use of a multiple imputation technique offers a more sophisticated form of data perturbation, one that could potentially improve data security (and, hence, allow greater accessibility) without seriously compromising modeling utility. Several workshop participants expressed support for the idea of exploring this type of approach to gain a clearer idea of how models might perform using imputed data and to assess the promise of the technique in terms of data protection. There is now a large multiple imputation apparatus in the Survey of Consumer Finances (SCF); Kennickell has shown it can be done. What remains to be seen is how effective and how useful the technique will be. Kennickell's research is moving toward answering that question.

The SCF is conducted by the Federal Reserve Board, with survey information collected by the National Opinion Research Center at the University of Chicago. The data include sensitive and detailed information about respondents' assets and liabilities, as well as extensive demographic and geographic information. The survey, which oversamples wealthy households, is derived from statistical records based on tax returns maintained by the Statistics and Income Division of the Internal Revenue Service. To gain access to this information, the Fed agrees to a disclosure review similar to that for the public Statistics of Income files.

Because the SCF is subject to legal constraints on data release, and because it contains sensitive information on a sample that includes a high-wealth population, the survey is a logical candidate for the multiple imputation experiment as a means of disclosure limitation. Because missing data have always been an important problem in the SCF, substantial resources have been devoted to the construction of an imputation framework that can be used to simulate data to replace those originally reported.

For the public-release version of the SCF, the survey applies standard

perturbation and data limitation techniques—rounding, collapsing categories, providing geographic detail only at the Fed division levels, truncating negative values, withholding variables, and a variety of more minor changes that cannot be disclosed. In addition, the full final internal version of the data is used to estimate models that are, in turn, used to simulate data for a subsample of observations in the public version of the data set.

Kennickell's multiple imputation system deals with all variables, imputing them one at a time. It is iterative and generates, as the name implies, many imputations. The models, which are different for binary, continuous, and categorical variables, involve essentially relaxing data around reported values. The method requires a full probability specification at the outset; the notion behind the multiple imputation is to then sample from the full posterior distribution. It is this sampling that generates the variability needed for disclosure limitation.

Kennickell described the application of multiple imputation for the SCF as a type of structured blurring: "a set of cases that are identified as unusual plus another set of random cases are selected, and selected variables within those cases are imputed subject to a range constraint (unspecified to the public), but they are constrained to come out somewhere in a broad neighborhood around the original values." The knowledge of which data values have been intentionally altered is also partially disguised. The cumulative effect of the process is to decrease a user's confidence that any given record represents an actual original participant.

The method is computationally intensive, but Kennickell described it only "as a modest step in the direction of generating fully simulated data." He argued that it is possible to simulate data that do a good job of reproducing all simple statistics and the distributional characteristics of the original reported data. The extent to which imputed data will be able to provide a satisfactory basis for the leading-edge research required for fully informed policy is not yet clear. It is not known how imputation affects error structures of complicated models; what sampling error means in a fully simulated data set; what happens to complex relationships among variables; and, more generally, how researchers will interpret modeling results. One way to begin addressing these questions is to create synthetic versions of existing data sets with known nonlinear relationships or complex interactions and see whether they could have been detected with the simulations. Many of the workshop participants agreed that these performance questions require serious attention and that the answers will ultimately determine the success of imputation methods. Quantitative assessments of the extent to which disclosure risks can be reduced using these methods are also needed.

At this point, social science researchers are skeptical about the accuracy of analyses not based on "original" data. Richard Suzman of the National Institute on Aging (NIA) added that all leading researchers currently supported by

NIA are opposed to the imposition of synthetic data.[1] Finis Welch and Suzman each noted that the value of synthetic data sets in longitudinal research is unproven; with the exception of the 1983–1989 panel, the SCF is cross-sectional. While complex longitudinal data increase disclosure risks, it is also more difficult to preserve key relationships and interactions among variables when this type of data is altered. Perturbation, therefore, may be more damaging to analyses that rely on longitudinal data than to those that rely on cross-sectional data.

These criticisms notwithstanding, Stephen Fienberg of Carnegie-Mellon University considers Kennickell's work a major success in the use of modern statistical methods and disclosure limitation research. Fienberg made the point that all data sets are approximations of the real data for a group of individuals. Rarely is a sample exactly representative of the group about which researchers are attempting to draw statistical inferences; rather, it represents those for whom information is available. Even a population data set is not perfect, given coding and keying errors, missing imputed data, and the like. Fienberg finds the argument that a perturbed data set is not useful for intricate analysis not altogether compelling. Yet researchers are more critical of controlled modifications to the data and the introduction of structured statistical noise than of sampling noise.

Thus two clear perspectives emerged among workshop participants. On one side are those who believe that, in addition to its role in statistical disclosure limitation, replacing real samples with records created from posterior distributions offers great potential in terms of maintaining fidelity to the original data goal (as opposed to the original data). On the other side are researchers who are concerned that synthetic data do not fit the model used by top researchers as they actually work with the data. Their position is that, in addition to delaying data release, imputation programs blur data in ways that create inaccuracies, such as those described earlier. Suzman expressed the need for the National Institutes of Health and others to advance empirical research that would address these issues. As these methods are advanced, it may become possible to provide researchers with clearer explanations of how imputation impacts the data. Moreover, data management programs may be developed that offer the option of choosing between using an altered public data set and submitting to additional safeguards to gain access to raw data.

RESTRICTED ACCESS

Presentations during Session I illustrated the benefits that can be derived from studies using complex research data files—that is, microdata files with

[1]Suzman did acknowledge a role for synthetic data in creating test data sets on which researchers could perform initial runs, thus reducing time spent in data enclaves.

longitudinal data, contextual information for small areas, or linked administrative data. Each research example cited was carried out under restricted data access arrangements. None could have been undertaken solely using microdata files that are available to the general public with no restrictions. Many of the analyses required the use of files that link survey data with administrative record information for the same persons.

A survey of organizations that released complex research data files as public-use files was described during Session III by Alice Robbin, Indiana University (for further details about this survey, see Chapter 5). Files with linked administrative data were seldom released in that format, primarily because of concerns that users with independent access to the administrative source files might be able to re-identify persons whose records were included in the research file. Restricted-access files were distinguished from public-use files by the inclusion of more detailed information on the geographic location of sample persons and contextual data, such as median income classes and poverty rates, for the communities where they lived.

While usually applied in a way that preserves basic statistics, masking procedures used to reduce disclosure risks associated with public-use files may introduce substantial biases when more complex methods of analysis are applied to the data. Therefore, arrangements for providing special or restricted access are often used to satisfy the needs of users for whom the available public-use data files are insufficient. Several such arrangements have been developed in recent years; primary among these are (1) use of the data by users at their own work sites, subject to various restrictions and conditions (commonly referred to as licensing); (2) controlled access at sites (often called research data centers) established for the purpose by custodians of the data; and (3) controlled remote access, in which users submit their analytical programs electronically to the custodian, who runs them and reviews the outputs for disclosure risk prior to transmission to the users.[2] The next chapter reviews workshop presentations that described current and planned restricted-access arrangements at various agencies.

ADVANTAGES AND DISADVANTAGES OF DATA ALTERATION VERSUS RESTRICTED ACCESS

The most frequently cited advantage of data alteration, as opposed to restricted access, is that it facilitates broader public release and simpler user acquisition. Steven Fienberg articulated the advantage of data perturbation

[2]There are other possible arrangements, such as the release of encrypted microdata on CD-ROMs with built-in analytical software, but these methods are not widely used at present and were not discussed at the workshop.

concisely: "These methods (particularly if they can be developed to produce more acceptable statistical properties than people are willing to admit) address a very compelling public need, which is the sharing of data collected at public expense that are a public good and would otherwise not be broadly accessed."[3] Proponents argue that sophisticated perturbation methods offer one of the few tools that may help meet simultaneously the need of researchers to access better data and the need to protect respondents who supply information.

The primary disadvantage of data alteration generally and advanced perturbation specifically is researchers' decreased confidence in modeling output; use of such data is believed by some to limit modeling flexibility as well. Researchers at the workshop expressed concern that data alteration inhibits complex modeling, particularly when the relationships that have the greatest policy relevance are nonlinear or when causal modeling requires correct ordering of temporal events. For example, it may be difficult to accurately model real-world retirement behavior, which is thought to be driven by jumps in eligibility and benefit rules faced by workers, if blurring techniques are used to smooth such spikes in the data. Moreover, even if key statistical properties are preserved, researchers must be convinced that this is the case before they will use the data; they must also learn how to interpret and report the results of models estimated from altered data. These are real costs associated with data perturbation.

The challenge for proponents of data imputation approaches is to determine how accurately relationships among data fields can be preserved and to communicate their findings to researchers. The extent to which this challenge can be met is, as of now, uncertain. Robert Boruch articulated a strategy for addressing this need. He suggested that it is important to monitor the performance of increasingly complex models, when data used in estimation procedures are altered in various ways, by building a knowledge base of calibration experiments.

The advantage of restricted access is that those granted permission have fuller access to primary data. On the other hand, costs are incurred in enforcing access rules and in operating data enclaves and remote programs. Restricted access arrangements also impose an operational burden on researchers. These operational costs can be significant; for instance, Kennickell reported that the Fed does not have the research budget to establish data centers or even licensing agreements. While their multiple imputation program is a major undertaking, it is a less costly method of providing broad

[3]Ivan Felligi and others believe that if data linkage continues to increase, it may not be possible to safely offer public release files at all. While this view may be extreme, it does point to the tradeoff: given more linking possibilities and richer native data, more restriction is required to hold disclosure risk constant.

access to researchers. Kennickell also believes that, among users of the SCF, unrestricted access to the data is a higher priority than is access to unrestricted data.

While researchers at the workshop did express a general preference for limited access to full data, as opposed to public access to limited data, they also noted that the on-site requirement can be burdensome. Thus, they voiced enthusiasm for the idea of developing flexible remote access systems. Researchers also want assurances that restricted access data sets at centers would not replace the types of data now publicly available. Most of the researchers indicated a preference for the licensing option, which is viewed as least burdensome (since they plan to follow access rules). Agency representatives noted that the sanctions in existing laws could be strengthened (see the next section). However, it is impossible to ensure that all users are familiar with the laws, and federal agencies are ultimately responsible for data safety. Licensing is effective because it transfers a portion of that responsibility to users, allowing agencies greater latitude in dissemination (see also the discussion of licensing in Chapter 5).

Ultimately, if different types of users can be identified reliably, appropriate levels of access can be established for each. Researchers are probably willing, for selected studies, to go through the required steps to use less-altered data under restricted access arrangements. In some cases, the existence of legal penalties for misuse will provide a sufficient deterrent, and access to full raw data may be allowed. Participants voiced the view that a one-size-fits-all approach to data access is unsatisfactory, since it would likely produce data of insufficient detail for cutting-edge research while perhaps unnecessarily disclosing information not needed for more general research. Similarly, marketers or the general public who want fast Web access likely cannot be granted the same access to the highest-quality data as those who undergo security precautions.

Participants were also optimistic about the ability to use technology to obtain a proper balance between confidentiality and accessibility. Latanya Sweeney and others described evolving approaches that may eventually advance confidentiality protection within both data alteration and restricted access frameworks. For example, there may be ways to improve remote access using rapidly evolving net-based foundations that would allow researchers to run interactive programs externally (see also the discussion of remote access in Chapter 5). More sophisticated linking may also be possible, particularly if methods can be developed to monitor the combinations of variables regularly used by researchers. Once a clear research need to link certain variables or data sources has been established, it may be safer to link at the source instead of having copies of both data sets go out in their entirety each time a researcher needs to make the link. Similar approaches may enhance the ability to establish joint data centers or centers with multiple sources of data.

ROLE OF LEGAL SANCTIONS

Stricter legislative protections offer another potentially efficient means of improving confidentiality—efficient because the probability of disclosure can be decreased without imposing costs on rule-abiding researchers. Indeed, several participants, including Richard Suzman, suggested that perhaps this method of data protection should be given the highest priority. These participants cited a recommendation from the report *Private Lives and Public Policies* that there should be "legal sanctions for all users, both external and agency employees, who violate requirements to maintain the confidentiality of data" (National Research Council and Social Science Research Council, 1993:7) and added that existing penalties should be stiffened.[4]

J. Michael Dean pointed out that the potential for unintended disclosure exists at the data source as well as at the user stage. A primary reason that agencies are able to maintain confidentiality is their ability to impose a high cost penalty for misbehavior (violators can lose their jobs). Dean argued that this regulation should be expanded across agencies and institutional lines, thereby creating more linking opportunities; in other words, the regulatory approach to native databases could be extended to linked data. Again, such approaches are generally applauded by researchers, who prefer regulation of the people using databases over alteration of the databases themselves.

Presenters from the HRS noted that, in part because of a lack of confidence in the adequacy of sanctions, the funding agency (NIA) demands that the University of Michigan provide linked data only to individuals working under federal grants so that disregard for the confidentiality guidelines will be subject to federal rules. The situation is different for agencies that have a licensing mechanism tied to the data, which allows for more options. Other agencies must operate purely within the realm of the Privacy Act. Many participants believe that increased harmonization of the legal framework is needed, if for no other reason than to allow researchers to know roughly what is expected without that expectation shifting from context to context.

[4]Donna Eden pointed out that there is considerable room for increasing penalties. For instance, in HIPAA there exists no private right of action for subjects. The act sets forth broad criminal prohibition, but only minor criminal sanctions for disclosure of data in violation of regulations. A $100 civil penalty is unlikely to be effective against corporate or even most individual abuses. Also, it should be noted that the size of a penalty is of limited importance if it is rarely imposed.

5

Current Agency and Organization Practices

This chapter reviews presentations that described current practices aimed at preserving the confidentiality of microdata. Practices related to release of public-use files and restriction of data access are reviewed in turn.

RELEASE OF PUBLIC-USE FILES

More than 20 years ago the Federal Committee on Statistical Methodology (1978) recommended that all federal agencies releasing statistical information formulate and apply policies and procedures designed to avoid unacceptable disclosures. In releasing public-use files—typically made available with no restrictions other than, in some cases, imposition of a user fee—agencies generally comply with this recommendation through the use of various forms of statistical disclosure limitation. Two workshop presentations described policies and procedures used for the release of microdata. Alice Robbin provided an overview of results from a survey on the statistical disclosure limitation (SDL) practices used by government agencies and research organizations that distribute public-use microdata files with longitudinal, linked administrative, or contextual data for small areas. Alvan Zarate of the National Center for Health Statistics (NCHS) presented an overview of the Interagency Confidentiality and Data Access Group's (ICDAG) Checklist on Disclosure Potential of Proposed Data Releases, developed to help ensure that

principal safeguards are in place when electronic data files are released for public use.[1]

Statistical Disclosure Limitation Practices

Alice Robbin, Thomas Jabine, and Heather Koball conducted a survey of organizations that produce and distribute complex research microdata files. The survey was intended to contribute empirical evidence on how knowledge about SDL procedures has been applied by these organizations in the production of public-use microdata. Information was gathered on the types of microdata that are released publicly, current SDL practices applied to public-use data, and organizations' demarcation between public-use and restricted access data. Several themes emerged from this information, including (1) the extent of variation in SDL practices across organizations, (2) special risks to data confidentiality, and (3) the tension between the data needs of researchers and data confidentiality.

Variation in Organizational SDL Practices. Survey respondents conveyed familiarity with the broad issue of data confidentiality. All respondents knew that direct identifiers of respondents should not be released and expressed concern about protecting respondents' identities. Furthermore, because of concerns about data confidentiality, few organizations release public-use geographic/contextual data for small areas. Similarly, linked administrative data are generally confined to a restricted access format.

On the other hand, the survey revealed considerable variation across organizations in terms of knowledge about SDL techniques. This variation is a function of the extent of practitioners' knowledge about deductive disclosure, the type of organization, and the timing of decisions related to release of public-use files. Some respondents appeared to be unfamiliar with terminology and concepts associated with data confidentiality, while others were well versed in these matters.[2] The treatment of special "at-risk" variables, such as age and income, varies widely by organization.

Most organizations appear to base their SDL decisions for public-use longitudinal files on a cross-sectional model. That is, they assess the risks of disclosure for a given cross section, with little consideration of longitudinal effects. One factor that may contribute to the relatively liberal policies ap-

[1]The ICDAG was recently renamed the Committee on Data Access and Confidentiality.

[2]This generalization—that there is a wide variety in knowledge and practice of SDL techniques—was corroborated by Erik Austin of the Inter-University Consortium for Political and Social Research. Austin has been involved with this issue for three decades; his organization has examined thousands of files and reviewed SDL plans so that data can be released publicly.

plied to longitudinal data is the fact that follow-up data are often released several years after earlier panels. Decisions about previous data releases may or may not play a role in decisions pertaining to the release of longitudinal files. A second factor that appears to influence decisions about release policies for longitudinal data is knowledge of user preferences. Staff are sensitive to the fact that longitudinal data are deemed particularly useful when the data contain the same variables over time.

The survey responses indicated greater variation among the standards of universities than among those of government agencies.[3] Government agencies, particularly the Bureau of the Census and NCHS, have standards on which they base decisions about release of public-use data and SDL techniques. Many of the Census Bureau's standards have had a significant influence on other organizations that distribute microdata. The Census Bureau also has a Disclosure Review Board, which reviews microdata sets prior to release. NCHS has an IRB; a data confidentiality committee; and a data confidentiality officer, who makes final decisions about SDL techniques for public-use data.

Special Risks. In general, issues related to deductive disclosure have been brought to the attention of organizations only in recent years; as a result, the SDL techniques applied to microdata sets have changed. Moreover, older longitudinal microdata sets are at particular risk for deductive disclosure as they contain more detailed information about respondents than would be released under current practices. The data sets also follow respondents over long periods of time, so they contain a wealth of detailed information, some of which is revealed only as a result of the longitudinal structure. The combination of changing SDL standards and the compilation of data on respondents over time may make older longitudinal data sets particularly vulnerable. At the same time, it is often the longitudinal structure that makes these microdata sets particularly useful to researchers.

Needs of Researchers Versus Data Confidentiality. Organizations appear to be keenly aware that their microdata sets are more useful if they provide greater detail. One respondent stated that his organization increased data availability because of the demands of users. The organization increased the level at which income was top-coded in response to complaints about the lack of data detail. Other respondents indicated that their decisions to release

[3]Austin agreed with this point as well, noting that academic data producers typically have less knowledge about SDL techniques than agency counterparts. He recommended establishing venues for communicating SDL techniques more effectively.

data are based, in part, on providing as much data as possible to researchers. They "advertise" public-use data on Web sites by highlighting the data's detailed and longitudinal nature. This advertising is designed to increase use of the data and to demonstrate to program funders that the data distribution function is being performed. Ease of access for researchers was also cited by respondents as an aspect of the tradeoff between data utility and the need for confidentiality. The Internet makes access to public-use data increasingly easy; this ease of access facilitates the research process, while also increasing the risk of deductive disclosure.

The survey findings are generally consistent with the body of empirical evidence that has accumulated on organizational decision making. They reveal that members of organizations are sensitive to the external environment, and that structural and political factors influence decisions related to release of public-use data. The findings also reveal that structures within organizations can be fragmented. Organizational units are governed by different policies, some of which are contradictory. Further, staff turnover contributes to loss of institutional memory, and historical records about data release decisions are often not maintained to compensate. Survey design, data collection, and preparation of public-use files may be overseen by different units of the same organization or by different organizations, and this can affect information flows. Management control also differs across organizations and units; some project managers are familiar with the nuts and bolts of data release decisions, while others are not.

Agency staffs are sensitive to the consequences of releasing data that could identify individuals, particularly in light of legislative initiatives responding to public concerns about confidentiality. One agency respondent noted that his organization's dissemination of a public-use file of survey data had ceased because of a recently enacted statute that was interpreted as preventing distribution of the data. Regardless of whether the statute was properly interpreted, what is important is that a perceived threat from the external environment resulted in the restriction of important data. Furthermore, policies governing data access and confidentiality are subject to change. Institutional interpretations of these policies influence decisions about the release of public-use microdata, how data will be prepared, and the conditions under which access will be permitted.

Robbin, Jabine, and Koball offered a number of policy recommendations regarding the release of public-use files. These recommendations were directed to producers, distributors, and analysts of large-scale microdata files, as well as to funding agencies and project managers.

Communicating and Educating About Statistical Disclosure Risk and Limitation Procedures. Appropriate policy and policy compliance require improved communication about current research on disclosure risk, as well

as education of professionals about good practice. Research into disclosure risk has been conducted for more than 20 years. Statistical agencies have published documents analyzing the risk and providing guidelines for good practice. Peer-reviewed journals have published articles on the subject. The American Statistical Association's Committee on Privacy and Confidentiality has prepared informational materials that are available at its Web site, and ICDAG has disseminated its Checklist on Disclosure Potential of Proposed Data Releases (discussed below). Yet agency and organization staffs appear inadequately aware of current SDL practices. The result is that, in some cases, statutory confidentiality requirements go unmet, while in others, data are overly restricted. To facilitate dissemination of information about good SDL practices and standards, Robbin, Jabine, and Koball recommended producing and circulating a bibliography of key publications that describe evaluative deductive disclosure methods. The American Statistical Association's Committee on Privacy and Confidentiality has prepared bibliographies on the subject, and the committee's work can serve as the basis for selecting additional informational resources.

Documents on the subject need to be available at a basic technical level to be useful for staff who may have less statistical expertise but are on the front lines of data production. Responsibility for ensuring that data organizations employ good SDL practices should not lie only with data processing staff (the survey results indicated that, in many cases, programmers were given nearly sole responsibility for preparing public-use files). Internal review units should be available to evaluate proposed releases of microdata files; outside of government, IRBs and other groups can incorporate experts on deductive disclosure.

The survey revealed that a number of respondents, while familiar with the general issues of data confidentiality, were not knowledgeable about disclosure risk and SDL techniques. Thus, there exists a clear opportunity to achieve advances through further education. Workshops and panels at annual professional meetings offer an appropriate forum for launching such efforts. Interactive environments such as IRBs and data centers represent additional ongoing opportunities.

Institutionalizing Communication to Improve SDL Practices. A general set of rules governing data release is not possible because virtually every proposed release is unique in some way, even within the same agency and program. It is important to obtain expertise on the subject in the initial planning stages of statistical programs and research projects, and then later during evaluation and testing to prepare public-use files. Improved documentation is also an essential aspect of communicating survey objectives and methods. Detailed documentation can minimize the loss of institutional memory that results from staff turnover and other factors.

Data User Participation in Data Release. There are multiple approaches to developing good SDL practices. Data users have important knowledge to contribute during the early stages of organizational decision making on the practices to be employed.

Checklist on Disclosure Potential of Proposed Data Releases

The introduction to the Checklist (Interagency Confidentiality and Data Access Group, 1999:1) clearly describes its function:

> Federal statistical agencies and their contractors often collect data from persons, businesses, or other entities under a pledge of confidentiality. Before disseminating the results as either public-use microdata files or tables, these agencies should apply statistical methods to protect the confidentiality of the information they collect. . . . [The Checklist] is one tool that can assist agencies in reviewing disclosure-limited data products. This Checklist is intended primarily for use in the development of public-use data products. . . . The Checklist consists of a series of questions that are designed to assist an agency's Disclosure Review Board to determine the suitability of releasing either public-use microdata files of tables from data collected from individuals and/or organizations under an assurance of confidentiality.

Zarate's overview of the Checklist was presented within the broader theme of how agencies operate "caught between the twin imperatives of making usable data available, while also protecting the confidentiality of respondents." Zarate noted that, while disclosures can occur, it is not justifiable to withhold valuable data for legitimate research purposes.

Zarate explained that the Office of Management and Budget's Statistical and Policy Office helped form the ICDAG in 1991 to coordinate and promote research on the use of statistical disclosure methods and to catalog related developments at agencies and among academic researchers. The Checklist is intended for use in risk assessment by agency statisticians in charge of data release. Although it is not a formal regulatory document, its widespread visibility should motivate a closer look at organizational methods.

Though the Checklist does offer nontechnical discussion and advice on all basic SDL techniques, users must be familiar with survey design and file content. Zarate pointed out that none of the rules can be followed blindly. There are real constraints on any attempt to standardize data protection; for instance, rules may be very different when applied to data for a demographically unusual group or for a survey topic that involves especially sensitive information. The Checklist is not a substitute for knowing the data that are to be released.

Researchers at the workshop voiced the concern that, if users are not adequately knowledgeable about the data and the associated risks and benefits, they may misuse documents such as the Checklist as a rationale for

overprotection. With institutions' reputations on the line, standardization can lead to conservatism in release policies, which researchers worry will inevitably limit the availability of data required for the most important and innovative research. In this context, the Checklist must be an evolving document. Users must be educated enough to adapt it to fit their specific requirements, reflecting, as Zarate put it, "that there is an art as well as a science to disclosure analysis."

Currently, the Checklist emphasizes proper handling of geographic information. "Small areas" are defined as 100,000 people by the Census Bureau; previously, geographic information was available only at the 250,000 person sampling unit level. The more recent definition reflects a rule of thumb without a real quantitative basis. Zarate argued that there is a real research need to develop empirical evidence to justify recommendations regarding geographic specificity. In fact, the disclosure risk posed by geographic delimitation can be assessed only in the context of other variables that are available in data records, as well as information about ease of external linkage. Adding detailed geographic identifiers to specific age, race, and other contextual variables makes data more useful, but also increases the probability of disclosure. The Checklist directs attention to the variety of available external files (e.g., voter registration, birth and death records) that could be linked to disclose record identity. It may also help guide decisions when data are being issued in a format that is easily manipulated.

Finally, Zarate suggested that the Checklist needs to be developed to direct special attention to longitudinal data. At the time of the workshop, the Checklist had only scratched the surface in terms of alerting data disseminators about additional risks that arise when records are followed through time. If an agency is locked into certain procedures, it can become clear over time that appropriate levels of security are not in place. For instance, characteristics that can be predicted from one period to the next may not be masked by top-coding or other techniques. The Checklist does not currently address these issues, but is expected to do so in the future.

RESTRICTION OF DATA ACCESS

Several of the workshop presentations described existing and planned restricted access arrangements for managing complex research data files. Paul Massell of the Census Bureau provided a comparative overview of licensing arrangements used by six U.S. agencies and two university-based social science research organizations. Marilyn McMillen of the National Center for Education Statistics (NCES) gave a detailed description of licensing procedures used by NCES, with emphasis on inspection procedures used to monitor observance of the conditions of access. Mark McClellan of Stanford University described procedures employed to protect confidentiality at an

academic research center that is using microdata files from multiple sources under licensing arrangements. Several presentations addressed research data centers: J. Bradford Jensen of the Census Bureau's Center for Economic Studies and Patrick Collins of the recently established California Census Research Data Center at the University of California, Berkeley, described their centers. John Horm reported on NCHS's Research Data Center, which offers both on-site and remote access to the agency's non-public-use data sets. And Garnett Picot of Statistics Canada outlined his agency's current restricted access procedures, as well as its plans to establish several research data centers.

The following sections summarize the features of the three principal kinds of restricted access arrangements as presented and discussed at the workshop, as well as one special procedure involving respondent consent that has been used by Statistics Canada. The main features of interest of these arrangements are the adequacy of the data for the desired analyses, eligibility requirements, the means used to provide adequate protection for individually identifiable data, the costs of obtaining and providing access, and the way these costs are shared by users and custodians of the files.

Licensing

NCES was one of the first organizations to issue licenses to researchers that allow them to receive and use nonpublic microdata sets at their own work sites. Nearly 500 licenses for files from several different NCES surveys have been issued since 1991. There are no specific restrictions by type of organization; licenses have been issued to government agencies at all levels, universities, research corporations, and associations. Applicants must provide a description of their research plan and demonstrate that it requires the use of restricted data, identify all persons who will have access to the data, and prepare and submit a computer security plan. They must execute a license agreement signed by an official with authority to bind the organization legally to its conditions and submit affidavits of nondisclosure signed by all persons who will have access to the data. They must agree to submit to unannounced inspections of their facilities to monitor compliance with security procedures; an NCES contractor carries out a systematic program of inspections. Licensees are subject to severe criminal penalties for confidentiality violations, as specified in the National Education Statistics Act of 1974.

During the past decade, several other agencies and organizations have developed and used licensing agreements for access to restricted data sets. Specific conditions vary. Some licensors provide access only to institutions certified by the National Institutes of Health as having met procedural criteria for IRBs or human-subject review committees. The duration of license agreements varies, with extensions being available in most instances. Some licensors require that publications based on the data be submitted to them for

disclosure review; others leave this responsibility to the licensee. Most agreements allow for unannounced inspections of facilities, but not all licensors have a systematic inspection program such as that conducted for NCES. Every licensee must cover the costs of going through the application process, which generally requires a significant amount of paperwork, and of establishing the physical and other security safeguards required to obtain approval of their computer security plans. Unlike NCES, which uses agency funds to cover the costs of processing applications and conducting inspections, some licensors charge user fees to cover these costs fully or partially.

Potential penalties for violations vary substantially. Federal agencies other than NCES that release files may be able to impose penalties under the Privacy Act or other legislation; however, these penalties would be less severe than those available to NCES. Penalties available to universities and other licensing organizations are generally of a different kind: immediate loss of access and denial of future access to data, forfeiture of a cash deposit, notification of violations to federal agencies that fund research grants, and possible liability to civil suits for violating contract provisions.

Research Data Centers

The Census Bureau pioneered the distribution of public-use microdata files from the decennial census and household surveys. However, microdata from establishment censuses and surveys cannot be publicly released because of higher associated disclosure risks. The Census law does not permit release of restricted data to users under licensing arrangements. Thus, the only viable option is to provide for access to such files at secure sites maintained by the Census Bureau. Access is allowed only to persons who are regular or special sworn Census employees and would be subject to penalties provided in the law for violations of its confidentiality provisions.

The Census Bureau's Center for Economic Studies, established in the mid-1980s at Census headquarters in Suitland, Maryland, initially constructed longitudinal files of economic data that were used for research by Census staff and by academic research fellows working at the Center as special sworn employees. Since then, additional research data centers have been established in the Bureau's Boston regional office; at Carnegie-Mellon University in Pittsburgh; and at the University of California at Los Angeles and the University of California, Berkeley. Another center is scheduled to open at Duke University in 2000. To date, only files of economic data for firms or establishments have been available, but the centers are planning to add restricted data sets from the decennial census and major household surveys, as well as linked employer–employee data sets.

All researchers desiring to use the research data centers' facilities must submit proposals that are reviewed for feasibility, scientific merit, disclosure

risk, and potential benefits to the Census Bureau.[4] The applicant must explain why the research cannot be done with publicly available data files. To minimize disclosure risks, projects are limited to those that emphasize model-based estimation, as opposed to detailed tabulations. To ensure that no confidential data are disclosed, all research outputs are reviewed by center staff and may not be removed from the center without the administrator's approval. Fees are charged for use of the centers' facilities, but some fellowships are available on a competitive basis to partially defray these costs; grantees of the National Science Foundation and the NIA are exempted.

NCHS recently established a research data center that provides both on-site and remote access to nonpublic data files from several NCHS surveys (see below for discussion of the center's remote access arrangements). The main requirements and conditions for on-site access are similar to those of the Census Bureau's research data centers. Research proposals must be submitted and are reviewed by a committee for disclosure risk, consistency with the mission of NCHS, and feasibility given the availability of the center's resources. All outputs are subject to disclosure review before being taken off site. Users are charged a basic fee for use of the center's facilities and an additional fee for any programming assistance provided by the center staff.

Statistics Canada recently decided to establish six to eight research data centers to provide access to data from five new longitudinal surveys of households and persons, including one with linked employer data. Other restricted data sets will be added as needed. The features of the centers will be similar in most respects to those of the Census Bureau's regional centers. They will be located at secure sites that have stand-alone computing systems and are staffed by Statistics Canada employees. They will operate under the confidentiality requirements of the Canadian Statistics Act, and only "deemed employees" will be allowed access to the data. Proposed research projects will be subject to a peer review process led by Canada's Social Science and Humanities Research Council. For users to be considered "deemed employees," they must produce an output or service for Statistics Canada. To meet this requirement, each user must produce a research paper that will be part of a series sponsored by the agency. The research paper will not include policy comments, but after meeting the requirement to produce the paper, researchers will be free to publish their results anywhere, accompanied by their interpretation of the policy implications.

[4]Access to identifiable data by special sworn employees is permitted only when such access is deemed to further the agency's mission, as defined by law.

Remote Access

Arrangements for provision of remote access to NCHS restricted data files, which preceded the establishment of the research data center, have now been taken over by the center. Center staff construct the data files needed for various analyses, a process that can include merging user-supplied files with the appropriate NCHS data files. The center provides a file of pseudo-data in the same format as the real data so users can debug their programs. In this manner, the number of back-and-forth iterations is substantially reduced. SAS is the only analytical software that can be used, and some of its functions, such as LIST and PRINT, cannot be used. Users submit their programs to the center by e-mail; following disclosure analysis, output is returned to them by e-mail. Charges for access to a file for a specified time period depend on the number of records included in the file. There is also a charge for file construction and setup services provided by center staff.

Statistics Canada has used remote access procedures on an ad hoc basis. An example is the provision of access to nonpublic data from a longitudinal survey of children and youth. The results have been mixed; a small evaluation survey and informal contacts with researchers have indicated that the system is judged by some to be cumbersome to use and not sufficiently interactive.

Respondent Consent Procedure

Section 12 of the Canadian Statistics Act permits Statistics Canada to share nonpublic survey data with an incorporated organization for statistical purposes, provided that survey respondents have given their permission to do so. The agency has used this procedure, known as a Section 12 agreement, to make such data sets available to other federal departments and to provincial statistical agencies. The respondent consent procedure must specify the organizations that will have access to the data. Typically, from 90 to 95 percent of respondents, who can be either persons or firms, give permission for their data to be used in this way.

Discussion

One trend that emerged from the presentations and discussion of restricted access was the rapid growth during the 1990s in the number of researchers obtaining access to complex research data files through all three of the principal methods of restricted access. The establishment of regional research data centers and the inclusion of demographic files at the Census Bureau's centers is likely to fuel a further expansion. NCES and the National Science Foundation are collaborating on the development of a manual of licensing procedures, which could potentially be used by other agencies and

organizations with the authority to employ this arrangement. Adoption of a more uniform set of procedures could reduce the cost and time required to submit proposals to licensors. There have been some negative user reactions to the controlled remote access approach because of limitations on software options, delays, and the relative difficulty of interaction between researcher and data source. However, a substantial research effort is now under way to develop more effective procedures for controlled access to microdata sets via the Internet.

Although expanded access to research files by a variety of methods is likely, there are some legal obstacles. Both the Census Bureau and NCHS, for example, have concluded that they do not have the legal authority to issue licenses that would allow researchers to use restricted data sets at their own facilities. Some workshop participants suggested that the Census Bureau should make restricted data files of other agencies available at its research data centers. Even if this were done, however, access to files from various agencies at a central point would probably still require different administrative procedures because of differences in the laws governing access to each agency's data. Under interagency agreements, NCES has undertaken distribution of files created by other agencies using licensing arrangements; it has the legal authority to do so as long as the files in question include data relevant to education.

The use of restricted access arrangements, which has been deemed necessary to provide adequate protection for confidential information about individuals and businesses, results in increased costs to conduct research. Custodians of the data files need additional resources to process applications, operate inspection systems, staff research data centers, and inspect outputs to ensure that disclosure does not occur. Researchers require resources to prepare applications for access, to provide appropriate physical security for the data, or to visit a secure site. At present, these costs are being covered partly by federal agency budgets and partly by user fees. The Census Bureau's research data centers have been supported in part by grants from the National Science Foundation and NIA, but may eventually have to recover more of their costs from users. Several workshop participants suggested that, if possible, graduate students should be exempted from such user fees.

Various restricted access arrangements offer different levels of protection for the confidentiality of individually identifiable information. Researchers working in research data centers under the supervision of agency employees are under closer supervision than those licensed to work with the data at their own facilities. Although there have been no known disclosures of individual information from NCES data files released under licenses, inspections have turned up numerous violations of the license requirements, such as failure to notify the agency of changes in personnel authorized to use the data. Protection of the data under controlled remote access arrangements depends prima-

rily on the effectiveness of automated screening systems and the vigilance of agency staff responsible for manual reviews of outputs prior to their release to users.

Most arrangements for restricted access are time-limited, and licensees are generally required to return or destroy their files and derived work files containing potentially identifiable records. It was pointed out that such provisions can make it difficult or impossible for other researchers to attempt to replicate research findings or for either the original or other researchers to pursue leads generated by the initial results.

References and Bibliography

Chapman, A.R., ed.
1997 *Health Care and Information Ethics: Protecting Fundamental Human Rights.* Kansas City, KS: Sheed and Ward.

Fanning, J.
1998 Privacy and Research: Public Policy Issues. Seminar presentation in Statistical Seminar Series on Confidentiality, Washington, DC, October 15. 1998. Office of the Assistant Secretary for Planning and Evaluation, U.S. Department of Health and Human Services.

Federal Committee on Statistical Methodology
1978 *Report on Statistical Disclosure-Avoidance Techniques.* Statistical Policy Working Paper 2. Subcommittee on Disclosure-Avoidance Techniques. Washington, DC: U.S. Department of Commerce.

Gostin, L.O.
1995 Health information privacy. *Cornell Law Review* 80:451-528.

Interagency Confidentiality and Data Access Group
1999 Checklist on Disclosure Potential of Proposed Data Releases, Statistical Policy Office, U.S. Office of Management and Budget.

Jabine, T.B.
1993a Procedures for restricted access data. *Journal of Official Statistics* 9(2):537-589.
1993b Statistical disclosure limitation practices of United States Statistical Agencies. *Journal of Official Statistics* 9(2):427-454.

Juster, F.T.
1991 Discussion. *American Statistical Association 1991 Proceedings of the Social Sciences Section.* Alexandria, VA.: American Statistical Association.

National Research Council
1979 *Privacy and Confidentiality as Factors in Survey Response.* Panel on Privacy and Confidentiality as Factors in Survey Response, Committee on National Statistics, Committee on Behavioral and Social Sciences and Education. Washington, DC: National Academy Press

National Research Council and Social Science Research Council
1993 *Private Lives and Public Policies: Confidentiality and Accessibility of Government Statistics.* Panel on Confidentiality and Data Access, George T. Duncan, Thomas B. Jabine, and Virginia A. de Wolf, eds. Committee on National Statistics. Washington, DC: National Academy Press.

Appendixes

APPENDIX A

Workshop Participants

Presenters

Norman Bradburn (*Workshop Chair),* National Opinion Research Center
Erik Austin, Inter-university Consortium for Political and Social Research
Mary Ann Baily, Institute for Ethics, American Medical Association
Robert Boruch, University of Pennsylvania
Richard Burkhauser, Cornell University
Patrick T. Collins, California Census Research Data Center, University of California, Berkeley
J. Michael Dean, University of Utah
George Duncan, Carnegie-Mellon University
Donna Eden, Office of the General Counsel, U.S. Department of Health and Human Services
Stephen E. Fienberg, Carnegie-Mellon University
Robert M. Gellman, Privacy and Information Policy Consultant, Washington, D.C.
Rachel Gordon, University of Illinois at Chicago
John Horm, National Center for Health Statistics
J. Bradford Jensen, Center for Economic Studies, Bureau of the Census
Sallie Keller-McNulty, Los Alamos National Laboratory
Arthur Kennickell, Federal Reserve Board
Paul Massell, Bureau of the Census
Mark McClellan, Stanford University
Marilyn M. McMillen, National Center for Education Statistics

Garnett Picot, Statistics Canada
Tom Puglisi, Office for Protection from Research Risks, National Institutes of Health
Alice Robbin, Indiana University
Latanya Sweeney, Carnegie-Mellon University
Robert Weathers, Cornell University
Finis Welch, Texas A&M University
Robert Willis, University of Michigan
Alvan Zarate, National Center for Health Statistics

Invited Guests

Paul P. Biemer, Research Triangle Institute
Lewis Berman, National Center for Health Statistics
Katharine Browning, Office of Juvenile Justice and Delinquency Prevention
Jay Casselberry, Energy Information Administration
Chris Chapman, Bureau of Labor Statistics
Stephen Cohen, Bureau of Labor Statistics
Virginia A. de Wolf, Office of Management and Budget
Cathryn Dippo, Bureau of Labor Statistics
Nancy Donovan, General Accounting Office
Patricia Doyle, Bureau of the Census
Judy Droitcour, General Accounting Office
John Eltinge, Bureau of Labor Statistics
Anastasia J. Gage, USAID
Gerald W. Gates, Bureau of the Census
Dan Gaylin, U.S. Department of Health and Human Services
Nancy Gordon, Bureau of the Census
Brian Greenberg, Social Security Administration
Easley Hoy, Disclosure Review Board, Bureau of the Census
Betsy Humphreys, National Library of Medicine
Thomas Jabine, Consultant, Committee on National Statistics
Nancy Kirkendall, Department of Education
Julia Lane, Center for Economic Studies, Bureau of the Census
Susan Lapham, Bureau of Transportation Statistics
Thomas Louis, University of Minnesota, and Member, Committee on National Statistics
David Mednick, Bureau of Transportation Statistics
Jay Meisenheimer, Bureau of Labor Statistics
Heather Miller, Office of Extramural Research, National Institutes of Health
Nancy Miller, National Institutes of Health
Joseph Moone, U.S. Department of Justice
Kristen Robinson, National Center for Health Statistics

Stuart Rust, Bureau of Labor Statistics
James Scanlon, U.S. Department of Health and Human Services
Lois Schein, Social Security Administration
Eleanor Singer, University of Michigan
Edward Spar, Council of Professional Associations on Federal Statistics
James Spletzer, Bureau of Labor Statistics
Richard M. Suzman, National Institute on Aging
Louise Wideroff, National Institutes of Health
Laura Zayatz, Bureau of the Census

NRC Staff

Jamie Casey, Committee on National Statistics
Barney Cohen, Committee on Population
Kevin Kinsella, Committee on Population
Heather Koball, Committee on National Statistics
Christopher Mackie, Committee on National Statistics
Terri Scanlan, Committee on National Statistics
Miron Straf, Committee on National Statistics
Barbara Boyle Torrey, Commission on Behavioral and Social Sciences and Education
Andrew White, Committee on National Statistics
Lee Zwanziger, Institute of Medicine

APPENDIX B

Workshop Agenda

Thursday, October 14, 1999

8:45 WELCOME AND OPENING REMARKS

Barbara Boyle Torrey, *Director, Commission on Behavioral and Social Sciences and Education*
Miron Straf, *Director, Committee on National Statistics*
Richard Suzman, *Behavioral and Social Research Program, National Institute on Aging*

INTRODUCTION AND WORKSHOP PREVIEW

Norman Bradburn, *Workshop Chair*

9:05 **SESSION I. Linked longitudinal databases: Achievements to date and prospects**

A. TOPIC 1: Case Studies

Paper 1: HRS application: "How Policy Variables Influence the Timing of Social Security Disability Applications"
– Richard Burkhauser, Cornell University
– Robert Weathers, Cornell University

Paper 2: NLS-Y application: "Confidential Data Files Linked to the National Longitudinal Survey of Youth, 1979: A Case Study"
– Rachel Gordon, University Illinois at Chicago

B. TOPIC 2: Risks and Rewards of Data Linking

Presentation: "Protecting Confidentiality of Linked Datasets: Don't Throw the Baby Out with Bathwater"
– J. Michael Dean, University of Utah

C. TOPIC 3: Report from the Conference on the Value of Linked Data

Presentation: Robert Willis, University of Michigan

D. GENERAL DISCUSSION OF PAPERS/PRESENTATIONS

11:10 **SESSION II. Legal and ethical requirements for dissemination: Recent and prospective developments**

A. TOPIC 1: Philosophical Issues

Paper: "Regulating Access to Research Data Files: Ethical Issues"
– Mary Ann Baily, George Washington University

Discussant: Robert Boruch, University of Pennsylvania

B. TOPIC 2: Recent and Prospective New Legislation

Paper: Donna Eden, U.S. Department of Health and Human Services

Discussant: Robert Gellman, Privacy and Information Policy Consultant

C. TOPIC 3: Problems and Issues of Institutional Review Boards

Paper: Thomas Puglisi/Jeffery Cohen, Office for Protection from Research Risks, National Institutes of Health

D. GENERAL DISCUSSION, led by Norman Bradburn

2:00 **SESSION III. Procedures for releasing public-use microdata files**

A. TOPIC 1: Agency Practices for Releasing Public-Use Micro-Data Files

Presentation: "Checklist on Disclosure Potential of Proposed Data Releases," developed by the Interagency Confidentiality and Data Access Group
– Alvan Zarate, National Center for Health Statistics

Paper: "A Survey of Statistical Disclosure Limitation (SDL) Practices of Organizations that Distribute Public Use Microdata"
– Thomas Jabine, Private Consultant
– Alice Robbin, Indiana University
– Heather Koball, Committee on National Statistics

Discussant: Erik Austin, Inter-university Consortium for Political and Social Research

Paper: "Multiple Imputation in the Survey of Consumer Finances"
– Arthur Kennickell, Federal Reserve Board

Discussant: Stephen Fienberg, Carnegie Mellon University

B. TOPIC 2: Quantification of Disclosure Risk

Paper: – Sallie Keller-McNulty, Los Alamos National Laboratory
– George Duncan, Carnegie Mellon University

Discussant: Latanya Sweeney, Carnegie Mellon University

4:30 GENERAL DISCUSSION

5:00 CLOSING COMMENTS AND ADJOURN

Friday, October 15, 1999

9:00 **SESSION IV. Procedures for restricting access: who gets access to micro-data, and under what conditions**

A. TOPIC: Use of restricted access procedures for major linked longitudinal databases

Paper: "Review of Data Licensing Agreements at U.S. Government Agencies and Research Organizations"
– Paul Massell, U.S. Census Bureau

Discussant: Mark McClellan, Stanford University

Presentation: Remote data access at the National Center for Health Statistics
– John Horm, National Center for Health Statistics

Discussant: Finis Welch, Texas A&M University

Presentation: Statistics Canada procedures; data liberation and associated Internet background information
– Garnett Picot, Statistics Canada

Presentation: Data licensing systems
– Marilyn McMillen, National Center for Education Statistics

Presentation: Data Centers—A critical assessment
– Patrick Collins, California Census Research Data Center
– J. Bradford Jensen, Carnegie Mellon Census Research Data Center

1:30 GENERAL DISCUSSION

2:00 **SESSION V. Workshop Summary**

Review and address questions posed at the opening of the workshop
- Thomas Louis, Member, Committee on National Statistics
- Norman Bradburn, Workshop Chair

General response
- Workshop participants

2:45 **CLOSING COMMENTS/ADJOURN**