8

Access and Confidentiality Issues with Administrative Data

Henry E. Brady, Susan A. Grand, M. Anne Powell, and Werner Schink

The passage of welfare reform in 1996 marked a significant shift in public policy for low-income families and children. The previous program, Aid to Families with Dependent Children (AFDC), provided open-ended cash assistance entitlements. The new program, Temporary Assistance for Needy Families (TANF), ended entitlements and provided a mandate to move adult recipients from welfare to work within strict time limits. This shift poses new challenges for both monitoring and evaluating TANF program strategies. Evaluating the full impact of welfare reform requires information about how TANF recipients use TANF, how they use other programs—such as child support enforcement, the Food Stamp Program, employment assistance, Medicaid, and child protective services—and how they fare once they enter the job market covered by the Unemployment Insurance (UI) system.

Administrative data gathered by these programs in the normal course of their operations can be used by researchers, policy analysts, and managers to measure and understand the overall results of the new service arrangements occasioned by welfare reform. Often these data are aggregated and made available as caseload statistics, average payments, and reports on services provided by geographic unit. These aggregate data are useful, but information at the individual and case levels from TANF and other programs is even more useful, especially if it is linked with several different sets of data so that the histories and experiences of people and families can be tracked across programs and over time. Making the best use of this individual level information will require major innovations in the techniques of data matching and linking for research and evaluation.

Even more challenging, however, are the complex questions about privacy and confidentiality that arise in using individual-level data. The underlying concern motivating these questions is the possibility of inappropriate disclosures of personal information that could adversely affect an individual or a family. Such fear is greatest with respect to disclosure of conditions that may lead to social stigma, such as unemployment, mental illness, or HIV infection.

In this paper we consider ways to facilitate researchers' access to administrative data collected about individuals and their families in the course of providing public benefits. In most cases, applicants to social welfare programs are required to disclose private information deemed essential to determining eligibility for those programs. Individuals who are otherwise eligible for services but who refuse to provide information may be denied those services. Most people forgo privacy in these circumstances; that is, they decide to provide personal information in order to obtain public benefits. They believe that they have little choice but to provide the requested information. Consequently, it is widely agreed that the uses of this information should be limited through confidentiality restrictions to avoid unwanted disclosures about the lives of those who receive government services.

Yet this information is crucial for evaluating the impacts of programs and for finding ways to improve them. Making the 1996 welfare reforms work, for example, requires that we know what happens to families as they use TANF, food stamps, the child support enforcement system, Medicaid, child protective services, and employment benefits such as the UI system. In this fiscally conservative political environment, many program administrators feel using administrative data from these programs is the only way to economically carry out the required program monitoring. Program administrators believe that they are being "asked to do more with less" and that administrative data are an inexpensive and reliable substitute for expensive survey and other primary data collection projects.

How, then, should we use administrative data? Guidance in thinking about the proper way to use them comes from other circumstances in which individuals are required to forgo a certain degree of privacy in order to collect important information. These situations include the decennial census, public health efforts to control the spread of communicable diseases, as well as the information collected on birth certificates. Underlying each of these situations is a determination that the need for obtaining, recording, and using the information outweighs the individual's privacy rights. At the same time, substantial efforts go into developing elaborate safeguards to prevent improper disclosures.

Administrators of public programs must, therefore, weigh the public benefits of collecting and using information versus the private harms that may occur from its disclosure. The crucial questions are the following: What data should be collected? Who should have access to it? Under what conditions should someone have access? Answering these questions always has been difficult, but the need

for answers was less urgent in the days of paper forms and files. Paper files made it difficult and costly to access information and to summarize it in a useful form. Inappropriate disclosure was difficult because of the inaccessibility of the forms. It was also unlikely because the forms were controlled directly by public servants with an interest in the protection of their clients.

Computer technology has both increased the demand for data by making it easier to get and increased the dangers of inappropriate disclosure because of the ease of transmitting digital information. Continued advances in computer technology are providing researchers and others with the capabilities to manipulate multiple data sets with hundreds of thousands (in some cases, millions) of individual records. These data sets allow for sophisticated and increasingly reliable evaluations of the outcomes of public programs, and nearly all evaluations of welfare reform involve the extensive use of administrative data. The benefits in terms of better programs and better program management could be substantial. At the same time, the linking of data sets necessitates access to individual-level data with personal identifiers or other characteristics, which leads to an increased risk of disclosure. Thus, the weighing of benefits versus harms must now contend with the possibilities of great benefits versus substantial harms.

The regulatory and legal framework for dealing with privacy and confidentiality has evolved enormously over the past 30 years to meet some of the challenges posed by computerization, but it has not dealt directly with the issues facing researchers and evaluators. There is a good deal of literature on the laws and regulations governing data sharing for program administration, much of which presupposes limiting access to these data for just program administration in order to avoid or at least limit unwanted disclosures. Unfortunately, little has been said in the literature regarding the use of such data for research and evaluation, particularly in circumstances where these analyses are carried out by researchers and others from "outside" organizations that have limited access to administrative data. Because research and evaluation capabilities generally are limited by tight staffing at all levels of government, researchers and evaluators from universities and private nonprofit research organizations are important resources for undertaking evaluations and research on social programs. Through their efforts, these organizations contribute to improving the administration of social welfare programs, but they are not directly involved in program administration. Therefore, these organizations may be prevented from obtaining administrative data by laws that only allow the data to be used for program administration.

The problem is even more complex when evaluations require the use of administrative data from other public programs (e.g., Medicaid, Food Stamp Program, UI) whose program managers are unable or unwilling to share data with social welfare program administrators, much less outside researchers. To undertake evaluations of social welfare programs, researchers often need to link individual-level information from multiple administrative data sets to understand

how people move from one situation, such as welfare, to another, such as work. But unlike program administrators, credit card companies, investigative agencies, or marketing firms, these researchers have no ultimate interest in the details of individual lives. They do, however, need to link data to provide the best possible evaluations of programs. Once this linking is complete, they typically expunge any information that can lead to direct identification of individuals, and their reports are concerned with aggregate relationships in which individuals are not identifiable. Moreover, these researchers have strong professional norms against revealing individual identities.

Problems arise, however, because the laws developed to protect confidentiality and to prevent disclosure do so by limiting access to administrative data to only those involved in program administration. Even though researchers can contribute to better program administration through their evaluations, they may be unable to obtain access to the data they need to evaluate a program.

Ironically, evaluations have become harder to undertake just as new policy initiatives—such as those embodied in federal welfare reform—require better and more extensive research to identify successful strategies for public programs. Evaluations have become more difficult because disclosures of individual information—fears driven by considerations having virtually nothing to do with research uses of the data—have led to legislation making it difficult to provide the kinds of evaluations that would be most useful to policy makers.

Against this background, this paper considers how researchers can meet the requirements for confidentiality while gaining greater access to administrative data. In the next section of the paper, we define administrative data, provide an overview of the concepts of privacy and confidentiality, and review current federal laws regarding privacy and confidentiality. We show that these laws have developed absent an understanding of the research uses of administrative data. Instead, the laws have focused on the uses of data for program administration where individual identities are essential, with lawmakers limiting the use of these data so that information about individuals is not used inappropriately. The result is a legal framework restricting the use of individual level information that fails to recognize that for some purposes, such as research, identities only have to be used at one step of the process for matching data and then can be removed from the data file.

After a relatively brief overview of the state regulatory framework for privacy and confidentiality in which we find a mélange of laws that generally mimic federal regulations, the paper turns to an extended discussion, based on information from a survey of 14 Assistant Secretary for Planning and Evaluation (ASPE)-funded welfare leavers studies, of how states have facilitated data matching and linkage for research despite the many obstacles they encountered. Based on our interviews with those performing studies that involve data matching, we identify and describe 12 principles that facilitate it. We show that states have found ways to make administrative data available to researchers, but these methods often are

ad hoc and depend heavily on the development of a trusting and long-term relationship between state agencies and outside researchers. We end by arguing that these fragile relationships need to be buttressed by a better legal framework and the development of technical methods such as data masking and institutional mechanisms such as research data centers that will facilitate responsible use of administrative data.

ADMINISTRATIVE DATA, CONFIDENTIALITY, AND PRIVACY: DEFINITIONS AND LEGAL FRAMEWORK

Administrative Data, Matched Data, and Data Linkage

Before defining privacy and confidentiality, it is useful to define what we mean by administrative data, matched data, and data sharing. Our primary concern is with administrative data for operating welfare programs—"all the information collected in the course of operating government programs that involve the poor and those at risk of needing public assistance" (Hotz et al., 1998:81). Although not all such information is computerized, more and more of it is, and our interest is with computerized data sets that typically consist of individual-level records with data elements recorded on them.

Records can be thought of as "forms" or "file folders" for each person, assistance unit, or action. For example, each record in Medicaid and UI benefit files is typically about one individual because eligibility and benefit provisions typically are decided at the individual level. Each record in TANF and Food Stamp Program files usually deals with an assistance unit or case that includes a number of individuals. Medicaid utilization and child protective services records typically deal with encounters in which the unit is a medical procedure, a doctor's visit, or the report of child abuse.

Records have information organized into data elements or fields. For individuals, the fields might be the name of the person, his or her programmatic status, income last month, age, sex, and amount of grant. For encounters, the information might be the diagnosis of an illness, the type and extent of child abuse, and the steps taken to solve the problem, which might include medical procedures or legal actions.

It is important to distinguish between statistical and administrative data. Statistical data are information collected or used for statistical purposes only. Data gathered by agencies such as the U.S. Census Bureau, Bureau of Labor Statistics, Bureau of Justice Statistics, and the National Center for Health Statistics is statistical data. Administrative data are information gathered in the course of screening and serving eligible individuals and groups. The data gathered by, for example, state and local welfare departments are an example of administrative data. Administrative data can be used for statistical purposes when they are

employed to describe or infer patterns, trends, and relationships for groups of respondents and not for directing or managing the delivery of services.

Administrative data, however, are used primarily for the day-to-day operation of a program, and they typically only include information necessary for current transactions. Consequently, they often lack historical information such as past program participation and facts about individuals, such as educational achievement that would be useful for statistical analysis. In the past, when welfare programs were concerned primarily with current eligibility determination, historical data were often purged and data from other programs were not linked to welfare records. Researchers who used these data to study welfare found that they had to link records at the individual or case level over time to develop histories of welfare receipt for people. In addition, to make these data even more useful, they found it was worthwhile to perform data matches with information from other programs such as UI wage data; vital statistics on births, deaths, and marriages: and program participation in Medicaid, the Food Stamp Program, and other public programs. Once this matching was completed, researchers expunged individual identities, and they analyzed the data to produce information about overall trends and tendencies. Matched files are powerful research tools because they allow researchers to determine how participation in welfare varies with the characteristics of recipients and over time. They also provide information on outcomes such as child maltreatment, employment, and health.

Matched administrative data are becoming more and more widely used in the evaluation and management of social programs. In February 1999, UC Berkeley's Data Archive and Technical Assistance completed a report to the Northwestern/ University of Chicago Joint Center for Poverty Research that provided an inventory of social service program administrative databases in 26 states[1] and an analysis of the efforts in these states to use administrative data for monitoring, evaluation, and research. Unlike other studies that have dealt with data sharing in general, this study was concerned primarily with the use of administrative data for research and policy analysis.

The UC study found that the use of administrative data for policy research was substantial and growing around the country. More than 100 administrative data-linking projects were identified in the study sample. Linkages were most common within public assistance programs (AFDC/TANF, Food Stamp Program, and Medicaid), but a majority of states also had projects linking public assistance data to Job Opportunities and Basic Skills, UI earnings, or child support data.

---

[1]The 26 states inventoried in the report included the 10 states with the largest populations plus a random selection of at least four states from the northeast, south, west, and midwestern regions of the nation. These states comprise four-fifths of the U.S. population and more than five-sixths of the welfare population. This report can be viewed at http://ucdata.berkeley.edu.

Approximately a third of the states had projects linking public assistance data to child care, foster care, or child protective services. Four-fifths of the states used outside researchers to conduct these studies, and about half of all the projects identified were performed outside of state agencies. The vast majority of projects were one time, but there is a small, and growing, trend toward ongoing efforts that link a number of programs.

Figure 8-1 indicates the likelihood of finding projects that linked data across eight programs. Programs that are closer on this diagram are more likely to have been linked. Arrows with percentages of linkage efforts are included between every pair of programs for which 35 percent or more of the states had linkage projects. Percentages inside the circles indicate the percentage of states with projects linking data within the program over time. AFDC/TANF, Food Stamp Program, and Medicaid eligibility are combined at the center of this diagram

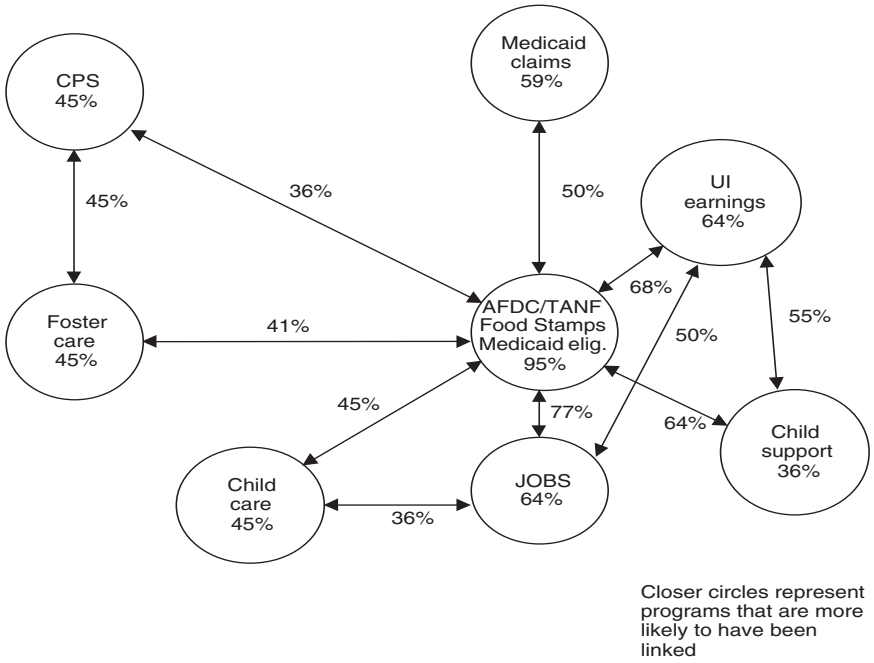

fig 8-1

FIGURE 8-1 Percent of states with projects linking data from social service programs. SOURCE: U.C. Data Archive and Technical Assistance (1999).

because they were the major focus of the study and because they are often combined into one system. The diagram clearly shows that there are many linkage projects across data sets from many different programs, frequently involving sensitive information.

Data Sharing

Matched data and data linkage should be distinguished from data sharing[2], which implies a more dynamic and active process of data interchange. Data sharing among agencies refers to methods whereby agencies can obtain access to one another's data about individuals, sometimes immediately but nearly always in a timely fashion. Data sharing offers a number of benefits. If different agencies collect similar data about the same person, the collection process is duplicative for both the agencies and the person. Data sharing therefore can increase efficiencies by reducing the paperwork burden for the government and the individual because basic information about clients only needs to be obtained once. Improved responsiveness is also possible. Data sharing enables agencies and researchers to go beyond individual program-specific interventions to design approaches that reflect the interactive nature of most human needs and problems, reaching beyond the jurisdiction of one program or agency. For example, providing adequate programs for children on welfare requires data about the children from educational, juvenile justice, and child welfare agencies. Data sharing is one way to ensure better delivery of public services and a "one-stop" approach for users of these services. Preis (1999) concluded, in his analysis of California efforts to establish integrated children's mental health programs, that data sharing is essential to good decision making and a prerequisite for service coordination. In fact, "if data cannot be exchanged freely among team members an optimal service and support plan cannot be created" (Preis, 1999:5).

Although data sharing has many benefits, it raises issues regarding privacy and confidentiality. Should data collected for one program be available to another? What are the dangers associated with having online information about participants in multiple programs? Who should have access to these data? How can confidentiality and privacy rights be protected while gaining the benefits of linking program data?

When agencies engage in data sharing, the technical problems of getting matched data for research and policy analysis are easily surmounted because information from a variety of programs is already linked. But matched and linked data sets for research and policy analysis can be created without data sharing, and data matching poses far fewer disclosure risks than data sharing because identifi-

[2]Note that we are using the term "data sharing" in a fashion that is much narrower than its colloquial meaning.

ers only need to be used at the time when data are merged. As soon as records are matched, the identifiers are no longer needed and can be removed. The merged data can be restricted to a small group of researchers, and procedures can be developed to prohibit any decisions from being made about individuals based on the data. Nevertheless, even data matching can lead to concerns about invasions of privacy and breaches of confidentiality.

Both data sharing and data matching require the careful consideration of privacy issues and techniques for safeguarding the confidentiality of individual level data. The starting place for understanding how to attend to these considerations is to review the body of law about privacy and confidentiality and the definitions of key concepts that have developed in the past few decades. After defining the concepts of privacy, disclosure, confidentiality, and informed consent, we then briefly review existing federal privacy and confidentiality laws.

Privacy

The right to privacy is the broadest framework for protecting personal information. Based on individual autonomy and the right to self-determination, privacy embodies the right to have beliefs, make decisions, and engage in behaviors limited only by the constraint that doing so does not interfere unreasonably with the rights of others. Privacy is also the right to be left alone and the right not to share personal information with others. Privacy, therefore, has to do with the control that individuals have over their lives and information about their lives.

Data collection can intrude on privacy by asking people to provide personal information about their lives. This intrusion itself can be considered a problem if it upsets people by asking highly personal questions that cause them anxiety or anguish. However, we are not concerned with that problem in this paper because we only deal with information that has already been collected for other purposes. The collection of this information may have been considered intrusive at the time, but our concern begins after the information has already been collected. We are concerned with the threat to privacy that comes from improper disclosure.

Disclosure

Disclosure varies according to the amount of personal information that is released about a person and to whom it is released. Personal information includes a broad range of things, but it is useful to distinguish among three kinds of information. *Unique identifiers* include name, Social Security number, telephone number, and address. This information is usually enough to identify a single individual or family. *Identifying attributes* include sex, birth date, age, ethnicity, race, residential address, occupation, education, and other data. Probabilistic matching techniques use these characteristics to match people across datasets when unique identifiers are not available or are insufficient for identification.

Birth date, sex, race, and location are often enough to match individual records from different databases with a high degree of certainty. Finally, there is information about *other attributes* that might include program participation status, disease status, income, opinions, and so on. In most, but not all cases, this information is not useful for identification or matching across data sets. But there are some instances, as with rare diseases, that this other information might identify a person. These three categories are not mutually exclusive, but they provide a useful starting place for thinking about information.

Identity disclosure occurs when someone is readily identifiable on a file, typically through unique identifiers. It can also occur if there are enough identifying characteristics. *Attribute disclosure* occurs when sensitive information about a person is released through a data file. *Inferential disclosure* occurs when "released data make it possible to infer the value of an attribute of a data subject more accurately than otherwise would have been possible" (National Research Council and Social Science Research Council, 1993:144). Almost any release of data leads to some inferential disclosure because some of the general facts about people are better known once the data are published. For example, when states publish their welfare caseloads, it immediately becomes possible to say something precise about the likelihood that a random person in the state will be on welfare. Consequently, it would be unrealistic to require "zero disclosure." "At best, the extent of disclosure can be controlled so that it is below some acceptable level" (Duncan and Lambert, 1986:10).

One fallback position might be to say that the publication of data should not lead to absolute certainty regarding some fact about a person. This would rule out the combination of identity and attribute disclosure to an unauthorized individual.[3] This approach, however, may allow for too much disclosure because data could be published indicating a high probability that a person has some characteristic. If this characteristic is a very personal matter, such as sexual orientation or income, then disclosure should be limited further.

Disclosure, then, is not all or nothing. At best it can be limited by making sure that the amount of information about any particular person never exceeds some threshold that is adjusted upward as the sensitivity of the information increases. In the past 20 years, statisticians have begun to develop ways to measure the amount of information that is disclosed by the publication of data (Fellegi, 1972; Cox, 1980; Duncan and Lambert, 1986). Many complexities have been identified. One is the issue of the proper baseline. If everyone knows some sensitive facts from other sources, should researchers be allowed to use a set of

[3]Bethlehem et al. (1990:38) define disclosure in this way when they say that "Identification is a prerequisite for disclosure. *Identification* of an individual takes place when a one-to-one relationship between a record in released statistical information and a specific individual can be established." It seems to us that this is a sufficient condition for improper disclosure to have occurred, but it is not clear that it is a necessary condition.

data that contains these facts? For example, if firms in some industry regularly publish their income, market share, and profit, should data files that contain this information be considered confidential? Another problem is the audience and its interest in the information. Disclosure of someone's past history to an investigative agency is far different from disclosure to a researcher with no interest in the individual. Finally, there is the issue of incremental risks. In many instances, hundreds and even tens of thousands of individuals are authorized to access administrative data. As such, access by researchers represents an incremental risk for which appropriate safeguards are available and practical.

Because disclosure is not all or nothing, we use the phrase "improper disclosure" throughout this paper.[4] Through this usage we mean to imply that disclosure is inevitable when data are used, and the proper goal of those concerned with confidentiality is not zero disclosure unless they intend to end all data collection and use. Rather, the proper goal is a balance between the harm from some disclosure and the benefits from making data available for improving people's lives.

Confidentiality

Confidentiality is strongly associated with the fundamental societal values of autonomy and privacy. One definition of confidentiality is that it is "a quality or condition accorded to information as an obligation not to transmit that information to an unauthorized party" (National Research Council and Social Science Research Council, 1993:22). This definition leaves unanswered the question of who defines an authorized party. Another definition of confidentiality is more explicit about who determines authorization. Confidentiality is the agreement, explicit or implicit, made between the data subject and the data collector regarding the extent to which access by others to personal information is allowed (National Research Council and Social Science Research Council, 1993:22). This definition suggests that the data subject and the data collector decide the rules of disclosure.

Confidentiality rules ensure that people's preferences are considered when deciding with whom data will be shared. They also serve a pragmatic function, encouraging participation in activities that involve the collection of sensitive information (e.g., medical information gathered as a part of receiving health care). Guarantees of confidentiality are also considered essential in encouraging

[4]Most of the literature on statistical data collection (e.g., National Research Council and Social Science Research Council, 1993) assumes that disclosure in and of itself is a bad thing. This presumption developed because most of this literature deals with a very specific situation where statistical agencies have collected data under the promise that they will not share it with anyone and where disclosure refers to information that can be readily attached to an individual. Because we deal with a much broader class of situations, we find it useful to distinguish between disclosure and improper disclosure where impropriety may vary with the circumstances of data collection and data use.

participation in potentially stigmatizing programs, such as mental health and substance abuse treatment services, and HIV screening programs.

Confidentiality limits with whom personal information can be shared, and confidentiality rules are generally found in program statutes and regulations. Varying levels of sensitivity are associated with different data. Accordingly, variations in privacy and confidentiality protections can be expected.

Confidentiality requires the development of some method whereby the limits on data disclosure can be determined. In most situations, the data collection organization (which may be a governmental agency) and the source of the information should be involved in determining this method. In addition, as the government, as the representative of the general public, has an obvious interest in regulating the use of confidential information. There are several ways that these parties can ensure confidentiality, including anonymity, informed consent, and notification.

Anonymity

Anonymity is an implicit agreement between an individual and a data collector based on the fact that no one can identify the individual. Privacy can be protected by not collecting identifying information so that respondents are anonymous. Anonymity is a strong guarantor of protection, but it is sometimes hard to achieve. As noted earlier, even without names, Social Security numbers, and other identifying information, individuals sometimes can be identified when enough of their characteristics are collected.

Informed Consent and Notification

The strongest form of explicit agreement between the data subject and the data collector regarding access to the personal information collected is informed consent. An underlying principle of informed consent is that it should be both informed and voluntary. In order for consent to be informed, the data subject must understand fully what information will be shared, with whom, how it will be used, and for how long the consent remains in effect. Consent requires that the subject indicate in some way that he or she agrees with the use of the information.

Consent can be written, verbal, or passive. Written consent occurs when a data subject reads and signs a statement written by the data collector that explains the ways information will be used. Verbal consent occurs when a data subject verbally agrees to either a written or verbal explanation of how information will be used. Verbal consent is often used when data subjects are contacted over the telephone, when they are illiterate, or when written consent might create a paper trail that might be harmful to the subject.

Passive informed consent is similar to, but distinct from, notification. Passive consent occurs when people have been notified about the intent to collect or

use data and told that their silence will be construed as consent. They can, however, object and prevent the collection or use of the data. With notification the elements of choice and agreement are absent. People are simply informed that data will be used for specified purposes. Notification may be more appropriate than informed consent when data provided for stated purposes are mandatory (such as information required for participation in a public program).

Some privacy advocates believe that conditioning program participation on the completion of blanket information release consent forms is not voluntary (Preis, 1999). Without choice, it is argued that the integrity of the client-provider relationship is compromised. As a result, many confidentiality statutes and regulations provide a notification mechanism so that the subjects of data being released can be informed of the release (e.g., Privacy Act), or they provide a mechanism for data subjects to decide who will be allowed access to their personal information (e.g., Chapter 509, California Statutes of 1998).

One of the difficulties facing data users in attempting to gain informed consent is that it is often very hard to describe the ultimate uses to which information will be put, and blanket descriptions such as "statistical purposes" are often considered too vague by those who regulate the use of data. It is also possible that data users may want to use the data for reasons not previously anticipated when the data were originally collected and, hence, not described when informed consent was initially granted from data providers. In such cases, data users may need to recontact data providers to see if providers are willing to waive confidentiality or data access provisions covering their data for the new uses of the data. However, the legality of these waivers is still being sorted out. See NRC (1993) for an example of a case where such waivers were not considered sufficient to cover the public release of collected data.

Confidentiality and Administrative Data

Administrative data are often collected with either no notification or some blanket notification about the uses to which the information will be put. As a result, legislatures and administrative agencies are left with the problem of determining the circumstances under which program participation records, drivers' license data, or school performance data should be considered private information and treated confidentially. One solution is to release only anonymous versions of these data through aggregation of the data or removal of identifying information. Anonymity, however, is not always feasible, especially when researchers want to link individual-level data across programs. In this case, should the collecting agency regulate the use of the information to ensure confidentiality when the individual has not been notified or has not provided informed consent? Can the government or some other regulatory body regulate the use of information and substitute for informed consent? What constitutes notification or informed consent? In the next section, we provide a quick overview of how the federal government has dealt with some of these issues.

FEDERAL PRIVACY AND CONFIDENTIALITY LAWS

Fair Information Practices

Several important bodies of federal law and regulation protect privacy and confidentiality of individuals served by one or more government programs, and about which government collects information. These laws reflect the Fair Information Practice Principles that were voluntarily developed and adopted by several government groups and privacy sector organizations in the 1970s. In 1973, the U.S. Department of Health, Education, and Welfare's (HEW's) Advisory Committee on Automated Personal Data Systems, Records, Computers and the Rights of Citizens published these principles in the report, *Records, Computers, and the Rights of Citizens*. These principles have served as the basis for formulation of the federal Privacy Act of 1974, the Freedom of Information Act, and subsequent federal laws and regulations. The Committee recommended five basic information principles for governing the use of personal information:

1. There must be no personal data record-keeping systems whose very existence is secret.
2. There must be a way for a person to find out what information about the person is in a record and how it is used.
3. There must be a way for a person to prevent information about the person that was obtained for one purpose from being used or made available for other purposes without the person's consent.
4. There must be a way for a person to correct or amend a record of identifiable information about the person.
5. Any organization creating, maintaining, using, or disseminating records of identifiable personal data must assure the reliability of the data for their intended use and must take precautions to prevent misuses of the data.

These principles were clearly developed to regulate situations where data would be used to learn about individuals or to make decisions about them.[5] Rules

[5]Other commissions and organizations developed similar codes of fair information practice that appear to limit severely the availability of data. Hotz et al. (1998) summarizes the common themes as follows:

- Promote openness.
- Provide for individual participation.
- Limit the collection of personal information.
- Encourage accurate, complete, and current information.
- Limit the use of information.
- Limit the disclosure of information.
- Ensure the information is secure.
- Provide a mechanism for accountability.

1, 2, and 4 require that individuals know about databases and can correct faulty information. These are important principles for agencies that collect information, but they have little relevance for researchers who want to use these data. Rules 3 and 5, however, propose strict ground rules for researchers' use of data. Under the strictest construction, they might require researchers to get prior consent from subjects for the use of administrative data. In reality, federal law has been somewhat less restrictive than this construction might imply.

Numerous federal privacy and confidentiality laws have been enacted in recent decades that elaborate on the Fair Information Practices. These include the Privacy Act of 1974 and the Data Matching and Privacy Protection Act of 1988.[6]

Privacy Act of 1974

The Privacy Act of 1974 was born out of the Watergate scandal in response to public outcry against the many invasions of privacy that occurred in that case. The concern was focused on the government's collection and disclosure of personal information. The Privacy Act places information disclosure limitations on the federal government, providing that certain records cannot be disclosed without the permission of the individual who is the subject of the record.

The act establishes certain responsibilities and conditions for information collection, maintenance, use, and dissemination. The information gathered must be relevant and necessary to the agency's mission. It should be collected directly from the individual to the greatest extent possible. The individual subjects of the data have to be informed of (1) the purpose of data collection, (2) whether participation in the collection of data is voluntary or mandatory, (3) the planned uses for the data, and (4) the consequences to an individual who does not provide the information.[7]

Third-party disclosure by a federal agency is also regulated by the Privacy Act. Data may be disclosed only when (1) the data subject has provided written consent authorizing the disclosure and (2) the disclosure in question is altogether exempted by the Act or it falls within an exception that allows for certain types of disclosures without consent.

The Fair Information Practices and the requirements of the Privacy Act of 1974 would seem to make research use impossible in the typical case where data are used by researchers in unanticipated ways after they have been collected and where contacting individuals at that point is nearly always prohibitively expensive. Research has, however, proceeded by using the "routine use" exemptions of

---

[6]Other important laws include the Freedom of Information Act (enacted in 1966), the Family Education Rights and Privacy Act of 1974, the Confidentiality of Alcohol and Drug Abuse Patient Records Act, the Right to Financial Privacy Act of 1978, and the Drivers Privacy Protection Act of 1994.

[7]U.S.C.S. §552a(e).

the Privacy Act and similar legislation that serve as the legal basis for disclosing information to a state agency that operates a parallel benefits program.[8] This exemption requires that (1) the use is compatible with the purposes for which the information was collected, and (2) the agency places notices about its information disclosure plans in the Federal Register and provides a 30-day opportunity for interested persons to comment on any new or intended use of the agency's data. The act also provides that consent is not required when the recipient of data provides the agency with written assurance that the data will be used solely as a statistical record and will be transferred in a form that is not identified individually.

The Privacy Act establishes limitations on what can be done with personal information collected by federal agencies, but the act itself is not the primary source of protection at the agency level. Separate federal laws and regulations have been promulgated that govern federally funded programs, and the provisions of the Privacy Act frequently have been included in them, thus extending its protections down to the state and local governments and other nongovernmental entities that administer and deliver these federally funded services. Thus, the Privacy Act provides a good starting place for understanding the legal issues associated with data sharing, but a thorough understanding requires examining informational privacy, confidentiality, and consent provisions for each specific federal program and agency.

Data Matching and Privacy Protection Act of 1998

In response to concerns about computer matching and perceived attempts by government agencies to circumvent the Privacy Act, Congress passed the Computer Matching and Privacy Protection Act of 1988 (and amendments to this new Act in 1990). Although no new standard is established by this Act, it creates procedural requirements for agencies that engage in computer matching. Matching agreement contracts are required between source and recipient agencies in a data-sharing program. The agreement must specify the purpose, justification, and procedures for the intended matching program. Although there are no criteria for determining when matching is appropriate, these agreements do provide notice and regulate the behavior of each party to the agreement. Matching agreements must describe the procedure by which applicants and recipients of federal assistance will be notified that information they provide may be subject to verification via a matching program. In addition, there must be procedures for verification and the opportunity of data subjects to contest findings.

The Computer Matching and Privacy Protection Act also adds new oversight provisions to the Privacy Act. Specifically, Data Integrity Boards are required for

[8]U.S.C.S. §552a(b)(3).

federal agencies that are involved in computer matching activities. These boards, composed of senior agency officials, have responsibility for reviewing matching agreements and programs for compliance with federal privacy laws. They also serve a clearinghouse and reporting function.

These acts and practices create a regulatory framework for the collection and use of data. For researchers, there are exemptions from requiring informed consent in which recipients did not give their consent when the data were collected initially. Agencies, for example, can forego informed consent when the use of the data is compatible with the purposes for which the information was collected and when the agency provides notice of its intentions in the Federal Register. They can also use data when the data will be used solely as a statistical record and will be transferred in a form that is not individually identifiable. In most cases, these procedures were not designed specifically to facilitate research, but they have been used for that purpose.

Common Rule—Institutional Review Boards

Concerns about the conduct of research have led to the development of Institutional Review Boards (IRBs) at universities, at government agencies, and at private organizations that conduct federally sponsored research involving human subjects. IRBs play an increasingly important role in the regulation of organizations that undertake social policy research using administrative data.

The federal "Common Rule," adopted in 1991, governs nearly all research involving human subjects that is conducted or supported by any federal department or agency.[9] Researchers and their institutions must comply with safeguards that ensure that individuals freely consent to participate in such research. Researchers also must ensure that the research employs procedures that are consistent with sound research design and that do not pose unnecessary risk to the research subjects. Finally, there must be adequate provisions to protect the privacy of research subjects and to maintain the confidentiality of individually identifiable private information.

The review of all federally funded research by IRBs is the principal mechanism by which these safeguards are implemented, and informed consent is the primary way that IRBs ensure that human subjects are protected. However, an IRB may waive some or all elements of informed consent under a number of circumstances.[10] Research involving the use of educational testing, surveys, and interviews is entirely exempt from review if individual identities cannot be established from the information so obtained. Research involving analysis of existing data is exempt if the information is either publicly available or recorded in a

[9]45 CFR Part 46.

[10]In an effort to simplify the complex regulations governing IRBs, we conflate waiver of informed consent (which does not necessarily mean exemption from IRB review) with exemptions.

manner such that individuals cannot be identified either directly or through identifiers linked to individuals. Also exempt from the rule is research that is designed to evaluate public benefit or service programs and that is conducted by or subject to the approval of federal department or agency heads. Finally, a waiver of informed consent may be given if the research involves no more than minimal risk to the subjects, the waiver will not adversely affect the rights and welfare of the subjects, and the research could not practicably be carried out without the waiver.

As with the Privacy Act, IRBs place a great emphasis on informed consent, although there are some provisions for waiving consent when anonymity can be assured, when risk is minimal, or when public benefit programs are being evaluated. The emphasis on informed consent is not surprising because IRBs were established initially to oversee medical research which often involves medical procedures. The need for informed consent regarding the procedure to be performed is obvious in this case because of the great potential for harm. Moreover, there may be no other way to protect subjects except through informed consent.

The role of informed consent is somewhat different in the conduct of most social science research, which involves acquiring information about subjects. It is possible, of course, to do harm through the collection of social science data by asking questions that provoke great anxiety or consternation, but the major danger is undoubtedly the possibility that private information will be revealed. In this case, confidentiality may be the primary concern, and some method for controlling the *use* of the data may be much more important than informed consent regarding its *collection*. Informed consent is one way to control the use of data, but it is not the only way. Anonymity potentially provides even better protection than informed consent. Other methods for protecting confidentiality also might provide the protections that are needed. For example, the confidentiality of administrative data might be protected without informed consent through the development of procedures such as the Data Integrity Boards and other mechanisms created by the Privacy Act and the Computer Matching and Privacy Protection Act. At the moment, however, IRBs rely heavily on informed consent, and they typically have only a limited understanding of the intricacies of matching administrative data and the laws regarding confidentiality.

Summary of Federal Legislation

Federal legislation has been built on a concern about disclosure of information about individuals. It has been done without much thought about the needs of researchers who only care about individual identities when they match data sets. At the moment, the federal regulatory environment for data is characterized by a multiplicity of laws, cross-cutting jurisdictions (e.g., Data Integrity Boards and IRBs), and some incoherence. The emphasis on informed consent in many laws would appear to limit severely the use of administrative data, but agencies have

used the provisions for statistical analysis and for "routine use" to allow researchers to use administrative data. All in all, the legal situation is highly ambiguous for researchers, and no one has come to grips with what should be done with data when informed consent is not possible and when researchers need identities solely for the interim stage of data matching.[11]

STATE PRIVACY AND CONFIDENTIALITY CONSIDERATIONS

It would be useful to conduct a state-by-state analysis of how privacy, confidentiality, and consent laws affect research and to compare the results with the impacts of federal laws and regulations. This analysis would contribute significantly to achieving a more complete and substantial understanding of how state and federal requirements interact with one another. However, this task is far beyond what we can do here. Instead, we make some comments based on the secondary literature.

State constitutional privacy protections are very diverse. For example, in California, privacy protections are expressly mentioned in the constitution, while Washington state's constitution requires that certain information—such as who receives welfare—be publicly available. In addition to state constitutional provisions regarding privacy and confidentiality, every state has enacted numerous privacy protection laws principally drafted in response to a specific perceived problem. The result is many narrow prescriptions, rather than a coherent statement of what information is private, when it can be collected, and how it can be used. Consequently, it is hard to know exactly what information is protected, and how it is protected. In addition, many privacy laws have exceptions and exemptions that make them hard to understand, hard to apply, and subject to divergent interpretations (Stevens, 1996). The resulting laws have been described as "reactive, ad-hoc, and confused" (Reidenberg and Gamet-Pol, 1995).

There are two broad classes of laws, those dealing with privacy in general and those that mention privacy and confidentiality in the process of establishing programs. The general privacy laws deal with computer crime, medical records, the use of Social Security numbers, access to arrest records, and other issues. Table 8-1 indicates the presence of general state privacy protections for the states in which there are ASPE welfare leavers studies (Smith, 1999).[12] It shows that state privacy laws cover a broad range of issues from arrest records to wiretaps,

[11]The recent National Academy Press publication, *Improving Access to and Confidentiality of Research Data* (National Research Council, 2000) is directed to this exact set of concerns.

[12]Basic information about state privacy laws in all states is available in *Compilation of State and Federal Privacy Laws* (Smith, 1999). We have focused on states with ASPE leavers studies to complement the survey described later.

TABLE 8-1 Privacy Laws in States with Welfare Leavers Studies

	AZ	CA	DC	FL	GA	IL	MD	MA	MO	NJ	NY	NC	OH	PA	SC	TX	WA	US
Arrest records	x	x	x	x	x	x	x	x	x	x	x		x	x	x		x	
Computer crime	x	x		x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Credit	x	x		x	x		x	x		x	x		x			x	x	x
Criminal justice	x	x		x	x	x	x	x		x		x	x		x		x	x
Employment		x	x	x	x	x	x	x		x	x	x	x	x			x	x
Government data banks	x	x	x	x		x	x	x	x	x	x	x	x	x	x		x	x
Insurance	x	x	x	x	x	x	x	x	x	x		x	x		x			x
Medical	x	x	x	x	x	x	x	x	x	x	x		x	x		x	x	x
Polygraphing	x	x	x		x	x	x	x			x			x	x	x	x	x
Privacy statutes	x	x		x	x	x		x		x	x	x		x	x	x	x	x
Privileges					x		x	x	x	x	x	x	x			x	x	
School records	x	x		x		x	x	x		x	x		x			x	x	x
Social Security numbers		x		x	x							x	x				x	x
Tax records	x				x		x	x			x	x	x		x		x	x
Testing				x			x	x				x	x	x			x	
Wiretaps	x	x	x	x	x	x	x	x		x	x	x	x	x		x	x	x
Miscellaneous		x		x		x		x			x							x

NOTE: An x indicates that the state law covers the subject (but not necessarily that the law affords a great deal of privacy protection).

and that some topics, such as arrest records, computer crime, medical records, and wiretaps, have led to more legislative activity by states than other topics such as the uses of Social Security numbers, credit information, or tax records. Moreover, some states, such as California, Florida, Maryland, Massachusetts, Ohio, and Washington, have laws that cover many more areas of concern than other states such as Missouri, South Carolina, or Texas. These laws affect researchers when they seek to utilize Social Security numbers for matching or to obtain school, arrest, or tax records.

Programmatic laws regulate the collection and uses of information as part of the social program's legislation at the federal and state levels. Harmon and Cogar (1998) found that federal program statutes and regulations provide substantial privacy protections similar to that in the federal Privacy Act. Explicit limits on disclosure within the statutes authorizing federal programs and agencies are common, as is the imposition of informational privacy protections on states via federal program regulations. Harmon and Cogar (1998) also found that—as with the provisions of the Privacy Act—federal regulations do not clearly specify penalties or the consequences of violating the regulations by state or local personnel or contractors. Their study of five states found state information privacy laws to be similar to federal protections.

Most of the state and federal laws regarding the collection and use of data for programs are quite restrictive, but they typically have a clause, similar to the "routine use" provisions in the federal Privacy Act, that allows agencies to use data to achieve the "program's purpose." Researchers and others who want access to the data use this clause in the same way as the "routine use" clause of the Privacy Act. Harmon and Cogar (1998) suggest that federal agencies often label their data uses as "routine" without determining if the use is consistent with the purpose for which the information was collected. Some state agencies follow a similar practice, although standards vary dramatically from state to state and agency to agency.

In their report about experiences in five states, "The Protection of Personal Information in Intergovernmental Data-Sharing Programs," Harmon and Cogar (1998) describe the complexity of the information protection provisions that apply to individuals under the U.S. Department of Agriculture (USDA) Food Stamp Program's Electronic Benefit Transfer (EBT) project and the HHS Child Support Enforcement Program's Federal Parent Locator Service/National Directory of New Hires project. None of the states reported major violations of privacy in the operation of the Child Support Enforcement and EBT programs, but the significant variation in regulation of information across the states could prove a significant barrier to the overall data-sharing responsibilities of the systems and for researchers who want to use the data. Moreover, most of the states, with the exception of Maryland, paid little heed to researchers' needs. Maryland's statutes specifically authorize public agencies to grant researchers access to personal information under specified conditions. This statute appears as Appendix 8-A as an example of model legislation that authorizes researcher access to data.[13]

UC Berkeley's Data Archive and Technical Assistance also explored confidentiality issues in its inventory (UC Data Archive and Technical Assistance, 1999) of social service administrative databases in 26 states. This study found that researchers and administrators from other programs who seek access to social service data must negotiate with the owners of the data, and they must demonstrate that they meet the legal criteria for access. Legislation and regulations were characterized as generally requiring the party petitioning for access to the data to identify: (1) the benefits associated with release of the data, (2) how the research will benefit administration of the programs, and (3) how confidentiality of the data will be protected from unauthorized disclosure.

In most cases, a formal contract or interagency agreement was required, and often these agreements are required because of legislative mandates. Apart from the legal issues of gaining access to confidential data, there are often coordination issues that affect the transfer of information from one agency to another. Only

[13]We also include Washington state's statute, which provides for researchers having access to administrative data.

about half of the states surveyed for this report had specific, well-outlined policies and procedures for sharing confidential administrative data.

The use of administrative data for research purposes has not been considered in the development of most federal and state legislation. The major purpose of most federal and state confidentiality and privacy legislation has been to regulate the use and disclosure of information about individuals.[14] As a result, a strict interpretation of most laws might preclude research uses that require data matching even though identifiers are removed before data analysis and researchers have no interest in individual information. This outcome would be mostly inadvertent. In their desire to protect individuals, lawmakers typically have written legislation that makes no distinction between research uses and disclosure of information about individuals. State and federal agencies sometimes have overcome restrictions on research by accommodating researchers through the use of the routine use and program purpose clauses. This accommodation is fitful and uncertain because it depends on each agency's interpretation of these clauses and its overall interest in allowing researcher access to administrative data.

ACCESS TO CONFIDENTIAL DATA IN PRACTICE: INTERVIEWS WITH RESEARCHERS CONDUCTING WELFARE LEAVERS STUDIES

The legal basis for the use of social program administrative data by nongovernmental researchers is ambiguous. Consequently, governmental agencies that are inclined to provide data to researchers usually can find a legal way to do so through a broad interpretation of the statutory "routine use" or "program purposes" clauses, while agencies that are inclined to block researcher uses can also do so by interpreting these clauses narrowly. From the research perspective, the best solution to this problem would be that privacy and confidentiality legislation take into account the significantly fewer risks posed by research uses of data and develop clearcut regulatory mechanisms tailored to the needs of researchers. We discuss this possibility later (Guiding Principle 12), but it is worth knowing that in the absence of a favorable regulatory environment, many researchers and program administrators have found ways to undertake research with administrative data. Because it may be difficult to get better legislation, the methods used by these program administrators and researchers deserve careful consideration.

To identify these methods, we interviewed researchers and state administrators working in federally funded welfare leavers projects. Because of the complexity of the lives of individuals leaving welfare, these studies require diverse

[14]Basic information about state privacy laws is available in a recent publication, *Compilation of State and Federal Privacy Laws* (Smith, 1999).

BOX 8-1
Welfare Leavers Studies: States/Localities Interviewed

- Arizona
- California (San Francisco Bay Area Counties; Los Angeles County)
- Florida
- Georgia
- Illinois
- Missouri
- New York
- Ohio (Cuyahoga County)
- Massachusetts
- South Carolina
- Texas
- Washington State
- Washington, D.C.
- Wisconsin

types of data, including multiple sources of confidential administrative data. In this section, we discuss information from 14 welfare leavers studies.[15] These include projects that received fiscal year 1998 ASPE grants to study the outcomes of individuals and families who left the TANF program, and Texas.[16] (We refer to this group of projects as "Welfare Leavers Studies".)

This research began by reviewing the findings from the inventory of research uses of social services administrative data in 26 states that UC DATA completed in 1999. A series of questions then was developed as the basis for telephone interviews with the state officials and researchers conducting ASPE-funded Welfare Leavers Studies. Officials and researchers working on these studies were queried about their experiences with confidentiality and data access. More than 20 individuals in the 14 locations listed in Box 8-1 were interviewed in winter 1999/2000.

In the course of our interviews with Welfare Leavers Studies representatives, we identified 12 guiding principles or practices we believe to be at the heart of successfully overcoming issues of data confidentiality and privacy. We found repeated examples of these principles or practices being put into action across the country in varying ways. They are listed in Box 8-2. The principles, the keys to data collaboration, fell naturally into four categories that are discussed in more detail later: the characteristics of the requesting organization, the characteristics of the organization providing the data, the characteristics of the requesting organization, the "contract" process itself, and the legal framework.

[15]Fall 1999.

[16]Texas was not an ASPE Fiscal Year 1998 welfare leavers study grantee.

BOX 8-2
Twelve Guiding Principles of Data Access and Confidentiality

The Characteristics of the Organization With the Data

1. Strong political or administrative leadership
2. Designation of a "Data Steward" in the department and structuring staffing levels and responsibilities to cover data access requests.
3. Develop a written confidentiality and security procedure—keep a catalog of written documents: contracts, memorandums of understanding (MOU's), personal security agreements.
4. The agency architecture encompasses all "providing" agencies as in "super agencies."
5. A central clearinghouse negotiates or assists in legal and/or technical issues.
6. Plan for data access in the development of information systems.

The Characteristics of the Requesting Organization

7. The reputation and integrity of the requesting organization engenders trust.
8. Trust between organizations, a history of working together, and strong personal relationships.
9. Develop a confidentiality/security procedure and keep a catalog of exemplary written contracts, MOUs, and personal security agreements.

The "Contract" Process

10. Put in writing mechanisms for monitoring confidentiality and security and for sanctioning breaches.
11. Congruence of research agency goals: demonstrated benefits to participating organizations.

The Legal or Statutory Authority

12. Statutory language authorizes or is broadly interpretable to authorize data access for researchers.

The specific principles range from the obvious—"Put Procedures and Contracts in Writing"—to the sublime—Find Strong Leadership." We discuss each of the principles in detail and give illustrative examples of these principles. See Table 8-2 for a complete listing of examples of the principles in the Welfare Leavers Study sites.

Data Access Principles Regarding the Organization with the Data

Principle 1: Strong Political or Administrative Leadership

We found that many new and established data-matching projects were successful because they had the interest or patronage of well-connected or inspiring

TABLE 8-2 Twelve Guiding Principles of Data Access and Confidentiality Examples from Interviews with Welfare Leavers Study Researchers (Fall 1999)

The Characteristics of Donor Organization	Examples
1. Strong leadership	California: California Department of Social Services (CDSS), Employment Development Department (EDD) Illinois Missouri: Governor Mel Carnahan, Missouri Training & Employment Commission New York: Federal Department of Labor Texas: Federal Department of Labor
2. Staff levels or responsibilities	California: Labor Market Information Division Illinois: Bureau of Program Design & Evaluation Missouri: "Administrative Data Guardian" Washington State: Office of Planning & Research Wisconsin: Data Stewardship
3. Written confidentiality/ security procedure	California Illinois: Dept. of Human Services Wisconsin: Data Stewardship
4. Agency architecture	Arizona: Arizona Department of Economic Security (ADES) Illinois: Dept. of Human Services
5. Central clearinghouse	Arizona: ADES Data Mart Florida: Florida Education & Training Program Placement Information Program Illinois: Chapin Hall, University of Chicago South Carolina: Budget & Contracts Board Texas: State Occupational Information Coordinating Committee Washington State: Internal Review Board
6. Plan for data sharing in development of information systems	California: Family Health Outcomes Project
The Characteristics of Requesting Organization	
7. Reputation and/or integrity	California: RAND Illinois: Chapin Hall, University of Illinois Massachusetts: Center for Survey Research at University of Massachusetts-Boston Ohio: Manpower Demonstration Research Program (MDRC)

TABLE 8-2 Continued

8. History of working together, personal relationships	California: UC Data & CDSS Georgia: Georgia State University & Department of Children and Family Services (DFCS) Illinois: Chapin Hall at University of Chicago, Illinois & Department of Children and Family Services, Department of Employment Security & Illinois Department of Human Services Missouri: University of Missouri & state agencies New York: Office of Transitional and Disability Assistance (OTDA) and Department of Labor (DOL) Ohio: Case Western University (CWRU) and Bureau of Employment Services (BES), CWRU and DSS, and CWRU and MDRC Washington, DC: Urban Institute & Department of Human Services
9. Written confidentiality/ security procedure	California: UC Data, RAND Ohio: Case Western Reserve University
The "Contract" Process	
10. Put in place mechanisms for monitoring confidentiality and security and/or sanctioning breaches. contracts in writing	California Georgia Illinois Missouri New York Ohio South Carolina Washington State Washington, DC Wisconsin
11. Congruence of research to agency goals– demonstrated benefits to participating organizations	Arizona California CalWORKs California Leavers Studies Florida Georgia Illinois Massachusetts Missouri New York Ohio South Carolina Washington State Washington, DC Wisconsin

continues

TABLE 8-2 Continued

The Legal or Statutory Authority	
12. Statutory language authorizes or is broadly interpreted to authorize data access	California Georgia Illinois Missouri New York Ohio South Carolina Washington State

leaders. This, in and of itself, comes as no surprise. However, the sources of this leadership are diverse.

In some cases, this leadership was political in nature. For example, the University of Missouri at Columbia Department of Economics began its long collaboration with the Missouri Department of Social Services at the request of Governor Mel Carnahan. In January 1997, the university was asked to begin an analysis of the workforce development system for the Governor's Training and Employment Council. Because of the high-profile support for this project, the agencies providing data were forthcoming so as not to appear to be hindering the effort. A governor's directive can be powerful.

Another example of political leadership can be found in the moving force behind the Texas State Occupational Information Coordinating Committee (SOICC). The SOICC was mandated by the U.S. Congress via the federal Job Training Partnership Act (JTPA) and the Carl D. Perkins Vocational Education Act of 1976. The Texas SOICC receives no state general revenue funding and is supported by the U.S. Department of Labor through the national network organization National Occupational Information Coordinating Committee.

Data linking is facilitated when those at the top make it clear that they want to know about the impacts governmental programs are having on clients. Governors can provide this kind of leadership. More commonly, and perhaps most effectively, this leadership can be found among program administrators, bureau chiefs, and agency heads. For example, California found valuable leadership in the California Department of Social Services (CDSS) Research Branch. Staff in the Research Branch made use of many years of experience in service to the state to forge data-sharing coalitions between CDSS and the California Employment Development Department. In Illinois, the decisions to link data were made by Department of Human Services administrators who were supporting the Welfare Leavers Study.

Principle 2: Designation of a "Data Steward" in the Department and Structuring Staffing Levels and Responsibilities to Cover Data Access Requests

Adequate staffing is essential for ironing out the issues of data access. Data-linking requests require extensive administrative and analytic effort. In fact, as the rapid growth of information technology makes privacy and security policies de rigueur, information security officers in many states are requiring the completion of more and more complicated data security and confidentiality procedures for data linking.

Information security offices are not solely responsible for the time and effort it takes to get a data-linking project approved. Each state department often requires approval by a contracts office, a legal office, and the program with the data. In addition, many projects are required to submit their project for review by the state's human subjects committee. Each of these approvals can take from a few days to a few weeks, or even months in some cases.

Success in data-linking projects requires staff dedicated to shepherding data requests through the complexities of confidentiality requirements and data access issues. Although lawyers are often assigned these tasks because of their knowledge of statutorily defined notions of confidentiality, experienced government staff with a research bent must be involved as well in order to explain the technical aspects of data linking. In fact, agency staff with a strong investment in data linking and a belief in the benefits of research can overcome exaggerated fears about data linking and overly narrow interpretations of the law.

A delicate balance must be reached here. The law regarding the use of administrative data is typically sufficiently ambiguous that beliefs about the usefulness of a research project, about the risks from data matching, and about the trustworthiness of researchers can determine the outcome of a data request. It is easy for lawyers to assume that research is not very useful, that the risks of data matching are great, and that researchers cannot be trusted with the data. Yet we found in our interviews that research staff believe data matching provides extraordinary opportunities for high-quality and relatively inexpensive evaluations. Moreover, researchers can make the case that the risks from data matching for research purposes typically are quite low—certainly much lower than the risks from many other kinds of data matching projects. What is needed is a balance of agency staff committed to both the appropriate protection of data and the appropriate sharing of data for research and evaluation. We were told in our interviews that there are plenty of staff people, legal and otherwise, who are zealously "protecting" data in the name of confidentiality, but there are not enough with strong investments in data linking and a belief in the benefits of research to their department to make the case for data matching.

Our interviews provide examples. One respondent in Missouri referred to himself as the administrative data "guardian." He saw himself as the data shep-

herd, the person who saw that the data got to where it needed to go and got there safely. He facilitated data access, safeguarded data confidentiality, and educated researchers about the complexities of the data. Other Missouri respondents reported this administrator to be knowledgeable and helpful. In the Washington State Department of Social and Health Services, staff in the Office of Planning and Research blazed new trails of data access through state divisions that were unfamiliar with, if not uncomfortable with, providing data to researchers. One respondent from Wisconsin reported an environment of data "stewardship" coming about in the state, an environment of making data available in a responsible manner. The California Employment Development Department, Labor Market Information Division has designated a Confidential Data Coordinator. In Illinois, the Bureau of Program Design and Evaluation in the Department of Human Services frequently negotiates data access arrangements.

Principle 3: Develop a Written Confidentiality and Security Procedure—Keep a Catalog of Written Documents: Contracts, Memorandums of Understanding (MOU's), Personal Security Agreements.

A written policy of confidentiality and security is a must. This document should make explicit the data security procedures required of the data requesting organizations by the agencies with the data. This written policy should include detailed standards to maintain the privacy of individual data subjects. Another necessary document is a written guideline to obtaining data. This document can be provided to data requesters to assist them in applying for access to confidential data. The confidentiality and security manual and the guideline to obtaining data can provide assurance to data-providing agencies that proper consideration will be given to maintaining the confidentiality of their data in advance of the data being requested of them. They will also reassure data-providing organizations that their staff will not waste precious staff time fielding fly-by-night data requests.

In addition to these documents, there should be an archive of exemplary memorandums of understanding, letters of understanding, contracts for goods and services, data access agreements, and confidentiality agreements for use among state agencies or between state agencies and nongovernmental organizations. These documents should have explicit sections on the maintenance of data security and confidentiality, similar to the protocol described. The archive should also contain statements regulating individual behaviors, commonly known as "personal security agreements" or "statements of confidentiality". These documents require each individual staff person on the project to acknowledge procedures required for maintaining confidentiality and penalties for a breach of these procedures. An archive promotes quick and thorough contract negotiations, and it avoids the nuisance of having to start from scratch with every data request.

The California Department of Social Services Research Branch has prepared two such model documents: "The CDSS Confidentiality and Security Policy" and "The Guidelines for the Preparation of A Protocol." Also, in the new environment of "Data Stewardship," Wisconsin is developing templates and exemplar agreements.

Principle 10, "Put in Writing Mechanisms for Monitoring Confidentiality and Security and for Sanctioning Breaches," discusses briefly which confidentiality and security procedures one might want to include in a contract and therefore in the archive of documents.

Principle 4: The Agency Architecture Encompasses All "Providing" Agencies as in "Super Agencies"

In some cases, a "super agency" organization can facilitate sharing of data among departments within the agency. For example, in response to the latest welfare reforms, some states combined state agencies under an umbrella organization. In most cases, administrative data are considered to be owned by this overarching agency. Although this does not eliminate the need for appropriate bureaucratic negotiation on data access, in most cases it makes the process easier.

One respondent referred to the Illinois Department of Human Services as a "super agency." The department handles data for AFDC/TANF, the Food Stamps Program, Substance Abuse, Mental Health, Special Supplemental Nutrition Program for Women, Infants, and Children (WIC) (family case management), Medicaid, and Child Care programs (and their data). Gaining access to some of these data was reported to be easier because of the "super agency" structure. It was reported that gaining access to data from Substance Abuse and WIC (family case management), although by no means easy, would have been even harder had not the agencies been part of this "super agency."

The Arizona Department of Economic Security (ADES) also can be considered a super agency. ADES covers a broad range of programs, including AFDC/TANF, Food Stamp Program, Medicaid, Child Welfare, Child Care, *and* Child Support Enforcement *and* Unemployment Insurance. A respondent reported that no interagency data access agreements were necessary with *any* of these programs because of this all-encompassing administrative structure.

Principle 5: A Central Clearinghouse Negotiates or Assists in Legal and/or Technical Issues

A centrally located institution or center can help facilitate data access. This center can be placed in the state government or outside, and it can serve a number of purposes.

First, a central organization can serve as a *data archive or data warehouse* that actually stores data from multiple state agencies, departments, and divisions.

In some cases, data archives match the data and provide data requesters with match-merged files. In other cases, data archives provide a place where data from multiple agencies are stored so that data requesters can obtain the data from one source and match the data themselves.

Second, a central organization can serve as a *data broker*. This organization does not actually store data from other agencies but "brokers" or "electronically mines" data from other agencies on an ad hoc or regular basis. This organization then performs analyses on these data and reports results back to the requesting agency. The data are stored only temporarily at the location of the data broker, before they are returned to the providing agency or destroyed.

Third, a central organization can serve another very important purpose, as a *clearinghouse for legal issues around confidentiality*. Organizations like this are sometimes called internal review boards. They maintain exemplar or template agreements, contracts, documents, as described earlier.

For example, the South Carolina Budget and Control Board (SC BCB) serves all three functions—data archive, data broker, and internal review board. The SC BCB plays a key role in the general management of state government. This institution is unique to South Carolina and oversees a broad array of central administrative and regulatory functions. In our interview with staff from the Welfare Leavers Study grantee in South Carolina, we learned of the office of Research and Statistics in the SC BCB. The office gathers, analyzes, and publishes data vital to the social, and economic well-being and health of residents of South Carolina. These data are used by other state agencies and by local governments to guide planning, management, and development decisions. The office also works with other agencies to prevent overlap and duplication of data-gathering activities. The Welfare Leavers Study grantee (South Carolina Department of Social Services) negotiated data access through the SC BCB and conducted their analysis inhouse. However, one South Carolina respondent noted that despite the central location of this clearinghouse, it was still necessary to obtain legal authorization to data access on an agency-by-agency basis.

The Arizona Department of Economic Security is in the process of building a data warehouse, referred to as the "data mart." The data mart will automatically receive and link data from all the programs covered by ADES. The Welfare Leavers Study researchers used this resource to access data. At this point, the data mart provides only data-archiving and data-matching functions. However, eventually the data mart will include front-end data analysis functions.

The Texas State Occupational Information Coordinating Committee (SOICC) serves as a data broker. SOICC does not archive or store data at all. Our respondent reported that SOICC "mines data electronically" from relevant agencies, conducts analysis, and provides requesters with results of these analyses.

In Florida, the Florida Education and Training Placement Information Program (FETPIP) serves a data brokerage role by archiving data and providing

analysis. However, our respondent reported that FETPIP did not archive data or provide analyses for the Florida Welfare Leavers Study grantee.

The Chapin Hall Center for Children at the University of Chicago has developed an extensive archive of child welfare and family welfare data. The center uses these data to assess the impacts of welfare reform and other programs on child well-being. Chapin Hall's archive of data on children's welfare is called the Integrated Database on Children's Services in Illinois (IDB). Built from administrative data collected over two decades by Illinois human services programs, the IDB allows researchers to create a comprehensive picture of the interactions children and their families have with social programs offered by the state. One respondent cited this database as an absolutely invaluable resource.

The University of Missouri at Columbia Department of Economics is another example of a center that archives and analyzes data. Here data from multiple state agencies are matched, merged, and analyzed. The archive contains data from five state agencies: the Department of Economic Development, the Department of Social Services, the Department of Labor and Industrial Relations, the Department of Elementary and Secondary Education, and the Department of Higher Education. The staff provide research and analysis for many of the separate agencies on an ad hoc and a contractual basis.

In Washington State, the Institutional Review Board serves a role as a central place to resolve legal issues of data access. The IRB assisted the Welfare Leavers Study grantee in ironing out legal issues. The IRB serves as a human subjects review board and maintains exemplar documents.

Principle 6: Plan for Data Access in the Development of Information Systems

It would be difficult to include all the requirements for the development of information systems in a single principle. The development of information systems requires a set of its own guiding principles, including, but not limited to, adoption of common identifiers and establishment of standardized data definitions.

Rather than try to list all of the relevant principles, we cite the following example from California: The Family Health Outcomes Project (FHOP). It is a joint project of the Department of Family and Community Medicine and the Institute for Health Policy Studies (both at the University of California at San Francisco). Initiated in 1992, FHOP is a planning and training effort to streamline and standardize the administrative aspects of state child and family health programs in California.

FHOP has developed an information structure for an extremely fragmented and difficult-to-access system—health care and health-related services for women and children in California. California has many categorical health and social

service programs serving women, children, and families. Each has a separate application and eligibility process, although all require similar application information. Clients must complete an application for each service they wish to receive, often at different times and in different locations. To bring these programs "together," FHOP has developed CATS, a "Common Application Transaction System." CATS addresses the need for a uniform, accessible application and eligibility determination process and provides aggregate data for state and local planning and management.

CATS is a methodology for integrating registration and eligibility determination across numerous state-funded family health programs. CATS establishes unique client identification through the use of core data elements (birth name, birth date, birth place, mother's first name, and gender) and confirmatory data elements (social security number, other client number, father's name, mother's maiden name, current name/client alias/nickname, county of client's residence, and zip code of client's residence). Utilizing probabilistic matching and relative weighting of the core data elements, CATS can uniquely identify clients and find duplicate records for the same client.

Health care providers can link local automated registration systems to the state CATS hub, which can then return eligibility and demographic information. The CATS goal is to simplify the eligibility process so that the necessary demographic and self-declared financial information need only be collected and entered once.

In summary, CATS includes a standardized approach to collecting demographic, race, ethnic, and financial eligibility information; standardized confidentiality procedures and informed consent for sharing information; information on client eligibility status for Medi-Cal, Family Planning, Healthy Families Children's Health Insurance Program (CHIP), and Children's Medical Services; methods for the discovery of duplicate client records for tracking and case management; and a secure Internet connection option for community clinics and private providers. By providing a common method for collecting information on participation in state child and family health programs, CATS makes it possible to identify clients across programs, track them over time, and monitor outcomes. From a researcher's perspective, systems such as CATS make matching data across data systems much simpler.

Data Access Principles That Have to Do With the Characteristics of Requesting Organization

Principle 7: The Reputation and Integrity of the Requesting Organization Engenders Trust

In many cases, we found that the reputation of the requesting agency was a major factor in successfully obtaining approval for the use of administrative data.

This reputation can be technical, academic, or professional. We found that some of our respondents were reassured by the sheer prominence of the requesting organization.

However, in most cases, feelings of confidence were firmly based on the earned substantive reputation of the requesting organization. Most of the examples we found were organizations that had established a reputation through extensive experience with similar types of research and therefore provided key expertise. For example, Chapin Hall has a well-deserved reputation for its extensive technical expertise in the complex issue of matching administrative data from child and family welfare systems. In fact, Chapin Hall's reputation is so great that the Illinois Department of Human Services believes that it could do no better than subcontract with Chapin Hall when doing any matching of children's and families' services data.

Another example comes from Massachusetts, where the Department of Transitional Assistance contracted with the Center for Survey Research at the University of Massachusetts, Boston, to field the survey of former TAFDC (Transitional Aid to Families with Dependent Children) households. The Department of Transitional Assistance provided the Center for Survey Research with confidential data necessary for developing a sample of welfare leavers. It was reported that the department chose the center in large part because of the center's local reputation for expertise and competence.

Principle 8: Trust Between Organizations, a History of Working Together, and Strong Personal Relationships

Of all the guiding principles, trust between organizations appears to make the most wide-ranging contribution to successful data access. In our interviews with Welfare Leavers Study grantees, and in discussions with other researchers and state and nongovernmental staff, we learned of countless longstanding relationships between departments, between organizations, and between individuals. These relationships played a large and very important role in establishing the trust and confidence necessary for smooth contract negotiation and productive collaboration in the Welfare Leavers Studies.

The separation of Principle 7, "Reputation," and Principle 8, "Trust," does not mean these two are mutually exclusive, but it is meant to imply they are somewhat different. Past projects may have been established because of the requesting organization's reputation, but future projects depend heavily on the development of trust. In many cases, the established association was continued because the projects went well. In some cases, however, it was reported that the past project was not entirely successful, but that the association was continued, it seems, merely based on personal friendships or the force of one or more personalities (not necessarily friendships). Whatever the case, these prior relationships were a major factor in the success of the majority of the data access efforts we

examined, including projects in California (San Francisco Bay area counties), Washington, DC, Georgia, Illinois, Missouri, New York, Ohio, and South Carolina.

Obviously, this phenomenon is not limited to welfare leavers projects. California, New York, Missouri, Arizona, and Texas respondents all reported knowledge of data access projects that were facilitated because of personal relationships. Indeed, it should be noted here that many of these longstanding relationships were the result of Principle 1, Strong Leadership.

Principle 9: Develop a Confidentiality/Security Procedure and Keep a Catalog of Exemplary Written Contracts, MOU's, and Personal Security Agreements.

This principle is parallel to Principle 3 except that it applies to the requesting organization. Every data-requesting organization should maintain a file of data access and confidentiality documents. Such a resource provides reassurance to the providing agency that the requester has given appropriate consideration to the issues of data access. In fact, one state administrator said they do not take seriously organizations that do not have a written procedure. Furthermore it allows the requesting agency to respond quickly to data access opportunities without having to reinvent the wheel. UC Data Archive and Technical Assistance at the University of California, at Berkeley has a Manual on Confidentiality and Security, which includes exemplar contracts, personal security agreements, and description of extensive data security procedures.

Principle 10 discusses briefly what confidentiality and security procedures one might want to include in a contract and therefore in the archive of documents.

Data Access Principles That Have to Do with the "Contract" Process

Principle 10: Put in Writing Mechanisms for Monitoring Confidentiality and Security and for Sanctioning Breaches

A contract between the requesting organization and the department providing the data makes accessing administrative data much easier. In a contract, confidentiality and security measures or requirements are clarified and put in writing. Written provisions to uphold confidentiality and security provide a vehicle for action if a breach of confidentiality occurs. Nearly all our respondents reported that their collaboration was governed by a written contract.

Contracts should include clauses that contractually provide for data security and maintenance of confidentiality. The following list provides examples of provisions that should be specified in any written contract governing access to confidential data. This list, although not intended to be exhaustive, illustrates most of

the procedures requested by state agencies for protecting the confidentiality of individuals in research projects using administrative microdata files:

- Prohibition on redisclosure or rerelease.
- Specification of electronic data transmission (e.g., encryption methods for network access).
- Description of storage and/or handling of paper copies of confidential data.
- Description of storage and/or handling of electronic media such as tapes or cartridges.
- Description of network security.
- Requirement for notification of security incidents.
- Description of methods of statistical disclosure limitation.
- Description of the disposition of data upon termination of contract.
- Penalties for breaches.

Furthermore, contracts should include references to statutes that provide for explicit sanctions of breaches of confidentiality. For example, California State Penal Code, Section 502, included in contracts, states that:

> …(c) Except as provided in subdivision (h), any person who commits any of the following acts is guilty of a public offense:
>
> (1) Knowingly accesses and without permission alters, damages, deletes, destroys, or otherwise uses any data, computer, computer system, or computer network in order to either (A) devise or execute any scheme or artifice to defraud, deceive, or extort, or (B) wrongfully control or obtain money, property, or data.
>
> (2) Knowingly accesses and without permission takes, copies, or makes use of any data from a computer, computer system, or computer network, or takes or copies any supporting documentation, whether existing or residing internal or external to a computer, computer system, or computer network.
>
> ….
>
> (4) Knowingly accesses and without permission adds, alters, damages, deletes, or destroys any data, computer software, or computer programs which reside or exist internal or external to a computer, computer system, or computer network.
>
> (5) Knowingly and without permission disrupts or causes the disruption of computer services or denies or causes the denial of computer services to an authorized user of a computer, computer system, or computer network.
>
> ….
>
> (d) (1) Any person who violates any of the provisions of paragraph (1), (2), (4), or (5) of subdivision (c) is punishable by a fine not exceeding ten thousand dollars ($10,000), or by imprisonment in the state prison for 16 months, or two

> or three years, or by both that fine and imprisonment, or by a fine not exceeding five thousand dollars ($5,000), or by imprisonment in a county jail not exceeding one year, or by both that fine and imprisonment.

All staff members who have access to the confidential data should sign a document agreeing to uphold the required confidentiality measures. This is sometimes called a "personal security agreement," a "confidentiality agreement," or a "disclosure penalty document." This agreement should notify the employee of the penalties for disclosure of the personal identities of the individuals of the data and requires that the employee acknowledge and understand the penalties. This task can be time consuming, but it is worth the effort. It is simplified if files of exemplar documents are maintained (Principles 3 and 9).

If money cannot flow between the requesting organization and the providing organization, then a no-cost contract can be put in place, which puts the requesting agency under the confidentiality constraints.

Principle 11: Congruence of Research Agency Goals: Demonstrated Benefits to Participating Organizations

Successful collaborations occur when all the parties perceive benefits for themselves. Requesting organizations should make sure that the goals of their research contribute to the goals of the organization providing the data. All our respondents reported that this was an important factor is easing the data access process. The importance of studying welfare leavers and the federal funding of the studies helped to facilitate data access. More generally, researchers find that they have greater access to data when there is obvious congruence between their research goals and the agency's need to comply with federal or state requirements, e.g., waiver demonstrations, reporting of performance measures, or completing of specified grant-related evaluations.

But the benefits are not always obvious and can come in many forms. For example, researchers can provide briefings, presentations, or technical assistance on special analyses to state administrators and staff on research completed with the administrative data. Researchers who have successfully obtained administrative data with confidential identifiers can return merged, cleaned, and enhanced databases to their state colleagues. As part of completing their research with the administrative data, researchers often clean and enhance the data. They may eliminate questionable outliers, identify likely biases, develop ways of dealing with the biases, and enhance the data by geocoding addresses. Researchers often develop high levels of expertise with certain types of administrative data. Sometimes researchers develop software applications to do their own analyses of the data which, if provided to the agency, would allow the agency to conduct their own analyses more efficiently. When this expertise comes back to the agency, in the form of briefing, technical assistance, software applications, or other format, the agency sees the benefits to them of sharing these data.

Also, researchers from the academic and nonprofit research fields can serve on and can often provide great benefit to the agencies' ad hoc or standing expert panels. These panels give guidance to the agencies on research methodologies, data analysis, software development, reports or products produced by contractors, development of information systems, public policies, technical administrative procedures, and legislative solutions.

Requesting organizations must seriously consider including services like these to the data-providing agencies in their contracts and requests to state departments. They not only provide state officials with something for the trouble of making data available, but they also provide proof to legislators and the general public that data access provides substantial public benefits.[17]

Data Access Principles That Have to Do with the Legal or Statutory Authority

Principle 12: Statutory Language Authorizes or Is Broadly Interpretable to Authorize Data Access for Researchers

As discussed earlier, lawmakers have written legislation that protects the privacy of individuals but makes no distinction between research uses and disclosure of personal information. State agencies sometimes have overcome the legislative restrictions by accommodating researchers through a broad interpretation of the statutory "routine use" and/or "program purposes" clauses.

In our interviews, we learned that many state agencies interpret evaluation and research to be an integral part of the performance of their duty. Respondents reported knowledge of statutory language that was being broadly interpreted to allow diverse data-linking projects, such as "administration of programs under this title," "eligibility determination," "performance of the agency's duty," "implementation of state policy," "routine use," "direct benefit to the public," and "research into employment and training programs." For example, one contract stated that data were "being shared pursuant to Section 1704 of the Unemployment Insurance Act which states in pertinent part that 'The Director shall take all appropriate steps to assist in the reduction and prevention of unemployment…and to promote the reemployment of unemployed workers throughout the State in every feasible way…" (820 ILCS 405/1704). It was reported that the New York Department of Labor has had access to welfare data and employment-related data for 5-6 years under statutorily approved language to "monitor employment and training programs."

---

[17]The U.S. Census Bureau "Research Data Centers Programs" is entirely based on the strong belief that researchers can help the Bureau improve the quality of its data, and researchers are required, by law and regulation, to develop strong rationales for why their work will improve census data.

Conclusions

In a very ambiguous and unclear legal environment, states nevertheless have found ways to provide researchers with data, but it is a difficult process requiring strong leadership, adequate staff, extensive negotiations over confidentiality and security, and trust between the data-requesting and data-providing organizations. It also requires that data-providing organizations believe that they are obtaining substantial benefits from providing their data to researchers. In some cases, the benefits follow because the state has contracted with the researchers, but in other cases researchers must find ways to convince agencies that their research will be helpful to the agency itself.

All in all, the situation for research uses of administrative data is precarious. The laws are unclear about whether data can be used for research. Agencies are only sometimes convinced that research is in their best interests. Coordinating and convincing many different agencies is a difficult task. An obvious solution would be to develop a better legal framework that would recognize the smaller risks of data disclosure from datalinking for research, but before this can be done, researchers have to develop a menu of technical and institutional solutions to the problems of data confidentiality.

TECHNICAL AND INSTITUTIONAL SOLUTIONS

There are two basic ways to limit disclosure, data alteration, and restricted access to data. The recent National Research Council (2000) report on "Improving Access to and Confidentiality of Research Data" notes the strengths and weaknesses of each method:

> 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 (29).

"Anonymizing" data by removing identifying information is one method of data alteration, but this procedure may not limit disclosure enough. Data alteration can be thought of as a more versatile and thorough collection of methods for reducing the risk of disclosure.

Requiring informed consent for the use of data can be thought of as an institutional method for restricting access, but it may be impractical or it may be inadequate in many cases. Once data have been collected in an administrative system, it is nearly impossible to go back and obtain informed consent, but perhaps more importantly, informed consent might not really serve the purposes of individuals who cannot easily judge the costs and benefits of the various ways data might be used. We discuss some institutional methods such as Confidential

Research Data Centers that can protect individual privacy and ensure confidentiality while making data available to researchers.

Data Alteration

Cross tabulations. One way to avoid unwanted disclosure is to present only aggregate data in the form of tables. In many cases, this amply limits disclosure, although at the cost of losing the analytical power that comes from being able to analyze individual-level data. Moreover, in some cases, the identification of individuals, families, firms, or other specific units can still be inferred from the tables themselves. One way to guard against this is to require a minimum number of reporting units, for example, five individuals in each cell of the table. This goal can be achieved starting with tables developed from unadjusted microdata through aggregation, suppression, random rounding, controlled rounding, and confidentiality edits (see Cox, 1980; Duncan and Pearson, 1991; Office of Management and Budget, 1994, 1999; Jabine, 1999; Kim and Winkler, no date).

Aggregation involves reducing the dimensionality of tables such that no individual cells violate the rules for minimum reporting. For example, data for small geography such as census block groups might be aggregated to census tracks for sparsely represented areas.

Suppression is the technique of not providing any estimate where cells are below a certain prespecified size. As row and column totals generally are provided in tabular data, there is a further requirement when suppressing cells to identify complementary cells that are also suppressed to ensure that suppressed data cannot be imputed. The identification of complementary cells and ensuring that suppressed cells cannot be imputed generally requires judgments of which potential complementary cells are least important from the vantage of data users. It also requires statistical analyses to ensure that suppressed cells cannot be estimated.

Random rounding is a technique whereby all cells are rounded to a certain level, such as to multiples of 5. The specific procedure provides that the probability for rounding up or down is established on the initial cells value. For example, the number 2 would not automatically be rounded to 0 but instead would be assigned a 60-percent probability of rounding down and a 40- percent probability of rounding up, and the final rounded value would be based on these random probabilities. Similarly, 14 would have an 80-percent probability of rounding to 15 and a 20-percent probability of rounding to 10. A problem with random rounding is that row and column cell totals will not necessarily equal reported actual totals.

Controlled rounding is a process using linear programming or other statistical techniques to adjust the value of rounded cells so that they equal published (actual) totals. Potential problems with this approach include (1) the need for

more sophisticated tools, (2) for some situations there may not be any solution, and (3) for large tables the process may be computationally intensive.

Confidentiality edit is a process whereby the original microdata are modified. One confidentiality edit procedure called "swapping" is to identify households in different communities that have a certain set of identical characteristics and swap their records. The Census Bureau used this procedure in developing some detailed tabulations of the 100-percent file. Another edit procedure called "blank and impute" involves selecting a small sample of records and blanking them out and refilling with imputed values.

Tables of magnitude data. An additional problem arises with magnitude data such as total employees or revenue for a firm. For example, where a single firm is dominant, the publication of data on the industry may allow a fairly accurate estimate of the firm's data. In this case rules need to be established, for instance that no single firm can account for more than 80 percent of the cell total, to provide protection. This rule can be generalized in the form of "no fewer than n (a small number) of firms can contribute more than k percent of the cell total." These rules are used to identify "sensitive cells" that require suppression. The process of suppression requires complementary suppression, as discussed.

Unfortunately, all of these methods lead to a loss of significant amounts of information. Published tables, because they generally only provide cross-tabulations of two or three data elements, often do not provide the precise analysis that a researcher needs, and they are usually not useful for multivariate analysis. In these cases, researchers need to obtain microdata.

Masking public use microdata[18] Although microdata provide extraordinary analytical advantages over aggregated data, they also pose substantial disclosure problems for two reasons. Microdata sets, by definition, include records containing information about individual people or organizations, and micro-datasets often include many data elements that could be used to identify individuals. Although it is very unlikely that an individual could be identified on a data set by age group, size category, the combination of these three items might be enough to identify at least some people (Bethlehem et al, 1990:40). In fact:

> In every microdata set containing 10 or more key variables, many persons can be identified by matching this file with another file containing the key and names and addresses (disclosure matching). Furthermore, response knowledge (i.e., knowing that the person is on the file) nearly always leads to identification (disclosure by response knowledge), even on a low-resolution key. Finally, analysis showed that on a key consisting of only two or three identifiers, a considerable number of persons are already unique in the sample, some of them "rare persons" and therefore also unique in the population" (p. 44).

[18]See "Report on Statistical Disclosure and Limitation Methodology" prepared by the Subcommittee on Disclosure Limitation Methodology and published by the Office of Management and Budget in 1994.

A variety of methods can be used to mask the identity of individuals or households in microdata, although it is harder to mask the identities of firms because of the small number of firms and the high skew of establishment size in most business sectors. Units can be masked by providing only sample data, not including obvious identifiers, limiting geographical detail, and limiting the number of data elements in the file. High-visibility elements can be masked by using top or bottom coding, recoding into intervals or rounding, adding noise, and swapping records.

- *Sampling* provides a means of creating uncertainty about the uniqueness of individuals or households.
- *Eliminating obvious identifiers* involves removing items such as name, address, and Social Security number or other variables that would allow for identification of individuals or households.
- *Limiting geographical detail* creates a greater pool and reduces the chance of identification of records with unique characteristics. For example, the Census Bureau restricted the geography for the Public Use Microdata Sample for the 1990 Census to areas with populations of at least 100,000.
- *Limiting the number of data elements* in a file reduces the probability that an individual can be uniquely identified.
- *Top and bottom coding* provide a means of eliminating disclosure risk. Top coding establishes an upper bound on continuous data, for example, 85 years and older would be coded as 85. Bottom coding is similar and might be used for old housing units.
- *Recoding into intervals and rounding* are a means of taking continuous data and grouping the data. In each case unique information can be modified to mask identity. For example, data of birth might be transformed into age groups.
- *Random noise* can be added to microdata by adding or multiplying values by a randomly determined factor. This process can be useful in preventing individuals from attempting to match the public use database with other databases where identity is known.
- *Swapping, blanking and imputing, and blurring* are techniques used to modify the original data but not significantly change the statistical properties of the database. Swapping is identifying matching records based on key fields and swapping the detailed data. Blanking and imputing is to blank out certain data on selected records and statistically impute new values. Blurring is to replace exact values with mean values of all records meeting certain profiles.

Many of these methods are now commonly used when microdata are released to the public. Researchers, however, worry that the loss of information from data alteration may make it difficult or even impossible to do many kinds of analysis, and some statisticians have suggested that these methods do not provide

sufficient disclosure protection (Bethlehem et al., 1990). These worries have led some to propose even more radical alterations of the data that would amount to creating "simulated data."

Simulated data can be created from the original microdata by using variants of imputation methods (see Rubin, 1987, 1993; Little and Rubin, 1987, Kennickell, 1997, 1998) to impute entirely new values of every variable for every case. The resulting data set is composed entirely of "made-up" people, and it may be possible to do analysis that is almost as good with these data as with the original information. Developing these methods is an active research area.

Some researchers, however, are wary of these methods, and in a recent seminar run by the Committee on National Statistics, Richard Suzman of the National Institute on Aging (NIA) reported that "all leading researchers currently supported by NIA are opposed to the imposition of synthetic data" (National Research Council, 2000:32). The solution may be to turn to institutional solutions, as suggested by Bethlehem et al. (1990:45):

> Therefore, if microdata are released under the conditions that the data may be used for statistical purposes only and that no matching procedures may be carried out at the individual level, any huge effort to identify and disclose clearly shows malicious intent. In view of the duty of a statistical office to disseminate statistical information, we think disclosure protection for this kind of malpractice could and should be taken care of by legal arrangements, and not by restrictions on the data to be released.

Institutional Methods for Restricted Access

If data alteration is not the final answer (and there is substantial disagreement about this given some of the technical possibilities), then some new institutional forms need to be created to protect confidentiality. Many approaches are possible, but we shall discuss two especially useful ones, research data centers and licensing combined with substantial penalties for misuse.

Research data centers. The U.S. Census Bureau has been working with other federal agencies for the past few years to create Census Research Data Centers (CRDCs) in several locations around the country (Boston, California, Chicago, Pittsburgh, and North Carolina) where researchers can go to work with nonpublic census data under strict supervision and after a stringent application process. The goal of the CRDCs is to improve the quality of census data by getting researchers to use the data in new ways that push the data to their limits. The centers are locked and are secure facilities where researchers can come to work on microdata, but only after they have developed a proposal indicating how their work will help to improve the data and signed a contract promising to meet all the obligations to protect it required of Census Bureau employees. Once they have passed these hurdles, they can work with the data in the CRDC facility, but they can only

remove output once it has undergone disclosure analysis from an on-site Census Bureau employee.

The CRDC model has worked well for some innovative projects, but it has its drawbacks. It is costly, requiring several hundred thousand dollars a year to cover space, equipment, the Census Bureau employee salary, and other needs, and it is not clear how these costs can be covered in the long run even though fees have been charged to researchers. Although the CRDCs have improved access for some researchers, others still must travel some distance to the nearest site. The approval process takes time, and the outcome is uncertain. Data availability often depends on the ability of Census Bureau employees to devote time to projects that may not be their first priority. There is some concern on the part of the Census Bureau about having microdata located away from the Census Bureau itself. Universities have concerns about storing confidential data on-site.

Despite these problems, something like these centers seems to be an inevitable result of researchers' desires for data and the confidentiality concerns of the governmental agencies that own the data. In our discussion of principle 5, "A central clearinghouse negotiates or assists in legal and technical issues," we noted that organizations such as the University of Chicago's Chapin Hall, the South Carolina Budget and Control Board, and the University of Missouri at Columbia's Department of Economics are developing variants of these centers. We can imagine many different approaches to these centers depending on where they are located (state governments or universities), how they are funded, how they determine access to data, and what types of responsibilities and limitations are placed on researchers.

Licensing and increased penalties for misuse—The great drawbacks of the RDC model are the costs and the need to travel to specific locations to do research. For some data sets, another approach might make more sense. Since 1991, the National Center for Educational Statistics (NCES) has issued nearly 500 licenses for researchers to use data from NCES surveys (National Research Council, 2000:44). As part of the licensing process, researchers must describe their research and justify the need for restricted data, identify those who will have access to the data, submit affidavits of nondisclosure signed by those with this access, prepare and execute a computer security plan, and sign a license agreement binding the institution to these requirements. Criminal penalties can be invoked for confidentiality violations. This model easily could be extended to other data, and it would work especially well for discouraging disclosure matching in cases where unique identifiers, but not all key identifiers, have been removed from the data.

Summary of Alternatives for Ensuring Confidentiality

Both data alteration and institutional restrictions hold promise for making data accessible while protecting confidentiality. Both approaches are still in their

infancy, and much needs to be learned. It is possible that combinations of the two will work best. Simulated data sets might be released to the public to allow researchers to learn about the data and to test preliminary hypotheses. When the researcher feels ready, he or she could go to a research data center for a relatively short period of time to finish the analysis.

SUMMARY AND RECOMMENDATIONS

Summary

Matching and linking administrative data can be a great boon to researchers and evaluators trying to understand the impacts of welfare reform, but researchers sometimes find that they cannot access administrative data because of concerns about individual privacy, the ambiguity of statutory authority, and agency fears about public scrutiny.

Concerns about individual privacy and the desire to protect confidential data have grown dramatically in the past decade. Data matching often raises the Orwellian threat of a big brother government that knows all about its citizens' lives. The result has been a welter of laws that have often reacted to the worst possibilities that can be imagined rather than to realistic threats. Researchers, we have argued, do not pose the worst threats to data confidentiality, but they have had to cope with laws that assume data users will try to identify individuals and use sensitive information in inappropriate ways. In fact, researchers have only a passing interest in individual identifiers and microlevel data. They want to be able to do analysis that employs the full power of individual level data and to link data using identifiers to create even more powerful data sets. But as researchers they have no interest in information about individuals.[19] At worst, researchers pose only a moderate risk of disclosure.

Nevertheless, agencies with data must deal with an ambiguous legal environment that makes it hard to know whether and under what circumstances information can be shared with another agency or with researchers. Many agencies are hesitant to share information because of the lack of clear-cut statutory authority about who can access and use data. Others prefer the current situation, viewing ambiguous laws as providing greater flexibility and latitude. The downside of this ambiguity is that much is left to the individual judgments of agency managers who must deal with fears of legislative and public scrutiny. Although providing greater access to information potentially increases public knowledge and understanding about the agency, this information may cause others to second-

[19]The exception is when researchers want to contact individuals listed in an administrative file. The human subject risks are greater here, and they require greater scrutiny.

guess the agency. The result is a skeptical and suspicious posture toward researchers' requests for data.

Overcoming these obstacles requires experience, leadership, the development of trust, and the availability of resources.[20]

Most data requesters and potential data providers are just beginning to gain experience with the rules governing research uses of administrative data. Most requesters are unfamiliar with the relevant laws and with agencies' concerns about confidentiality. Many agencies with administrative data have not had much experience with researchers, and they lack the relatively long time horizon required to wait for research to pay off. This is especially true of those parts of the agency that control administrative data. As a result, data requestors are impatient with procedures and find it hard to proceed. Agencies, faced with the unknown, delay providing data because they prefer to attend to their day-to-day problems. Leadership is essential for overcoming these problems.

Trust is also important. Trust may be hard to establish because of fears about how the data will be used and worries about whether the data will be protected against inappropriate disclosure. The "providing" agency must trust that the "receiver" will both protect confidentiality and not use the information in a way that compromises the basis on which the providing agency collected the information. The data provider also must believe it will receive some payoff for it from providing the data.

Even with experience, leadership, and trust, enough resources may not be available to overcome the many obstacles to providing data. Requesters may run out of steam as they encounter complicated requirements and seemingly endless meetings and negotiations. Providers may balk at the requester's requests for documentation and technical assistance in using the data. Adequate resources, also are essential for successful projects. There must be staff members who can help prepare data requests and the data themselves. There must be resources to fund the facilities (such as data archives or research data centers) that facilitate data access.

We found many instances where administrative data were used successfully, but the legal, technical, and institutional situation is parlous. Laws and regulations continue to be enacted with virtually no consideration of the needs of researchers. Technical advances offer some hope of making data available while protecting confidentiality, but technical advances such as the Internet and powerful computers also threaten data security. Institutional arrangements are precarious, often perched on nothing more than the leadership and trust developed by a few individuals.

---

[20]There are also technical obstacles to using administrative data, but we do not believe these are the major difficulties faced by most researchers. These obstacles include hardware and software incompatibility and lack of common standards. Fortunately, technological advances increasingly are addressing these issues, and they are less and less important compared to other difficulties.

Recommendations

Against this backdrop, our recommendations fall naturally into three categories: legal, technical, and institutional. Interestingly, in our interviews and in those reported in another study[21] we found differences of opinion about the proper set of prescriptions. One perspective is that the only way that data access will work is if there is a specific legislative mandate requiring it. Otherwise, it is argued, agencies will have no incentives to solve the many problems posed by efforts to make data more accessible. The other perspective suggests that just requiring public agencies to engage in making data available does not mean they will have the capacity or the ability to actually implement it. Rather, the priority should be on providing the tools and resources necessary to support research access to administrative data, with sparing use of statutory mandates. There seems to be some truth in both perspectives, and we make recommendations on both sides.

Legal Issues

Two sets of legal issues seem most pressing to us:

1. *Develop model state legislation allowing researchers to use administrative data.* Although we have some models for legislation that would help researchers gain access to data, we do not have a thoroughgoing legal analysis of what it would take to facilitate access while protecting confidentiality. We strongly suspect, for example, that such legislation must carefully distinguish research from other uses by developing a suitable definition of what is meant by research. In addition, it must describe how researchers could request data, who would decide whether they can have access, how data would be delivered to them, and how the data would be safeguarded. At the federal level, H.R. 2885, "The Statistical Efficiency Act of 1999" appears to provide an important means for improving researcher access to confidential data.

2. *Clarify the legal basis for research and matching with administrative data, with special attention to the role of informed consent and Institutional Review Boards*—Most of the projects using administrative data have relied on "routine use" and "program purposes" clauses to obtain access to the data, but IRBs prefer to base permissions to use data on informed consent, which is typically not obtained for administrative data. These approaches are somewhat at

[21]Landsbergen and Wolken (1998) interviewed officials in five states about barriers to establishing, maintaining and evaluating informational data sharing policies and practices. Although this study focused on data sharing and these five states' experiences with regard to environmental programs, the conclusions clearly extend to data access in other topical areas.

odds, and they have already started to collide in some circumstances where IRBs have been leery of allowing researchers access to data because of the lack of informed consent. Yet informed consent may not be the best way to protect administrative data because of the difficulty of ensuring that subjects are fully informed about the benefits and risks of using these data for research. At the same time, "routine use" and "program purpose" clauses may not be the best vehicle either. Some innovative legal thinking about these issues would be useful. This thinking might provide the basis for implementing our first recommendation.

Technical Issues

New techniques may make it easier to protect data making the data accessible to researchers:

3. *Develop better methods for data alteration, especially "simulated" data*. Although there are differences of opinion about the usefulness of simulated data, there is general agreement that simulated data would at least help researchers get a "feel" for a data set before they go to the time and trouble of gaining access to a confidential version. It would be very useful to develop a simulated dataset for some state administrative data, then see how useful the data are for researchers and how successfully they protect confidentiality.

4. *Develop "thin-clients" that would allow researchers access to secure sites where research with confidential data could be conducted.* Another model for protecting data is to provide access through terminals—called "thin-clients"—that are linked to special servers where confidential data reside. The linkages would provide strong password protection, and ongoing monitoring of data usage. All data would reside on the server, and the software would only allow certain kinds of analysis. As a result, agencies would have an ongoing record of who accessed what data, and they would be able to block some forms of sensitive analysis such as disclosure matching.

Institutional Issues

The primary lesson of our interviews with those doing Welfare Leavers Studies is that institutional factors can contribute enormously to the success or failure of an effort to use administrative data:

5. *Support agency staff who can make the case for research uses of administrative data*. There is a large and growing infrastructure to protect data, but there is no corresponding effort to support staff who can make the case for research uses of administrative data. Without such staff, agencies may find it much easier to reject data requests, even when they are justified on legal and practical grounds.

6. *Support the creation of state data archives and data brokers who can facilitate access to administrative data.* One way to get a critical mass of people who can help researchers is to develop data archives and data brokers whose job is to collect data and make the data available within the agency and to outside researchers. In our presentation of Data Access Principle 5, we described several models for what might be done to create central clearinghouses that negotiate and assist in legal and technical issues related to data access. A *data archive or data warehouse* stores data from multiple state agencies, departments, and divisions. In some cases, an archive matches the data and provides data requesters with match-merged files. In other cases, data archives provide a place where data from multiple agencies are stored so that data requesters can obtain the data from one source and match it for themselves. *Data brokers* do not actually store data from other agencies but "brokers" or "electronically mines" data from other agencies on an ad hoc or regular basis. These organizations then perform analyses on the data and report results back to the requesting agency. The data are stored only temporarily at the location of the data broker, before being returned to the providing agency or destroyed.

7. *Support the creation of university-based research data centers.* Another model worth exploring is university-based research data centers modeled after the Census Bureau's Research Data Centers. These centers, located around the country, provide a site where researchers can use nonpublic Census data to improve the quality of census data by getting researchers to evaluate new ways to push the data to their limits. The centers are locked and secure facilities where researchers can come to work on microdata, but only after they have developed a proposal indicating how their work will help to improve the data and signed a contract promising to meet all the obligations to protect it required of Census Bureau employees. Once they have passed these hurdles, they can work with the data in the CRDC facility, but they can only remove output once it has undergone disclosure analysis from an on-site Census Bureau employee. A similar model could be developed for administrative data.

8. *Use contract law to provide licenses and criminal and civil law to provide penalties for misuse of data.* Licensing arrangements would allow researchers to use data at their own workplace. Researchers would describe their research and justify the need for restricted data, identify those who will have access to the data, submit affidavits of nondisclosure signed by those with this access, prepare and execute a computer security plan, and sign a license agreement binding themselves to these requirements. Criminal penalties could be invoked for confidentiality violations. This model would work especially well for discouraging matching in cases where unique identifiers, but not all key identifiers, have been removed from the data.

REFERENCES

Bethlehem, J.G., W.J. Keller, and J. Pannekoek
1990 Disclosure control of microdata. *Journal of the American Statistical Association* 85(March):38-45.

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

Duncan, G.T., and D. Lambert
1986 Disclosure-limited data dissemination. *Journal of the American Statistical Association* 18 (March):10-18.

Duncan, G.T., and R.W. Pearson
1991 Enhancing access to microdata while protecting confidentiality: Prospects for the future. *Statistical Science* 6(August):219-232.

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

Harmon, J.K., and R. N. Cogar
1998 *The Protection of Personal Information in Intergovernmental Data-Sharing Programs: A Four-Part Report on Informational Privacy Issues in Intergovernmental Programs*. Electronic Commerce, Law, and Information Policy Strategies, Ohio Supercomputer Center, Columbus, OH, June.

Hotz, V. Joseph, Robert George, Julie Balzekas, and Francis Margolin.
1998 *Administrative Data for Policy-Relevant Research: Assessment of Current Utility and Recommendations for Development*. Chicago: Joint Center for Poverty Research.

Jabine, Thomas B.
1999 Procedures for restricted data access. *Journal of Official Statistics* 9(2):537-589.

Kennickell, Arthur B.
1997 *Multiple Imputation in the Survey of Consumer Finances*. Washington, DC: Federal Reserve Bank.
1998 Multiple Imputation in the Survey of Consumer Finances. Unpublished paper Prepared for the Joint Statistical Meetings, Dallas, Texas.

Kim, Jay J., and W.E. Winkler
no date Masking Microdata Files. Unpublished Bureau of the Census discussion paper.

Landsbergen, D., and G. Wolken
1998 *Eliminating Legal and Policy Barriers to Interoperable Government Systems*. Electronic Commerce, Law, and Information Policy Strategies, Ohio Supercomputer Center, Columbus, OH.

Little, Roderick, and Donald B. Rubin
1987 *Statistical Analysis with Missing Data*. New York. John Wiley and Sons.

National Research Council
2000 *Improving Access to and Confidentiality of Research Data: Report of a Workshop,* Christopher Mackie and Norman Bradburn, eds. Commission on Behavioral and Social Sciences and Education, Committee on National Statistics. 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. G.T. Duncan, T.B. Jabine, and V.A. de Wolf, eds. Commission on Behavioral and Social Sciences and Education, Committee on National Statistics. Washington, DC: National Academy Press.

Office of Management and Budget
1994 Report on Statistical Disclosure and Limitation Methodology. Statistical Policy Working Paper 22. Prepared by the Subcommittee on Disclosure Limitation Methodology, Federal Committee on Statistical Methodology, May.
1999 Checklist on Disclosure Potential of Proposed Data Releases. Prepared by the Interagency Confidentiality and Data Access Group: An Interest Group of the Federal Committee on Statistical Methodology, July.

Preis, James
1999 *Confidentiality: A Manual for the Exchange of Information in a California Integrated Children's Services Program.* Sacramento: California Institute for Mental Health.

Reamer, F.G.
1979 Protecting research subjects and unintended consequences: The effect of guarantees of confidentiality. *Public Opinion Quarterly* 43(4):497-506.

Reidenberg and Gamet-Poll
1995 The fundamental role of privacy and confidence in the network. *Wake Forest Law Review* 30(105).

Rubin, Donald B.
1987 *Multiple Imputation for Nonresponse in Surveys*. New York: John Wiley and Sons.
1993 Discussion of statistical disclosure limitation. *Journal of Official Statistics* 9(2):461-468.

Smith, R. E.
1999 *Compilation of State and Federal Privacy Laws with 1999 Supplement*. Providence: Privacy Journal.

Stevens, D.
1996 *Toward an All Purpose Confidentiality Agreement: Issues and Proposed Language*. Baltimore, MD: University of Baltimore.

UC Data Archive and Technical Assistance
1999 *An Inventory of Research Uses of Administrative Data in Social Service Programs in the United States 1998*. Chicago: Joint Center for Poverty Research.

U.S. Department of Health, Education, and Welfare, Advisory Committee on Automated Personal Data Systems, Records, Computers and the Rights of Citizens
1973 *Records, Computers, and the Rights of Citizens*. Washington, DC: U.S. Department of Health, Education, and Welfare.

APPENDIX 8-A

State Statutes Providing Researcher Access to Data

MARYLAND:

This Maryland statute is a model for what might be done in other states.

Government Code. §10-624. Personal records

(c) Access for research.—The official custodian may permit inspection of personal records for which inspection otherwise is not authorized by a person who is engaged in a research project if:

(1) the researcher submits to the official custodian a written request that:

(i) describes the purpose of the research project;

(ii) describes the intent, if any, to publish the findings;

(iii) describes the nature of the requested personal records;

(iv) describes the safeguards that the researcher would take to protect the identity of the persons in interest; and

(v) states that persons in interest will not be contacted unless the official custodian approves and monitors the contact;

(2) the official custodian is satisfied that the proposed safeguards will prevent the disclosure of the identity of persons in interest; and

(3) the researcher makes an agreement with the unit or instrumentality that:

(i) defines the scope of the research project;

(ii) sets out the safeguards for protecting the identity of the persons in interest; and

(iii) states that a breach of any condition of the agreement is a breach of contract.

WASHINGTON:

The following statute from Washington state also provides language for model legislation that authorizes researcher access to data.

Revised Code of Washington (RCW). Chapter 42.48. Release of Records for Research

RCW 42.48.010 Definitions.

For the purposes of this chapter, the following definitions apply:

(1) "Individually identifiable" means that a record contains information which reveals or can likely be associated with the identity of the person or persons to whom the record pertains.

(2) "Legally authorized representative" means a person legally authorized to give consent for the disclosure of personal records on behalf of a minor or a legally incompetent adult.

(3) "Personal record" means any information obtained or maintained by a state agency which refers to a person and which is declared exempt from public disclosure, confidential, or privileged under state or federal law.

(4) "Research" means a planned and systematic sociological, psychological, epidemiological, biomedical, or other scientific investigation carried out by a state agency, by a scientific research professional associated with a bona fide scientific research organization, or by a graduate student currently enrolled in an advanced academic degree curriculum, with an objective to contribute to scientific knowledge, the solution of social and health problems, or the evaluation of public benefit and service programs.

This definition excludes methods of record analysis and data collection that are subjective, do not permit replication, and are not designed to yield reliable and valid results.

(5) "Research record" means an item or grouping of information obtained for the purpose of research from or about a person or extracted for the purpose of research from a personal record.

(6) "State agency" means: (a) The department of social and health services; (b) the department of corrections; (c) an institution of higher education as defined in RCW 28B.10.016; or (d) the department of health.
[1989 1st ex.s. c 9 § 207; 1985 c 334 § 1.] NOTES: Effective date — Severability — 1989 1st ex.s. c 9: See RCW 43.70.910 and 43.70.920.

RCW 42.48.020 Access to personal records.

(1) A state agency may authorize or provide access to or provide copies of an individually identifiable personal record for research purposes if informed written consent for the disclosure has been given to the appropriate department secretary, or the president of the institution, as applicable, or his or her designee, by the person to whom the record pertains or, in the case of minors and legally incompetent adults, the person's legally authorized representative.

(2) A state agency may authorize or provide access to or provide copies of an individually identifiable personal record for research purposes without the informed consent of the person to whom the record pertains or the person's legally authorized representative, only if:

(a) The state agency adopts research review and approval rules including, but not limited to, the requirement that the appropriate department secretary, or the president of the institution, as applicable, appoint a standing human research review board competent to review research proposals as to ethical and scientific soundness; and the review board determines that the disclosure request has scientific merit and is of importance in terms of the agency's program concerns, that

the research purposes cannot be reasonably accomplished without disclosure of the information in individually identifiable form and without waiver of the informed consent of the person to whom the record pertains or the person's legally authorized representative, that disclosure risks have been minimized, and that remaining risks are outweighed by anticipated health, safety, or scientific benefits; and

(b) The disclosure does not violate federal law or regulations; and

(c) The state agency negotiates with the research professional receiving the records or record information a written and legally binding confidentiality agreement prior to disclosure. The agreement shall:

(i) Establish specific safeguards to assure the continued confidentiality and security of individually identifiable records or record information;

(ii) Ensure that the research professional will report or publish research findings and conclusions in a manner that does not permit identification of the person whose record was used for the research. Final research reports or publications shall not include photographs or other visual representations contained in personal records;

(iii) Establish that the research professional will destroy the individual identifiers associated with the records or record information as soon as the purposes of the research project have been accomplished and notify the agency to this effect in writing;

(iv) Prohibit any subsequent disclosure of the records or record information in individually identifiable form except as provided in RCW 42.48.040; and

(v) Provide for the signature of the research professional, of any of the research professional's team members who require access to the information in identified form, and of the agency official authorized to approve disclosure of identifiable records or record information for research purposes.
[1985 c 334 § 2.]

RCW 42.48.030 Charge for costs of assistance.

In addition to the copying charges provided in RCW 42.17.300, a state agency may impose a reasonable charge for costs incurred in providing assistance in the following research activities involving personal records:

(1) Manual or computer screening of personal records for scientific sampling purposes according to specifications provided by the research professional;

(2) Manual or computer extraction of information from a universe or sample of personal records according to specifications provided by the research professional;

(3) Statistical manipulation or analysis of personal record information, whether manually or by computer, according to specifications provided by the research professional.

The charges imposed by the agency may not exceed the amount necessary to reimburse the agency for its actual costs in providing requested research assistance.

RCW 42.48.050 Unauthorized disclosure—Penalties.

Unauthorized disclosure, whether wilful [sic] or negligent, by a research professional who has obtained an individually identifiable personal record or record information from a state agency pursuant to RCW 42.48.020(2) is a gross misdemeanor. In addition, violation of any provision of this chapter by the research professional or the state agency may subject the research professional or the agency to a civil penalty of not more than ten thousand dollars for each such violation.

RCW 42.48.060 Exclusions from chapter.

Nothing in this chapter is applicable to, or in any way affects, the powers and duties of the state auditor or the joint legislative audit and review committee.
[1996 c 288 § 34; 1985 c 334 § 6.]

RCW 42.48.900 Severability — 1985 c 334.

If any provision of this act or its application to any person or circumstance is held invalid, the remainder of the act or the application of the provision to other persons or circumstances is not affected.
[1985 c 334 § 8.]