Journal of Law and Policy

Volume 14
Issue 1
SCIENCE FOR JUDGES V:
Risk Assessment Data: Disclosure and Protection

Article 5

2005

Access to Research Data: Reconciling Risks and Benefits

Eleanor Singer

ACCESS TO RESEARCH DATA: RECONCILING RISKS AND BENEFITS

*Eleanor Singer, Ph.D.**

INTRODUCTION

In 2003, the National Research Council of the National Academy of Sciences established the Panel on Data Access for Research Purposes to "assess competing approaches to exploiting the research potential of microdata . . . while preserving confidentiality."[1] The panel was asked to consider the tradeoffs between the risks and benefits of access to research data and to make recommendations about "how microdata should optimally (from a societal standpoint) be made available to researchers."[2] As the panel's chair, I drew heavily on its final report for this article, but the views expressed are my own, not those of the panel or of the Academy.

In the present context, "research data" refers to information collected from individuals, households, firms, and other organizations for exclusively statistical purposes under a pledge of confidentiality. Confidentiality means that the information will not be disclosed in identifiable form to an unauthorized party.[3]

Examples of information collected exclusively for statistical purposes include: (1) the information collected in the decennial census; (2) the monthly employment data collected by the U.S. Census Bureau for the Bureau of Labor Statistics; and (3) the

* Research Professor, Institute for Social Research, University of Michigan.

[1] PANEL ON DATA ACCESS FOR RESEARCH PURPOSES, NAT'L RESEARCH COUNCIL, EXPANDING ACCESS TO RESEARCH DATA: RECONCILING RISKS AND OPPORTUNITIES 2 (2005) [hereinafter PANEL ON DATA ACCESS].

[2] *Id.*

[3] SYS. SEC. STUD. COMMITTEE, NAT'L RESEARCH COUNCIL, COMPUTERS AT RISK: SAFE COMPUTING IN THE INFORMATION AGE 289 (1991).

information obtained in the Health Interview Survey, also collected by the Census Bureau. The purpose of such data collections is to generate information about *categories* of persons or organizations, such as children under the age of 18 or households with only one parent present. Although policy decisions based on that information may affect individual members of the category, no *direct* action is taken for or against a specific individual or organization on the basis of the information collected.

In contrast to data collected for statistical purposes, information gathered for administrative purposes is expressly designed to facilitate a course of action affecting a particular person or organization.[4] Examples of administrative data include information required from individuals seeking to obtain a driver's license or to qualify for Medicare or Medicaid. Such data are not collected under a pledge of confidentiality,[5] and individuals have no reasonable expectation of the data's confidentiality.

Although administrative data are not the subject of the present discussion, they become relevant to it when, as happens increasingly often, they are linked to data collected for statistical purposes under a pledge of confidentiality. To perform such linkages the consent of the individual is ordinarily required, and the linked administrative data become subject to the assurance of confidentiality provided by the researcher.[6]

This article discusses the benefits and potential costs of access to research data, ways of reconciling the two, and why it is important to do so. I want to stress two main points. First, there are competing claims in this arena between those who, for good reasons, want easier access to research data, and those who, for

---

[4] "Administrative purposes" may include regulatory, legislative, or judicial purposes. PANEL ON CONFIDENTIALITY AND DATA ACCESS, NAT'L RESEARCH COUNCIL, PRIVATE LIVES AND PUBLIC POLICIES: CONFIDENTIALITY AND ACCESSIBILITY OF GOVERNMENT STATISTICS 24 (1993) [hereinafter PANEL ON CONFIDENTIALITY AND DATA ACCESS].

[5] *Id*. ("[T]he Privacy Act of 1974 defines a statistical record to be[:] a record in a system of records maintained for statistical research or reporting purposes only and not used in whole or in part in making any determination about an *identifiable individual*.") (emphasis added).

[6] *Cf. infra* Part I.B.

equally good reasons, are concerned that granting such access risks harming not only individuals but the research enterprise itself. Increasingly, the courts are going to be asked to adjudicate these competing claims.

The second, and perhaps even more important point, is that there are legal, technical, and administrative devices for managing the tension between these competing claims. These devices may not totally satisfy either privacy advocates or those who want fast and free access to detailed information about individuals. But given the benefits of both access and confidentiality protection, it is important to make optimum use of all available tools.

Part I of this article provides an example of why issues of data access are of interest to judges, and discusses recent legislation affecting access to research data. Part II begins with some examples of the uses of research data, and then discusses the benefits and potential costs of expanded access as well as technical, legal, and social changes that have increased the tension between benefits and costs. Part III reviews the threats to confidentiality posed by expanded access to research data, and discusses a variety of technological, legal, and administrative solutions.

I. WHY DATA ACCESS IS OF INTEREST TO JUDGES

A. Southern Illinoisan v. Department of Public Health

One reason judges should be concerned with issues of data access is the recent Illinois Appellate Court decision, *Southern Illinoisan v. Department of Public Health*.[7] The Illinois Department of Public Health appealed the Jackson County trial court decision, which had ordered the disclosure of certain Illinois Cancer Registry (Registry) information to the *Southern Illinoisan,* a newspaper in Carbondale, Illinois, based upon a Freedom of

[7] 812 N.E.2d 27 (Ill. App. Ct. 2004).

Information Act (FOIA) request.[8] This was the second appeal from the trial court's decision; in the first, the Appellate Court held that the phrase in the Illinois Health and Hazardous Substances Registry Act,[9] forbidding public inspection or dissemination of any group of facts that would tend to identify persons in the Registry, referred to any group of facts that "reasonably" would tend to identify specific persons.[10]

On remand, the district court held hearings to determine if the information sought by the newspaper would reasonably tend to identify specific persons in the Registry. Latanya Sweeney, a professor at Carnegie Mellon University, testified that using only the information sought by the plaintiffs in the case—information stripped of obvious identifiers like name and address—she was able to identify 18 of 20 individuals exactly. She was also able to supply two plausible alternative names for each of the remaining two individuals.[11]

The Appellate Court nevertheless ordered defendants to turn over the requested information, affirming the second trial court decision. The court found that Dr. Sweeney's knowledge and analytical skill are "beyond that of the average person."[12] Thus, it was not reasonable to believe that anyone "with less knowledge, education, and experience than Dr. Sweeney"[13] would be able to identify individuals.[14] But this kind of knowledge, while not universal, is by no means arcane. Students at the Massachusetts Institute of Technology, for example, performed a similar analysis for individuals in Chicago's Homicide Database.[15]

---

[8] *Id*.

[9] Illinois Health and Hazardous Substances Registry Act, 410 ILL. COMP. STAT. 525/(d) (1998).

[10] S. Illinoisan v. Dep't of Pub. Health, 747 N.E.2d 401 (Ill. App. Ct. 2001).

[11] *S. Illinoisan*, 812 N.E.2d at 29.

[12] *Id.*

[13] *Id.* at 30.

[14] *Id.*

[15] S. OCHAS, ET AL., IDENTIFICATION OF INDIVIDUALS IN CHICAGO'S HOMICIDE DATABASE: A TECHNICAL AND LEGAL STUDY (2001) (unpublished

B. Recent Federal Legislation Affecting Access To Research Data

Another example of why judges should be interested in the issues involved in access to research data is the relatively recent passage of a federal law, the Confidential Information Protection and Statistical Efficiency Act (CIPSEA).[16] Subpart A of CIPSEA creates equal confidentiality protection for all data collected by federal agencies for exclusively statistical purposes under a pledge of confidentiality, and it raises the level of protection for many of them to that enjoyed previously only by the Census Bureau and the National Center for Health Statistics. The law also protects such information against FOIA requests.

The law's protections also extend to contractual agents of the federal agencies, who may be organizations or individual researchers. Currently, the U.S. Office of Management and Budget (OMB) is preparing regulations to implement the safeguards under CIPSEA. These regulations are expected to define more precisely both the reach of protection for confidential statistical records and the opportunities for research access.

Under CIPSEA, willful disclosure of identifiable information collected under a promise of confidentiality may result in a fine of $250,000, five years imprisonment, or both. The law's provisions, however, have not yet been tested in court and the extent to which they will prevail if they conflict with other laws and interests—for example, national security concerns—remains to be seen.

Several other recently enacted laws—the USA PATRIOT Act of 2001 (Patriot Act), the Shelby Amendment, and the Information Quality Act—are relevant to the issue of data access.[17]

The Patriot Act[18] overturned the strict confidentiality

---

manuscript, on file with the author).

[16] *See* E-Government Act of 2002, 44 U.S.C. § 3501 (2006).

[17] The implications of an additional statute, the Health Insurance Portability and Accountability Act of 1996, 18 U.S.C. § 3486 (2000), are beyond the scope of this paper.

[18] Uniting and Strengthening America by Providing Appropriate Tools Required to Intercept and Obstruct Terrorism Act (USA PATRIOT Act) of

protection of education records gathered and maintained by the National Center for Education Statistics (NCES), a change in protection that was later reflected in corresponding amendments to the statute governing NCES. The Patriot Act allows the Attorney General to petition a court for access to identifiable education records, including research records, for use in the investigation and prosecution of terrorist activities.

The Shelby Amendment also permits access to federal research information for non-research purposes.[19] The Amendment requires OMB to set forth regulations to ensure that all data that are supported by a federal grant to colleges, universities, hospitals and other nonprofit institutions "will be made available to the public through procedures established under the [FOIA]."[20] The resulting OMB guidelines restrict access to published or cited research that has been used by the federal government to develop legally binding regulations and rulings. The guidelines also provide an exemption to access under FOIA for information that would result in a "clearly unwarranted invasion of personal privacy, such as records that could be used to identify a particular person in a research study."[21] CIPSEA supports the OMB's interpretation of the Amendment because it limits disclosure of confidential information under FOIA as well. However, the validity of the OMB guidelines and the CIPSEA restrictions have yet to be tested through litigation.

Federal statistical agencies also confront increased scrutiny of the quality of information that they disseminate to the public, even if the data have not been used as part of the regulatory process. The

---

2001, Pub. L. No. 107-56, 115 Stat. 272 (codified in scattered sections of the U.S.C.).

[19] Omnibus Consolidated and Emergency Supplemental Appropriations Act, Pub. L. No. 105-277, 112 Stat. 2681 (1998).

[20] *Id*. § 2681-495.

[21] PANEL ON DATA ACCESS, *supra* note 1, at 24. For a detailed discussion of this issue, see AD HOC COMMITTEE ON ENSURING THE QUALITY OF GOVERNMENT INFORMATION, NAT'L RESEARCH COUNCIL, ENSURING THE QUALITY OF DATA DISSEMINATED BY THE FEDERAL GOVERNMENT: WORKSHOP REPORT (2003).

Information Quality Act,[22] also known as the Data Quality Act, directs OMB to issue guidelines for "ensuring and maximizing the quality, objectivity, utility, and integrity of information disseminated . . . by federal agencies"[23] to the public, and requires all federal agencies to establish administrative procedures for correcting disseminated information that does not meet those standards. The resulting OMB guidelines[24] define "scientific information" to include agency distribution of public use and restricted use statistical datasets.[25] "Influential scientific information," which is defined as information reasonably expected to "have a clear and substantial impact on important public policies or important private sector decisions,"[26] must meet even higher information quality standards. Agencies that disseminate influential scientific information must first conduct a peer review and reveal enough about the data and methods used to facilitate independent reanalysis, while taking into account privacy, confidentiality, and intellectual property rights. There is concern among some researchers that the Data Quality Act, like the Shelby Amendment, may be used by those opposed to certain policy initiatives to challenge the findings and quality of research data as a means of impeding agency regulatory activities. In 2003, some 19 federal agencies received data correction requests under the Data Quality Act.[27]

---

[22] Consolidated Appropriations Act of 2001 (commonly known as the Information (or Data) Quality Act), Pub. L. No. 106-554, 515, 114 Stat. 2763, 2763A-153 (2000).

[23] *Id.* § 515(a).

[24] *See* OMB Final Information Quality Bulletin for Peer Review, 70 Fed. Reg. 2664-02 (January 14, 2005).

[25] *Id.* at 2667.

[26] *Id.*

[27] *See* Draft 2005 Report to Congress on the Costs and Benefits of Federal Regulations, June 2005, *at* www.whitehouse.gov/omb/inforeg/2005_cb/draft_2005_cb_report.pdf.

II. COSTS AND BENEFITS OF EXPANDED ACCESS TO RESEARCH DATA

A. The Uses of Research Data

It is hard to find a sector of society that does not make use of research data in some way. First, of course, there is use by the government itself, in its policy-formulating role. Public policies often focus on population groups defined in terms of one or more characteristics: low-income families, veterans, Medicare patients, preschool children, drug addicts, and homeowners, to name a few from a long list. Policy design proceeds on the basis of knowing how many people there are in these groups; how they are geographically distributed; and how they differ in other characteristics. For example, how will changing the age of eligibility for Social Security affect retirement decisions across different occupations and regions of the country? Information about the potential impact of such policy changes can influence the legislation eventually adopted.[28]

Other public policies focus on public and private establishments such as public schools, military bases, hospitals, prisons, small businesses, health care providers, and financial institutions. For establishments, too, complex policy-making requires access to complex, multivariate microdata derived from large-scale surveys of individuals and establishments. Access to microdata—that is, individual-level information, as distinct from aggregated summary data—provides the analytic flexibility needed for sophisticated policy analysis and planning. At the same time, access to microdata, especially microdata linked to administrative records, such as Social Security earnings statements or Medicare records, increases the probability that the confidentiality of the data can be breached.[29]

The most important sources of the information used for policy design, evaluation, and planning purposes are the more than 70

---

[28] For other examples, see PANEL ON DATA ACCESS, *supra* note 1.

[29] *Id*.

federal agencies that carry out statistical activities of at least $500,000 per year. In fiscal year 2004, these agencies were authorized to spend about $4.8 billion for statistical programs to serve the Nation's informational needs.[30] The agencies either collect data themselves or, very commonly, contract with survey organizations such as Westat or the National Opinion Research Center (NORC) to collect the data for them.[31] In addition, state agencies also collect data needed to carry out state governmental functions.

Social scientists, of course, are heavy producers as well as consumers of such data, often with the aid of government grants or contracts. Such surveys as the University of Michigan's Health and Retirement Study, a national longitudinal survey of some 12,000 Americans aged 50 and over, or the NORC's General Social Survey, a cross-sectional national survey of 3,000 Americans 18 and over, are done under the leadership of social scientists with grants from the federal government, and are made available for use by the research community and the general public. This last statement—that research data are made widely available for reanalysis by others—is key to the tension faced by data collection agencies.

B. Benefits of Expanded Access to Research Data

There are many benefits of increasing access to research data; three are discussed below.

First, primary data collection is increasingly expensive, difficult, and burdensome and, thus, sharing data may reduce the cost of collecting the information and alleviate some of the burden on respondents. Because ours is a mobile society, whose members carry out many activities outside the home, individuals are

---

[30] U.S. OFFICE OF MANAGEMENT AND BUDGET, BUDGET OF THE UNITED STATES GOVERNMENT FISCAL YEAR 2004, p. 7 (2003).

[31] The list of federal agencies that gather data includes: the Census Bureau, the Bureau of Labor Statistics, the National Center for Health Statistics, the Centers for Disease Control, the National Institute of Drug and Alcohol Use, and the National Center for Education Statistics.

becoming more difficult to reach and, when reached, they are often too busy to answer the researcher's questions. But because interviewing all persons who have been selected for a sample is crucial for the validity of survey results, difficulties in reaching potential respondents and persuading them to be interviewed mean escalating survey costs.

Furthermore, the kinds of surveys carried out or sponsored by government have become longer over time, and often inquire into topics, such as health and financial status, that are considered sensitive by respondents. As a result, surveys have become more burdensome, further increasing their difficulty and cost. Sharing research data by giving many researchers access to a single data set increases the return on this increased investment.

Second, there is a benefit from sharing research data that accrues to the statistical agencies themselves. When data are shared with external researchers along with study results, the agencies can improve their own data collection methods and analytic capabilities. Faulty techniques that might not otherwise have been discovered can be identified, and techniques that are effective will be promoted. Researcher access also ensures that additional information about statistical procedures, which might otherwise not be completely documented by agencies, is archived.

Third, democratic societies require multiple perspectives brought to bear on research data. Facts do not speak for themselves. They are interpreted by analysts with different points of view and, sometimes, different axes to grind. If access to data is limited to those with only one point of view, there is no opportunity for a critique of the policies that governments or private industries develop. Broader access does not guarantee better policy, but it does make possible an informed critique and evaluation of whatever policies are adopted and their outcomes. The current debate over Social Security is a case in point. Competing analyses have produced significantly different estimates of when the Social Security trust fund will run out of money, how much is required to assure its solvency, and how much various proposals, such as taxing incomes over $90,000 or raising the retirement age, would contribute to reducing or

eliminating the deficit.[32]

Finally, science, as well as policy, benefits from replication because replication guards against both good-faith error and deliberate fraud. It is often impossible, for a variety of reasons, to produce a data set identical to one that was used to formulate a theory or a practical policy. Because of this, it is essential that scientists as well as policy makers have access to the original data on which the policies or theories were based in order to replicate the analyses that were carried out. Cyril Burt, for example, was able to perpetuate his theory of the heritability of intelligence for many years because only he had access to the disputed data set of 53 pairs of identical twins on which he based his claims.[33]

C. Potential Costs of Expanded Access to Research Data

One cost of providing unrestricted access to fully detailed microdata is a potential breach of the data's confidentiality. This section considers the various ways in which confidentiality can be breached. But first, it is important to consider why confidentiality matters.

Confidentiality matters, in the first place, for the individual. Much of the information requested by government statistical agencies and their agents is sensitive for a variety of reasons. The Health and Retirement Study, for example, obtains detailed information from individuals about earnings and assets in order to analyze the relationship between health, wealth, and retirement decisions, which is the primary aim of the survey. But if the information became known to unscrupulous outsiders, it might

---

[32] *See*, *e.g.*, CENTER FOR AMERICAN PROGRESS, PROGRESSIVE GUIDE TO THE SOCIAL SECURITY DEBATE, http://www.americanprogress.org/site/pp.asp?c=biJRJ8OVF&b=289148; THE SOCIAL SECURITY NETWORK, A CENTURY FOUNDATION PROJECT, http://www.socsec.org/; FEDERAL RESERVE BANK OF SAN FRANCISCO, FRBSF ECONOMIC LETTER 99-20 (June 25, 1999), http://www.frbsf.org/econrsrch/wklyltr/wklyltr99/el99-20.html.

[33] J. A. Plucker, *Human intelligence: Historical Influences, Current Controversies, Teaching Resources* (2003), *avilable at* http://www.indiana.edu/~intell (last visited June 16, 2005).

make respondents liable to fraud, theft, or other kinds of abuse.

To take another example, the National Institute for Drug and Alcohol conducts detailed surveys of adolescent drug use for epidemiological purposes. If adolescents honestly admit the use of illegal drugs and if that information became known to law enforcement agencies, they would be subject to legal sanctions. There are numerous examples but the general point has been made: if the information collected in many surveys became publicly known, together with identifying information about the respondents, respondents would be subject to a variety of harms ranging from embarrassment to employment discrimination, criminal victimization or imprisonment. A breach of confidentiality and its potential consequences, in other words, is probably the most serious risk to which participants in social research—as opposed to biomedical research—are subject.

Individual harm, however, is not the only consequence of a confidentiality breach—decreased participation in social research may also result. Unlike much biomedical research, social research depends on the voluntary cooperation of those selected, usually by probability sampling methods, to participate. Biomedical researchers assume that the individuals they study are homogenous enough so that one person can be substituted for another. This is not true in social research, where individual variation is precisely the object of study.[34] If significant elements of the population who are designated for measurement refuse to cooperate, the inferences drawn from the responses of those who do cooperate are likely to be erroneous and misleading. There is evidence from a variety of studies, most of them carried out in connection with the U.S. decennial census, that public concerns about privacy and confidentiality have increased significantly over the last decade,[35]

---

[34] *See, e.g.,* PANEL ON INSTITUTIONAL REVIEW BOARDS, SURVEYS, AND SOCIAL SCIENCE RESEARCH, NAT'L RESEARCH COUNCIL, PROTECTING PARTICIPANTS AND FACILITATING SOCIAL AND BEHAVIORAL RESEARCH 102 (2003).

[35] ELEANOR SINGER, US CENSUS BUREAU, CENSUS 2000 TESTING, EXPERIMENTATION, AND EVALUATION PROGRAM TOPIC REPORT NO. 1, TR-1, PRIVACY RESEARCH IN CENSUS 2000 4-7, (2003), *available at*

and that they significantly reduce cooperation. For example, concerns about privacy and confidentiality significantly reduced participation in the decennial censuses of 1990 and 2000.[36] Small-scale experiments also demonstrate that the likelihood of confidentiality breaches is perceived as an important risk by survey respondents, and is significantly correlated with their decision about whether or not to participate in nongovernmental surveys.[37] This reluctance extends to other behaviors,[38] for example, providing one's Social Security Number (SSN).[39] Thus, there is ample evidence that confidentiality breaches potentially harm both the individual and the larger society.

---

http://www.census.gov/pred/www/rpts/TR-1.pdf.

[36] *See* Eleanor Singer, et al., *The Impact of Privacy and Confidentiality Concerns On Survey Participation: The Case of the 1990 Census*, 57 PUB. OPINION Q. 465 (1993); Eleanor Singer, et al., *Attitudes and Behavior: The Impact of Privacy and Confidentiality Concerns on Participation in the 2000 Census*, 65 PUB. OPINION Q. 368 (2003); NAT'L RESEARCH COUNCIL, 2000 CENSUS: COUNTING UNDER ADVERSITY 2004; S. Hillygus, et al., *Civic Mobilization and Privacy Concerns in the 2000 Census* (2006).

[37] Eleanor Singer, *Exploring the Meaning of Consent: Participation in Research and Beliefs about Risks and Benefits*, 19 J. OF OFFICIAL STAT. 273 (2003).

[38] N. A. Bates, *Development and Testing of Informed Consent Questions to Link Survey Data with Administrative Records*, AAPOR Annual Conference, May 13, 2005.

[39] Census Bureau efforts to obtain SSNs have become increasingly difficult over the last few years. For example, in the Survey of Income and Program Participation (SIPP), SSN refusals increased from 12 percent to 35 percent between the 1996 and 2001 panels, respectively. Email from Donna Ricinni, Chief Income Surveys Programming Branch, U.S. Census Bureau to N. Bates (Jan. 23, 2003) (on file with author); D. LEWIS, U.S. CENSUS BUREAU, FINAL PROJECT REPORT FOR IMPROVING SOCIAL SECURITY NU,MBER RESPONSE STUDY, 2005 (on file with author). SSN refusals in the Current Population Survey (CPS) increased from approximately ten percent in 1994 to almost 23 percent by 2003. R. TUCKER, U.S. CENSUS BUREAU, RECENT CURRENT POPULATION SURVEYS AND SOCIAL SECURITY NUMBER REPORTING EXPERIENCE, 1999; T. MARSHALL, U.S. CENSUS BUREAU, RECENT CURRENT POPULATION SURVEYS SOCIAL SECURITY NUMBER REPORTING EXPERIENCE, 2004 (on file with author).

D. Technical, Legal and Social Changes That Have Increased the Tension

The tension between access and confidentiality is not new. In 1985, a National Research Council (NRC) Panel issued a report that focused primarily on the benefits of access.[40] In 1993, another NRC Panel issued a comprehensive report titled *Private Lives and Public Policies,* which explicitly examined the various tensions involved in making research data more widely available while maintaining their confidentiality.[41] Why, then, take another look at the problem now? The reasons lie in the substantial changes, primarily technological but also legal and social, that have taken place in the last decade.

Undergirding all the other changes are the vast increases in computing power and storage capacity that have taken place. The information world now captures enormous numbers of records of personal and organizational information, stores them in data warehouses, analyzes them through sophisticated statistical and data mining techniques, and disseminates the results instantaneously through electronic communications media. The explosion in information technology is evident at each stage of the process of data capture, storage, integration, and dissemination, and can be measured by the reduced cost of each of these activities. For example, the NRC noted in 2005:

> [O]ne terabyte of storage can hold the contents of 2,000 file cabinets of documents. Ten years ago, such storage would have cost $1 million; now it can be obtained for less than $800. . . . Data integration—that is, consolidating information from heterogeneous databases—is no longer a horrendously complex task, but one that is facilitated by data standards (such as XML), the growth of the Web, and fast and inexpensive data transmission capability. Correspondingly, data dissemination through the Web and

---

[40] COMMITTEE ON NAT'L STAT., NAT'L RESEARCH COUNCIL, SHARING RESEARCH DATA (1985).

[41] PANEL ON CONFIDENTIALITY AND DATA ACCESS, *supra* note 4.

electronic mail is now free for all practical purposes.[42]

Such decreased costs benefit legitimate researchers, but they also provide opportunities for "data snoopers"—individuals or organizations that seek to identify individual respondents for a variety of purposes that include curiosity, mischief, and marketing as well as searching for criminals or terrorists. Data that are most useful to legitimate researchers also have properties that make individuals especially vulnerable to data snoopers (examples of such data include detailed geographic information, repeated data collections from the same subjects, and complete censuses rather than small probability samples). All of these properties make it easier to identify individuals even in a data file from which direct identifiers, such as names and addresses, have been removed.

The second change is the development of new ways of collecting information about individuals. For example, the use of scanners at supermarkets permits stores to collect and share a wealth of information about purchasing behavior and to link that to the information provided on consumers' credit applications and to census data for small areas. The use of keycards in businesses and government offices enables employers to track employee movements precisely.[43] So-called "black boxes" in new model automobiles record and store information on seat belt use and speed, but buyers are not always aware of these features.[44] Other kinds of information, such as biomarkers for various diseases and even unique genetic information, are increasingly collected as part of social surveys that also inquire into many details of individuals'

---

[42] *See* PANEL ON DATA ACCESS, *supra* note 1. *See also* George T. Duncan, *Exploring the Tension Between Privacy and the Social Benefits of Governmental Databases, in* A LITTLE KNOWLEDGE: PRIVACY, SECURITY AND PUBLIC INFORMATION AFTER SEPTEMBER 11 72 (Peter. M. Shane et al. eds., 2004) (stating that "advances in information technology have sharply lowered the costs of data capture, storage, integration and dissemination").

[43] EDWARD BALKOVICH, TORA K. BIKSON & GORDON BITKO, 9 TO 5: DO YOU KNOW IF YOUR BOSS KNOWS WHERE YOU ARE? 11 (RAND 2005) (discussing the use of Radio Frequency Identification access cards in the workplace).

[44] N.D. Bismarck, *States Seek to Regulate 'Black Boxes' in Autos,* N.Y. TIMES, Mar. 27, 2005, at 16.

behavior.[45]

A great deal of care and thought is needed when such data are released. For example, a map pinpointing the location of sample members can, potentially, serve to identify those individuals; knowing who is in the sample, in turn, greatly increases the ability of intruders to locate those records in the data file, and thus to learn things about them that the individual believed were available only to the researchers.

The third change is the existence and increasing availability of large databases containing information on hundreds of thousands, sometimes millions, of individuals.[46] These databases—such as those maintained by Experian,[47] for example—contain names and addresses, and sometimes Social Security Numbers, as well as information on a wide variety of individual characteristics, such as income, education, race, marital status, and much more. Together with sophisticated software for matching records electronically, these databases provide the tools for matching data records from which direct identifiers have been removed with other records that have some identical data elements along with the direct identifiers. This is essentially the method Dr. Sweeney used in the Illinois Cancer Registry case discussed earlier.[48] Among the technical changes, too, are new ways of disguising data to maintain their confidentiality, such as masking and multiple imputation, which are discussed in a later section.

The final two developments that have increased the tension between increasing access to research data and protecting confidentiality have already been discussed. Changes in the legal framework have placed a greater strain on the research system by increasing the risk of confidentiality breaches (e.g., Patriot Act,

---

[45] PANEL ON DATA ACCESS, *supra* note 1, at 60.

[46] Latanya Sweeney, *Information Explosion*, *in* CONFIDENTIALITY, DISCLOSURE, AND DATA ACCESS: THEORY AND PRACTICAL APPLICATIONS FOR STATISTICAL AGENCIES 43 (P. Doyle et al. eds., 2001).

[47] Experian, a credit bureau, maintains credit information on approximately 215 million U.S. consumers. *See* Experian Corporate Fact Sheet, http://www.experian.com/corporate/factsheet.html (last visited Dec. 3, 2005).

[48] *See supra* notes 11-12 and accompanying text.

Shelby Amendment, and Data Quality Act) while simultaneously increasing confidentiality protections (e.g., CIPSEA). Furthermore, increased public concerns about privacy and confidentiality have led to an increased reluctance to participate in statistical surveys or to provide Social Security Numbers, which are often used to link data from several surveys or to combine information from a survey and administrative records.

III. PROTECTING CONFIDENTIALITY

A variety of threats to the confidentiality of research data exist. Probably the most common is simple carelessness—not removing names, addresses, or telephone numbers from questionnaires or electronic data files, leaving cabinets unlocked, or not encrypting files containing identifying information. Increased access to research data is likely to heighten the risks stemming from carelessness and ignorance unless adequate precautions are taken.

Less common but potentially more serious threats to confidentiality are legal demands for identified data, either in the form of a subpoena or as a result of a FOIA request. Also of concern are instances of intrusion into government statistics by other government agencies for law enforcement purposes. Anderson and Seltzer, for example, have recently documented a number of attempts to use Census Bureau data for such purposes between 1910 and 1965.[49]

In addition to the legal attempts to obtain confidential information described above, confidentiality may also be breached as a result of illegal intrusions into the data. Such instances of identity theft have become more prominent in the news in the last ten years. For example, in 2005, the ChoicePoint Corporation, a data warehouse, was duped by thieves posing as businessmen into

---

[49] *See* Margo Anderson & William Seltzer, *The Challenges of "Taxation, Investigation, and Regulation:" Statistical Confidentiality and U.S. Federal Statistics, 1910-1965* (2004) (paper prepared for Census Bureau Symposium, America's Scorecard: The Historic Role of the Census in an Ever-Changing Nation, Woodrow Wilson International Center for Scholars, March 4-5, 2004), *available at* http://www.uwm.edu/~margo/govstat/Challenges.pdf.

selling hundreds of thousands of confidential records containing sensitive personal information.

A final threat to data confidentiality is the possibility of "statistical disclosure," which refers to the re-identification of individuals (or their attributes) as a result of an outsider's matching of survey data that has been stripped of explicit identifying information, such as names and addresses, with information available outside the survey. Although there are no known instances of the confidentiality of research data having been breached as a result of statistical disclosure, this is the risk that government data collection agencies and other survey organizations are currently most concerned about, and they are increasingly taking steps to protect against this possibility.

What can researchers do to protect data confidentiality against these threats? We discuss this under three headings: Development of norms and best practices; protections against legal intrusions; and protections against illegal intrusions.

A. Development of Norms and Best Practices

As noted above, the most likely reason for a breach of confidentiality involving research data is ignorance or carelessness. Laws and procedures designed to prevent confidentiality breaches and punish their occurrence are not enough to combat breaches caused by carelessness; they must be accompanied by internalized norms of research ethics and fair information practices, as well as practical knowledge of how to implement such policies.[50] The Panel on Data Access for Research Purposes has made three specific recommendations in this area: [51]

> Rec. 16: Statistical agencies and survey organizations that collect individually identifiable data should provide written guidelines for confidentiality protection and training in confidentiality practices and data management that guard

[50] For some practical suggestions along these lines, see R. M GROVES, ET AL., SURVEY METHODOLOGY, 358-59, 368-70 (2004).

[51] PANEL ON DATA ACCESS, *supra* note 1, at 82-84.

against disclosure for all staff who work with or have access to such data.

Rec. 18: Training in ethical issues related to research, including fair information practices, as well as principles and practices related to research with human subjects, should be part of the professional training of all those involved in the design, collection, distribution, and use of data obtained under pledges of confidentiality. Such training should be updated at intervals after the end of formal schooling.

Rec. 19: Professional associations should develop strong codes of ethical conduct that reflect the need to protect the confidentiality of personal data and make adherence to these codes an integral part of their educational activities.

B. Protections Against Legal Intrusions

An example of a legal intrusion into research data is a subpoena in a legal proceeding. Contingent valuation surveys, for example, value a public good, such as clean air or clean water, by asking respondents what they would be willing to pay for it. Such surveys are sometimes done in order to establish the damages that should be assessed in a man-made disaster, as in the case of the Exxon Valdez oil spill. Defendants in such cases have an interest in making sure that the survey was properly done, that the interviews were actually conducted, and that no distortions took place in the reporting of the results. At times, they have subpoenaed the actual survey records.[52] In one instance, the case was settled before the data were to be produced; in two other instances, the courts ordered the researchers to turn over raw data, including respondent identifiers, to the defendants.[53] Stanley

[52] For a review of several such cases, see Stanley Presser, *Informed Consent and Confidentiality in Survey Research,* 58 PUB. OPINION Q. 446, 446-59 (1994).

[53] Eliot Marshall, *Court Orders 'Sharing' of Data Science,* SCIENCE, July 16, 1993, at 284-86.

Presser, in his article *Informed Consent and Confidentiality in Survey Research,* discusses several actions that might be taken by professional survey organizations, such as the American Association for Public Opinion Research, to enhance protections of confidentiality in such cases—for example, mandating the destruction of identifiers in surveys designed for adversarial proceedings.[54] Another possibility is to verify the procedures used without disclosing the identity of the respondents, for example by allowing a disinterested survey expert to examine the interviews and the procedures to make sure that they meet professional standards.

FOIA requests are another source of legal intrusions into the data; an example of this is the case of the Illinois Cancer Registry, previously discussed. [55]

At times, law enforcement agencies have requested data collected under a promise of confidentiality by another agency. In 1917, for example, the Department of the Army asked the Census Bureau for information about men between the ages of 21 and 30 who were suspected of not having registered for the recently instituted draft. After a ruling by the Census Bureau's Solicitor General, the names of several hundred young men were turned over to the courts, draft boards, and the Justice Department by the Census Bureau, despite President Taft's proclamation of confidentiality prior to the decennial census.[56] In 1947, the

---

[54] For discussion on applicable judicial requirements, see Presser, *supra* note 52.

[55] S. Illinoisan v. Dep't of Pub. Health, 812 N.E.2d 27 (Ill. App. Ct. 2004). *See supra* note 8 and accompanying text.

[56] Anderson & Seltzer, *supra* note 49. As Anderson and Seltzer tell the story:

> [O]fficials in the Provost Marshal General's office and the local draft boards wrestled with administrative procedures to counter resistance to the draft registration and the draft. It soon became clear that the returns from the 1910 census could provide information to confirm the names, addresses and ages of individuals who might be suspected of not complying with the draft registration. The possibility of such requests frames the conflict between the commitment in Taft's proclamation [of confidentiality] and the requirements of modern war. Why should one

Attorney General made a similar request for information about suspected Communist sympathizers, but was denied.[57]

Two main defenses exist in law against the kinds of legal intrusions discussed above: Certificates of Confidentiality, and the recently enacted Confidential Information Protection and Statistical Efficiency Act of 2002.

Certificates of Confidentiality are issued prospectively by the Secretary of the Department of Health and Human Services to

---

> agency of government prevent another agency of government from doing its job? Why shouldn't the individual level data be made available to aid the war effort?
>
> However, in seeking to supply such information, the Census Director, Sam Rogers, knew he faced an obstacle. On June 22, 1917 he asked for guidance from the Secretary of Commerce and explained:
>
> > I have received numerous requests from registration officials in various parts of the United States to furnish them with information from the census records, showing the ages of men who they believe have failed to registered, although between the ages of 21 and 30.
>
> He also knew that Taft's proclamation guaranteed the individual data would not be used for enforcement purposes. Nevertheless, he saw a higher standard that outweighed the earlier pledge:
>
> > I believe that every branch of the Government, including this bureau, should assist at the present time, so far as possible, in securing a full registration. Accordingly, it is recommended that the matter be taken up with the President, with the view to having an order issued waiving the rigid rule laid down in Ex-President Taft's proclamation, and authorizing this Bureau to supply the proper officials (both registration and Federal) who are in control of the registration and prosecution of individuals who have failed to register, with data from the census schedules, which may show the ages of such individuals.
>
> On June 25, the Commerce Department Acting Solicitor issued the requested opinion. It gave the Census Director the authority to provide names and ages to the registration authorities . . . . Vincent Barabba noted in the 1970s that as a result of this opinion "personal information for several hundred young men was released to courts, draft boards, and the Justice Department."

Id. at 7-8.

[57] *Id.* at 29.

researchers on sensitive topics, such as sexual behavior or illegal drug use, whether federally funded or not.[58] They may also be issued by the National Institute of Justice for research supported by the Department of Justice. Under a Certificate of Confidentiality, a researcher may not be compelled "in any Federal, State, or local civil, criminal, administrative, legislative, or other proceedings to identify . . . the names or other identifying characteristics" of research participants.[59] The Certificate protects the identity of the respondents, not the data themselves. To the best of my knowledge, however, the effectiveness of a Certificate of Confidentiality has not been tested in the courts.

The other main legal protection for confidential information, CIPSEA,[60] has already been discussed. It is rooted in the distinction between administrative and statistical uses of data, and it protects information collected by statistical agencies or their agents under a pledge of confidentiality for exclusively statistical purposes from being disclosed in identifiable form except with the permission of the respondent.[61] Like the Certificate of Confidentiality, however, this law has not been tested in the courts.

C. Protections Against Illegal Intrusions

Identity Theft. The most visible example of illegal intrusions, by far, is identity theft, estimated in one survey to have affected more than 11 million Americans in 2003.[62] For example, as previously mentioned, confidential records at ChoicePoint, a data broker, were recently sold to thieves posing as legitimate businessmen. LexisNexis, another data broker, suffered similar

---

[58] Public Health Service Act § 301(d), 42 U.S.C. § 241(d), as amended by Pub. L. No. 100-607, sec. 163 (1988).

[59] Effect of Confidentiality Certificate, 42 CFR § 2a.7 (2006).

[60] *See* E-Government Act of 2002, 44 U.S.C. § 3501 (2006). *See supra* note 16 and accompanying text.

[61] *See* E-Government Act of 2002, 44 U.S.C. § 3501 (2006).

[62] Diane Hirte, *Identity Theft Numbers are Skyrocketing,* SHESHUNOFF.COM (Feb. 2003), *at* http://www.sheshunoff.com/email/archive/0203/oper_new1.html.

losses when thieves stole legitimate passwords and login names. ChoicePoint estimated that approximately 145,000 U.S. customers' files were compromised, and CBS News reported on March 5, 2005, that at least 750 people were defrauded as a result. LexisNexis said information on 32,000 U.S. customers was stolen.[63] Bank of America disclosed in February 2005 that it lost data tapes containing the Social Security Numbers and home addresses of the holders of 1.2 million government charge card accounts, and more recently Citibank "lost" data tapes containing similar information for some 3.9 million card holders when the tapes disappeared during transfer by UPS to a secure storage site.[64] A few days later, Mastercard International reported that more than 40 million credit card accounts of all brands might have been exposed to fraud through a computer security breach at a payment processing center.[65]

Improper Disclosure. Less visible, but equally disturbing, is the improper release of confidential data. For example, in a 1993 Harris telephone survey, 27% of respondents said that medical information about them had been improperly disclosed.[66] A recent Harris survey reported that only 14% of the sample gave the same response in 2005,[67] but drastic changes in the mode of the survey and the sample make comparison difficult.

In terms of the number of people whose confidential information was improperly disclosed, two airlines—Jet Blue and

---

[63] Paul Roberts, *Hackers Grab LexisNexis Info on 32,000 People*, PC WORLD (March 9, 2005), *available at* http://www.pcworld.com/news/article/0,aid,119953,00.asp.

[64] CNN MONEY, *Info on 3.9M Citigroup Customers Lost* (June 6, 2005) http://money.cnn.com/2005/06/06/news/fortune500/security_citigroup.

[65] Eric Dash & Tom Zeller, Jr., *Mastercard Says 40 Million Files Are Put at Risk*, N.Y. TIMES, June 18, 2005, at A1.

[66] Eleanor Singer, et al., *Privacy of Health Care Data: What Does the Public Know? How Much Do They Care?*, *in* HEALTH CARE AND INFORMATION ETHICS 401 (Audrey Chapman, ed., 1997).

[67] PRIVACY & AMERICAN BUSINESS, Conference (testimony of Dr. Alan F. Westin, Director of the Program on Information Technology, speaking on Privacy and Health Information Technology), Washington, D.C. (Feb. 23, 2004), http://www.pandab.org/WestinHHS.ppt#342,5ImproperDisclosureof.

Northwest—probably get the prize. In violation of their privacy policies, the two airlines disclosed confidential information, including names, addresses, and Social Security Numbers belonging to millions of their passengers, in order to help the National Aeronautics and Space Administration and the Pentagon develop systems designed to profile potential terrorists. The systems were subsequently said to have been abandoned because of privacy concerns.[68] In March, 2005, the Inspector General of the Department of Homeland Security blamed the Transportation Security Administration for failing to monitor adequately at least six airlines' transfer of sensitive passenger information to private companies and federal agencies in 2002 and 2003.[69]

Because none of these examples of illegal intrusion or improper disclosure involve research data, the concern shown by researchers and statistical agencies for protecting confidentiality may seem excessive. Nevertheless, the visibility of security breaches and other violations of privacy policies in the private sector have made the public more skittish about cooperating in legitimate research. Hence, government and academic researchers, who are dependent on the public's voluntary cooperation, take the protection of confidential data very seriously.

There are two main ways of protecting research data against illegal intrusion: one is to *restrict the data*—that is, to limit the detail of the information that is released in order to reduce its identifiability. The other is to *restrict access* to data whose original detail has not been altered. Neither of these, of course, protects against the other sources of confidentiality breaches discussed above: carelessness, legal intrusions, and improper disclosure. And so ironically, the very steps statistical agencies and other data collection organizations *can* take to protect confidentiality—which have the unintended consequence of making access to research data by legitimate researchers more difficult—are unlikely to

---

[68] Sara Kehaulani Goo, *Confidential Data Used for Air Security Project*, WASHINGTON POST, Jan. 17, 2005, available *at* http://www.washingtonpost.com/ac2/wp-dyn/A26037-2004Jan17?language=printer.

[69] Eric Lipton, *Agency Partly to Blame in Misuse of Passenger Data, Report Says*, N.Y. TIMES, Mar. 26, 2005, at 14.

reduce the confidentiality breaches that are known to occur most often.

1. Restricting Data

The first step in restricting data is to remove obvious identifiers—names, addresses, telephone numbers, Social Security numbers and any other information that *uniquely* identifies an individual. But removal of obvious identifiers is only the first step in what has come to be known as a "disclosure limitation" review.[70]

A review of data for purposes of disclosure limitation aims to eliminate from the data file those records with unique values on variables, as well as those with values that occur very infrequently. For example, a black female judge between the ages of 40 and 45 in Pinellas County, Florida would be readily identifiable in a survey even if her name and address were not part of the data record. More generally, date of birth, together with relatively small geography (such as county) and gender, is often enough to make identification possible if the data are matched against publicly available electronic files that contain names and addresses.[71] Such identifications are even easier if the intruder knows that the person is actually a part of the sample. Hence, date of birth and small area identifiers are routinely removed from data files intended for public release.

In addition to removal of direct identifiers, there are two main techniques for restricting data. One of these is called data masking; the other uses statistical techniques to create synthetic data that

---

[70] *See*, *e.g.*, SUBCOMM. ON DISCLOSURE LIMITATION METHODOLOGY, OFFICE OF MGMT. & BUDGET, EXECUTIVE OFFICE OF THE PRESIDENT, WORKING PAPER 22: REPORT ON STATISTICAL DISCLOSURE LIMITATION METHODOLOGY (1994).

[71] NAT'L RESEARCH COUNCIL, SUMMARY OF A WORKSHOP ON INFORMATION TECHNOLOGY RESEARCH FOR FEDERAL STATISTICS 36 (National Academy Press 2000), available at http://www.nap.edu/html/itr_federal_stats/ch2.html.

retain most of the properties of the original data set.[72]

Data Masking. The goal of data masking is to reduce the number of low-frequency cases in a data set (e.g., people earning more than $250,000 a year) and/or to create ambiguity about them. There are many ways of masking data. All of them make identification more difficult but also reduce the usefulness of the information disseminated. A general framework for analyzing the joint impact of various disclosure limitation techniques on disclosure risk and data utility is the risk-utility (R-U) confidentiality map,[73] which incorporates quantified measures of disclosure risk as well as measures of data utility.[74] More research, however, is needed on how to optimize the disclosure limitation-utility tradeoffs. *Synthetic Data.* The second technique for restricting data is to create an alternative data set by statistically modeling the original data records. In the most extreme version, none of the new records correspond to a real individual, and in that sense the statistically modeled "synthetic" data provide complete protection for confidential information, although they do not necessarily prevent an intruder from *believing* that he or she has identified a real person.[75]

Research on synthetic data creation and modeling is proceeding

---

[72] For excellent discussions of both of these techniques, see P. DOYLE, ET AL., CONFIDENTIALITY, DISCLOSURE, AND DATA ACCESS: THEORY AND PRACTICAL APPLICATIONS FOR STATISTICAL AGENCIES (2001); *Special Issue: Disclosure Limitation Methods for Protecting the Confidentiality of Statistical Data* 14, J. OF OFFICIAL STAT. (Stephen E. Fienberg & Leon C. R. J. Willenborg, eds., 1998). *See also* Jerome P. Reiter, *New Approaches to Data Dissemination: A Glimpse into the Future?*, 17 CHANCE 12 (2004).

[73] G.T. Duncan, et al., *Disclosure Risk vs. Data Utility through the R-U Confidentiality Map in Multivariate Settings* 1-2 (Working paper, 2003), http://www.heinz.cmu.edu/wpapers/retrievePDF?id=2005-16.

[74] *See* G.T. Duncan & D. Lambert, *Disclosure-Limited Data Dissemination*, 81 J. OF THE AM. STAT. ASS'N 10 (1986); G.T. Duncan & D. Lambert, *The Risk of Disclosure for Microdata*, 7 J. OF BUS. & ECON. STAT., 207 (1989); D. Lambert, *Measures of Disclosure Risk and Harm*, 9 J. OF OFFICIAL STAT. 313 (1993).

[75] D.B. Rubin, *Statistical Disclosure Limitation*, 9 J. OFFICIAL STAT. 461 (1993); T.E. Raghunathan, et al., *Multiple Imputation for Statistical Disclosure Limitation*, 19 J. OF OFFICIAL STAT. 1 (2003); Reiter, *supra* note 72.

rapidly. Researchers are interested in relationships among variables or attributes, not the identity of individuals who happen to exhibit those attributes. At present, synthetic data are capable of reproducing many of the relationships between data elements in the original observations—for example, those between education and occupation or education and income. But higher-order interactions—for example, among race, gender, education, and income—are modeled with less accuracy and precision. To improve the ability of synthetic data to reproduce accurately these complex relationships among variables, researchers are experimenting with imputing only some, rather than all, of the variables in the original data set. Understandably, such partially synthetic data sets are also subject to some, albeit small, disclosure risk.[76] To date, various studies of the usefulness of simulated data suggest that it is a promising approach for various kinds of inferential analysis.[77]

One advantage of synthetic data over data masking methods is their potential for estimating various sources of error. Some true relationship exists in the population between, say, education and income. The observed relationship in a given sample is just one random sample from that population distribution. Synthetic data created from the observed data can be thought of as simply another random sample from the population. Therefore, as Reiter puts it, "the user analyzing these synthetic samples is essentially analyzing alternative samples from the population,"[78] and the synthetic data on average will have similar characteristics as the observed data. Because this is true on average, and not for any particular synthetic data set, researchers using such data must use multiple data sets (hence, "multiple imputation") both to estimate the true values of the variables of interest and—equally important—to estimate the

---

[76] Reiter, *supra* note 72, at 16.

[77] *See*, *e.g.*, J.M. Abowd & S.D. Woodcock, *Disclosure Limitation in Longitudinal Linked Data*, *in* CONFIDENTIALITY, DISCLOSURE, AND DATA ACCESS: THEORY AND PRACTICAL APPLICATIONS FOR STATISTICAL AGENCIES 215-78 (Pat Doyle et al. eds., 2001).

[78] Reiter, *supra* note 72, at 15.

sources of error associated with them.[79]

2. Restricting Access

To cope with the loss of accuracy and precision that results from masking and imputation, research data are also made available in a variety of modes that retain the original identifying detail but restrict access to those who can meet certain criteria. Some of the ways in which confidential data are made available are (1) in special research data centers, to which researchers must travel;[80] (2) through licensing agreements, which permit access in the researcher's home institution;[81] and (3) through remote electronic access.[82] All of these modes of access stipulate that the researcher must meet certain requirements, which vary in their stringency. Most of these alternatives require approved research plans and license agreements, which permit researchers to work with the data at their own institution, require a data protection plan and, ordinarily, an agreement to permit auditing the researcher's adherence to this plan. Research data centers and remote access involve scrutiny of researchers' output for possible breaches of data confidentiality. All access modes require researchers to sign a confidentiality agreement, and they all entail penalties for the willful violation of such an agreement. NCES, for example, imposes a fine up to $250,000 and/or five years imprisonment for such a violation (a Class E felony), as does CIPSEA. The Health and Retirement Study's penalties for willful disclosure include

---

[79] *See* Reiter, *supra* note 72, at 15; Raghunathan, *supra* note 75, at 2-3.

[80] For example, the Census Bureau currently sponsors eight Research Data Centers (RDCs), with a ninth scheduled to open in late 2005. The National Center for Health Statistics (NCHS) and the Agency for Health Research and Quality each maintain one RDC, and the Bureau of Labor Statistics (BLS) maintains three. PANEL ON DATA ACCESS, *supra* note 1, at 29.

[81] Licensing was first established in 1989 by the NCES. *Id.* at 33. Other agencies that have licensing procedures include BLS and the Division of Science Resources Statistics of the National Science Foundation. *Id*.

[82] Monitored remote access is currently implemented in four federal statistical agencies: NCES, NCHS, the Census Bureau, and the Economic Research Service in the Department of Agriculture. *Id.* at 31.

forfeiture by the investigator—and possibly the investigator's institution—of all current federal funding, and denial of future funding by the sponsoring agency.[83]

Like methods of restricting data, which curtail the usefulness of the information made available, methods of restricting access also impose certain costs on the investigator. Gaining access to confidential data may involve inconvenience, delay, and financial costs, since remote access modalities as well as research data centers require the payment of a fee to defray the expense of maintaining the service or facility. In an effort to provide better access to such data, the Panel on Data Access for Research Purposes has proposed a number of recommendations designed to improve methods of restricting data, on the one hand, and of facilitating restricted access to detailed confidential data, on the other. For example, with respect to access, the panel has recommended that the Census Bureau broaden the interpretation of the criteria used to give researchers access to its confidential data;[84] that the statistical agencies sponsor research on cost-effective means of providing secure access through remote data access mechanisms;[85] that the use of licensing agreements for confidential data be expanded;[86] and that such agreements include provision for auditing compliance to security procedures and penalties for their violation.[87] With respect to improving methods of restricting data to limit its identifiability, the panel has recommended that agencies responsible for data collection should sponsor or conduct research on (1) developing measures for quantifying disclosure risk; (2) estimating the effect on disclosure risk of adding selected variables from confidential files to public use files; (3) estimating and improving the utility-disclosure limitation tradeoffs of alternative disclosure limitation methods, including synthetic data; and (4) developing disclosure limitation

---

[83] For further discussion of restricted access modes, see PANEL ON DATA ACCESS, *supra* note 1, at 31.

[84] *Id.* at 77 (Recommendation 9).

[85] *Id.* at 78 (Recommendation 10).

[86] *Id.* at 79 (Recommendation 11).

[87] *Id.* at 80 (Recommendation 13).

methods for establishment (business) data.[88]

CONCLUSION

This article has tried to communicate a number of points. First, there is an inherent tension between providing easy access to research data and protecting the confidentiality of those data. Second, increasingly, the courts will be called on to adjudicate the competing claims of those who want broader access, and those who want greater protection for privacy and confidentiality. Some of these claims may involve questions of national security, which will only complicate matters further. Third, a democratic society demands wide access to high-quality research data, but good data require the public's continued willingness to provide information, and to do so honestly and accurately. Fourth, such cooperation, in turn, depends on the public's confidence that their privacy and confidentiality will be respected and that they will not be harmed as a result of their voluntary cooperation in research. Finally, managing the tension between access to research data and protecting the confidentiality of those data requires recourse to technical, administrative, and legal solutions, some of them not yet invented. But even with the tools we have now, it is possible to do a great deal, if not everything, to satisfy both sides of the controversy.

[88] *Id.* at 72 (Recommendation 5).