United States General Accounting Office

GAO

April 2001

RECORD LINKAGE AND PRIVACY

Issues in Creating New Federal Research and Statistical Information

GAO-01-126SP

Preface

Our nation has an increasing ability to accumulate, store, retrieve, cross-reference, analyze, and link vast numbers of electronic records in an ever faster and more cost-efficient manner. These advances bring substantial federal information benefits as well as increasing responsibilities and concerns.

Record linkage—a computer-based process that combines multiple sources of existing data—is a case in point.[1] Federally sponsored linkage projects conducted for research and statistical purposes have many potential benefits, such as informing policy debates, tracking program outcomes, helping local government or business planning, or contributing knowledge that, in some cases, might benefit millions of people.[2]

Despite these benefits, concerns about personal privacy are relevant: Linkages often involve data on identifiable persons. Indeed, because "the whole is greater than the sum of the parts," linking independent data on individuals creates new information about them.

Linkage benefits, privacy issues,[3] and privacy-protection strategies are being discussed in federal agencies, professional workshops, and academic literature. But for many policymakers and others outside these professional circles, there is no

[1]For a more specific definition of record linkage, see pp. 41-2.

[2]These statistical and research linkage projects are undertaken to produce information on populations or large groups of people. The purpose is not to take any government action or make any judgment with respect to any individual data subject; the principle of "functional separation" (discussed in chap. 3) emphasizes the importance of guarding against such uses of these linkage projects.

[3]In this study, we use the term "privacy issues" to refer to personal privacy, confidentiality, and security (see app. I).

overview or "roadmap" to key issues in this new and still developing field. This study is intended as a first step toward filling this gap.

Our overall goals are to stimulate discussion, inform the general public, and provide a context for policymakers and others whose knowledge or experience may be limited to certain aspects or types of linkage. To this end, we present examples of (1) different types of linkage projects, (2) major privacy issues, (3) privacy-protection techniques, and (4) strategies for enhancing "data stewardship."[4] In addition, we lay the groundwork for more comprehensive inquiry by providing questions for further study.

Throughout, our focus is on linkage projects that involve person-specific data, are conducted under federal auspices (or with federal funding), and produce new research or statistical information.[5] (See fig.1.)

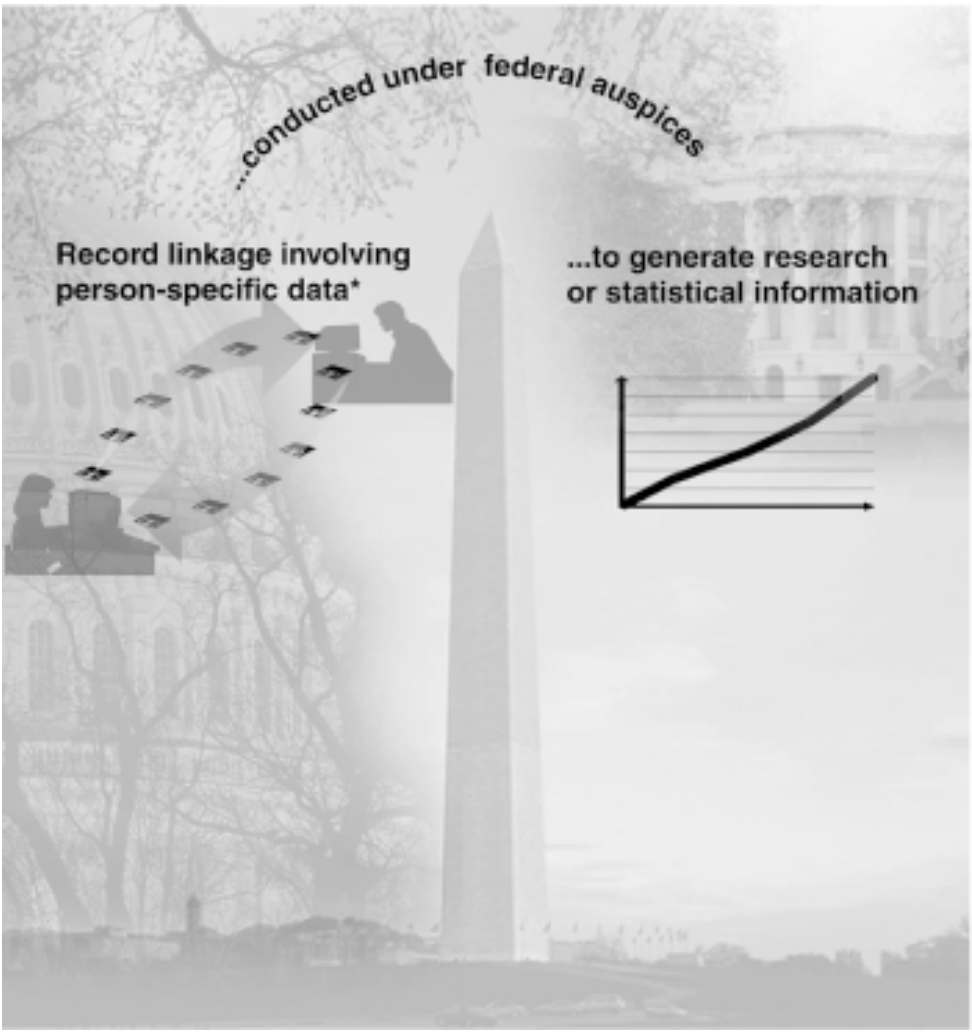
This study finds that

- Linkage projects tap survey data, existing records on individuals, and "contextual data" to provide new kinds of information. For example, one project links individual teens' survey responses to those of their best friends, thus providing new information on peer influences. Another links personnel records on chemical exposures to death records to help identify cancer-causing substances.

[4]This study does not provide a detailed legal analysis, assess specific agency practices, or develop recommendations.

[5]We exclude projects that link data (1) on organizations or business establishments; (2) without federal involvement (e.g., private-sector linkages); or (3) to facilitate actions toward or judgments about individuals (e.g., checking eligibility for benefits or loans).

Figure 1: Focus of This Report



*Defined as a computer-based process that combines (1) existing person-specific data with (2) additional data that refer to the same persons, their family or friends, school or employer, or geographic environment. (See chap. 2, pp. 41-2 for further definition.)

Such projects may also link each person's data to characteristics of the area where that person lives, the school attended, or other contextual information.

- Record linkage projects like these raise privacy issues, such as whether consent to linkage was obtained; whether linkages required sharing identifiable data with other organizations; and whether "deidentified" linked data are subject to reidentification risks when released for research or other uses.
- Various techniques that may help address these privacy issues include signed consent forms, tools for masked data sharing (such as list inflation, third-party linkage, or grouped linkage), and secure data centers where researchers analyze linked data under controlled conditions.
- Strategies for enhancing data stewardship could include, among others, developing agency systems for accountability and fostering or supporting an organizational culture that emphasizes the values of personal privacy, confidentiality, and security.

We provided a draft of this study to the following agencies for comment: U.S. Census Bureau, Department of Health and Human Services (HHS), Internal Revenue Service (IRS), Office of Management and Budget (OMB), and Social Security Administration (SSA). Agencies responding[6] generally supported the findings of our study; pointed to the importance of our work; and in some cases, volunteered to collaborate with GAO on future work in this area. They also provided technical comments, which we incorporated as appropriate.[7]

[6]Officials at SSA said they had no comments.

[7]Some agencies pointed to a need for comprehensive information on laws and agency practices. While this is beyond the scope of this study, we delineated a number of questions for further study.

The Census Bureau noted that because of this study's organization (i.e., the later chapters discuss privacy protections and stewardship), readers of the earlier chapters may not realize the kinds of protections or strategies that are being explored or, in some cases, are in use at agencies such as Census. We therefore added statements to earlier chapters alerting readers to material covered later in this volume.

We believe there is a recognition—at the Census Bureau and other agencies we talked with—that maintaining and improving privacy protections is key to achieving the public's cooperation in providing accurate records and participating in surveys and studies. Some of the privacy-protection techniques and stewardship strategies we discuss are in use at various federal agencies, but we did not assess the adequacy of such protections in any agency.

Readers with questions or comments are invited to contact me or Judith Droitcour at (202) 512-2700.[8] Other key staff include Nancy Donovan, Eric Larson, Patrick Mullen, and Theodore Saks. We are grateful to several experts for their contributions to our work.[9]

Nancy R. Kingsbury
Managing Director
Applied Research and Methods

[8]Copies of this study may be ordered; Web access is also provided. (For details, see last page.)

[9]Appendix II lists experts not currently with the federal government.

Contents

Preface		1
Chapter 1 Introduction and Background		10
	Background on Linkage and Privacy	14
	Objectives, Scope, and Methodology	29
Chapter 2 Generating New Information		32
	Main Data Sources	34
	Main Types of Linkages	39
	Linkage Results	47
	Next Steps and Questions for Further Study	52
Chapter 3 Privacy Issues		54
	Consent to Linkage	57
	Data Sharing to “Make the Link”	63
	Deidentification and Reidentification	68
	Potential Sensitivity of Linked Data	72
	Security of Linked Data	74
	Next Steps and Questions for Further Study	74
Chapter 4 Building a Privacy Protection Toolbox		76
	Techniques for Masked and Secure Data Sharing	78
	Procedures for Reducing Reidentification Risks	85
	Techniques to Reduce Sensitivity	93
	Consent Forms and Alternatives	96
	Security Measures (Stored Data)	98

Next Steps and Questions for Further Study 99

Chapter 5 Some Strategies for Enhancing Data Stewardship 100
Project-by-Project Decisions 102
Systems for Accountability 110
Organizational Culture 111
Next Steps and Questions for Further Study 115

Appendixes
Appendix I: Privacy Concepts 116
Appendix II: Experts Consulted 122
Appendix III: Conferences and Workshops Attended 124
Appendix IV: Selected Laws and Regulations Relating to Record Linkage and Privacy 126
Appendix V: Toward a More Complete Representation of Federal Record Linkage 136
Appendix VI: Abbreviations 138

List of References 140

Tables
Table 2.1: Examples of Person-Specific Datasets, by Data Source and Agency Type 38
Table 2.2: Examples of Record Linkage 40
Table 3.1: Examples of Person-by-Person Linkage, by Consent Procedure 60

Table I.1: Primary Privacy Concept or Category Associated With Each Issue 119

Figures

Figure 1: Focus of This Report 3
Figure 1.1: Record Linkage Topics 12
Figure 1.2: Combining Priorities on Information Gains and Privacy Issues 28
Figure 4.1: The Third-Party Model for Masked Data Sharing 80
Figure 4.2: Sample HRS Consent Form 97
Figure 5.1: Managing Security Risks 112

Introduction and Background

Record linkage can provide research and statistical information relevant to complex decisions about programs or policies. For example, information about peer influences on teen behavior, achieved through record linkages, can help people decide what kinds of programs would discourage early pregnancy, teenage drinking, and delinquency. While record linkage has been variously defined, this study uses the following broad definition.

Definition of Record Linkage

For purposes of this study, record linkage is defined as combining (1) existing person-specific data with (2) additional data that refer to the same persons, their family and friends, school or employer, area of residence or geographic environment.

Our focus is on linkage projects that involve person-specific data, are conducted under federal auspices (or with federal funding), and produce new research or statistical information concerning populations or large groups.[1] Privacy issues are important because person-specific data are involved and because actual linkages typically occur at the individual level, multiplying the quantity of data recorded on each individual.[2] The linked data are sometimes accessed by many researchers.

[1]As explained in chapter 2 (see pp. 41-2), we include (1) person-by-person linkages (both exact matches or probabilistic matches) and (2) person-by-context linkages. We exclude linkages intended to match similar persons based on, for example, demographic characteristics.

[2]Scheuren (1997). Some view the results as an “explosion of facts.”

In discussing privacy issues, this study refers to personal privacy (which concerns an individual's status and rights), confidentiality (a status accorded to information, requiring that its disclosure be controlled), and security (safeguards, such as encryption, for data and related systems). While these concepts have been variously defined, we use the working definitions given in appendix I. The concept of data stewardship is also relevant.[3] Individuals entrust information about themselves to agencies or research organizations that then assume the stewardship role.[4]

As illustrated in figure 1.1, our study describes (1) how record linkage can create new research and statistical information, (2) why linkage heightens certain privacy issues, (3) what kinds of techniques might help address privacy issues, and (4) how data stewardship might be enhanced. This study excludes projects that link data on organizations or business establishments but not individuals;[5] lack federal involvement (i.e., state-level, private-sector, and other linkages);[6] or are intended to result in actions toward data subjects (e.g., federal compliance audits).[7]

[3]A steward manages another's property, affairs, or in this case, data. For agencies, stewardship includes functions of officials and staff, such as privacy officers and advocates, disclosure officials, and survey managers. Stewardship carries responsibility for data subjects' personal privacy, confidentiality of data, and data security.

[4]GAO (1997). See also George T. Duncan et al. (1993).

[5]A large U.S. Census Bureau project links records on organizations for the quinquennial economic census; it involves records from Census, IRS, SSA, and Bureau of Labor Statistics (Census, 2000a).

[6]Private-sector linkages include credit checks and other linkages conducted for commercial reasons, such as marketing.

[7]The term "computer matching" often refers to linkages that check the eligibility for benefits or loan programs. See GAO (2000a).

Figure 1.1: Record Linkage Topics

Building a privacy protection toolbox
What tools might help ensure privacy
in record linkage, for example
• techniques for masked data sharing
• safer data and safer settings
• techniques to reduce sensitivity
• consent forms
• security measures
Data stewardship strategies
How data stewardship strategies
might enhance linkage privacy by improving
• project-by-project decisions
• accountability systems
• organizational culture

Background on Linkage and Privacy

Record linkage emerged decades ago, with early work aimed at developing new information in areas such as health research.[8] It was early recognized that linkage contained the potential for reducing data collection costs and respondent burden as well as improving data quality.[9]

To use a simple analogy from everyday life, when you balance your checkbook, matching your stack of checks to the bank statement check by check, you are carrying out a matching process analogous to linking records.[10] In linkage for research and statistics, the matching process adds new information. The records may be matched using names, addresses (or geocodes[11]), Social Security numbers (SSN), other identifiers, or some combination of these. The matched data are then preserved in a new, enhanced dataset. The linked dataset will be used to generate new, fuller information on the aggregate population. That is, it will be used to describe or make inferences about a population of individuals, analyze patterns in the data, and evaluate or inform programs or policies.

Record linkage has flourished, apparently for two key reasons:

- The first is the development of computer technology and the increasing tendency to maintain large-scale

[8]See Newcombe et al. (1959). For bibliographies of early linkages, see Kilss and Alvey (1985). See also, Jabine (1993).

[9]U.S. Department of Commerce (1978).

[10]Dean and Olson (1999).

[11]Geocodes are location codes, ranging from postal codes (e.g., zip plus four) to latitude and longitude (which can be determined by handheld Global Positioning System (GPS) devices).

sets of records in the public, as well as the private, sector.[12]

- The second reason is the recognition of the potential power of the linkage approach—the value added and the richness of datasets achieved by combining diverse data sources that, taken by themselves, are subject to various limitations—for areas such as health care delivery and outcomes, education, and economic policy. Various benefits of record linkage (including reduced cost, relative to new data collections) are described in recent reports from the National Academy of Sciences.[13]

Of course, the accuracy of linkage varies because, for example, some names are very common, the digits in some SSNs or other key numbers may be inaccurate or reversed, or these numbers may be missing for some data subjects.[14]

Background material presented below includes the need for transparent (open) government policies and practices, the use of record linkage in a wide variety of federal agencies, and the role of laws and values in information privacy issues.

[12]Fellegi (1997).

[13]These reports include the National Research Council (NRC 2000); the Institute of Medicine (IOM 2000); National Cancer Policy Board, IOM, NRC (2000).

[14]Indeed, HHS told us that, "There are many obstacles [to successful linkage] (principally from poor or non-reporting of the key variables in one or the other data set) and there are sometimes a number of records that cannot be matched or that are matched with only a low probability of accuracy."

Need for Transparency

A recent poll shows that many Americans perceive government as potentially a threat to their privacy.[15] But how knowledgeable are Americans about relevant issues in the area of research and statistics? How transparent (i.e., open and clear to everyone) is federal involvement in record linkage?

Federal statistical and research policies and practices may not be well known among members of the general public.[16] Some policymakers are no doubt aware that the Congressional Budget Office (CBO) has sought access to a recent version of the "linked dataset" that the Census Bureau first created over a decade ago by combining large-scale surveys it conducted with records on survey respondents obtained from the Internal Revenue Service and the Social Security Administration.[17] Other policymakers may have become aware of record linkage issues through congressional discussions about a proposal to allow data sharing among statistical agencies.[18]

[15]According to a nationally representative poll of 1,017 adults conducted in May 2000 by Opinion Research, 43 percent believed government is the biggest privacy threat, compared with 24 percent for the media and 18 percent for corporations (Purdy, 2000). An earlier 1995 Equifax survey similarly indicated that "82% of respondents are concerned about threats to their privacy [and] their uneasiness is more focused on the government than business" (American Demographics Marketing Tools Supplement, 1996, p. 31). Other surveys have indicated a relatively low level of trust in government (Singer et al., 1997; Panel on Civic Trust and Citizen Responsibility, 1999).

[16]For example, a survey in the mid-1990s indicated that only about one-fourth of adults knew that the Census Bureau is forbidden by law to give other agencies census information that includes a person's name and address (Singer et al., 1997).

[17]CBO sought the linked dataset, stripped of personal identifiers, for its long-term models of the Social Security and Medicare programs. These models are intended to help the Congress evaluate proposed changes to those programs.

[18]Since 1996, a number of bills have been introduced to allow statistical agencies greater flexibility to share data among

Despite discussions of record linkage in professional forums, much record linkage likely remains invisible to the general public and some policymakers as well.

Although privacy issues stemming from new technology have received considerable media attention during the past year, the main focus has been on uses of data that target individuals for action—rather than on statistical or research uses.

Thus, outside the statistical and research communities, few may be aware that

- A variety of federal agencies, contractors, and grantees use record linkage technology to produce new information that might help millions of persons or improve government programs (through, e.g., health research, improvements in the quality and efficiency of federal statistical programs, measuring government performance, or evaluating social programs).
- Statistical agencies envision streamlining linkages of government records (which are stored in separate "silos").
- Recommendations have been made for expanded support of linkage projects in cancer research.[19]
- Federal agencies and private-sector experts are debating the privacy issues raised by record linkage for statistics and research—as well as discussing a

themselves. More recently, in the 106th Congress, the House passed The Statistical Efficiency Act of 1999, H.R. 2885. This bill generally would have permitted the disclosure of data to a specified set of statistical agencies for exclusively statistical purposes and prohibited the disclosure of these data in identifiable form, for any purpose other than a statistical purpose, without informed consent. Data-sharing legislation involving statistical agencies was also discussed in a recent report to Congress (U.S. Department of the Treasury, 2000).

[19]National Cancer Policy Board, IOM, NRC (2000).

variety of techniques and strategies aimed at protecting privacy, which may be relevant to linkage.[20]

A pioneering linkage researcher (Ivan Fellegi) has pointed to the need for broad discussion or debate involving the general public,[21] and we believe that greater transparency is desirable for two reasons. First, open decisions about whether—and how—linkages should be conducted might foster or help support responsible data stewardship in federal agencies. Second, without greater transparency, "a single … error or accident might … put a [sudden] spotlight on the extent of linkage going on in government [and] … the incident might balloon out of control."[22] In the wake of such an incident, there might be a risk of "throwing out the baby with the bath water" or a loss of trust in government.

Two incidents—each involving very extensive linkages of personal data in other countries—illustrate the need for greater openness and participation to support sound linkage decisions and prevent unwise ones and to avoid possible perceptions of government secrecy.

- Just last year, an audit by Canada's Privacy Commissioner disclosed that a large and essentially unknown, though not secret, government database linked decades of records on more than 33.7 million persons.[23] Human Resources Development Canada (HRDC) had developed this dataset for "research,

[20]The discussions have occurred in academic journals as well as workshops and conferences (see app. III).

[21]Fellegi (1997).

[22]Fellegi (1997), p. 9.

[23]Privacy Commissioner of Canada (2000a, 2000b).

evaluation, policy and program analysis"—but it was described as one of "near Orwellian" proportions. The ensuing public outcry[24] led to an official statement that the database was being destroyed and to other changes at HRDC, which were directed at better protecting personal information. Nevertheless, some Canadians remained skeptical "that the giant information network was taken apart" or have otherwise expressed a lack of trust in HRDC.[25]

- A somewhat similar incident occurred in Sweden in 1986.[26] There, the database in question covered only 15,000 persons, but linkages extended from childhood into adulthood without the data subjects' knowledge and included detailed arrest records and for some, questionnaire responses on political attitudes.[27] Although the database had not been secret, it was unknown to the general public and data subjects alike. When a leading Swedish newspaper highlighted this "secret" database, there was strong reaction from the public. The database was then reportedly stripped of identifiers to prevent any further linkage.[28]

U.S. Involvement in Record Linkage

In the United States, many different kinds of federal agencies conduct or sponsor record linkage. The agencies perhaps most heavily involved in linkage to produce statistical and research information include the following.

[24]A Canadian official reportedly received more than 18,000 letters, phone calls, and e-mail messages from people demanding to know what was in their files (Washington Times, 2000). See also Toronto Star (2000); Ottawa Sun (2000a; 2000b); Toronto Sun (2000).

[25]Washington Times (2000); Ottawa Sun (2000a).

[26]Flaherty (1989).

[27]Respondents apparently were not told that their answers would be part of the linked database (New York Times, 1986).

[28]New York Times (1986).

Agencies Conducting or Sponsoring Record Linkage

- Statistical agencies, such as the Bureau of the Census and the National Center for Health Statistics (NCHS), which are charged, respectively, with providing comprehensive data on the U.S. population and the economy and with tracking trends in health and disease.
- Research agencies, such as the National Cancer Institute (NCI) and the National Institute for Occupational Safety and Health (NIOSH), which study causes of disease, assess the impact of treatments, and conduct research on work-related diseases and injuries, among other activities.
- Statistical or research offices of program agencies with large datasets, ranging from IRS to agencies charged with ensuring the security of the elderly or the vulnerable (e.g., SSA or Health Care Financing Administration (HCFA)).

Other agencies and offices conduct or sponsor linkage projects to help evaluate programs or measure performance. For example:

- Agencies administering block grants, such as the Substance Abuse and Mental Health Services Administration (SAMHSA), are funding state efforts to measure program outcomes using record linkage. For example, program data on persons treated for drug addiction are paired with records on their employment or other outcomes (e.g., involvement with law enforcement).
- GAO has estimated the long-term impact of a Department of Labor (DOL) training program by linking trainees' records from a DOL study with their SSA records on employment and earnings spanning several years.[29]

[29]GAO (1996a).

Groups Working to Enhance Privacy Techniques and Strategies

A variety of techniques to protect an individual's privacy and strategies to enhance data stewardship have been developed and may be useful in record linkage. Many individual statisticians and researchers, as well as federal agencies, have contributed to these efforts.

Various groups have taken a leadership or coordinating role in efforts to improve techniques and stewardship strategies. These include the

- OMB and its Interagency Council on Statistical Policy, Federal Committee on Statistical Methodology (Federal Committee), and Confidentiality and Data Access Committee;[30]
- HHS Data Council[31] and the HHS Office for Human Research Protections;[32] and
- The National Research Council and its Committee on National Statistics, as well as the Institute of Medicine, within the National Academy of Sciences (NAS), among others.

To cite two relatively recent examples: Within HHS, the Agency for Healthcare Research and Quality (AHRQ) and the Office of the Assistant Secretary for Planning and Evaluation (ASPE) sponsored an IOM workshop on data privacy in health services research.[33] Multiple agencies supported a workshop

[30]This committee was formerly known as the Interagency Confidentiality and Data Access Group (ICDAG).

[31]The HHS Data Council consists of HHS officials who have a direct reporting relationship to the Secretary, the HHS Privacy Advocate, and the Senior Advisor on Health Statistics. The Council coordinates HHS data collection and analysis activities, including privacy policy activities.

[32]This office oversees research involving human subjects that is funded by HHS.

[33]IOM (2000).

convened by the Committee on National Statistics to study the interface between data access and confidentiality.[34]

The activities of many professional organizations and committees are also relevant.

The Role of Laws

Laws governing the collection, administration, and disclosure of records and data maintained by federal agencies are relevant to record linkage.[35] Notably, agencies must follow the limits and conditions imposed by governmentwide laws, such as the Privacy Act of 1974,[36] as well as any applicable agency-specific laws. These laws extend varying levels of protection to records maintained by federal agencies.

The Privacy Act

The Privacy Act governs the responsibility of federal agencies concerning the content, access, and disclosure of records concerning individuals.[37] It

[34]NRC (2000). The primary sponsor was the National Institute on Aging (NIA) but additional funding was provided by at least five other agencies.

[35]For more discussion of selected laws and regulations relating to record linkage, see appendix IV.

[36]5 U.S.C. 552a. OMB has issued guidelines for implementing the Privacy Act, which are published in the Federal Register (40 Fed. Reg. 28948 (July 9, 1975)). In addition, OMB Circular No. A-130 (revised, Nov. 30, 2000) establishes policies for the management of federal information resources, including the protection of personal privacy by the federal government.

[37]The record linkage described in this study does not include activities covered by the Computer Matching and Privacy Protection Act of 1988. In that year, Congress amended the Privacy Act to regulate the use of computer matching conducted by federal agencies or using federal records subject to the statute. These amendments generally define computer matching as the computerized comparison of two or more automated systems of records or a system of records with nonfederal records for the purpose of (1) establishing or verifying eligibility for a federal benefit program or (2) recouping payments or delinquent debts under such programs. Matches performed to support any research

establishes governmentwide policies for the disclosure of data by federal agencies and requires agencies to safeguard identifiable information.[38] Under the act, agencies are not to disclose identifiable information to third parties without the individual's prior consent. The act contains 12 categories of exceptions to the consent requirement. These are intended to accommodate legitimate needs for identifiable information, such as conducting research and statistical activities that involve record linkage.

For example, under the act, an agency may disclose a record

- to officers and employees of the agency maintaining the record who have a need for the record in the performance of their duties;
- for a "routine use," that is, a use for a purpose that is compatible with the purpose for which it was collected;
- to a recipient who has provided advance written assurance that the record will be used solely as a statistical research or reporting record and that the record is to be transferred in a form that is not individually identifiable; and
- to the Bureau of the Census for purposes of planning or carrying out a census or related activity, according to the provisions of title 13.

or statistical project—the specific data of which may not be used to make decisions concerning the rights, benefits, or privileges of specific individuals—are not subject to the act.

[38]Generally, an officer or employee of an agency who willfully discloses material covered by the Privacy Act to any person or agency not entitled to receive it can be found guilty of a misdemeanor and be fined up to $5,000.

Several of these exceptions have implications for research and statistics.[39] For example, information disclosed to Census is used for statistical activities. Agencies, such as HHS and component agencies, have established research as a routine use of certain records, thus allowing disclosure outside the agency.[40] Concerns have been expressed about agency use of this exception.[41]

Other Relevant Statutes and Guidance

In addition to governmentwide statutes, many agencies are also subject to other laws that specify the confidentiality and data access policies they must follow. Some of these laws may limit record linkage activities. Notably, statistical information is protected by various agency-specific statutes, as illustrated below:

- The Census Bureau's activities with regard to confidentiality are governed by section 9 of title 13 of the U.S. Code, which requires that information furnished to the Bureau be kept confidential and be used exclusively for the statistical purposes for which it was supplied.[42]
- The NCHS records are protected by the following basic legal requirement in the Public Health Service Act, as amended.[43] No information obtained in the course of NCHS' activities may be used for any

[39]OMB (1975), George T. Duncan et al. (1993), Cecil and Griffin (1985).

[40]See, for example, Fanning (1998).

[41]See appendix IV.

[42]13 U.S.C. 214 provides serious penalties for wrongful disclosure by Census employees.

[43]42 U.S.C. 242m. A similar statute protects information collected by AHRQ, which conducts research, demonstration projects, and evaluations (42 U.S.C. 299, 299c-3).

purpose other than that for which it is supplied unless there has been consent. Also, such information may not be published or released in an identifiable manner unless there has been consent.

Furthermore, OMB issued an order establishing government policy to protect the privacy and confidentiality interests of individuals and organizations who furnish data for federal statistical programs.[44] This order establishes standards regarding the disclosure and use of information acquired for exclusively statistical purposes.[45] Agencies are to comply with the order to the extent permissible under their statutes.[46]

Various other agencies have restrictive provisions concerning disclosure of certain information. For example, 26 U.S.C. 6103 prohibits IRS from disclosing any return or return information except as authorized by title 26. A key exception, contained in 26 U.S.C. 6103(j), authorizes the furnishing of certain specific return information to the Census Bureau "for the purpose, but only to the extent necessary in the structuring, of censuses and … conducting related statistical activities authorized by law."

[44]For OMB's order concerning confidentiality of statistical information, see the Federal Register (62 Fed. Reg. 35044 (June 27, 1997)).

[45]If the agency collecting the information proposes to disclose data in identifiable form for purposes other than statistical, then prior to disclosure, it is to fully inform the affected respondents of the facts regarding such disclosure.

[46]For an earlier discussion of statutes specific to statistical agencies, see our report concerning authorizing statutes and confidentiality provisions for statistical agencies (GAO, 1996c). An extensive discussion of pertinent agency-specific statutes relating to federal statistical programs may be found in George T. Duncan et al. (1993).

The "Common Rule"

In addition, there are certain federal regulations, most notably the Federal Policy for the Protection of Human Subjects, known as the Common Rule, that govern certain research projects that involve human subjects or personal information on them; these projects may include record linkage. Under the Common Rule, research supported or regulated by any of 17 federal agencies is subject to certain federal oversight requirements.[47] In accordance with the Common Rule, organizations have established local institutional review boards (IRB), made up of both scientists and nonscientists, to approve or disapprove research projects depending on such factors as whether researchers minimize the risks to research subjects and obtain their informed consent.[48]

The Role of Values

Within the framework provided by current laws and regulations relevant to record linkage, there is room for interpretation, stewardship decisions, and thus, value judgments.

Opinions and values on information privacy issues might be conceptualized as positions on a one-dimensional spectrum ranging from those perspectives that put the highest priority on privacy issues to those that put the highest priority on information gains. Toward one end of this spectrum would be various statements by advocates of privacy, human rights, and vulnerable populations or by those concerned about the possibility of government

[47]HHS regulations codified at part 45, Part 46, Subpart A of the Code of Federal Regulations. In addition, 16 other agencies have adopted regulations incorporating the substance of the HHS regulations.

[48]IRBs can be associated with organizations ranging from universities to government agencies. They must implement the Common Rule, including provisions to protect the privacy of human subjects and the confidentiality of data that identify individual persons. Some record linkage projects are reviewed by IRBs.

misuse.[49] Toward the other end of the spectrum are statements by researchers that place a particularly high value on generating new information. This would include, for example, statements of health researchers cautioning against overemphasizing privacy issues because limiting researchers' access to data might discourage the generation of needed information.[50] However, some experts view a one-dimensional conceptualization as a zero-sum approach that does not capture the full set of positions.

As shown in figure 1.2, a two-dimensional depiction of relevant values is also possible. Here, the horizontal axis represents the value placed on privacy issues, which can range from low to high, and the vertical axis represents the value placed on information gains, which can also range from low to high.

On this graph, opinions from the zero-sum spectrum described above are depicted by the icons positioned along the diagonal from top left to bottom right. The crucial additional position on this graph is in the upper right quadrant: the combination of a high value placed on information gains and a high value placed on privacy issues.

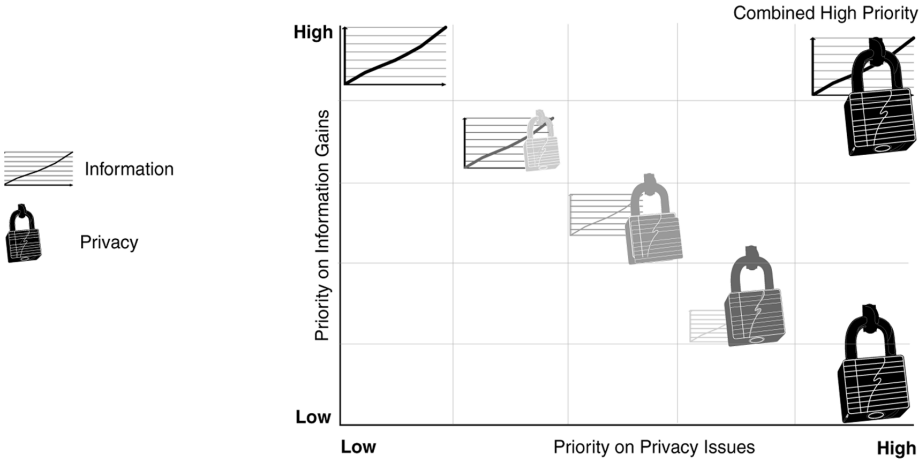
Some of the experts we spoke with emphasized that the upper right quadrant is where their perspective would be represented, rather than at any point on the one-dimensional spectrum. Considering this, our own position is to place a high priority on both values.[51]

[49]Chapman (1997); Berman and Goldman (1989); Thompson (2000); Seltzer (1998, 1999).

[50]See Dean and Olson (1999); Korn (2000).

[51]Although "no single solution" would be appropriate for all federal agencies, many may see opportunities to emphasize both values (see George T. Duncan et al., 1993, p. 19). And as noted by NRC

Figure 1.2: Combining Priorities on Information Gains and Privacy Issues



Those who prioritize both information gains and privacy issues may be more likely to

- champion techniques designed to build in personal privacy, confidentiality, or security while still allowing information gains and
- work to foster improved stewardship or decision-making processes that better balance or, where possible, maximize both personal privacy and information gains.

Those who prioritize both values may also be more likely to recognize the complexities involved. For example, enhancements of personal privacy, confidentiality, and security may improve data quality

(2000, p. 5), "most federal agencies are accountable for … ensur[ing] appropriate standards of privacy and confidentiality, and facilitating responsible dissemination to users."

by encouraging persons to provide more accurate personal information.[52] Also, record linkage may sometimes support personal privacy by allowing statistical agencies and researchers to avoid new, and perhaps intrusive or burdensome, data collections.[53]

Objectives, Scope, and Methodology

This study's objectives are to (1) show how record linkage generates new statistical and research information, (2) review a number of relevant privacy issues, (3) illustrate the kinds of techniques that might be included in a "privacy protection toolbox" for record linkage, and (4) explore a sampling of strategies for data stewardship.

Our focus is on privacy issues rather than technical topics such as potential problems in the quality of linked data and methods for analyzing linked data.[54]

The scope of this study is limited to record linkage for statistics and research. Statistics involves developing quantitative information through enumeration, measurement, analysis, and dissemination. This information is developed by federal agencies to describe the social, economic, and general conditions of the nation. Research refers to the use of a systematic, objective process to discover and analyze relationships between variables. Both statistics and research use individual data during the analytical process but present findings in aggregate form.[55]

[52]Goldman (1998). A similar point is made by the Health Privacy Working Group (1999).

[53]See Prevost and Leggieri (1999).

[54]For a discussion of technical issues, see Fair (1999); Newcombe, Fair, and Lalonde (1992); Winkler (1995).

[55]These definitions are based in part on those provided by the Privacy Protection Study Commission (1977).

The scope of this study is further limited to linkages that involve data on individual persons and that are conducted under federal auspices (or with federal funding).

To address objective 1, we identified a set of examples of linkages conducted under federal auspices. We limited this set of linkage examples to those involving (1) health data or (2) data on income, earnings, or wealth.[56] We identified specific examples by reviewing the literature, attending linkage conferences or workshops, and talking with various agencies as well as experts outside the federal government (see apps. II and III).[57] The information drawn from this set of examples is intended to be illustrative rather than representative of federal practice. We also developed a list of questions for further study of the scope, extent, and benefits of record linkage.

To address objectives 2, 3, and 4, we sought the knowledge and views of a variety of researchers, privacy experts, and staff from several agencies.[58] Additional information was obtained through reviewing literature (including reports from conferences and earlier GAO reports) and attending conferences and workshops.[59] With respect to

[56]We recognize that linkage is also conducted in many other important areas, including, for example, education statistics and crime research.

[57]Some of the experts were previously with federal agencies that conducted record linkage.

[58]See appendix II for a list of experts consulted. While some of the experts we talked with are active in professional organizations, we did not formally coordinate with these organizations.

[59]For a list of conferences attended, see appendix III. Proceedings from two other conferences or meetings were reviewed. (See National Committee on Vital and Health Statistics, 1998; HHS Task Force on Privacy, 1993.)

objective 2 (privacy issues), we developed some further information using the set of linkage examples developed for objective 1. In addressing objective 3, we targeted a general readership rather than the professional statistician as our primary intended audience. We also developed questions for further study of relevant privacy issues, techniques appropriate for a privacy toolbox, and relevant stewardship strategies.

Throughout, the discussion is intended to be illustrative. We did not conduct audits to determine agencies' compliance with privacy law or confidentiality requirements, and we did not attempt to analyze different agencies (or types of agencies) in terms of possible variations in their privacy protection policies and practices or their legal frameworks. In considering strategies for enhancing data stewardship, we focused on managerial rather than legislative approaches.

We conducted our data collection and analysis from December 1999 through December 2000. The agencies we visited in the course of our work include the Office of Management and Budget, the Department of Commerce (Census), IRS, SSA, and HHS. Within HHS, we talked with officials at ASPE, AHRQ, HCFA, SAMHSA, the Centers for Disease Control and Prevention (NCHS and NIOSH) and the National Institutes of Health—NIA, NCI, and the National Institute for Child Health and Human Development (NICHD).

Generating New Information

As highlighted on the opposite page, this chapter addresses the first of four key record linkage topics: how record linkage generates new information. Specifically, this chapter presents examples of federal record linkage, tracing each case from the original data sources through the point of linkage and, lastly, to the statistical or research results. The examples involve data or records from two substantive areas: (1) health and (2) income, earnings, and wealth.

Subsequent chapters deal with privacy issues, as well as outline various privacy protection techniques and strategies for data stewardship that are either being used now by various federal agencies or might be used in the future.

When researchers and statisticians link records, they put together "pieces of a puzzle." Once linked, diverse data sources that by themselves have limited meaning can generate different and potentially more valuable information. The varied examples we selected indicate how linkages may

- provide new data on the quality of health care or on the ways that aging persons interact, over time, with benefit programs, thus potentially informing decisions about major federal programs, such as Medicare and Social Security;
- assess the accuracy or improve the timeliness of data that are relevant to key government policies or to private-sector business decisions (e.g., by updating estimates of the population for local areas); and
- add to basic knowledge about key topics, such as cancer-causing substances or peer influences on delinquency, which may inform personal decisions as well as policy directions and program design.

Generating new statistical and research information
How record linkage does this by combining
• sample survey data
• archival records
• contextual information
Privacy Issues
Why record linkage heightens concerns about
• consent
• data sharing
• reidentification risks
• potential sensitivity
• security of linked data
Building a privacy protection toolbox
What tools might help ensure privacy in record linkage, for example
• techniques for masked data sharing
• safer data and safer settings
• techniques to reduce sensitivity
• consent forms
• security measures
Data stewardship strategies
How data stewardship strategies might enhance linkage privacy by improving
• project-by-project decisions
• accountability systems
• organizational culture

Perhaps most importantly, new information generated by linkage might not be obtainable in any other way.

Main Data Sources

Record linkage draws on a variety of data. For this discussion of record linkage and privacy, we define three main sources of data as

- sample surveys and other studies of individual persons, based on a sample of the target population;
- full sets of existing records on individual persons (archives); and
- contextual data, such as characteristics of geographic areas where individuals live.

Sample Surveys and Other Studies of Individual Persons

Many sample surveys and other studies conducted under federal auspices have a research or statistical purpose. Typically, participation is voluntary.[1] Sample surveys often cover thousands of persons selected from a population of millions.

Four examples of sample surveys are as follows: The Add Health survey, conducted under a grant from a research agency, asks high school and middle school students in a nationwide sample of 80 communities to fill out a pencil-and-paper survey while in school.[2] The Survey of Income and Program Participation (SIPP), conducted by a statistical agency, involves personal and telephone interviews with a sample of the household population aged 15 and older. The Health

[1]Note that typically, data from a sample of individuals cannot be linked to data on the same persons collected in a different sample. The reason is that in most cases, the same persons would not be included in both samples.

[2]Add Health refers to the National Longitudinal Study of Adolescent Health. Some students are later interviewed in their homes on sensitive topics not covered in the in-school questionnaire (e.g., drug abuse and sexual experience), using a self-interview technique in which the respondent wears earphones and silently interacts with a voice-assisted computer.

and Retirement Study (HRS, conducted under a grant from a research agency) is a personal interview survey of middle- and retirement-aged persons. The Longitudinal Study of Aging (LSOA, conducted jointly by a statistical agency and a research agency) focuses on elderly persons. (Often in this study, we will refer to sample surveys simply as surveys.)

Other sample studies that in some instances utilize record linkage include, for example, randomized field studies.[3]

Full Sets of Existing Records on Individual Persons (Archives)

For purposes of convenience, we use the term "archive" to refer to a full set of existing records.[4] A full set of records is intended to cover all relevant individuals. Coverage of the full set of persons in a target population or group means that linkage to a sample survey or to other existing records is possible.[5] Thus, for linkage, full sets of records represent a crucial data source. Unlike sample surveys, for these records, participation in data collection is typically mandatory or not optional.[6] We distinguish between

[3]See Boruch et al. (2000).

[4]This follows the earlier use of this term by Boruch and Cecil (1979) and Webb et al. (2000). We do not use the term archive to refer to information stored at the National Archives.

[5]By contrast, as noted previously, two unrelated sample surveys (e.g., LSOA and HRS) would usually not be candidates for person-by-person linkage. The reason is that one would expect that, on the whole, different persons would be included in each survey.

[6]By mandatory we mean that participation is legally required or that nonparticipation is associated with some negative consequence, or both. By "not optional," we mean that the records are collected as a matter of course without regard to personal preferences.

two categories of datasets, based on their original purpose.[7]

- The first category consists of **administrative datasets**, which federal program agencies have created to operate their programs. For example, Medicare health insurance datasets cover all participants (nearly 40 million beneficiaries) and allow HCFA, a program agency, to reimburse providers. Other examples would be datasets of persons' earnings and benefits kept by SSA (essentially all regularly employed U.S. workers), data from federal estate tax returns kept by IRS, and personal address or other data from federal income tax returns (which cover about 85 percent of the population). Yet another example would be datasets of information provided by participants in benefit programs to demonstrate eligibility for those programs. These various administrative datasets may be used in statistical or research analyses.
- The second category consists of **"records-research datasets."** These datasets contain records generated in multiple settings and locations. For example, clinical records are generated by various hospitals; personnel records are generated by different employers. Records from multiple locations are compiled in a dataset for research use. For example, NCI compiles the Surveillance, Epidemiology, and End Results (SEER) database—a registry of data from clinical records. NIOSH extracts data[8] from personnel records to create databases for research on workplace health risks. (This was done, e.g., at chemical plants to

[7]Of course, some datasets are created for dual purposes (Scheuren, 1995).

[8]HHS told us that the data extracted concern relevant aspects of employees' work histories and workplace exposures to health risks.

identify workers who were exposed to dioxin, the contaminant in the defoliant known as Agent Orange.)

Full sets of existing records also include high school grades, birth and death records,[9] and many others.

Examples of person-specific datasets—sample surveys and full sets of existing records—are provided in table 2.1, by type of data-generating agency.[10]

Contextual data

Contextual data are used "to provide information on the context in which individual attitudes, behavior, or other experiences take place."[11] Contextual data describe entities larger than individuals; included here are characteristics of (1) the geographic areas in which people live, (2) the employers and schools where individuals work or study, and (3) relevant state and local governments. For example, a geographic area where specific persons reside may be described in terms of its crime rate, unemployment rate, average income level, or its air quality or pollution level. Each area may also be described in terms of the number of businesses, churches, or other organizations located there. Employers may be characterized by type of pension plan provided; and

[9]Birth and death records are maintained by state governments. Birth records can be used for such diverse purposes as obtaining passports; proving age; demonstrating citizenship; or obtaining insurance or governmental benefits. When compiled as vital statistics, birth and death records can provide important sources of data for research. They are also used for administrative purposes.

[10]We use the term "person-specific dataset" to distinguish surveys of individuals and records on individuals from contextual data. Person-specific datasets may or may not include explicit identifiers. For example, some person-specific datasets include code numbers that may be linked to identifiers in carefully guarded "cross-walk" files maintained either by the agency that maintains the dataset or by others.

[11]Piccinino and Mosher (1999).

schools, by average test scores. State and local governments may be described in terms of their policies or practices.

Table 2.1: Examples of Person-Specific Datasets, by Data Source and Agency Type

	Type of agency creating or funding dataset *(specific agency)*		
Data Source	**Statistical**	**Research**	**Program**
Sample survey 	Survey of Income and Program Participation, or SIPP *(Census)* Longitudinal Study of Aging, or LSOA[a] *(NCHS)*	National Longitudinal Study of Adolescent Health, or Add Health[b] *(NICHD)* Health and Retirement Study, or HRS[c] *(NIA)* LSOA[a] *(NIA)*	[d]
Full set of existing records[e] 	[d]	Surveillance, Epidemiology, and End Results, or SEER, dataset, based on clinical records* *(NCI)* Data extracted from personnel records *(NIOSH)**	Medicare records** *(HCFA)* Social Security earnings and benefits records** *(SSA)* Tax records** *(IRS)*

***Records-research database**.

****Administrative data**.

[a]LSOA is a joint project of a statistical agency (NCHS) and a research agency (NIA).

[b]Add Health is conducted under a grant to the University of North Carolina, funded primarily by NICHD.

[c]HRS is conducted by the University of Michigan and is funded primarily by a grant from NIA.

[d]No example.

[e]Full sets of existing records (here termed “archives”) include administrative data and records-research databases.

Contextual data are obtained from a wide variety of public- and private-sector sources, including for example, religious or charitable organizations (e.g., National Council of Churches) and various types of federal, state, and local agencies.[12]

A Note on Decennial Census Data

Data from the decennial census are a special case. Individual-level data from the short form might be categorized as an archive because the entire population is canvassed and participation is mandatory, although the records were not preexisting. (According to the Census Bureau, individual-level "microdata" from current and recent censuses have not been used in linkages, with the exception of internal statistical studies undertaken by Census.) Other microdata from the long form of the census describe a sample and thus might be categorized as sample survey data. Publicly available local area data from the census represent an important form of contextual data and are routinely used in linkage.

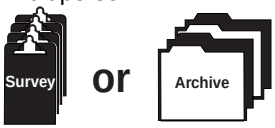
Main Types of Linkages

Linkages fall in two main types—person-by-person linkages (including multiperson, survey-archive, and multiarchive links) and person-by-context linkages. Examples of each type are provided in table 2.2.

We note that, once linkage has occurred, the resulting linked dataset may be stripped of explicit identifiers or further "deidentified" as discussed in chapter 3.

[12]In some cases, contextual data are publicly available, organized in a convenient and accessible fashion. In other cases, much effort may be needed to obtain data on, for example, neighborhood crime levels. In still other cases, contextual data may be developed in a special study of schools or employers.

Table 2.2: Examples of Record Linkage

Type of record linkage or data combination	Substantive area: Health data	Substantive area: Income or wealth data
PERSON BY PERSON		
Multiperson 	Add Health survey data on teens linked across best friends *(NICHD*)*	Estate tax returns linked for beneficiaries and decedents from 1916 to 1981[a] *(IRS*)*
Survey-archive[b]	LSOA survey data on elderly persons linked to Medicare insurance records and death records *(NCHS*)*	HRS survey data on middle- and retirement-aged persons linked to SSA records *(NIA*)* SIPP survey data on youths and adults aged 15 and older linked to SSA records *(Census*)*
Multiarchive	SEER clinical data on cancer patients linked to Medicare insurance records *(NCI,* HCFA*)* Personnel records from chemical plants linked to death and other records[c] *(NIOSH*)*	Individuals' addresses from income tax returns linked to SSA records[d] *(Census*)*
PERSON BY CONTEXT	Add Health survey data on teens linked to data on neighborhoods and schools *(NICHD*)* SEER clinical data and Medicare insurance records linked to census-tract information, based on patient address data *(NCI,* HCFA*)*	Survey data on middle- and retirement-aged persons (HRS or linked HRS-SSA data) linked to state-level data *(NIA*)*

See next page for footnotes.

Key linking or sponsoring agency. (See also table 2.1.)

[a]See Wahl (1997). The linked data were limited to federal estate tax returns filed in Wisconsin.

[b]This type is termed "survey-archive" for convenience; it would logically include links of archives to sample studies other than surveys, such as randomized field trials based on a sample of individual persons.

[c]See Fingerhut et al. (1991); Steenland et al. (1999). NIOSH has authority to obtain personnel records under the Occupational Safety and Health Act (29 U.S.C. 657 and 669).

[d]This linkage is conducted to produce information on migration from one state or county to another, specifically, migration rates by age, sex, race, and Hispanic origin. These migration rates are used in the production of intercensal estimates that are required by law (13 U.S.C. 181). No person-level linkage to census data is involved in the intercensal estimates process.

Some linked datasets are not made available outside the linking agency. Some datasets are stripped of explicit identifiers; these may made available to researchers under controlled conditions. In some cases, linked data are more fully "deidentified" and made available for public use (see chap. 3).

- **Person-by-person linkages** generally fit Newcombe and colleagues' basic definition of linkage as "the bringing together of two or more separately recorded pieces of information concerning a particular individual or family."[13] We extend this definition to include not only the individual and the family but also the friendship group and other relationships.

  These linkages may be "exact matches" or "probabilistic matches." These two kinds of matches are similar in that (1) the goal is to match two or more records on the same unit (e.g., same person or same family) and (2) each match is achieved by means of identifying information (e.g., names, addresses, and SSNs). However, probabilistic linkage recognizes that

[13]Newcombe et al. (1959).

the accuracy of linkages varies and that a differing weight of evidence characterizes different matches.[14]

- **Person-by-context linkages** bring together information on a person and a larger entity related to him or her (e.g., a student and his or her school).

  While person-by-context linkages do not fit Newcombe and colleagues' definition, they are consistent with the definition provided at the outset of this study. Person-by-context linkages were suggested to us by some agencies, and some of the experts we talked with predicted that in the future, such linkages may become more extensive, owing to current advances in technology (e.g., greater use of GPS for geographically based linkages, greater availability of other contextual information on the Internet). Person-by-context linkages can generate new information that may be useful in planning programs or making other policy-relevant decisions; and they may, in some instances, heighten privacy issues. We include them for these reasons and also because we wish to present a broad overview of types of linkage that involve data on individual persons.

We believe that the types of linkages described above encompass most federal or federally-sponsored linkage projects that fit our definition of linkage (see chap. 1). We do not know how frequently each type of linkage is conducted.

[14]See Fellegi and Sunter (1969). By contrast to both exact matching and probabilistic matching, "statistical matching" has a goal of linking similar units, based on demographic or other characteristics, not identifying information. We excluded statistical matches from this study.

Person-by-Person Linkage

We provide examples of three kinds of person-by-person linkage: multiperson links, survey-archive links,[15] and multiarchive links. Each represents a different kind of data combination.

Multiperson Links

Multiperson links create dyads or small groups based on a specified relationship between persons (e.g., husband-wife) and on data indicating which respondents or data subjects have this relationship to each other. Multiperson linkages traditionally combine data on members of a nuclear family (e.g., creating parent-youth pairs from survey data, combining SSA earnings records for husbands and wives, creating "tax families" by combining household members' tax returns).[16] However, researchers may tap wider multiperson networks, as illustrated by the following examples.

- Best friends are linked in the Add Health survey based on the questionnaire item that asks each teen respondent to identify his or her five best male friends and five best female friends, using code numbers and a school roster. The best-friends linkages are meaningful because the in-school survey covers most teens in each sampled community. (In addition, the subsample of teens who were also interviewed at home includes most teen residents in some communities. This subsample provides the basis for teen-neighbor linkages, using each student's home address and geocode.[17])

[15]For purposes of convenience, we use the term "survey-archive links" to describe linkages involving sample studies of individuals and full sets of existing records. Sample studies may be surveys or other kinds of studies, such as randomized field studies.

[16]See Kandel (1973); Rittenhouse and Miller (1984); Mitchell et al. (1996); Scheuren and Petska (1993).

[17]Interviewers visiting students' homes used a hand-held GPS device that indicates latitude and longitude. Thus, data can be linked for

- Generations are linked in an analysis of federal estate tax records. The linked data consist of pairs of estate tax returns. That is, an earlier year return in which a specific person is named as the beneficiary is paired with a later year return in which that person is named as the decedent (based on SSNs or other information). This allows inherited wealth to be compared to bequests. Some information actually spans three generations, that is, when data on the beneficiary for the more recent return are included in an analysis.

Survey-Archive Links

Survey-archive links match information provided by survey respondents with existing records on these same individuals. For example:

- Survey respondents' reports (from SIPP) of income and participation in programs, such as Food Stamps or Temporary Assistance for Needy Families, are linked to SSA records on these respondents' earnings and Social Security benefits. The links or matches are based on the individual's Social Security number.[18]
- Middle-aged and older persons' survey responses (on HRS) concerning their health and decisions about retirement or application for Social Security Disability Insurance (SSDI) are linked to their SSA records (with their written consent to the transfer of SSA records outside the government for purposes of linkage).[19]
- Elderly respondents' reports of health status, residential status, hospitalizations, and other variables (from the LSOA survey) are linked to Medicare insurance records (on hospitalization, home health

students who live close to each other. Other Add Health linkages pair teen data with parent and sibling interviews.

[18]SIPP linkages are made only for those respondents who provide an SSN (78 percent).

[19]HRS data are also linked to information on pension plans, obtained from respondents' employers (Juster and Suzman, 1995).

services, and hospice care) as well as, eventually, death records.[20]

Multiarchive Links

Multiarchive links combine two or more full sets of existing records. For example:

- Data from personnel records that identify workers exposed to Agent Orange are linked to death records to determine whether each worker has died, and if so, the date and cause of death.[21] Workers not identified as having died are confirmed to be alive by (1) linkage to IRS data on the most recent address from which the individual filed taxes and (2) checking with the local postmaster to confirm that the individual is still receiving mail at that address.[22]
- Data from clinical records on older persons' cancer diagnoses (compiled in NCI's SEER database) are linked with Medicare insurance records, which indicate type of insurance, health care usage, and cost.[23]

[20]See Kovar et al. (1992). HHS told us that "92 percent of the LSOA sample was considered eligible for matching to Medicare records. A successful match was completed for 81 percent of LSOA respondents."

[21]NIOSH has legal authority to obtain personnel records under the Occupational Safety and Health Act (29 U.S.C. 657 and 669). The National Death Index (NDI) provides access to causes of death for all U.S. deaths from 1979 through the most recent year stored on the NDI file (death records are added to the NDI file annually, approximately 10 months after the end of a particular calendar year). For earlier years, vital status and state where deceased can be identified using Social Security records and death certificates obtained from the states.

[22]NIOSH is permitted access to IRS records for the purposes of locating individuals who are, or may have been, exposed to occupational hazards in order to determine the status of their health or to inform them of the possible need for medical treatment (26 U.S.C. 6103(m)(3)).

[23]Selected data from clinical records (originally created by hospitals or laboratories) are initially maintained by registries at the state or metropolitan level. (See Hankey et al., 1999.) When stripped of

- Address data from income tax returns for 2 consecutive years are linked to each other to determine whether individuals changed their place of residence. These person-level "migration data" are linked to SSA data that indicate each individual's age, sex, and race, as well as whether he or she is of Hispanic origin.[24] The linked migration data are used in a key statistical program (required by law), which is described later in this chapter.

Person-by-Context Links

Person-by-context links are unlike the foregoing in that they combine person-specific data with contextual information on larger entities (geographic areas, political subdivisions, schools, and employers). Linkages are based on knowing where each respondent or data subject lives, what school he or she attends, or who his or her employer is. For example:

- Teen survey responses (Add Health) are linked to information on relevant neighborhoods. A contractor compiles data on the geographic areas covered by the survey and links these to specific respondents, based on their addresses or GPS indications of precise latitude and longitude.
- These same survey responses (again, Add Health) are also linked to school characteristics, based in part on interviews with principals or school administrators.

personal identifiers, these data are transferred to and subsequently maintained by NCI in the SEER database. The linkage process, which involves identifying data from the registries, is described in materials made available for the SEER-Medicare Data Users Workshop (Bethesda, MD, June 24, 1998).

[24]Census receives these data from IRS and SSA. The Internal Revenue Code (26 U.S.C. 6103(j)(1)(A)) permits the Census Bureau to acquire federal tax data. No personally identifiable results are transferred from Census back to IRS or SSA; in addition, the data are not available in public-use files.

These data can be combined with the multiperson links described above.

- Survey data on individuals' retirement and disability decisions (from HRS) are linked to state-level contextual data.[25] Researchers compile state-by-state contextual data on unemployment rates and rates of approval for SSDI disability applications. Information on each respondent's place of residence allows linkage of these contextual data to the linked HRS-SSA data described above.[26]
- SEER clinical records and Medicare insurance records on the same patients (linked as described above) are linked to published census-tract information, such as the median income of the tract where the patient lives and the percent of residents in that tract who are high school graduates.[27] This linkage is based on information derived from the patient's home address.

Linkage Results

A wide variety of information is achieved through person-by-person linkages and person-by-context linkages.

New Information Gained Through Person-by-Person Linkage

This section describes results achieved through multiperson, survey-archive, and multiarchive links.

Logically, multiperson linkages have the potential to provide new information on patterns of interpersonal influence (which can have import for the design or refinement of a variety of government programs). To illustrate this:

[25]Burkhauser et al. (1999a).

[26]Information on state of residence is not included in generally available HRS datasets. (In generally available HRS datasets, the geographic data identify only major divisions of the United States—Northeast, Mid-Atlantic, etc. For HRS datasets with context links, analyses require special arrangements and privacy protections.)

[27]Potosky et al. (1997).

- Influences on teen behavior are assessed through Add Health linkage of self-reported behaviors for teen best friends. Other analyses examine teens who are neighbors. The early results indicate that while family is the most important, best friends are more influential than neighborhood peers.[28] Potentially, information like this can help improve programs to prevent early pregnancy, teenage drinking, drug abuse, and delinquency.
- A comparison of dollars inherited to dollars bequeathed, which is a key aspect of the transfer of wealth across generations, is possible because of the linked estate tax data. The as-yet unpublished results seem to indicate that inherited wealth is systematically dissipated, but the slowest rate of dissipation occurs for the wealthiest families; this information is relevant to a body of work in economic studies.[29]

Turning to survey-archive links, these are intended to produce data that are more accurate, more detailed, longer term, and generally more extensive than would be possible from either source alone.[30] When the existing records (archives) pertain to government programs, the results may help indicate how individuals interact with programs. For example:

- Data accuracy checks are based on comparing each respondent's report of past-year hospitalizations

[28]Greg J. Duncan et al. (1999).

[29]Wahl (1998).

[30]Surveys ask about diverse topics but are limited by respondent recall. By contrast, existing records may include data on a limited set of events but include the recorded details of those events, even if long in the past.

(LSOA survey) with information from his or her Medicare insurance records.[31]

- Estimates of use of health care services in the last year of life are also based on linked LSOA-Medicare data.[32]
- Analyses exploring the work, retirement, and disability decisions that people make as they age have been conducted based on the HRS-SSA linked data.[33]
- The SIPP-SSA linkage is intended to expand and improve data available from the SIPP statistical program. Studies involving these linked data have been used by the Census Bureau to evaluate and potentially improve the quality of the SIPP data; they have also been used by SSA and other policy analysts to project the effects of proposed changes in the Social Security law and project future earnings subject to the Social Security tax.[34]

Because of many persons' limited recall, the information provided by these survey-archive links might not have been obtainable in any other way.

Finally, multiarchive links can bring together the "different pieces of a puzzle" needed to test policy-relevant hypotheses or produce new statistical estimates. Three examples follow.

- The carcinogenicity of dioxin in humans is evaluated by the International Agency for Research on Cancer (IARC) and others using the NIOSH research based on

[31]Stearns et al. (1996a).

[32]Stearns et al. (1996b).

[33]See, for example, Burkhauser et al. (1999a and 1999b).

[34]Feldstein and Liebman (2000); Iams and Sandell (1997). These analyses were completed at Census Bureau data centers, which as explained in chapter 4, are controlled "safer settings" in which the data are protected.

linked personnel records and death records, described above.[35]

- Analyses of how type of insurance (Health Maintenance Organization (HMO) or fee-for-service) relates to early detection of cancer, as well as treatment outcomes, are based on linked clinical-insurance records (the SEER-Medicare link described above, which may be augmented by HMO data). For example, stage of cancer at diagnosis has been compared for aged Medicare beneficiaries receiving care in HMO versus fee-for-service settings.[36]
- In the development of annual intercensal population estimates for states and counties by subgroup (i.e., age group, sex, race, and Hispanic origin), a key component is state/county migration rates. Specific migration rates for areas and demographic groups are obtained through the SSA-IRS links described above.[37] The resulting population estimates are used by federal, state, and local governments, as well as many businesses, for a variety of planning uses (such as deciding where to build roads or schools or where to locate new businesses). In addition, Census told us that these estimates are used as a basis for allocating federal funds, constructing per capita rates for important health and economic indicators, and in refining results in many federally-sponsored surveys.

[35]Bailar (1991); Hoover (1999); IARC (1997). The Environmental Protection Agency is also using these results, together with various other data, to estimate potential ranges for wider population health risks from dioxin exposure.

[36]For analyses of early detection, see Riley et al. (1994, 1999). For analyses of cancer treatments received and outcomes by type of insurance, see Potosky et al. (1997, 1999).

[37]Census (1999, 2000b). Note that intercensal estimates are required by law (13 U.S.C. 181).

The information provided by these multiarchive linkages may not have been obtainable in any other way.

New Information Gained Through Person-by-Context Linkage

Person-by-context links—here, linkages of person-specific data and geographic data—provide new information on how community factors may influence individual behavior. Person-by-context links often build on other kinds of linkage. For example:

- The relative influence of contextual and interpersonal factors on teen delinquency are assessed by combining the Add Health multiperson links (discussed above) with contextual data.[38] This secondary linkage yields new information for improving youth programs.
- Linking HRS-SSA (survey-archive) links with contextual data generates information on how economic factors and policies may influence elderly persons' decisions concerning disability applications.[39]

Alternatively, person-by-context linkages are sometimes used to provide the additional "surrogate" or "proxy" data that are needed for special analyses. For example, the SEER-Medicare insurance records described above do not include data on individual patients' income or education levels, but such factors may affect patient outcomes. By linking publicly available census tract data on income and education to individual patients' clinical and insurance records, researchers were able to use a more complete model to examine patient outcomes in fee-for-service vs. HMO settings.[40]

[38]Greg J. Duncan et al. (1999).

[39]Burkhauser et al. (1999a).

[40]Potosky et al. (1997).

Next Steps and Questions for Further Study

Rather than attempting to develop a comprehensive representation of federal linkages, this chapter has drawn on a set of examples that illustrate how new information is generated. Questions relevant to a comprehensive or representative depiction would concern (1) the scope of federal record linkage efforts, (2) their goals and impacts, and (3) current federal agency plans, likely future directions, and barriers to linkage. We believe that addressing questions in these areas would represent a logical next step if further study were undertaken; more specific questions are outlined in appendix V.

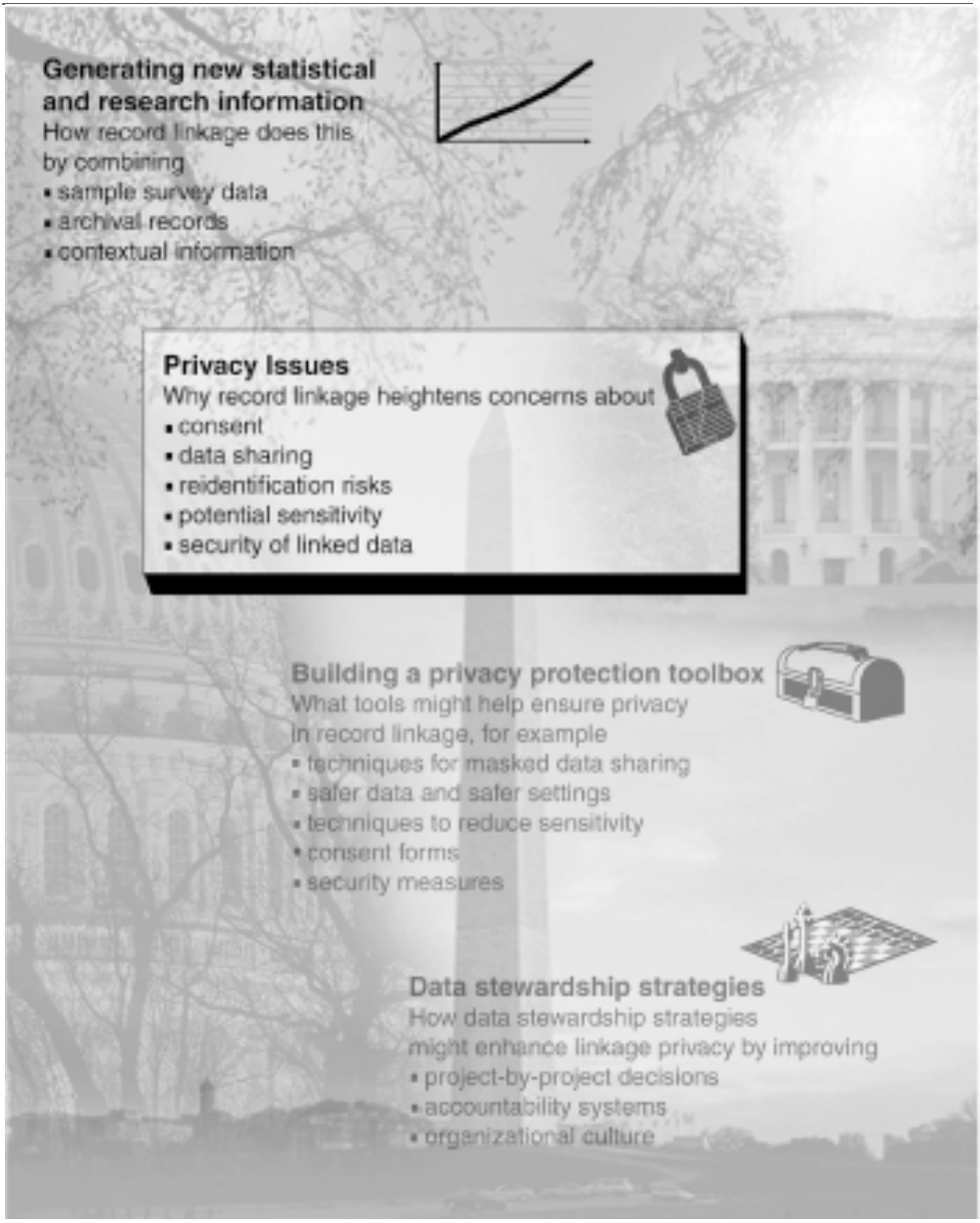
Privacy Issues

As highlighted on the opposite page, this chapter deals with privacy issues. (Subsequent chapters present various privacy protection techniques and strategies for data stewardship that are either being used now by various federal agencies or that might be used in the future.) Among the privacy issues that may arise with respect to record linkage for research and statistics are five examples:

- **Consent to linkage**. Although data subjects' consent to linkage is sometimes obtained, in other instances, data subjects may be unaware that, in essence, new information about them is being created.[1] Some linkages require data sharing between agencies, and when this occurs, certain laws and policies concerning disclosure and consent are relevant. Notably, the Privacy Act generally requires consent for disclosure from one agency to another, but there are exceptions (see chap. 1).[2]
- **Data sharing**. In order to compile the information needed for record linkage and "make the link," agencies must often share identifiable person-specific data. But traditionally, data have been kept separately, and various statutes have been enacted to prohibit or control certain kinds of data sharing. Privacy concerns stem from a desire to control information about oneself and a perceived potential for inappropriate government use, as explained below. Security risks could also arise during data transfer.

[1]Although various forms of notification do exist (see app. IV), in some cases, these may not inform most data subjects (Relyea, 2001). (Informed consent is discussed in app. I.)

[2]Somewhat similarly, OMB's Order Providing for the Confidentiality of Statistical Information limits sharing of personally identifiable survey data without respondent consent. However, while this may be relevant to specific linkages, we found that the examples of survey-archive linkage discussed in the previous chapter do not involve sharing survey data. Rather, in each instance, administrative data were transferred to the survey agency.


Generating new statistical and research information
How record linkage does this by combining
• sample survey data
• archival records
• contextual information
Privacy Issues
Why record linkage heightens concerns about
• consent
• data sharing
• reidentification risks
• potential sensitivity
• security of linked data
Building a privacy protection toolbox
What tools might help ensure privacy in record linkage, for example
• techniques for masked data sharing
• safer data and safer settings
• techniques to reduce sensitivity
• consent forms
• security measures
Data stewardship strategies
How data stewardship strategies might enhance linkage privacy by improving
• project-by-project decisions
• accountability systems
• organizational culture

- **Reidentification risks**. Some datasets are linked using a code-number procedure or are stripped of explicit identifiers as soon after the linkage as possible; nevertheless, reidentification of at least some data subjects may be possible through a deductive process, so only controlled use would be appropriate. To facilitate broader access to statistical and research data, agencies have created more fully "deidentified" public-use datasets.[3] Although many linked datasets are not made available for public use, some are—and concerns about the reidentification risks associated with these datasets are increasing.[4]
- **Potential sensitivity**. The potential sensitivity of data (risk of harm to data subjects) cuts across all other privacy issues. This is true for linked data as well as for single-source datasets. However, as explained below, linkage may heighten the sensitivity of data that, taken by themselves, appear to be relatively innocuous.
- **Security of linked data**. Security is crucial to protecting stored data. For linked data, this is especially true because a linked dataset may be more detailed or more sensitive than its components.

These various privacy issues are not unique to linked datasets, but may be more complex or challenging when linkage is involved. Most are relevant to each type of record linkage discussed in the previous chapter.

[3]Public-use datasets are released without restrictions on eligibility of data users or intended use (ICDAG, 1999).

[4]Deidentified datasets consist of microdata from which explicit identifiers have been stripped and to which other changes have been made to minimize the potential for reidentification of data subjects. By contrast, restricted access files, which may be made available to researchers through a variety of arrangements, include more detailed information—although in many cases the data are stripped of explicit identifiers.

Each of the five privacy issues can potentially be addressed through tools and techniques such as those discussed in Building a Privacy Protection Toolbox (chap. 4). Also relevant are strategies for effective data stewardship discussed in chapter 5.

Consent to Linkage

The issue of consent to linkage derives from a core concept of personal privacy: the notion that each individual should have the ability to control personal information about himself or herself.

Perceptions about the need for consent may vary according to type of linkage. For example, consent requirements have been advocated for multiarchive links (because full sets of existing records often do not have a voluntary component) and for linkages that are not closely related to the original purpose of the data collection. Consent requirements have also been advocated when vulnerable populations are involved or when risks appear to be higher.[5]

When consent to linkage is not obtained, data subjects might not know that their records are being linked. General notices may be provided, as discussed in appendix IV, but such notices may not specifically mention linkage.

Differences in opinion and practice may occur for the two major categories of linkage examined in this report: (1) person-by-person linkage and (2) person-by-context linkage.

[5]Scheuren (1997); George T. Duncan et al. (1993); Baily (1999); Thompson (2000).

Consent to Person-by-Person Linkage

There is a spectrum of opinions as to whether data subjects' consent is needed for person-by-person linkage.[6] For example

- *Consent to linkage should be obtained.* Meaningful consent to linkage may, in some instances, require careful description of anticipated benefits as well as confidentiality and security risks. Over time, check-backs with data subjects may be needed if research objectives—and data subjects' preferences—change.[7]
- *Consent from every data subject may not be needed*—at least if many data subjects are asked and the overwhelming majority consents.[8]
- *Consent to linkage may not be necessary if certain safeguards are in place*, such as review by a group with the interests of the data subjects in mind or use of appropriate confidentiality and security protections.[9]
- *Consent to linkage should not be required,* because that would be overly burdensome and potentially biasing or not practicable in some situations (e.g., it might be very difficult to recontact persons years after a survey or other data collection).[10]

The concept of informed consent emphasizes the need to adequately inform the data subject or other data provider (see app. I). We did not assess the set of linkage examples described in chapter 2 in terms of

[6]We believe these opinions pertain to survey-archive and multiarchive linkage; they may or may not extend to multiperson linkage.

[7]Gastwirth (1986), Scheuren (1997).

[8]Melton (1997).

[9]Scheuren (1997); Wallman and Coffey (1997).

[10]Related opinions have been expressed by Melton (1997); HHS (1999); Al-Shahi and Warlow (2000).

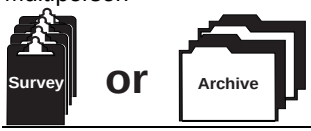
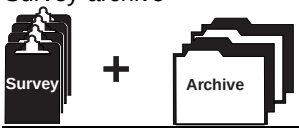
whether or not informed consent to linkage was obtained. However, for this small set of examples (which may not be generalizable), we did ask agencies whether or not consent was obtained, and if so, what procedures were used.

As shown in table 3.1, agencies told us that consent was obtained in four of our eight examples of person-by-person linkage. In one instance, the consent procedure consists of a signed consent form (see col. 1). In the other three, the survey interviewer or questionnaire asks respondents for a number that would be used to make the linkage; provision of that number was taken as consent (see col. 2).

Where provision of a number was taken as consent, practices varied. For example, in one survey (LSOA), the questionnaire provides a general statement concerning the need for the respondent's SSN.[11]

[11]According to the LSOA questionnaire, the interviewer says: "The Social Security Number allows Medicare records to be easily and accurately located and identified for statistical research purposes …. What is your Social Security number?" The questionnaire also states that providing the Social Security number is voluntary and will not effect the respondent's benefits in any way. (Kovar et al., 1992, p. 150.)

Table 3.1: Examples of Person-by-Person Linkage, by Consent Procedure

	Agency indicated consent was		
Linkage type	**Obtained—signed consent form**	**Obtained—other procedure**	**Not obtained[a]**
Multiperson 	[b]	Add Health best friends[c] *(NICHD*)*	Estate tax records, beneficiaries and decedents *(IRS*)*
Survey-archive 	HRS+SSA records *(NIA*)*	LSOA+Medicare records[d] *(NCHS*)* SIPP+SSA records[d] *(Census*)*	[b]
Multiarchive	[b]	[b]	SEER+Medicare records *(NCI,* HCFA*)* Personnel records+ death/address records *(NIOSH*)* Address (tax)+SSA records *(Census*)*

*Primary agency conducting or sponsoring the linkage.

[a]In each case where consent was not obtained, agencies explained their view that the link was not prohibited—and in some cases was authorized—by law.

[b]No example.

[c]Respondents were asked to use a school roster and report code numbers of best friends. If not provided, linkages were not made.

[d]Each respondent in LSOA and SIPP was asked for his or her SSN (and in some cases, those of other household members). In the LSOA, caregivers or others were asked to provide the SSN for incapacitated persons. If not provided, linkages were not made. Both NCHS and Census told us that they informed respondents about the reason for asking for the SSN.

In another (SIPP), the questionnaire itself does not include an explanatory statement, although the interviewer's manual spells out language for the interviewer to use in explaining why the SSN is needed.[12] There is also a notice on the back of the introductory letter sent to SIPP respondents. The LSOA and SIPP materials include a statement that provision of the SSN is voluntary.

In all four instances where agencies said that consent was obtained, a survey was involved.[13] By contrast, for all four linkages where agencies told us that consent to person-by-person linkage was not obtained (see col. 3), no survey data were involved.[14]

In each case where consent was not obtained, agencies explained their view that the link was not prohibited—and in some cases was authorized—by law. For example, in the case of intercensal estimates, IRS and SSA are authorized by statute to disclose data to the Census Bureau for statistical estimates.

On one hand, there may not be a viable mechanism for obtaining consent to person-by-person linkage at the

[12]The SIPP interviewer manual states that the survey "collects social security numbers so we can obtain information that was provided to other government agencies. This helps us avoid asking questions for which information is already available and helps ensure the accuracy and completeness of the survey results. We protect administrative records information that we obtain from these agencies from unauthorized use …. Providing your social security number is voluntary." (U.S. Census Bureau, 1997, p. 4-8.)

[13]We do not know how generalizable this pattern is.

[14]This is consistent with an earlier finding that "To the extent that statistical uses of administrative data are permitted by statutes and regulations, the data subjects and providers are usually not asked for their consent, and … in some instances they are not even given any notification of such uses." (George T. Duncan et al., 1993, p. 72.)

time that administrative data are generated.[15] Administrative data may be created continually, and agency willingness to change the basic forms to include items on consent to linkage would likely vary across agencies. Records-research databases may be compiled years after the original records were generated, and in these instances, it could be very difficult to locate many of the persons involved; others might be deceased. Full sets of existing records may be so large as to discourage efforts to recontact data subjects.[16]

On the other hand, there may be a heightened need for consent to linkage in cases where the original data collections were mandatory or not optional.

Consent to Person-by-Context Linkage

Turning to the issue of consent to person-by-context linkage, for the three examples presented in chapter 2, agencies told us that consent was not obtained. Some believe that consent is not necessary for such linkages because they do not bring together two separate person-specific records. This seems a reasonable position for some person-by-context links (e.g., those that link state-level data to person data and are carefully controlled).

But if a public use dataset will be created, person-by-context links might significantly increase reidentification risks. There may also be instances where person-by-context links create new and

[15]Some records are generated under circumstances where it is difficult to achieve informed consent or to assure data subjects that consent is voluntary. For example, when clinical medical records are generated in emergency or other serious or potentially life-threatening medical circumstances, patients might not be capable of giving full attention to the consent-to-linkage issue; also, they might fear (even if erroneously) that not consenting could affect the speed of their treatment.

[16]See, for example, George T. Duncan et al. (1993).

potentially sensitive information about an individual. And, as explained below, more extensive context links may be made in the future. (See sections below on reidentification risks and potential sensitivity.) For these reasons, we believe that the issue of consent to person-by-context linkage might be relevant for some current applications—or heightened in the future.

Data Sharing to "Make the Link"

For purposes of this report, we use the term data sharing to refer to the transfer of personally identifiable data across agency (or other organizational) lines, including "one-way" sharing. A particular record linkage project may or may not require data sharing. For example, data sharing is not required for the Add Health linkage of teen best friends (because only one dataset is involved). By contrast, data sharing is required for the linkage of addresses from IRS income tax records to SSA records (because these datasets are maintained at different agencies and because the actual linkage is conducted by Census).

Data sharing can be limited by the Privacy Act and by various agency statutes; for example, sharing of IRS data is prohibited by statute, with certain exceptions.[17] Based on these exceptions, two of the examples discussed in the previous chapter involve transfer of IRS data to Census and NIOSH.[18]

While data sharing has many legitimate uses and potential benefits,[19] privacy issues stem from perceptions about the possibility of government misuse. Historically, government officials intent on

[17]26 U.S.C. 6103.

[18]26 U.S.C. 6103(j)(1)(A) and 6103(m)(3).

[19]See NRC (1985) for a discussion of the benefits of data sharing for research purposes.

conscripting soldiers (World War I), interning persons of Japanese origin (World War II era), or identifying tax data for selected individuals for political purposes (Watergate era) have made requests for transfers of information.[20] Despite the 1974 passage of the Privacy Act (which can, under some circumstances, limit data sharing), negative perceptions may persist.

Currently, there are two sets of data-sharing issues relevant to record linkage: (1) functional separation and (2) risks to confidentiality and security.

Functional Separation

Functional separation concerns the issue of when it is proper for an agency to share data with another agency.[21]

Principle of Functional Separation

Data collected for research or statistical purposes should not be made available for administrative action toward a particular data subject.

According to this principle, individually identifiable information collected or compiled for research or statistical purposes, which logically would include survey data as well as records-research databases, may enter into administrative and policy

[20]In 1917, Census provided names and addresses of draft-eligible persons to other federal officials (GAO, 1998a). In 1942, Census identified neighborhoods with concentrations of Japanese to the War Department (GAO, 1998a). In the 1970s, tax information on a number of individuals was disclosed to the White House (George T. Duncan et al., 1993, p. 48).

[21]In response to the need for data-sharing safeguards, the Privacy Protection Study Commission and the National Academy of Sciences' Panel on Confidentiality and Data Access emphasized the principle of "functional separation."

decisionmaking only in aggregate or completely anonymous form.

Although it is generally agreed that research and statistical data should be protected from being used in government actions taken with respect to specific individuals,[22] record linkage sometimes involves sharing research and statistical data with specific individuals or units within program agencies. For example, records-research databases may be shared with a statistical or research office within the program agency. For such situations, data-sharing arrangements and agreements have been developed to protect against improper uses.[23] Still, some experts believe that linkage involving statistical and research data and program data should be limited to "one-way" sharing.[24] That is, they believe that the program agency should transfer the data to the research or statistical agency, and that there should be no reverse flow of identifiable data.

Since the 1970s, various groups have indicated that the use of individually identifiable administrative records for research or statistics should be permitted based on demonstrated need to achieve an important research objective and assurance of stringent safeguards.[25] However, issues have recently been

[22]This issue of misuse does not apply to linkages that involve only data developed for administrative purposes and subjected to administrative uses (e.g., to check eligibility for benefits or loans).

[23]For example, Jabine (1993) indicates that such arrangements may include swearing in all employees of the program agency with access to a shared file as employees of the statistical agency, among other protections.

[24]With one-way sharing, it seems more obvious that the confidentiality of survey data can be protected.

[25]Privacy Protection Study Commission (1977); Fanning (1998). See also Wallman and Coffey (1997).

discussed regarding whether, or under what conditions, administrative health records should be available to researchers.[26]

Risks to Confidentiality and Security

Other privacy issues in data sharing include risks to confidentiality (more organizations and more persons are privy to identifiable data) and certain security risks (e.g., risks during transfer). Some see data sharing as inherently risky and believe that increased safeguards may be needed—especially during transfer. (Safeguards for secure transfer are discussed in chap. 4.)

Because different data are stored at different agencies, data sharing may be required for many survey-archive and multiarchive links. As noted at the outset of this study, we did not develop a compendium of linkage activities. However, each of the six examples of survey-archive and multiarchive links discussed in the previous chapter involves transfer of personally identifiable information across units.[27] They illustrate how data sharing may work, as follows.

[26]Notably, in the health privacy debate, it has been emphasized that when consent is not obtained for the research use of personally identifiable medical records, there are higher requirements for confidentiality and security protections. This was recognized in a federal regulation for privacy standards that was issued on December 28, 2000, and was to go into effect on February 26, 2001. For example, without consent, information may not be disclosed for research unless a review body finds that the proposed research has adequate plans for protecting identifiable information and meets other criteria ("Standards for Privacy of Individually Identifiable Health Information," 65 Fed. Reg. 82462, to be codified at 45 CFR Parts 160 through 164). HHS has changed the effective date to April 14, 2001, and has provided a new comment period to consider revisions. For related GAO work, see GAO (2000f) and GAO (2001a, 2001c).

[27]Because different agencies collect different kinds of data and because the combination of different kinds of data yields new information, this result seems logical. However, it is entirely possible, for example, for a research agency to conduct a survey

- Each of the six data-sharing linkages involves cooperation either between a statistical agency (Census or NCHS) and a program agency (SSA or HCFA) or between a research agency (NCI or NIOSH) and a program agency.[28]
- In some cases, the data transfers are essentially "one way." For example, various kinds of IRS data are transferred to Census, but linked data are not sent back to IRS.[29]
- In other cases, there is mutual sharing. For example, the SSA transfer of data to the HRS team at the University of Michigan is based on consent statements signed by each respondent or data provider. The agreement between the University and SSA provides that the linked HRS-SSA data will be shared with SSA after removing explicit identifiers and that SSA will not attempt to reidentify survey respondents.

Data sharing necessarily involves physical or electronic transfer, or both. Special security risks could arise during transfer or transmission because of the potential for interception, loss, or delivery to parties other than the intended recipient.[30]

(We note that multiperson linkage typically occurs within a single dataset; in such cases, data sharing is not involved. Similarly, person-by-context links can be achieved by transferring the contextual data to the

and to link those data to a records-research database maintained by that same agency. We do not know how often the different patterns may occur.

[28]In two of the six cases, the data transfers were made across units within HHS; one between NCHS and HCFA and the other between NCI and HCFA.

[29]Legal provisions applicable to Census prohibit sending nonpublic linked data back to IRS.

[30]NRC (1997).

agency housing the survey or archive, thus avoiding the need to transfer person-specific data.[31])

Deidentification and Reidentification

Federal agencies have a long history of creating public-use datasets with deidentified information. These deidentified data have provided researchers and members of the general public with maximum access to information; they have also helped agencies maximize the return on their investments in data collection.

Growing concerns about reidentification risks have led to considerable efforts to develop methods aimed at minimizing these risks.[32] Such risks might be higher for linked datasets than for the component data. Agencies may face trade-offs between (1) attempting to meet the difficult challenges of minimizing the reidentification risks associated with wider access and (2) providing more restricted forms of access to confidential data.[33]

Many linked datasets are never made available for public use. For example, Census' linked data from IRS and SSA, which help produce intercensal estimates, are not made available to the public; but other linked datasets are (e.g., Add Health's person-by-context links and parent-child links—which the study deems low risk).[34]

[31]Of course, it might be that in some cases, the contextual data would involve confidential information on, for example, establishments where individuals are employed.

[32]Disclosure limitation ("safer data") and restricted access techniques ("safer settings") are described in chapter 4.

[33]See NRC (2000), chapter 2: "The Data Access, Confidentiality Tradeoff."

[34]Add Health provides a public-use dataset containing interview data from adolescents, test scores, data from in-school and parent

As background, the Privacy Act generally prohibits agencies and staff from disclosing or releasing identified data, subject to certain exceptions. Public-use datasets contain "anonymous" microdata. That is, the information is on individual persons or organizations but

- explicit identifiers have been stripped away and
- federal agencies are required to make sure that the identity of individuals cannot be "reasonably deduced."[35]

Thus, for example, potentially identifying information, such as birth date and zip code, are made less specific than in the original data (e.g., changed to year of birth and region of the country).[36] These seemingly small changes can considerably reduce reidentification risks—although "zero risk" may be an impossibly high standard.[37]

In recent years, concerns about reidentification have grown, however, because the proliferation of computer technology may enable "data snoopers" or

questionnaires and individual summary friendship network characteristics, but no "best friends" linkage and no teen-neighbors linkage. The CD-ROM containing the public-use dataset also includes community contextual variables. The public-use dataset consists of 50 percent of selected samples. (Best-friend linkages are not included in the public-use dataset.)

[35]According to OMB's 1975 guidance for implementing the Privacy Act (40 Fed. Reg. 28948, 28954 (July 9, 1975)), this means that the identity of the individual cannot be determined by combining various statistical records or public records or other available sources of information. This may be increasingly difficult for agencies to achieve because more information is becoming available to a greater number of persons through the proliferation of computers and the Internet.

[36]Other disclosure limitation techniques are discussed in chapter 4.

[37]See, for example, George T. Duncan et al. (1993).

"data detectives" to link deidentified public-use datasets to other information (e.g., voter registration lists) in order to reidentify specific persons.[38] When information is linked in this way, some believe that it can potentially provide "an electronic shadow of a person … that is as identifying and personal as a fingerprint even when the information contains no explicit identifiers."[39] Of course, misidentification is also possible. At least one agency (NCHS) warns users against attempting to reidentify persons in its public-use datasets, but no penalties are specified.

Reidentification risks may be higher for datasets with person-by-person linkages than for their components because of the following.

- The greater depth of linked data makes reidentification easier. To take an example from multiperson linkage, if data on husband and wife are linked, just knowing the age, occupation, and ethnicity of both partners may make some atypical couples very identifiable. A couple might be easily identified if, for example, one spouse is considerably younger than the other, both are physicians, and one is Asian, the other not—especially when this is combined with other information, such as rural residence in the Midwest, income category, and number of children.[40] (Of course, even in the absence of linkage, very detailed datasets would also be potentially at risk.)
- As has been pointed out, whenever data from two files are linked (as occurs in survey-archive or multiarchive links), persons or groups with access to one of those

[38]Sweeney (1997; in press (b)).

[39]Sweeney (in press (c)), p. 1.

[40]As explained in chapter 4, disclosure limitation techniques look for these unique cases and alter them to protect confidentiality in public use microdata.

files may be able to reidentify individuals in the linked dataset. For example, in survey-archive links, those with access to the archive at the source agency could, potentially, identify survey respondents in the linked dataset.[41] This risk could be serious when the archive in question is administrative data from a program agency.[42]

Turning to person-by-context links, these could also facilitate reidentification. Suppose, for example, that survey data on persons are linked to data on number and type of churches or other religious organizations in each neighborhood. Even if zip codes are stripped away before releasing the public-use dataset, anyone with access to data on the location of religious organizations might identify residential areas, which in turn would make reidentification of specific individuals much easier.

As indicated in the previous chapter, some believe that person-by-context linkage is increasing. One expert suggested that eventually an "avalanche" of contextual information—proximity to local programs, information on schools and employers (which may not be publicly available), crime and disease rates for neighborhoods, as well as median income for census tracts—might be assembled for linkage to a single set of person-level data. If this were to occur, the reidentification risks associated with person-by-context links might be raised considerably for public use datasets. Indeed, it might be very difficult to overcome these risks even when using the "safer data"

[41]Federal Committee (1994), p. 63.

[42]The reason is that the program agency has a mission that involves taking action with respect to individuals. (See chap. 4 for potential solutions.) The same risk is not present when the full set of existing records is a dataset compiled by a research agency for research purposes.

techniques described in the next chapter, and agencies generating linked person-by-context datasets might decide not to make them available for public use.

Of course, as noted above, many linked datasets are not currently made available for public use. Some are, however, and in those instances, the linked data may "pos[e] special risks for reidentification."[43] Public-use datasets maximize access to information, and for that reason, they are championed by some. Alternatives for making linked data accessible to researchers without resorting to public-use datasets are discussed in chapter 4.

Potential Sensitivity of Linked Data

The privacy issues discussed above—consent to linkage, data sharing to make the links, and the reidentification risk associated with dissemination of linked data—are all intensified when sensitive data are involved. And when sensitivity is increased, there is also a need for greater caution in releasing identifiable linked data to researchers outside the linking organization(s).

This is important because federal record linkage often involves sensitive information and we believe that the linkage itself can heighten sensitivity, as explained below.

Although sensitivity is a subjective and relative concept, certain laws provide protection for what could be considered sensitive information. For example, the Computer Security Act of 1987 defines sensitive information as including any unclassified information that, if lost, misused, or accessed or modified without authorization could adversely affect

[43]Robbin et al. (1999).

the privacy to which individuals are entitled under the Privacy Act.[44]

Even aside from the law, certain stigmatized or illegal behaviors would seem to clearly qualify as sensitive.[45] Some of the Add Health questions (teenage drinking, delinquency, and sexual behavior) would fit in this category. But other information—particularly, financial or medical information—might also be sensitive in that it could affect someone's ability to obtain a job or a mortgage. Under this widened definition, most of the examples discussed in the previous chapter would be at least somewhat sensitive.

Various data that appear to be of low sensitivity can become more sensitive when linked. For example, if a person's survey report of income is linked to his or her tax return—and the results indicate disparate income reports—the linked data would be more sensitive than the original independent data (because there is a new implication about the individual). Even some context links could create sensitivity by, for example, identifying persons associated with residential areas, schools, or places of employment with negative characteristics (e.g., high rates of stigmatized diseases). In instances where negative contextual

[44]See P.L. 100-235. For an example of a law that does not use the term "sensitive," but provides protections for information that could be considered sensitive, see the Comprehensive Alcohol Abuse & Alcoholism Prevention, Treatment and Rehabilitation Act of 1970. This law makes records on substance abuse patients confidential. These records relate to the identity, diagnosis, prognosis, or treatment of such patients (42 U.S.C. 290dd-2).

[45]Sensitive health information can include mental illness, sexual behaviors or related diseases, and illegal drug use. Other categories of sensitive information can include welfare payments, family fights and reputation, criminal history records, financial status, and so forth.

information is either not known to the public or difficult to access, linkage to a person-specific dataset might increase the sensitivity.[46]

Overall, it seems fair to say that sensitivity is potentially increased whenever the "whole is greater than the sum of the parts." And for a variety of reasons, certain questions—or linkages—may be perceived as sensitive by at least some data subjects even if there appears to be no risk of harm in the eyes of the researcher or other outside observer.

Security of Linked Data

Security is important for all personal data and crucial for sensitive personal data. As noted above, even data that appear to be of relatively low sensitivity may become more sensitive when linked. At the same time, security has become particularly challenging as access to computers and the Internet has spread through the population, and agencies rely more extensively on computerized systems and electronically available data. Therefore, although the basic mechanisms of security are the same for linked and component datasets, we briefly cover key security techniques in the next chapter, Building a Privacy Protection Toolbox.

Next Steps and Questions for Further Study

This chapter has outlined some key privacy issues relevant to record linkage. Delineating privacy issues in a more comprehensive manner would mean addressing at least three sets of questions. The first set of questions concerns more detailed information on the legal framework(s) within which different types of record linkages occur and the variation in legal and regulatory protections across agencies; potentially, questions about the effectiveness of current legal protections could also be addressed. The second set

[46]Of course, not all linkages increase sensitivity.

of questions concerns current agency policies and practices with respect to the privacy issues discussed in this chapter (including variation in those policies and practices across agencies). A third set of questions would address the issue of whether there are other relevant privacy concerns, that is, concerns additional to those outlined in this chapter.

Building a Privacy Protection Toolbox

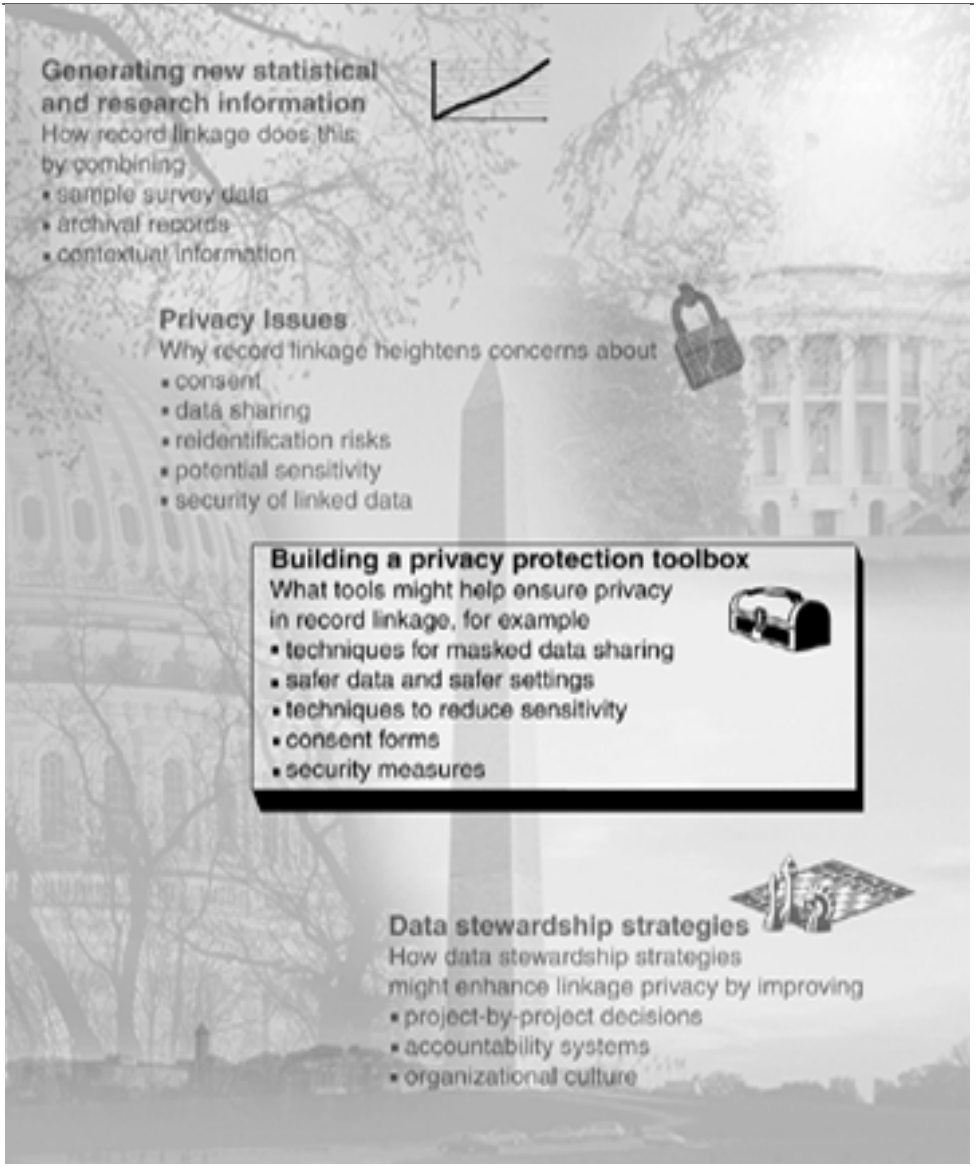
Is it possible to protect privacy while conducting record linkage? As highlighted on the opposite page, this chapter addresses the third key record linkage topic identified at the outset of this study: building a privacy protection toolbox. Specifically, we enumerate several techniques designed to address the privacy issues discussed in the previous chapter. Some of the techniques are uniquely relevant to linkage, while others have more general application.

We believe that a useful privacy protection toolbox would contain a variety of statistical, procedural, or other tools for protecting the privacy of data subjects or otherwise building in confidentiality and security. Among the tools relevant to linkage would be

- techniques for masked data sharing,
- procedures for reducing reidentification risks (including safer data and safer settings), and
- techniques to reduce the sensitivity of the data being linked.

Examples of such techniques are described below. Other issues discussed in this chapter include (1) the relevance of the toolbox to the consent issue and (2) security measures for stored data.

It is important to realize that some of the techniques described here are now in use at some agencies and would rightly be termed currently available “tools.” Others are procedural and may require some feasibility assessment. Still other techniques are best considered potential tools because they are statistical techniques that have not been assessed with specific reference to record linkage or are still in experimental form.


Generating new statistical and research information
How record linkage does this by combining
• sample survey data
• archival records
• contextual information
Privacy Issues
Why record linkage heightens concerns about
• consent
• data sharing
• reidentification risks
• potential sensitivity
• security of linked data
Building a privacy protection toolbox
What tools might help ensure privacy in record linkage, for example
• techniques for masked data sharing
• safer data and safer settings
• techniques to reduce sensitivity
• consent forms
• security measures
Data stewardship strategies
How data stewardship strategies might enhance linkage privacy by improving
• project-by-project decisions
• accountability systems
• organizational culture

To date, most attention has been focused on the development of techniques for reducing reidentification risks.

We describe a variety of techniques in hopes of stimulating increased efforts to build an effective privacy protection toolbox for record linkage. We further note that many techniques carry costs in terms of information loss or additional expense,[1] but at the same time, they may provide information gains. For example, some privacy protection techniques may result in a loss of precision, but may make it possible to obtain information that otherwise would not have been feasible (because of ethical concerns, restrictive laws, or lack of cooperation).[2] Given the potential advantages and disadvantages, we believe these techniques should be selected and applied with care.

Techniques for Masked and Secure Data Sharing

Techniques for masked sharing or linkage include list inflation, third-party models, and grouped linkage. Secure transfer is aided by techniques, such as encryption, as well as physically secure transfer vehicles (e.g., secure data lines). Safeguard reviews can help ensure that security measures are being followed in another agency. These various approaches illustrate how a privacy protection toolbox might help safeguard data, even when linkage requires sharing data across organizations.

List Inflation

List inflation might be used to protect confidentiality when, for example, a statistical agency requests administrative data for survey respondents from a

[1]See, for example, Ruggles (2000).

[2]One instance where this may apply is the "three-card method," an indirect technique initially designed for asking foreign-born persons about their immigration status but potentially applicable in a variety of other areas. This method is described later in this chapter (see p. 93).

program agency.[3] Suppose, for example, that Agency 1 has conducted a survey and requests administrative data from Agency 2. To keep Agency 2 from knowing, with certainty, who was in the survey,[4] Agency 1 could inflate the basic list of survey respondents (e.g., adding random SSNs or SSNs from other surveys to those of the survey respondents) before transferring the list to Agency 2. This approach, commonly called "comingling" by some or "salting" by others, is always used by NCHS.

Using list inflation is essential to preserving confidentiality when the basic list is, in itself, sensitive. For example, as described to us by officials at SAMHSA, researchers in the State of Washington use linkage to evaluate drug treatment. Here, the basic list consists of persons who have received treatment for drug addiction. This list is inflated before it is sent to the state unemployment office with a request for administrative records on these persons' employment or unemployment.

Third-Party Models for Masked Data Sharing

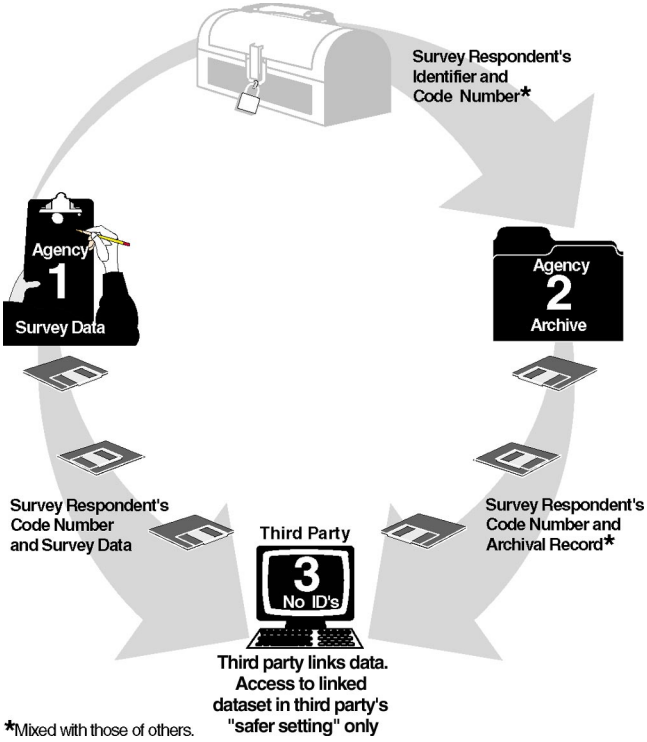
Third-party models for masked data sharing use a three-way linkage procedure to ensure that no one—not even the third party or the agencies supplying the data—will ever have access to both personal identifiers and linked data.[5] To illustrate, suppose that a key analysis requires Agency 1 survey data to be linked with existing records that are maintained at Agency 2 (a program agency). The third-party model (shown in fig. 4.1) would work as follows.

[3]Boruch and Cecil (1979), p. 122.

[4]Knowing who is in a survey potentially increases reidentification risks.

[5]Third-party models are included in the procedural models described by Boruch and Cecil (1979).

Figure 4.1: The Third-Party Model for Masked Data Sharing



Source: GAO, based on Boruch and Cecil (1979).

Agency 1 would send Agency 2 the personal identifiers of, for example, 10,000 respondents in the survey (e.g., SSNs) plus code numbers assigned to each respondent by Agency 1, but no substantive survey data. (For this step, list inflation could be used, as described above, to prevent Agency 2 from knowing who was in the survey.)

Agency 2 would sort through its administrative records, which might cover millions of persons, and identify the 10,000 that apply to the survey respondents in question (or all those that apply when

using an inflated list). Agency 2 would send these records—stripped of personal identifiers, but bearing the code numbers assigned by Agency 1—to a trusted third party.

Agency 1 would send the trusted third party the substantive survey data with each respondent identified only by code number.

The third party would receive survey data with code numbers only (from Agency 1) and administrative records for survey respondents (and others), again with code numbers only (from Agency 2). The third party would proceed to link the two sets of data, person by person, based on the code numbers (discarding the additional data from Agency 2). No personal identifiers would ever be seen by the third party.

After linking the datasets, the third party would remove the code numbers. In some situations, it might be appropriate for the third party to analyze the linked dataset, publish results without further disseminating the linked data, and then to destroy all data received. Alternatively, there might be restricted access to the linked data through a data center or "safer setting" maintained by the third party (see section below on safer settings and other techniques to reduce reidentification risks).[6]

If Agency 1 and Agency 2 were allowed to access the linked data outside a "safer setting," reidentification risks might be high. To maintain assurance of confidentiality—which would be particularly important for Agency 2, in order to maintain

[6]In providing linked data or results, the third party might have to assume some of the responsibilities (e.g., data quality assurance) that normally would have been carried out at Agency 1 or Agency 2.

functional separation—such risks must be minimized. One approach would be to limit all use to a safer setting.

Essentially, third-party models are a privacy-oriented variation on the "honest broker" concept.[7] Despite the masking of data, the privacy protections of this model depend in part on the trustworthiness of the third party.[8] Doubtless most agencies would prefer the ease of dealing directly with each other, but third-party models have the potential to generate new information while greatly reducing risks to personal privacy.

Grouped Linkage

Grouped linkage, or grouped data sharing, avoids transferring individual-level data.[9] To illustrate, suppose Agency 1 has survey data on, for example, test scores and is now interested in the relationship between test scores and certain administrative data (e.g., income) maintained at Agency 2. Agency 1 would form groups of approximately 1,000 individuals each based on their test scores (e.g., those with the lowest scores in the first group, highest in the last group). Agency 1 then sends Agency 2 identifying information (e.g., names, SSNs) for the persons in each group — asking Agency 2 for aggregate information on each group. Thus, for example, Agency 2 would calculate average family income for each group, based on its administrative data, and transfer these averages to Agency 1. Agency 2's administrative data on specific

[7]An "honest broker" is a neutral intermediary.

[8]The Census Bureau said that "Some people may not view third party models as reducing risks to personal privacy, especially if the third party is outside government."

[9]Spruill and Gastwirth (1982). See also Boruch et al. (2000), citing Schwartz and Orleans (1967).

individuals' incomes would never be transferred to Agency 1.

A variant on this approach—incorporating the list inflation technique—is needed to prevent Agency 2 from (1) knowing who is in the survey and possibly (2) deducing some substantive information about those individuals (e.g., deducing that some specific persons have low test scores). In addition to the "real" groups described above, a number of additional, "sham" groups, designed to show a distribution of incomes but not based on test scores, could be transferred to Agency 2. These sham groups would be based on list inflation, using, for example, persons who participated in a different study. When Agency 1 receives the aggregate data for each of these groups back from Agency 2, Agency 1 would simply weed out the sham groups.

Secure Transfer of Data

Tools for secure transfer of data from one data-holder to another are useful even when other privacy protection tools are in use. For example, the third-party model discussed above provides protection by transferring different portions of the data separately. Nevertheless, if someone were to intercept two or three of the transmissions, confidentiality would be compromised. Thus, tools for secure transfer are crucial.

Tools for secure transfer include encrypting all materials sent (i.e., converting clear text to code) and a variety of other mechanisms. There are two main kinds of encryption. Secret key encryption involves a code known to both sender and receiver, which is not disclosed to others and is transmitted separately from the data. Here, the same key is used by sender and receiver. Alternatively, a two-key system, introduced in the 1970s, is designed to provide greater protection and authentication of the data. A "user" has a set of

two mathematically related keys—one key is used for encryption while the second key is used for decryption; one is "public" and given out to senders to encrypt data, while the other is private and known only to the user. Only this private key can decrypt transmissions coded with the "matching" public key.

Additional mechanisms for secure transfer include, for example, using secure dedicated lines (rather than the Internet) and separately transmitting substantive data, personal identifiers, and linking codes.[10]

Research to further develop these kinds of tools was recently recommended by an Institute of Medicine committee.[11]

Written Agreements and Safeguard Reviews

Written agreements (some of which are known as MOUs) and safeguard reviews can help ensure that security measures are taken when information is shared across organizational lines.[12] Written agreements between the agency or organization providing the information and the recipient organization may specify the security governing the use and storage of the information and incorporate periodic inspections to verify that appropriate measures are in place. For example, IRS requires internal security inspections by the recipient agency

[10]NRC (1997), Dam (1996), Keller-McNulty (1993), GAO (1993, 1995).

[11]See IOM (2000).

[12]MOUs refer to memorandums of understanding, which are legal documents that, when data are shared, specify the explicit uses and conditions for use of that data by the recipient agency. For example, an MOU may cover the purpose and scope of the data sharing, any research use of the data, and assurances that the data will be kept confidential.

and conducts its own safeguard reviews of the recipient agency under certain circumstances.[13]

Procedures for Reducing Reidentification Risks

As explained in the previous chapter, reidentification issues are not unique to linked datasets, but the issues are more complex for linked datasets.

When deidentifying data for public release, the first step is to strip obvious identifiers. This is necessary, but it is not sufficient because there are still reidentification risks. At an October 1999 workshop, researchers and agency representatives expressed optimism about "the possibilities for developing tools that would enhance … the ability to increase data access without compromising data protection or conversely, to increase confidentiality without compromising data access."[14]

Most tools or techniques for reducing reidentification risks can be categorized in one of the following groups:

- traditional "safer data" techniques; notably, data-altering techniques developed by statisticians at the Bureau of the Census and the federal interagency Confidentiality and Data Access Committee, among others;
- more radical synthetic or simulated data techniques, which are being discussed and debated; and
- "safer settings," such as the data centers or "data enclaves" pioneered at Census, NCHS, and some other

[13]IRS (2000); Federal Committee (1994).

[14]NRC (2000), p. 5.

agencies, and other procedures, such as licensing, that essentially control access to the data.[15]

The advantages and disadvantages of these tools are currently being discussed.[16] Some approaches combine two of them. All are potentially appropriate for use with linked data. In any case, when preparing microdata for public release, the amount of "noise" added has to be enough to ensure (with reasonable certainty) that individuals cannot be identified.

Traditional Techniques for Safer Data

In developing and selecting safer data techniques (sometimes called disclosure limitation techniques), the challenge is to retain the essential meaning and analytic utility of the data while introducing sufficient distortions or eliminating sufficient portions of the data to protect confidentiality. Safer data techniques are summarized in a key interagency working paper.[17] Basic techniques for microdata include limiting geographic detail; top-coding quantitative variables (e.g., coding all households with incomes of $200,000 or more in a single category) and, analogously, bottom-coding; recoding quantitative variables into intervals (e.g., creating income groups); and limiting the overall number of variables in the dataset to be released to avoid identifying "population uniques."[18] Among the more complex techniques are adding "noise" or random error to quantitative variables and

[15]The terms "safer data" and "safer settings" are derived from those used by Marsh et al. (1994)—"safe data" and "safe settings."

[16]NRC (2000).

[17]Federal Committee (1994), pp. 20-4.

[18]A population unique occurs when only one individual or couple or family in a population has a certain combination of characteristics, and these characteristics are described in the dataset.

"swapping" or switching the values of selected variables for similar respondents or data subjects.[19]

But which variables should be modified or eliminated? One prioritization scheme categorizes each variable in terms of its likely disclosure risk and its analytic utility, perhaps based on expert judgments.[20] Variables with low analytic utility would be eliminated from the microdata. Those with high analytic utility and low disclosure risk would not be eliminated or modified. Priorities for eliminating or modifying other variables would depend on their combined utility-disclosure risk category.[21]

(We note that other disclosure limitation or "safer data" techniques are relevant to detailed tabular data, e.g., suppression of cells that contain too few respondents or data subjects. Some agencies require at least five subjects in a cell, others require three.[22])

While some researchers and statisticians have championed these approaches, others have emphasized that such techniques cannot be demonstrated to protect microdata in a mathematical

[19]A different approach is represented by the ongoing development of computer programs to distort and protect data (e.g., μ-ARGUS for microdata, and Datafly and Datafly II Systems). See Hundepool et al. (1997); Hundepool et al. (1998a); Sweeney (in press (a)).

[20]A team of disclosure analysts are asked to categorize variables as high, medium, or low disclosure risks. Expert users of the data are then asked to judge the analytic utility of potential disclosure variables (Rasinski and Wright, 2000).

[21]When microdata represent an entire population or defined subgroup, an additional method of disclosure limitation is to release data for only a sample. Or, for some sample surveys, it may be advisable to release microdata for only a subsample of respondents.

[22]Federal Committee (1994). A computer program for altering and protecting tabular data is τ-ARGUS (Hundepool et al., 1998b).

or absolute sense.[23] Today, "safer data" may be increasingly vulnerable as a result of ongoing changes in computer or Internet access and the increasing availability of records in electronic form. Taken together, these factors heighten the potential for reidentification even when some safer data techniques have been employed.

Synthetic or Simulated Data

A more radical approach to ensuring anonymity in microdata is the creation of entire synthetic or simulated datasets. As proposed in 1993, not one unit or data subject in the simulated microdata would be the actual unit or data subject.[24] Rather, artificial units or data subjects would be created based on knowledge of the real data, using multiple imputation. This means that essentially, for a simulated data subject, numerous potential values would be projected or imputed for each variable; the imputations would be done in such a way that interrelationships between variables would be preserved.

This ambitious approach has been used in a limited and exploratory fashion with data from the Federal Reserve Board's Survey of Consumer Finances,[25] and the results have been interpreted as indicating that it can be done.[26]

[23]For example, George T. Duncan et al. (1993, p. 137) states, "Zero risk requirements for disclosure of statistical records are, in practice, impossibly high standards." Some feel that the issue may be one of trade-offs—that is, that some low level of risk may be justified by the potential benefits of increased access.

[24]Rubin (1993). The Census Bureau told us that "it is very difficult to develop a synthetic dataset that will act like a real one for analytic purposes."

[25]Kennickell (1997).

[26]NRC (2000).

Like the "safer data" approach described in the foregoing section, the synthetic data approach described here would provide access to many users—potentially filling the role carved out by the more traditional techniques but providing greater protection for data subjects. Still, it would not appear to meet the needs of everyone. At a recent conference, "the perception among leading researchers appear[ed] to be that altered … [or] synthetic data can solve some problems but are inadequate for the majority of cutting-edge work."[27] Some researchers therefore prefer the "safer settings" and controlled access approaches described below.

Finally, as one expert pointed out, key policy decisions are not likely to be based on simulated data. Perhaps one solution would be for key analyses to be repeated by the agency in question, using the real data. (NCHS uses this approach with "practice data," as described below.)

Safer Settings, Controlled Access, Penalties

Three key approaches to protecting confidentiality by controlling access to linked datasets (rather than by altering, distorting, or reducing the data) are

- providing safer settings,
- using practice or "sham" data, and
- entering into licensing agreements.

Safer Settings

Safer settings (or controlled settings in which researchers can access identifiable data) are sometimes called data research centers or "data enclaves." These have been established at some federal agencies, where applications for use are reviewed. At NCHS' data center, all obvious personal and geographic identifiers are removed, but otherwise,

[27]NRC (2000), p. 9.

the data are generally unchanged. No external data can be brought into the enclave.[28] Only tabulations examined for statistical disclosure risks can be removed from it.[29] (See also, "practice data," below.) At Census data centers, "access is allowed only to persons who are regular or special sworn Census employees and would be subject to penalties provided in the law for violations of its confidentiality provisions." Furthermore, "access to identifiable data by special sworn employees is permitted only when such access is deemed to further the agency's mission as defined by law."[30] Census also told us that there is now a "practice of removing personal identifiers from analytical linked files."[31]

Some, but not all, controlled settings require that "projects be approved by agencies that host the data." Although responsible stewardship may require agencies to weigh potential scientific benefits against privacy risks, one participant at a recent workshop pointed out a potential impact on researchers—that such review may create the potential "for censorship, as well as milder forms of restriction arising from a lack of familiarity with … scientific literature [that is relevant to the proposal]."[32]

[28]Horm (1999).

[29]HHS told us that because of these protections, persons using the data center "effectively do not have access to identifiable data." HHS also told us that users of the NCHS data center "are required to sign Non-Disclosure Statements, which state the penalties for violations of NCHS' confidentiality provisions."

[30]NRC (2000), pp. 45-6; Collins (1999).

[31]Census further told us that: "Any crosswalk files containing personal identifiers and their corresponding unique Census Bureau identifiers are maintained in specially secure areas with very limited access."

[32]NRC (2000), p. 10.

Practice Data

Practice, or "sham," data are part of an alternative, innovative approach that has been pioneered by NCHS for its National Survey of Family Growth. Two steps are involved. In step 1, NCHS provides researchers with a practice or sham dataset, which they can use to refine analyses (or debug programs). In step 2, the researchers either provide NCHS with the specifications for a computer run that can be carried out on their behalf (sometimes called "remote access") or conduct the research themselves at the NCHS Research Data Center.

The Census Bureau told us about a variation on this approach—the use of a query system in which the user never sees the data, but can ask for a table or graph that depicts aggregations of underlying data.[33]

Licensing and Data Use Agreements

Licensing and data use agreements allow use of identifiable data under written, contractual conditions (e.g., who can have access, at what locations, and what security safeguards will be implemented). These agreements may also forbid attempts to reidentify data subjects or make new record linkages. Such agreements are used by a variety of agencies.[34] Notably, the National Center for Education Statistics requires unannounced inspections; backs this up with a systematic inspection program; and requires an "affidavit of nondisclosure," specifying severe penalties for unlawful disclosure of confidential information.[35]

[33]Census also said it is developing a system of this type ("American FactFinder") to offer on its web site.

[34]See NRC (2000).

[35]McMillen (1999), NRC (2000). Penalties include a fine of not more than $250,000 or imprisonment for not more than 5 years. See also NRC (2000).

HHS told us that researchers using the SEER-Medicare files "must sign data use agreements, which include criminal penalties for violation of the agreement."

Combining Safer Data and Controlled Access

Developing customized datasets with different levels of protection is another approach. For example, the principal investigator for the Add Health study told us that there are three security levels of linked-data release, each with different levels of reidentification risk, as follows:

- The public-use level includes selected links deemed low risk.
- At the contract level, links with additional (but still low) risk levels are provided.
- Links that are more sensitive and higher risk are provided only to those researchers who are willing to work under direct Add Health supervision in a high-security environment.

Logically, an alternative approach is to customize datasets to fit different research hypotheses; that is, a dataset may be "cropped" to include only the specific data needed to answer a specific research question (eliminating other variables). However, as one expert pointed out, if one cropped dataset is potentially linkable to another from the same study (i.e., if secondary linkage can occur), this approach may be vulnerable to reidentification.[36]

[36]Another approach, developed in the computer science area, has been described as "query set restriction" (Fienberg, 1997), citing Adam and Wortmann (1989).

Use of a custom dataset is sometimes controlled by written agreements or licensing, which may forbid secondary linkage.[37]

Techniques to Reduce Sensitivity

When very sensitive information is needed from survey respondents (e.g., their immigration status), it may be appropriate to use special data collection techniques that are designed to reduce sensitivity but still allow estimation of the sensitive answer category and record linkage. Techniques discussed below include

- the three-card method,
- earlier indirect estimation techniques (randomized response, item-count, and "nominative" techniques), and
- grouped linkage.

The Three-Card Method

The "three-card method," which is designed for large-scale surveys and is still in the experimental stage, involves three separate samples of respondents. Each sample is randomly drawn from the same population and consists of completely different persons.[38] Respondents in each sample are shown a different 8-1/2" by 11" card with alternative answers. Each card is arranged so that respondents in each sample will provide a different piece of less sensitive information—essentially, a different "piece of the puzzle."

By combining the less sensitive information, estimation of the sensitive answer category for a population or large group is possible. (To use a simple analogy, when all pieces of a puzzle, but one, are in place, the outlines of the missing piece are apparent.)

[37]McMillen (1999).

[38]GAO (1999f).

Yet no sensitive characteristic can ever be attributed to a single individual or a small group. The three cards are used as indicated below:

- Respondents in sample 1 are shown **card 1**, which has three boxes: Box A contains only one answer category—a "less sensitive" answer category. Box B contains various other less sensitive answer categories as well as the sensitive category. Box C is for "some other category not listed in Box A or Box B." *Sample 1 respondents are instructed to "pick a box" and are told that "if you're in Box B, we do not want to know which specific category applies to you."*
- Sample 2 respondents are shown **card 2**, which also has three boxes; however, a different less sensitive category now appears in Box A (and the category that was in Box A on card 1 is now listed in Box B). They are given the same instruction.
- For sample 3, Box A of **card 3** contains the remaining less sensitive answer categories. The two that appeared in Box A on the other cards are shifted into Box B (which also contains the sensitive answer category). Again, the instruction is the same.

Sample 1 data are used to estimate the less sensitive category shown in Box A on card 1 (e.g., for a question on immigration status in a survey of foreign-born persons, this might be the percent having an official "green card"). Sample 2 data are used to estimate the different less sensitive category shown in Box A of card 2 (e.g., percent who are naturalized citizens). Sample 3 data are used to estimate the remaining less sensitive categories (e.g., percentages with temporary visas, as refugees, and with grants of asylum); these categories appear in Box A of card 3.

Each less sensitive answer category appears in Box A of one card. Each different sample yields information on the "less sensitive" answer category (or categories)

that appear in Box A of the card used with that sample.

If the various categories are mutually exclusive and exhaustive, it is possible to estimate the sensitive answer category (e.g., percent illegal immigrants) through subtraction.

The three-card method was designed for personal interview surveys but potentially could be adapted for mail surveys, group administration surveys, and Internet surveys. A similar approach might also be used to code administrative or records-research datasets before transfer to researchers or linking. Such coding would protect personal privacy while allowing statistically unbiased estimates.

Earlier Indirect Estimation Techniques

The randomized response, item-count, and "nominative" techniques, which represent earlier attempts to reduce question sensitivity in surveys, are also designed to be applied at the point of data collection. Although randomized response is somewhat controversial as a data collection technique, its logic might be used to code existing data before transferring it or linking it. (The version of randomized response first proposed asks each respondent to privately operate a random spinner, which may point to either of two alternate categories. The respondent indicates only whether the spinner pointed to his or her correct answer. But based on knowing the probability of the spinner's pointing to each answer—for example, 0.2 probability of pointing to answer 1; 0.8 probability of pointing to answer 2—the analyst can estimate the overall percentage of respondents whose correct answer is in each category. Analogously, estimation is possible for

example, for all males, all females, or other large subgroups.[39])

Grouped Linkage

The grouped linkage or grouped data-sharing method, which was mentioned above, limits the sensitivity of linked data. This can be important when the linkage, itself, creates the sensitivity.[40]

Consent Forms and Alternatives

As discussed in the previous chapter on privacy concerns, asking respondents whether they consent to the linkage allows them to maintain some degree of control over the use of their records (because linkage is not performed for those who withhold consent).

Obtaining consent or providing the ability to "opt out" may be necessary for at least some linkage projects. One approach, used in the Health and Retirement Study for survey-archive linkage, is an explicit consent form, which asks the respondent's permission for specific records to be transferred from SSA to the University of Michigan for the purpose of linkage. The consent form (sized 8-1/2" by 11" and reduced in fig. 4.2) explains in clear language which records will be transferred and linked, why linkage is needed, and conditions of release.

The HRS interviewers obtain signed permission forms, and the SSA data are transferred and linked to survey data only for those respondents who agree to the linkage (about 75 percent).

[39]Warner (1965). A subsequent version uses two unrelated questions: a sensitive question and an unrelated less sensitive question with the same set of answer categories. The respondent randomly selects one of the questions, and answers that. The item-count technique is similar but does not require respondents to operate a randomizing device (Droitcour et al., 1991). The nominative technique (Miller, 1985) asks respondents to report on anonymous friends.

[40]Spruill and Gastwirth (1982).

Figure 4.2: Sample HRS Consent Form

The University of Michigan

The University of Michigan Survey Research Center at the ISR
Health and Retirement Study Permission Statement

To the Respondent:

We would like to obtain a history of your past earnings and any information about Social Security benefits you may have applied for or received. Since most people cannot recall this information very well, we are asking your permission to obtain it from government records of:

1) Your past Social Security covered earnings and total taxable earnings, both of which appear on the W-2 forms that people get from their employers.

2) Any information about benefits from programs administered by the Social Security Administration you may have applied for or received.

The information we are requesting is protected by Federal law, and cannot be released to us without your written consent. The University of Michigan is committed to maintaining the privacy and confidentiality of all data obtained from or relating to our survey respondents.

If you give us your Social Security number along with your permission to collect this information from the Social Security Administration, we will combine it with the information you have provided in this interview.

We will remove your name, date of birth, and Social Security number, and release the resulting unidentified statistical information to interested researchers for research purposes only. Additional procedures will be adopted to protect the confidentiality of individuals participating in the survey.

To the Social Security Administration:
I voluntarily authorize the Social Security Administration to release to the University of Michigan, for use in the Health and Retirement Study, information about the amounts of any earnings in my Social Security records along with the industries in which I worked, and information about benefits I applied for or received under programs administered by the Social Security Administration for the years 1937 through 1997. It is my understanding that the University of Michigan will protect the privacy and confidentiality of these data.

Social Security Number: ___ - __ - ____
[PLEASE PRINT]

Full Name: _____
(First) (Middle) (Last)

Date of Birth: ____/____/____
(Month) (Day) (Year)

Maiden Name (if relevant): _____

_____ _____
(SIGNATURE) (DATE)

Source: University of Michigan Survey Research Center.

Whether or not consent is obtained, the various techniques described above may be relevant to specific linkage projects. When it is not practicable to obtain consent or when it is feared that a requirement for consent would bias the results, decisions about whether to conduct the linkage in the absence of consent may be swayed by whether effective privacy protection techniques are employed.

Security Measures (Stored Data)

The basic physical and electronic security approaches that are used to protect any data stored electronically also are relevant for information resulting from record linkage. These include access controls, audit trails, and storage strategies.[41]

Access controls can limit or detect inappropriate access to stored information and protect it from unauthorized disclosure, modification, and loss. These include physical protections, such as using secure rooms and buildings that can incorporate safes, strict key or other controls, gates, guards, and electronic intrusion detection devices. Access controls for electronically stored information can include logical controls built into software that require users to authenticate themselves through passwords or biometric identification (fingerprint, retina, etc.) and that limit the files and other resources accessible to authenticated users, as well as the actions that they can execute. Ways of controlling external electronic access include using firewall technologies and encrypting data.

Audit trails are an effective security tool because they create a continuous log of information about system activity. This includes the user's identity, location,

[41]NRC (1997), GAO (2000d), IRS (2000), Jabine (1993). For audits of information security at various federal agencies, see GAO (2001b, pp. 97-106; 2000c; 2000d; 1999d; 1998c).

date, time, information accessed, and the function performed. Some believe that audit trails are potentially "one of the strongest deterrents to abuse."[42]

Separate storage of substantive data, personal identifiers, and the key to relate them may provide additional protection. A comprehensive guide for information security controls is contained in GAO's Federal Information Systems Controls Audit Manual.[43]

Next Steps and Questions for Further Study

This brief review of techniques is intended to convey the state of the art and to be a step toward achieving a privacy protection toolbox. The next steps would include (1) addressing questions concerning the validity, utility, costs, and benefits of the techniques listed here; (2) identifying other relevant techniques; and (3) exploring issues in agency and researcher adoption of privacy-protection techniques and their application (singly or in combination) to actual linkages. An additional set of questions would concern gaps in the set of existing techniques; if key gaps are identified, efforts might be made to devise new techniques.

[42]NRC (1997), p. 97.

[43]GAO (1999a).

Some Strategies for Enhancing Data Stewardship

As highlighted on the opposite page, this chapter is based on the concept of "data stewardship" and identifies a number of management strategies that may enhance agencies' or research units' efforts to deal appropriately with privacy issues.

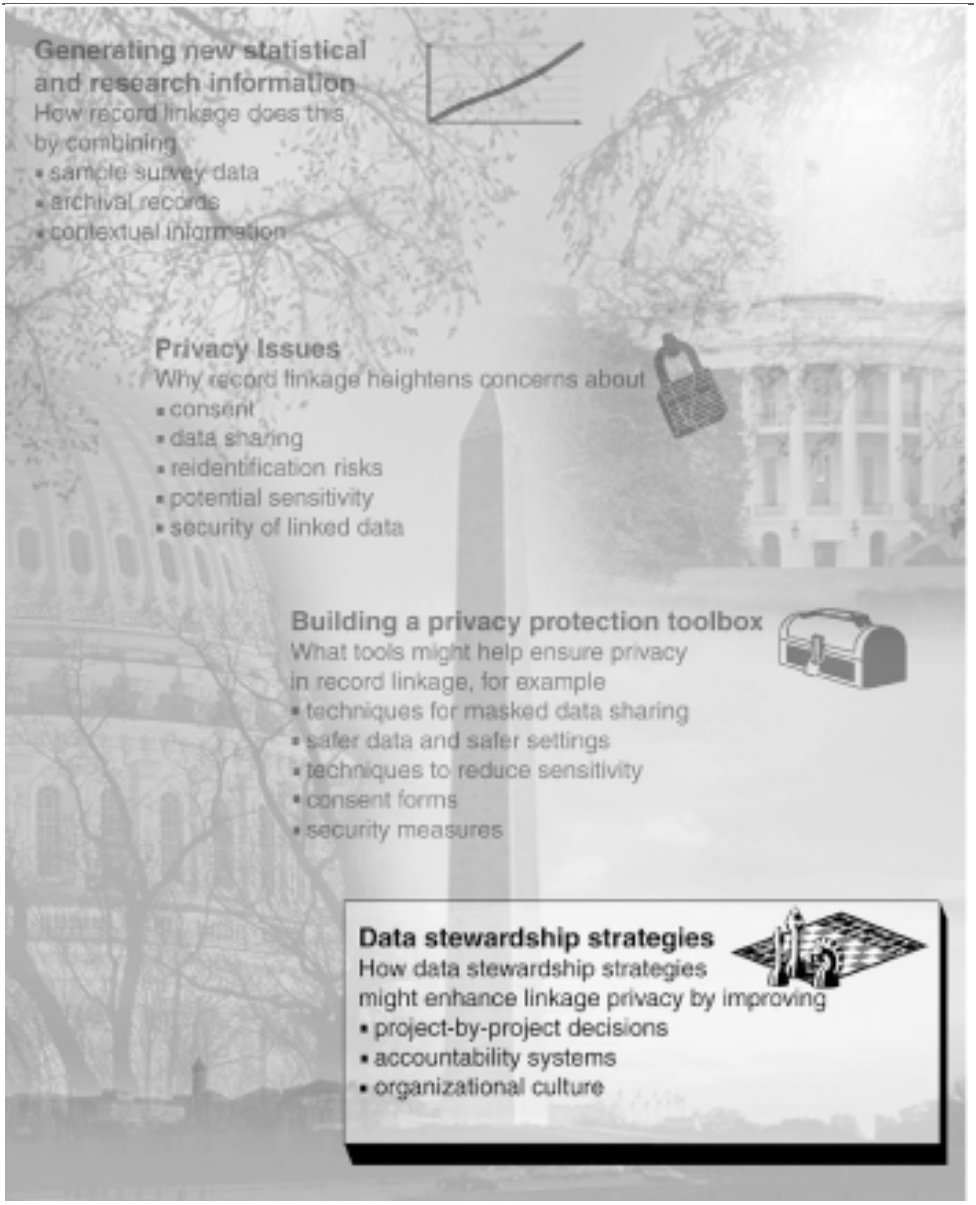
The overview presented in this chapter is intended mainly to illustrate strategies that may enhance privacy protection for linkage projects. Some of the strategies outlined here are currently used in federal statistical or research agencies, statistical offices of program agencies, or in universities or other organizations with federally funded projects.[1] Other strategies appear in the literature or have been suggested by experts with whom we talked; in some cases, these are potential strategies—that is, they have either not been assessed with specific reference to record linkage or not been tried in relevant settings.

We recognize that stewardship involves compliance with relevant laws and that data stewards may draw on the techniques described in the previous chapter. In addition, stewardship involves the coordination of

- project-by-project decisions, which may include whether or not to conduct a specific linkage or whether to release linked data;
- systems for accountability; and
- organizational culture.

These topics—project-by-project decisions, systems for accountability, and organizational culture—are discussed below, with a focus on privacy issues. (We realize that many of the strategies that are described here because they are relevant to record linkage also have more general application.)

[1]We did not conduct audits regarding which strategies are currently in use.


Generating new statistical and research information
How record linkage does this by combining
• sample survey data
• archival records
• contextual information
Privacy Issues
Why record linkage heightens concerns about
• consent
• data sharing
• reidentification risks
• potential sensitivity
• security of linked data
Building a privacy protection toolbox
What tools might help ensure privacy in record linkage, for example
• techniques for masked data sharing
• safer data and safer settings
• techniques to reduce sensitivity
• consent forms
• security measures
Data stewardship strategies
How data stewardship strategies might enhance linkage privacy by improving
• project-by-project decisions
• accountability systems
• organizational culture

Project-by-Project Decisions

In some instances, legal or other formal constraints determine whether a proposed linkage can be conducted. But there are many other instances where linkage decisions are made by a variety of officials and groups. These persons and groups must weigh risks and benefits, ideally from a neutral perspective. But they may be concerned with government program needs, agency strategic plans, the needs of a nonfederal research group (e.g., a group located within a university and receiving federal funds), or a host of other possible factors. For example:

- Decisions about whether to approve a proposed linkage that would involve person-specific data (and perhaps also related decisions on consent issues) may be made by federal officials at one or more agencies. In other instances, these decisions may be made by officials at universities receiving federal funds, or in some cases, a federal agency's institutional review board, grant review boards, other groups, or some combination of these.[2]
- Decisions about whether and how deidentified linked data will be released once the linkage has been completed are made by a variety of individuals and groups who must take account of reidentification risks.[3] These decisionmakers may require special controls, such as review of publications based on linked data.

[2]We recognize that some forms of linkage (e.g., multiperson linkage within a single dataset) could potentially be carried out without a formal agency decision focused on linkage.

[3]As explained in the previous chapter, "deidentification" refers to stripping explicit identifiers and modifying or eliminating other variables that could be used to identify individuals. Some have referred to those making disclosure decisions as "disclosure review boards." See ICDAG (1999); de Wolf et al. (1998).

In making these and other decisions that impact privacy, federal officials and others may be challenged by a need to consider potentially conflicting interests. This would include the interests of survey respondents and data subjects; organizational missions; and perspectives of a variety of other stakeholders, including those who might ultimately benefit from the information generated by linkage.[4]

Key Factors in Linkage Decisions

Where individuals or groups are called upon to exercise judgment in making decisions about proposed linkages, there is a need for

- sound ethics and values;
- expertise in protecting personal privacy, confidentiality, and security; and
- scientific expertise in the subject area and in research design.

Ethics and Values

Ethics and values are necessarily involved in many decisions about proposed linkages. That is, linkage decisions may involve weighing privacy issues or risks against anticipated benefits and need (e.g., the need for data to evaluate government programs).[5] For example, IRBs (which reviewed some of the linkages discussed in this study, when these were first proposed) have been charged, under federal regulation, with determining whether risks to research subjects are reasonable in light of the anticipated benefits. To cite a hypothetical example posed at a recent conference, an IRB deciding whether to approve linkage of children's survey responses with child abuse records would have to judge the potential benefits of the research to children in the long run versus the need to protect the personal privacy of the

[4]See Duncan and Lambert (1986).

[5]See, for example, Gastwirth (1986).

children who participated in the survey.[6] However, some believe that, in general, current federal regulations--taken by themselves--may not provide a sufficient conceptual framework for weighing risks and benefits of research.[7]

Technical Expertise

Technical expertise in privacy issues is needed to inform linkage decisions. For example, linkage decisions may involve judgments about whether planned privacy protections (such as the tools discussed in the previous chapter) are adequate, and this may require technical expertise. Another complication is whether the linkage requires data sharing. Some experts in computer security argue that once data cross organizational lines, the data steward no longer has control—at least not direct control.[8]

Methodological Expertise

Linkage decisions may also require judgments about whether a proposed linkage is, in fact, needed to accomplish the research purpose or whether an alternative linkage or other type of approach with lower privacy risk might be pursued. Such decisions would ideally be based on scientific or methodological expertise. To illustrate this, using a lower risk alternative to person-by-person linkage, one group of researchers used existing records to examine the relationship between (1) the use of community mental

[6]Thompson (2000). Risks may be more difficult to assess in some projects than others; for example, social science research risks may be more difficult to assess than physiological risks (McGough, 2000).

[7]See NBAC (2000a), which also proposes a model for analyzing risks and benefits.

[8]Computer security experts argue that MOUs, safeguard reviews, and other technical strategies aimed at ensuring the protection of data at the receiving agency may mitigate, but do not fully address, the issue of the data steward's loss of direct control. (See also Scheuren and Mulrow, 1999.)

health services and (2) incarceration. They did this by analyzing existing records that contained birth dates but not unique identifiers.[9] The overlap between the two populations was statistically estimated without linking individual records. Approaches such as this, or other methods, such as grouped linkage (see chap. 4), may provide an alternative to linkage at the individual level, which could be useful in some situations.[10]

Strategies to Enhance Linkage Decisions

As of this writing, we know of no governmentwide or interagency checklist, detailed model, or set of guidelines for reaching an overall judgment about a proposed linkage.[11] Such judgments would involve assessing a variety of subjective factors, such as the need to address the research question, the need to use linked data (vs. possible alternative research designs), the adequacy of proposed privacy protections, and overall risks and benefits of the proposed linkage. One possible basis for the development of guidance might be principles from "codes of fair information practices," including openness (nonsecrecy) and collection limitation (which implies that linkage should be limited in terms of the amount of information amassed or the percentage of the population covered), among others.[12] But even with detailed guidance, the independence and expertise of those making judgments would seem all important.

[9]Pandiani et al. (1998); Banks (1999).

[10]Grouped linkage (Spruill and Gastwirth, 1982; see chap. 4).

[11]Census told us it is developing a checklist to evaluate new linkage projects involving administrative data. Census also provided us with a basic checklist it is using for certain types of its projects.

[12]See the HEW Secretary's Advisory Committee (1973); Gellman (2000); OECD (1980, 1999).

Other potential strategies to enhance linkage decisions include obtaining

- independent assessments,
- input from data subjects,
- advice from privacy teams, and
- advice from scientific reviewers.

Independent Assessments

In judging risks and benefits, self-assessments by agencies and researchers who are involved in the linkage might be supplemented with independent assessments by autonomous review boards (such as IRBs) established to advise agencies "with the public's interest in mind."[13]

Input From Data Subjects

In judging risks and benefits, the deciding group or individuals should take into account the perspective of the data subject—especially if surveys of data subjects or focus groups have been conducted. For example, Census conducted focus groups regarding attitudes toward the possibility of linkage using income tax records in the decennial census.[14]

Privacy Teams

In assessing the adequacy of privacy protections, decisionmakers might be provided with access to "privacy teams" composed of three or more persons with expertise in different aspects of information privacy.[15]

[13]Scheuren (1999). Not all agencies have IRBs.

[14]Gates and Bolton (1998).

[15]RAND's IRB has access to a three-person privacy team (IOM, 2000).

Expert Reviewers

In judging the need for the proposed linkage, decisionmakers might be assisted by scientific or subject-matter reviewers and advisers, as needed.[16]

For the future, researchers and interagency groups might consider exploratory work aimed at developing appropriate checklists or other tools to support decisions about whether to conduct a proposed linkage and whether to require consent.

Key Factors in Decisions About Releasing Linked Data

Other decisions are made concerning proposed releases of linked data, whether in aggregated or microdata form. Essentially, a "data disclosure board" weighs the need to protect confidentiality against the need to preserve the usefulness of the data.[17] A decision to release may be more complex when two agencies both possess the same linked dataset (as could occur if the linkage is based on shared data). The problem is that each agency could release a different version of a masked dataset—either publicly or in restricted form to persons who conceivably might share or fail to protect the data. If this were to happen, then logically, it might be possible for a "data detective" to create identifying information by combining the two versions of the dataset.[18]

Generally, however, the assessment of reidentification risks and dissemination decisions may be less

[16]This possibility is suggested by the two-tier review system at the M.D. Anderson Cancer Center. As described to us by an expert we consulted, the two-tier system involves review by, first, a scientific committee and, second, by an IRB.

[17]Indeed, agencies making disclosure decisions have been seen as essentially "walking a tightrope," balancing "the agencies' public obligation to provide maximum information to society" and the need to protect the privacy of respondents or other data subjects (Fienberg and Willenborg, 1998), p. 338.

[18]Federal Committee (1994), pp. 75-6.

subjective than decisions about linkage and consent. At least, there have been efforts to develop stewardship strategies to help in this area.

Strategies to Enhance Decisions About Releasing Linked Data

Potential strategies to enhance dissemination decisions regarding linked data include using a checklist for disclosure review, building models to assess disclosure risks, using expert panels on disclosure risk, and coordination across agencies.

Using Checklists

A checklist was developed by a standing committee of the Federal Committee on Statistical Methodology for disclosure review boards to use in reviewing disclosure risks.[19] The Federal Committee's checklist, which was based on an earlier Census checklist, includes "rules of thumb" and key guidelines, as well as a section on matching.[20] More generally, the Federal Committee recommends that, where possible, disclosure review boards should follow auditable processes and use consistent practices for similar data.

Building Models

Models for assessing disclosure risk are the focus of an ongoing body of work.[21] A key step in assessing disclosure risk is identifying population uniques in the data to be released.[22] However, a model of disclosure

[19]This standing committee, formerly ICDAG, was recently renamed the Confidentiality and Data Access Committee. The purpose of this group has been described as coordinating and promoting research "on the use of statistical disclosure methods and … catalog[ing] related developments at agencies and among academic researchers." (NRC, 2000, p. 42.)

[20]ICDAG (1999).

[21]Federal Committee (1994), Fienberg and Willenborg (1998).

[22]See, for example, Skinner and Holmes (1992). To use a hypothetical example, if there were only one female Asian physician aged 60 or older living in South Dakota, she would be a "population unique." If she were represented in a dataset that included sex, race, occupation, age, and state of residence, she would be identifiable.

risk involves many other factors and should be based on a disclosure scenario.[23] Developing such a scenario involves judgments about such factors as which variables a data detective or snooper might have prior information on and what the possible motives or incentives for reidentification might be.[24] Logically, when linked datasets are involved, the disclosure scenario should also consider linkage elements that might heighten disclosure risks, notably, the existence of external files related to the linked file. If, for example, a linked survey-administrative dataset is held by the agency that conducted the survey, the release of even deidentified linked data would enable staff at the program agency to reidentify respondents.[25]

Using Expert Panels

Given that judgment is involved in developing a disclosure risk model, one strategy would be to establish standing panels of disclosure experts (separate from the disclosure review board).[26] Such a panel could provide input on specific elements of disclosure risks for a variety of linked or other datasets.

Coordination Across Agencies

For linked datasets that involve multiagency data sharing, disclosure decisions should be coordinated across agencies.[27] For example, research requests for linked SEER-Medicare data are submitted to NCI, but

[23]Fienberg and Willenborg (1998).

[24]These range from identifying any individual(s) in the dataset (which might be done for purposes of embarrassing a specific university or a researcher) to identifying many or all persons in a comprehensive dataset for marketing purposes.

[25]Federal Committee (1994).

[26]Rasinski and Wright (2000).

[27]Federal Committee (1994).

representatives from NCI, the SEER program, and HCFA are responsible for reviewing each request.

One strategy to ensure coordination (in advance) might be to specify coordination on dissemination in the MOU completed at the time when data-sharing agreements are reached. This would help avoid a situation in which different data released by two different agencies could be combined by "data detectives" or members of the general public to reidentify data subjects.

Systems for Accountability

When record linkage involves data sharing, and particularly when shared data are subject to legal restriction (e.g., IRS tax return and Census Bureau data), accountability has particular salience. But even if legal restrictions are not in play, accountability is crucial to data stewardship. The reason is that managerial review is needed in addition to physical and technical protections. Thus, at the agency level, systems for accountability are a key part of the agency management's data stewardship role.

Strategies to enhance accountability include

- centrally tracking record linkage projects via a management information system and
- assessing security risks for linked data (particularly projects that involve data sharing).

Developing an MIS

One strategy is to develop a computer-based management information system (MIS) designed to track linkage projects from inception to completion. Agencies sharing data with others (under, e.g., an MOU) might require such management information systems or develop a related system to enhance oversight.

Census staff briefed us on the development of a new pilot system that will cover all statistical projects involving administrative data. The new computerized system is designed to facilitate reviewing and auditing linkage projects, based on agreements with other agencies. Such a system is designed to build in procedures, such as annual reviews of all linkage projects—including inactive projects, until identifiers have been eliminated.[28] (Logically, security must be maintained for all datasets that are not cleared for public release, especially those with personal identifiers.)

Assessing Security Risks

Another strategy that seems relevant to record linkage is risk management. In the area of information security, we have recommended that agencies develop risk management strategies to prevent intruder attacks ("hacking in"), inappropriate access, and other lapses of security. Risk management strategies include processes based on the principles of assessing risks, implementing appropriate policies and controls, promoting awareness, and monitoring and evaluating the effectiveness of policies and controls.[29] These processes interact with a central focal point, as illustrated in figure 5.1.[30]

Organizational Culture

Social scientists have emphasized "the significance of culture as a major determinant of a population's beliefs, attitudes, and behaviors."[31] Following this view, the cultural values of staff, researchers, and indeed, all persons who come in contact with linked

[28]These annual reviews are routine at RAND (IOM, 2000).

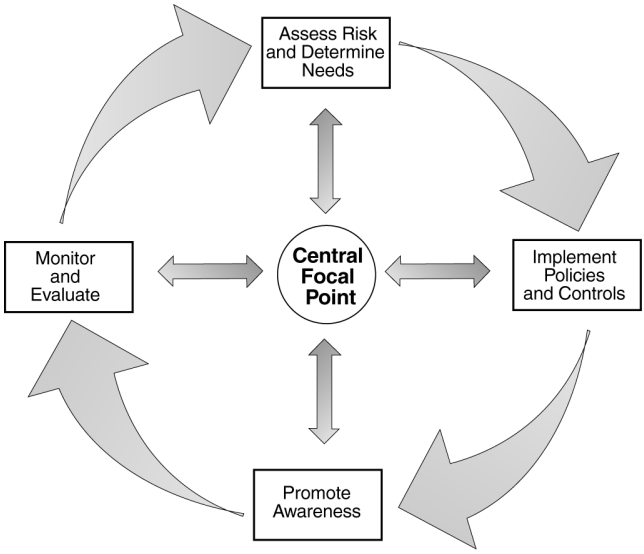
[29]GAO (1998b, 1999b, 2000d).

[30]Census commented that its Executive Steering Committee deals with security and other issues raised by record linkage.

[31]Bowditch and Buono (1985), p. 155.

data could potentially affect security, confidentiality, and thus, personal privacy. In fact, an expert on statistical ethics pointed to values as the ultimate ingredient in ensuring confidentiality and personal privacy.[32] Logically, the cultural dimension would be particularly important for staff and agencies that come in contact with sensitive data.

Figure 5.1: Managing Security Risks



Source: GAO (1999b), p. 6.

Like professional groups and other organizations, federal agencies and research groups that receive federal funds may be characterized by an "organizational culture"—that is, "shared patterns of beliefs, assumptions, and expectations held by

[32]William Seltzer of Fordham University, chair of the American Statistical Association's committee on professional ethics.

organizational members."[33] Both Census and NCHS staff have described a "culture of confidentiality," in which the protection of data is a key value.

Potential strategies for enhancing security, confidentiality, and personal privacy through organizational culture include

- basic strategies for heightening awareness,
- other management strategies, and
- change planning (both internally and externally).

Basic Strategies for Heightening Awareness

Various experts suggested that appropriate values can be supported by a number of strategies. For example, an IRS manual suggests conducting training programs and having discussions on security at group and managerial meetings; security bulletin boards installed throughout the workplace; articles on security in employee newsletters; pertinent articles from the press shared with managerial staff; and posters that display short, simple educational messages.[34] Other approaches include the use of agency intranets to facilitate sharing of information on privacy issues and related policies and efforts to facilitate interaction between the agency privacy officer and other staff. NCHS has issued a document covering similar material—the NCHS Staff Manual on Confidentiality.[35] Both the IRS and NCHS manuals cover relevant laws and penalties. Manuals such as these are designed to heighten staff awareness of federal requirements related to privacy and confidentiality.

[33]Bowditch and Buono (1985), p. 155.

[34]IRS (2000).

[35]NCHS (1984).

Other Management Strategies

Other ways of maintaining or improving organizational culture include displaying top management commitment and support for desired values and beliefs (e.g., through articulation of clear policies and procedures for data stewardship);[36] replacing or changing the responsibilities of employees who do not support desired values and beliefs; and assigning a manager or group the primary responsibility for perpetuating or changing culture.[37]

Research exploring how the public views privacy issues or relevant agency activities, which may be conducted for a variety of reasons, represents another potential tool for sensitizing staff.[38]

Self-Study and Change Planning (Internal)

In situations where a major change in direction is needed, an agency may start by studying unconscious, underlying values and assumptions (what has been called the "shadow side of espoused culture"[39]). This might be followed by a comparison of what is and what is desired (i.e., a gap analysis) and a plan for addressing gaps.[40]

Encouraging Change in External Environments

Educating various "stakeholders" in linkage projects (including those external to the project itself or its sponsoring agency) may also be relevant.[41] Other

[36]For example, IRS told us that their Privacy Advocate creates and implements privacy programs and policies for the agency. Census told us they have developed an Executive Steering Committee with coordinated standing committees to deal with privacy issues.

[37]GAO (1992).

[38]Census has undertaken research on how the public views certain record linkage and privacy issues (Gates and Bolton, 1998).

[39]Egan (1994), p. 87.

[40]IRS (1999), citing Egan (1994).

[41]A recent U.S. Department of Energy report, which discusses ethics in research on workers, was issued with the goal of educating

potential activities, although outside the federal government, might be encouraged by federal agencies. These include an increased emphasis on ethics and data privacy issues in graduate school training programs (especially those that may receive federal funding) as well as professional conferences and explicit journal review policies on confidentiality standards for acceptance of articles for publication.

Next Steps and Questions for Further Study

Perhaps the most obvious next step would be to delineate working models at federal agencies in order to (1) develop criteria for "best practices" for the stewardship strategies outlined in this chapter and (2) potentially identify other strategies. Where appropriate working models cannot be identified, additional work could develop new approaches. These varied efforts might foster comprehensive frameworks for stewardship, contribute to agency self-assessment guides, and support use of enhanced stewardship strategies.

various "stakeholders" about privacy principles, applicable laws, regulations, and codes of ethics.

Privacy Concepts

For purposes of this study, this appendix provides

- working definitions of three privacy concepts or categories—personal privacy, confidentiality, and security;[1]
- a brief discussion of the relationships among these three concepts; and
- a definition of informed consent.

Personal Privacy

Personal privacy includes an individual's status and rights. With respect to record linkage issues, the key factors may be

- whether or not the information about an individual, including his or her personal attitudes or experiences, are known to another (privacy status) and
- whether the individual has control over information about himself or herself being shared with anyone else (privacy rights).[2]

Many definitions of personal privacy emphasize the latter.[3] Logically, however, a person's privacy may be preserved, enhanced, or reduced by the choices that he or she makes as well as by the actions of others. We note that privacy status is also affected by freedom from excessive intrusion.[4] In this context, record linkage has sometimes been viewed as enhancing personal privacy because new data collections may be avoided (although it is sometimes

[1]These concepts have been variously defined. See Boruch and Cecil (1979); George T. Duncan et al. (1993); Fanning (1999); Hotz et al. (1998); and Lowrance (1997).

[2]Boruch and Cecil (1979).

[3]George T. Duncan et al. (1993); Fanning (1999); Goldman and Mulligan (1996).

[4]George T. Duncan et al. (1993).

seen as reducing privacy because "the whole may be greater than the sum of its parts").

Confidentiality

Confidentiality is a status accorded to information based on a decision, agreement, obligation, or duty such that the recipient of personal data must control disclosure. Confidentiality may be based on one or more of the following:

- promises, explicit or implicit, made to a data provider,[5]
- a legal requirement, or
- an inherent duty to avoid disclosure of personal data that would be harmful to a data subject.[6]

(A data provider is the person who supplies information on a data subject, whereas the data subject is the person or unit, e.g., family, described in recorded data. The data provider and the data subject may or may not be the same.)

We believe the third point listed above is particularly important if the data provider differs from the data subject (e.g., when a family member or employer provides information about the data subject)[7] or if the data subject is a member of a vulnerable population. (We note that a duty to avoid harming the data subject through disclosing personal information might extend

[5]George T. Duncan et al. (1993). Included here, for example, would be the obligation to honor pledges made in eliciting consent. See also Lowrance (1997), citing Penslar and Porter (1993).

[6]George T. Duncan et al. (1993), citing the 1971 President's Commission on Federal Statistics.

[7]Some believe that data should not be obtained from a data provider (e.g., physician) when the data provider has not received the consent of the data subject (e.g., patient) to transfer his or her data for research purposes.

to others in a personal relationship to the data subject, such as his or her family members.)

Security

Security refers to safeguards for data and related systems. Safeguards against unauthorized access, unauthorized disclosure, and misuse by internal or external parties may include

- physical controls (e.g., locks, guards);
- system hardware, software, and system access controls (such as passwords), and accountability controls (such as audit trails);
- special procedures for data transfer (e.g., encryption); and
- information practice policies.

Security may also include personnel policies (such as background checks), emergency preparedness, and other kinds of measures.[8]

Relationships Among Privacy, Confidentiality, and Security

The three concepts—personal privacy, confidentiality, and security—are closely interrelated.

Notably, if a breach in data confidentiality were to occur, this might impact personal privacy. And a breach in security could potentially impact both confidentiality and personal privacy. By the same token, protecting confidentiality helps protect personal privacy. And "through various security measures … confidential information … [can be shielded], thus protecting the [personal] privacy of individuals who are the subjects of the stored data."[9]

The five privacy issues discussed in chapter 3—consent to linkage, data sharing, reidentification risks,

[8]NRC (1997).

[9]Goldman and Mulligan (1996), p. 2.

sensitivity, and security of linked data—are related to personal privacy, confidentiality, and security as indicated in table I.1. That is, the checkmarks indicate which privacy concept or category is primary for each issue. For example, consent to linkage falls primarily under the concept of personal privacy, whereas reidentification risks come under the concept of confidentiality.

Table I.1: Primary Privacy Concept or Category Associated With Each Issue

	Privacy concept or category		
Five issues	**Personal privacy**	**Confidentiality**	**Security**
Consent to linkage*	✔	[a]	[a]
Reidentification risks	**	✔	[a]
Data sharing	**	✔	✔
Sensitivity	Sensitivity can heighten concerns in each category		
Security of linked data	**	***	✔

✔ = Primary privacy concept(s) associated with each issue.

*Obtaining consent to linkage vs. linking without consent.

**Personal privacy is potentially impacted by reidentification risks, data sharing, and the security of linked data.

***Confidentiality of linked data is potentially impacted by the security of linked data.

[a]Not applicable; the issue listed for this row does not fall primarily under the category for this column.

Source: GAO analysis.

The table also indicates that sensitivity of data can heighten concerns in each concept or category. Finally, as indicated at the outset of this section, confidentiality issues can impact personal privacy—and security can impact both confidentiality and personal privacy. The table, therefore, also makes note of these potential impacts.

It has been maintained that in the area of federal statistics and research, few, if any, breaches of security have resulted in actual violations of personal

privacy or harm to data subjects.[10] However, recent forums and reports on research data indicate a concern by federal agencies and the research community about privacy issues and potential risks.[11]

Informed Consent

The Federal Policy for the Protection of Human Subjects (known as the "Common Rule") emphasizes the concept of "informed consent," although waivers releasing the researcher from the need to obtain consent may be obtained from IRBs.[12] The definition of informed consent in the box below applies to the research or statistical use of personal data.[13]

Definition of Informed Consent

"[I]nformed consent refers to a person's agreement to allow…data to be provided for research and statistical purposes. Agreement is based on full exposure of the facts the person needs to make the decision intelligently, including any risks involved…. Informed consent describes a condition appropriate only when [there are no] … penalties for failure to provide the data…."

[10]NRC (2000). Some breaches of research data confidentiality and personal privacy have come to light. (See, e.g., GAO 1999c and 1999d). However, these few instances may not fairly represent the size, nature, or location of the problem.

[11]IOM (2000); NRC (2000).

[12]Many of the agencies involved in the examples in chap. 2 subscribe to the Common Rule, but others (e.g., IRS) do not. In a recent survey by NBAC, even when they subscribe to the Common Rule, agencies vary in their use of IRBs for intramural research, that is, research conducted within an agency. For example, NCHS and NIOSH research are reviewed within the IRB structures at the Centers for Disease Control and Prevention. Census and SSA do not apply IRB review to their intramural research. According to Census and SSA, the research is generally exempt from the Common Rule, and they have other mechanisms for review of research. (See NBAC, 2000b.)

[13]George T. Duncan et al. (1993), p. 23.

Appendix I
Privacy Concepts

Experts Consulted

This listing includes only experts not currently with the federal government. Additionally, we spoke with officials and staff at a number of federal agencies.

Richard Burkhauser, Cornell University

Joseph Cecil, Federal Judicial Research Center

George Duncan, Carnegie Mellon University

Robert Gellman, Consultant, Washington, D.C.

Janlori Goldman and Angela Choy, Health Privacy Project, Georgetown University

Thomas Jabine, Consultant, Washington, D.C.

Mary Grace Kovar, National Opinion Research Center, University of Chicago

Christopher Mackie, National Research Council

Patrice McDermott, OMB Watch

Fritz Scheuren, The Urban Institute

William Seltzer, Fordham University

Eleanor Singer, University of Michigan

Latanya Sweeney, Carnegie Mellon University

J. Richard Udry, University of North Carolina

Andrew White, National Research Council

Robert Willis, University of Michigan

Lee Zwanziger, Institute of Medicine

Leonard Zwelling, M.D. Anderson Cancer Center, University of Texas

Conferences and Workshops Attended

International Record Linkage Workshop and Exposition, Arlington, VA, Mar. 20-21, 1997.

National Conference on Health Statistics, Washington, D.C., Aug. 2-4, 1999.

Workshop on Confidentiality of and Access to Research Data Files, National Academy of Sciences, Washington, D.C., Oct. 14-15, 1999.

Federal Committee on Statistical Methodology Research Conference, Rosslyn, VA, Nov. 15-17, 1999.

Privacy and Confidentiality in Clinical and Social Science Research: Myth or Reality? Houston, TX, Feb. 10-11, 2000. (Cosponsored by NIH, U.S. Food and Drug Administration, University of Texas Health Science Center at Houston, and Prairie View A&M University.)

Virtual Government 2000: Digital Government—Making It Real, Washington, D.C., Feb. 22-23, 2000. (Cosponsored by AFCEA International and the Federal CIO Council.)

Workshop on the Role of Institutional Review Boards and Health Services Research Data Privacy, Institute of Medicine, National Academy of Sciences, Washington, D.C., Mar. 13-14, 2000.

Appendix III
Conferences and Workshops Attended

Selected Laws and Regulations Relating to Record Linkage and Privacy

Record linkage, as defined in this study, is a computer-based process that combines (1) existing data on identifiable persons with (2) additional data that may refer to the same persons, their family and friends, school, employer, or geographic environment. The report focuses on linkage projects that are conducted under federal auspices to produce new research or statistical information. As discussed in the text, some of these linkages occur within an agency (e.g., HHS) and others occur between different agencies (e.g., Census and IRS). Therefore, the legal framework that applies to a certain record linkage could vary widely depending on the agency or agencies involved and the type of data that are being linked. The following outlines selected laws and regulations that generally relate to record linkage and privacy protections in the federal government.

Federal agencies are required by law to protect an individual's right to privacy when they collect personal information. The Privacy Act of 1974 is the primary law regulating the federal government's collection, maintenance, and disclosure of personal information.[1] Other laws of general application that apply to the protection of personal information collected by the federal government are the Freedom of Information Act (FOIA), the Paperwork Reduction Act of 1995, and the Computer Security Act of 1987.

In addition to governmentwide statutes, some agencies are also subject to laws that specify the confidentiality and data access policies that they must follow. Lastly, there are certain federal regulations, most notably the Federal Policy for the Protection of Human Subjects (known as the Common Rule), that

[1]P.L. 93-579, 5 U.S.C. 552a.

govern certain research projects that involve human subjects.

Privacy Act

The Privacy Act places limitations on the collection, use, and dissemination of personally identifiable information maintained by an agency about an individual and contained in an agency's system of records.[2] The Privacy Act defines a "system of records" as any group of records under the control of an agency from which information is retrieved by the name of the individual or by some identifying number, symbol, or other identifying particular assigned to the individual. Under the act, an agency cannot disclose any information about an individual contained in a system of records to another person or agency without the prior written consent of the individual, unless the disclosure is authorized by law.

The Privacy Act authorizes 12 exceptions under which an agency may disclose information in its records without consent.[3] For example, the act authorizes an agency to disclose a record

[2]In 1988, Congress amended the Privacy Act to regulate the use of computer matching conducted by federal agencies or using federal records subject to the statute. These amendments, called the Computer Matching and Privacy Protection Act of 1988, generally define computer matching as the computerized comparison of two or more automated systems of records or a system of records with nonfederal records for the purpose of (1) establishing or verifying eligibility for a federal benefit program or (2) recouping payments or delinquent debts under such programs. Matches performed to support any research, or statistical project (the specific data of which may not be used to make decisions concerning the rights, benefits, or privileges of specific individuals) are not subject to the act. See P.L. 100-503, 5 U.S.C. 552a(o)-(s).

[3]5 U.S.C. 552a(b).

- for a "routine use," defined in the act as a use of a record for a purpose which is compatible with the purpose for which it was collected,[4]
- to those officers and employees of the agency which maintains the record who have a need for the record in the performance of their duties,[5]
- to the Bureau of the Census for purposes of planning or carrying out a census or survey or related activity under title 13, or
- to a recipient who has provided the agency with advance adequate written assurance that the record will be used solely as a statistical research or reporting record, and the record is to be transferred in a form that is not individually identifiable.

Several of these exceptions have implications for research and statistics.[6] For example, information disclosed to Census would be used for statistical activities. Agencies, such as HHS and component agencies, have established research as a routine use of certain records, thus allowing disclosure outside the agency.[7]

The Privacy Act also grants individuals the right of access to agency records maintained on themselves;

[4]Instead of obtaining individual consent prior to disclosure for such a routine use, the agency must publish a notice of the anticipated routine uses of the record in the Federal Register and accept comments from the public for a period of at least 30 days (5 U.S.C. 552a(e)(11)).

[5]It should be noted that the definition of "agency" for the purposes of the Privacy Act includes any "executive department." See 5 U.S.C. 552a(a)(1). Under this definition, HHS, for example, is a single "agency" for the purpose of Privacy Act restrictions (45 C.F.R. Part 5b).

[6]OMB (1975), George T. Duncan et al. (1993), Cecil and Griffin (1985).

[7]See, for example, Fanning (1998).

the right to amend that record if it is inaccurate, irrelevant, untimely, or incomplete; and the right to sue the government for violations of the act.[8]

When an agency is establishing or revising a system of records, the agency is required to publish in the Federal Register a notice including such information as the name and location of the system, the categories of individuals on whom records are maintained in the system, and each routine use of the records contained in the system. When collecting information on a form (that is to be entered in a system of records), agencies are required to notify individuals of the authority authorizing the solicitation of the information and whether disclosure of such information is mandatory or voluntary, the principal purposes for which the information is intended to be used, and the routine uses that may be made of the information.

Concerns have been expressed about agency use of the routine use exception. For example, in 1998, a Presidential memorandum noted the need for a reexamination of the federal government's role in personal privacy. OMB subsequently asked each agency to review its routine uses to identify any that are no longer justified, or no longer compatible with the purpose for which the information was collected.[9]

[8]Systems of records that are required by statute to be maintained and used solely as statistical records may be exempted from certain Privacy Act restrictions, such as the access and corrections provisions.

[9]OMB Instructions for Complying with the President's Memorandum of May 14, 1998, "Privacy and Personal Information in Federal Records." (OMB Memorandum 99-05, Jan. 7, 1999.)

Other Relevant Statutes

In addition to the Privacy Act, there are several other governmentwide statutes that relate to the protection of individually identifiable information. FOIA,[10] as amended, provides that the public has a right of access to federal agency records, except for those records that are protected from disclosure by nine stated exemptions. Two exemptions in FOIA protect personal privacy interests from disclosure. The first exemption allows the federal government to withhold information about individuals in personnel and medical files when the disclosure would constitute a clearly unwarranted invasion of personal privacy. The second exemption allows the federal government to withhold records or information compiled for law enforcement purposes, but only to the extent that the production of such law enforcement records or information could reasonably be expected to constitute an unwarranted invasion of personal privacy.

The Paperwork Reduction Act of 1995[11] requires the Office of Information and Regulatory Affairs within OMB to provide central guidance for and oversight of federal agencies' information management activities, including activities under the Privacy Act.[12] The Paperwork Reduction Act of 1995 also requires federal agencies to ensure compliance with the Privacy Act and coordinate management of the requirements of

[10]P.L. 89-487, 5 U.S.C. 552.

[11]P.L. 104-13, 44 U.S.C. 3501 et seq.

[12]OMB Circular No. A-130 establishes policies for the management of federal information resources. The Circular sets forth a number of general policies concerning the protection of personal privacy by the federal government, including the requirement that agencies limit the collection of information that identifies individuals to that which is legally authorized and necessary for the proper performance of agency functions.

FOIA, the Privacy Act, the Computer Security Act, and related information management laws.

The Computer Security Act of 1987,[13] as amended, provides for improving the security and privacy of sensitive information in federal computer systems. The act defines "sensitive information" to include any unclassified information that, if lost, misused, or accessed or modified without authorization, could adversely affect the national interest, conduct of federal programs, or the privacy to which individuals are entitled under the Privacy Act. The Computer Security Act requires federal agencies to identify their computer systems that contain sensitive information, establish training programs to increase security awareness and knowledge of security practices, and establish a plan for the security and privacy of each computer system with sensitive information. The Computer Security Act also requires the National Institute of Standards and Technology to develop standards and guidelines for the security and privacy of sensitive information in federal computer systems.

In addition to governmentwide statutes, some agencies are also subject to other laws that specify the confidentiality and data access policies they must follow. Some of these laws may limit record linkage activities. Notably, statistical information is protected by various agency-specific statutes, as illustrated below:

- The Census Bureau's activities with regard to confidentiality are governed by section 9 of title 13 of the United States Code, which requires that information furnished to the Bureau be kept

[13]P.L. 100-235, 15 U.S.C. 271 note, 272, 278g-3, 278g-4, 278h; 40 U.S.C. 1441 note.

confidential and be used exclusively for the statistical purposes for which it was supplied.[14]

- The National Center for Health Statistics' records are protected by the following basic legal requirement in the Public Health Service Act, as amended.[15] No information obtained in the course of NCHS' activities (from establishments or persons) may be used for any purpose other than the purpose for which it is supplied unless there has been consent. Also, such information may not be published or released in an identifiable manner unless there has been consent.[16]

Furthermore, OMB issued an order establishing government policy to protect the privacy and confidentiality interests of individuals and organizations who furnish data for federal statistical programs.[17] This order establishes standards regarding the disclosure and use of information acquired for exclusively statistical purposes.[18] Agencies are to comply with the order to the extent permissible under their statutes.

Various other agencies have restrictive provisions concerning disclosure of certain information. For example, 26 U.S.C. 6103 prohibits IRS from disclosing

[14]Generally, Census employees who willfully disclose information protected by section 9 are subject to a substantial fine or imprisonment of not more than 5 years, or both (13 U.S.C. 214).

[15]42 U.S.C. 242m.

[16]According to NCHS, unauthorized disclosure of confidential information is punishable under 18 U.S.C. 1905.

[17]OMB's Order Providing for the Confidentiality of Statistical Information (62 Fed. Reg. 35044 (June 27, 1997)).

[18]If the agency collecting the information proposes to disclose the information collected in identifiable form for purposes other than statistical purposes it is to, prior to disclosure, fully inform the affected respondents of the facts regarding such disclosure.

any tax return or return information except as authorized by Title 26 of the United States Code. A key exception, contained in 26 U.S.C. 6103(j), authorizes the furnishing of return information to Census "for the purpose, but only to the extent necessary in the structuring, of censuses and … conducting related statistical activities authorized by law."

The Federal Policy for the Protection of Human Subjects

Under the current Federal Policy for the Protection of Human Subjects, adopted in 1991 and known as the Common Rule, research projects that are supported or regulated by any of 17 federal agencies are subject to certain federal oversight requirements.[19] In accordance with the Common Rule, organizations have established local institutional review boards, made up of both scientists and nonscientists, to review whether researchers minimize the risks to research subjects and obtain their informed consent. When appropriate, IRBs are also supposed to consider whether the research projects under their review will protect the privacy of subjects and inform them of the extent to which their data will be kept confidential.[20]

Research using individually identifiable information may be permitted by an IRB with a waiver or

[19]HHS regulations (codified at 45 C.F.R. Part 46, Subpart A) apply to research involving human subjects that is conducted, supported, or regulated by HHS. In addition, the following agencies have adopted regulations incorporating the substance of the HHS regulations: Departments of Agriculture, Commerce, Defense, Education, Energy, Housing and Urban Development, Justice, Transportation, and Veterans Affairs; Agency for International Development; Central Intelligence Agency; Consumer Product Safety Commission; Environmental Protection Agency; National Aeronautics and Space Administration; National Science Foundation; and Social Security Administration.

[20]The Common Rule defines human subjects as "living individuals about whom an investigator conducting research obtains data through (1) intervention or interaction with the individuals or (2) their identifiable private information."

modification of informed consent if the IRB finds and documents that each of the following criteria has been satisfied: (1) the research involves no more than minimal risk to subjects; (2) the rights and welfare of subjects will not be adversely affected; (3) the research could not practicably be carried out without the waiver or alteration of the consent requirement; and (4) whenever appropriate, subjects will be provided with pertinent information after participation.

Appendix IV
Selected Laws and Regulations Relating to Record Linkage and Privacy

Toward a More Complete Representation of Federal Record Linkage

Developing a more complete representation of federal linkage efforts would involve addressing questions on (1) the scope of federal record linkage, (2) its goals and impacts, and (3) current federal agency plans and likely future directions. Study questions for each of these areas are described below.

The Scope of Federal Record Linkage

Major questions: How widespread is federal or federally sponsored record linkage for statistics and research (number of projects, numbers of data subjects in each project)? And what are the key characteristics of these efforts?

Specific questions about the scope of federal record linkage might include

- What kinds of person-specific data are involved in linkage for research and statistics? How long-term are the linkage projects?
- What types of linkage (multiperson, survey-archive, multiarchive, and context) are most widely used? What agencies are most heavily involved in each type?
- To what extent is record linkage used for program evaluation and performance measurement (which is of concern in all agencies, not only research and statistical units)?
- To what extent is each of the following involved in record linkage: (1) federal grantees and contractors; (2) state and local governments with federal funding; (3) universities, researchers, and others linking personally identifiable federal data?

Goals and Impacts of Federal Record Linkage

Major questions: What are the main goals of federal or federally sponsored record linkage projects conducted for statistics and research? How useful have these linkages been?

Specific questions on goals and usefulness might include

- What is the full range of substantive areas in which questions are being addressed with linked data?
- What are examples of linkage results that have been used by Congress—or otherwise impacted policy?
- What fields have been significantly advanced by use of linked data (according to leading researchers)?
- What program evaluations and agency performance measurements have relied on record linkage?

Agency Plans and Future Directions

Major questions: What kinds of linkages are being planned by federal agencies? What future directions can be anticipated?

Specific questions on agency plans and future directions might include the following:

- Do agencies anticipate an increase, decrease, or other changes in the direction of linkage activities?
- What are examples of major planned linkages (if any)?
- What limitations, constraints, or barriers have agencies experienced in attempting to plan needed linkages?

Abbreviations

AHRQ	Agency for Healthcare Research and Quality
ASPE	Assistant Secretary for Planning and Evaluation
CBO	Congressional Budget Office
DOL	Department of Labor
FOIA	Freedom of Information Act
GAO	General Accounting Office
GPS	Global Positioning System
HCFA	Health Care Financing Administration
HHS	Department of Health and Human Services
HMO	health maintenance organization
HRDC	Human Resources Development Canada
HRS	Health and Retirement Study
IARC	International Agency for Research on Cancer
ICDAG	Interagency Confidentiality and Data Access Group
IOM	Institute of Medicine
IRB	institutional review board
IRS	Internal Revenue Service
LSOA	Longitudinal Study of Aging
MIS	management information system
MOU	memorandum of understanding
NAS	National Academy of Sciences
NBAC	National Bioethics Advisory Commission
NCHS	National Center for Health Statistics
NCI	National Cancer Institute
NDI	National Death Index
NIA	National Institute on Aging
NICHD	National Institute on Child Health and Human Development
NIOSH	National Institute for Occupational Safety and Health
NRC	National Research Council
OMB	Office of Management and Budget
SAMHSA	Substance Abuse and Mental Health Services Administration
SEER	Surveillance, Epidemiology, and End Results Survey
SIPP	Survey of Income and Program Participation
SSA	Social Security Administration
SSDI	Social Security Disability Insurance
SSN	Social Security number

Appendix VI
Abbreviations

List of References

Adam, N.R., and Wortmann, J.C. "Security-control methods for statistical databases: A comparative study." ACM Computing Surveys, 21:515-556, 1989.

Alvey, Wendy, and Bettye Jamerson (eds.). Record Linkage Techniques, 1997: Proceedings of an International Workshop and Exposition. Washington, D.C.: Federal Committee on Statistical Methodology, Office of Management and Budget (OMB), 1997.

Al-Shahi, Rustam, and Charles Warlow. "Using Patient-Identifiable Data for Observational Research and Audit." British Medical Journal 321:1031-2, 2000.

American Demographics Marketing Tools Supplement. "Private Ayes," Jan.-Feb., 1996, pp. 31-2.

Bailar, John C., III. "How Dangerous is Dioxin?" New England Journal of Medicine 324(4):260-2, 1991.

Baily, Mary Ann. "Regulating Access to Research Data Files: Ethical Issues." In Workshop on Confidentiality of and Access to Research Data Files: Workshop Papers. National Academy of Sciences (NAS), 1999.

Banks, Steven M. (Bristol Observatory, Bristol, VT). "Probabilistic Measurement of Mental Health Treatment Outcomes." Presentation at the National Conference on Health Statistics, Washington, D.C.: Aug. 2-4, 1999.

Berman, Jerry, and Janlori Goldman. A Federal Right of Information Privacy: The Need for Reform. Washington, D.C.: The Benton Foundation, 1989.

Boruch, Robert F., and Joe S. Cecil. Assuring the Confidentiality of Social Research Data. Philadelphia: University of Pennsylvania Press, 1979.

Boruch, Robert F., et al. "Resolving Ethical and Legal Problems in Randomized Experiments." Journal of Crime and Delinquency 46(3): 330-53, 2000.

Bowditch, James L., and Anthony F. Buono. A Primer on Organizational Behavior. New York: John Wiley & Sons, 1985.

Burkhauser, Richard V., et al. "How Policy Variables Influence the Timing of Social Security Disability Insurance Applications." In Workshop on Confidentiality of and Access to Research Data Files: Workshop Papers. NAS, 1999a.

Burkhauser, Richard V., et al. "The Importance of Accommodation on the Timing of Disability Insurance Applications: Results from the Survey of Disability and Work and the Health and Retirement Study." Journal of Human Resources 34(3):589-611, 1999b.

Cecil, Joe Shelby, and Eugene Griffin. "The Role of Legal Policies in Data Sharing." In Sharing Research Data, by Stephen E. Fienberg et al. (eds.). NRC Committee on National Statistics, Commission on Behavioral and Social Sciences and Education, National Research Council. Washington, D.C.: National Academy Press, 1985, pp. 148-98.

Census. See U.S. Census Bureau.

Chapman, Audrey (ed.). Health Care and Information Ethics: Protecting Fundamental Human Rights. Kansas City, MO: Sheed & Ward, 1997.

Collins, Patrick. "The California Census Research Data Center: General Description and Research Opportunities." In Workshop on Confidentiality of and Access to Research Data Files: Workshop Papers. NAS, 1999.

Dam, Kenneth, and Herbert S. Lin (eds.). Cryptography's Role in Securing the Information Society. Washington, D.C.: National Academy Press, 1996.

Dean, J. Michael, and Lenora Olson. "Protecting Confidentiality of Linked Datasets: Don't Throw the Baby Out With the Bathwater." In Workshop on Confidentiality of and Access to Research Data Files: Workshop Papers. Washington, D.C.: NAS, 1999.

de Wolf, Virginia A., et al. "The 'Checklist on Disclosure Potential of Proposed Data Releases'." In 1998 Proceedings of the Section on Government Statistics and the Section on Social Statistics, pp. 97-100, American Statistical Association, 1998.

Droitcour, Judith, et al. "The Item Count Technique as a Method of Indirect Questioning: A Review of Its Development and a Case Study Application." In Paul Biemer et al. (eds.), Measurement Errors in Surveys. New York: John Wiley and Sons, 1991, pp. 185-210.

Duncan, George T., and Diane Lambert. "Rejoinder." Journal of the American Statistical Association 81(393):27-8, 1986.

Duncan, George T., et al. Private Lives and Public Policies: Confidentiality and Accessibility of Government Statistics. Washington, D.C.: National Academy Press, 1993.

Duncan, Greg J., et al. "Sibling, Peer, Neighbor and Schoolmate Correlations as Indicators of the Importance of Context for Adolescent Development." Paper prepared for presentation at research meetings of the Population Association of America. New York: Mar. 25-7, 1999.

Egan, Gerard. Working the Shadow Side: A Guide to Positive Behind-the-Scenes Management. San Francisco: Jossey-Bass, 1994.

Fair, Martha. "Recent Developments at Statistics Canada in the Linking of Complex Health Files." In Federal Committee on Statistical Methodology Research Conference. 5 vols. (Volume for Wednesday All Sessions, pp. 19-38). Springfield, VA: National Technical Information Service (NTIS PB99-166795), Nov. 1999.

Fanning, John P. "The Use of Administrative Records for Research and Statistical Purposes: Fair Information Practice Policy As A Guide." Presented at Seminar on Interagency Coordination and Cooperation, Council of Professional Associations on Federal Statistics, Bethesda, MD: Nov. 1998.

Fanning, John P. "Privacy and Research: Public Policy Issues." Unpublished discussion outline available from John Fanning, HHS Office of the Assistant Secretary for Planning and Evaluation, Washington, D.C., 1999.

Federal Committee on Statistical Methodology, Subcommittee on Matching Techniques. Report on Exact and Statistical Matching Techniques, Statistical Policy Working Paper 5. Washington, D.C.: Department of Commerce, 1980. Available at http://www.fcsm.gov/working-papers/wp5.html

Federal Committee on Statistical Methodology. Report on Statistical Disclosure Limitation Methodology. Statistical Policy Working Paper 22. Washington, D.C.: OMB, Office of Information and Regulatory Affairs, Statistical Policy Office, 1994. Available at http://www.fcsm.gov/working-papers/wp22.html

Federal Committee on Statistical Methodology Research Conference, 5 vols. Springfield, VA: NTIS (PB99-166795), Nov. 1999.

Feldstein, Martin, and Jeffrey Liebman. "The Distributional Effects of an Investment-Based Social Security System." Working Paper 7492. Cambridge, MA: National Bureau of Economic Research (NBER), Jan. 2000.

Fellegi, Ivan P. "Record Linkage and Public Policy—A Dynamic Evolution." In Alvey, Wendy, and Bettye Jamerson (eds.). Record Linkage Techniques, 1997. Washington, D.C.: Federal Committee, 1997, pp. 3-12.

Fellegi, Ivan P., and Alan B. Sunter. "A Theory of Record Linkage." Journal of the American Statistical Association 64:1183-210, 1969.

Fienberg, Stephen E. "Confidentiality and Disclosure Limitation Methodology: Challenges for National Statistics and Statistical Research." Draft paper. Pittsburgh: Carnegie Mellon Department of Statistics, Sept. 30, 1997.

Fienberg, Stephen, and Leon Willenborg. "Introduction to the Special Issue: Disclosure Limitation Methods for Protecting the Confidentiality of Statistical Data." Journal of Official Statistics 14(4):337-45, 1998.

Fingerhut, Marilyn, et al. "Cancer Mortality in Workers Exposed to 2,3,7,8-Tetrachlorodibenzo-p-dioxin." New England Journal of Medicine 324(4):212-18, 1991.

Flaherty, David H. Protecting Privacy in Surveillance Societies. Chapel Hill: University of North Carolina Press, 1989.

GAO. See U.S. General Accounting Office.

Gastwirth, Joseph L. "Comment." Journal of the American Statistical Association 81(393):23-5, 1986.

Gates, Gerald W., and Deborah Bolton. "Privacy Research Involving Expanded Statistical Uses of Administrative Records." In 1998 Proceedings of the Section on Government Statistics and the Section on Social Statistics, pp. 203-8, American Statistical Association, 1998.

Gellman, Robert. "Taming the Privacy Monster: A Proposal for a Non-Regulatory Privacy Agency." Government Information Quarterly 17(3):235-41, 2000.

Gill, Leicester E. "Ox-Link: The Oxford Medical Record Linkage System." In Alvey and Jamerson (eds.), Record Linkage Techniques, 1997. Washington, D.C.: Federal Committee, 1997, pp. 15-33.

Goldman, Janlori. "Protecting Privacy to Improve Health Care." Health Affairs 17(6):47-60, 1998.

Goldman, Janlori, and Deirdre Mulligan. Privacy and Health Information Systems: A Guide to Protecting Patient Confidentiality. Washington, D.C.: Center for Democracy and Technology, 1996.

Hankey, Benjamin, et al. "The Surveillance, Epidemiology, and End Results Program: A National Resource." Cancer Epidemiology, Biomarkers & Prevention 8:1117-21, 1999.

Health Privacy Working Group. Best Principles for Health Privacy. Washington, D.C.: Georgetown University Institute for Health Care Research and Policy, 1999.

HEW. See U.S. Department of Health, Education, and Welfare.

HHS. See U.S. Department of Health and Human Services.

Horm, John. "National Center for Health Statistics Approaches to the Release of Microdata: Data Perturbation and the Research Data Center." In Workshop on Confidentiality of and Access to Research Data Files: Workshop Papers. Washington, D.C.: NAS, 1999.

Hoover, Robert N. "Dioxin Dilemmas." Journal of the National Cancer Institute 91(9):745-6, 1999.

Hotz, V. Joseph, et al. (eds.). Administrative Data for Policy-Relevant Research: Assessment of Current Utility and Recommendations for Development. A Report of the Advisory Panel on Research Uses of Administrative Data of the Northwestern University/ University of Chicago Joint Center for Poverty Research, 1998.

Hundepool, Anco, et al. μ-ARGUS Users Manual, Version 3.0. Statistics Netherlands (1998a).

Hundepool, Anco, et al. τ-ARGUS Users Manual, Version 2.0. Statistics Netherlands (1998b).

Hundepool, Anco J. and Leon C. R. J. Willenborg. "μ- and τ-ARGUS Software Packages for Statistical Disclosure Control." In Alvey and Jamerson (eds.), Record Linkage Techniques, 1997. Washington, D.C.: Federal Committee, 1997, pp. 142-9.

Iams, Howard M., and Steven H. Sandell. "Projecting Social Security Earnings: Past Is Prologue." Social Security Bulletin 60(2):3-16, 1997.

IARC. See International Agency for Research on Cancer.

ICDAG. See Interagency Confidentiality and Data Access Group.

Institute of Medicine (IOM). Protecting Data Privacy in Health Services Research. Washington, D.C.: National Academy Press, 2000.

Interagency Confidentiality and Data Access Group. "Checklist on Disclosure Potential of Proposed Data Releases." Washington, D.C.: OMB Federal Committee on Statistical Methodology, July 1999.

Internal Revenue Service (IRS). Building a Foundation for Culture Change. Washington, D.C.: 1999.

IRS. Tax Information Security Guidelines for Federal, State, and Local Agencies: Safeguards for Protecting Federal Tax Returns and Return Information (Pub. 1075). Washington, D.C.: June 2000. Available at ftp://ftp.fedworld.gov/pub/irs-utl/pub1075.pdf

International Agency for Research on Cancer, World Health Organization. "Polychlorinated dibenzo-para-dioxins and polychlorinated dibenzofurans." IARC Monographs on the Evaluation of Carcinogenic Risks to Humans 69, Lyons, France: 1997.

Jabine, Thomas B. "Procedures for Restricted Data Access." Journal of Official Statistics 9(2):537-89, 1993.

Jabine, Thomas B., et al. Survey of Income and Program Participation Quality Profile. Washington, D.C.: U.S. Bureau of the Census, May 1990.

Juster, F. Thomas, and Richard Suzman. "An Overview of the Health and Retirement Study." Journal of Human Resources, 30(Supplement):S7-S56, 1995.

Kandel, Denise. "Adolescent Marihuana Use: Role of Parents and Peers." Science 181:1067-9, 1973.

Keller-McNulty, Sallie, and Elizabeth A. Unger. "Database Systems: Inferential Security." Journal of Official Statistics 9(2):475-99, 1993.

Kennickell, Arthur B. "Multiple Imputation and Disclosure Protection: The Case of the 1995 Survey of Consumer Finances." In Alvey, Wendy, and Bettye Jamerson (eds.). Record Linkage Techniques, 1997. Washington, D.C.: Federal Committee, 1997, pp. 248-67.

Kilss, Beth, and Wendy Alvey (eds.). Record Linkage Techniques—1985: Proceedings of the Workshop on Exact Matching Methodologies. Cosponsored with the Washington Statistical Society and the Federal Committee on Statistical Methodology (Pub. 1299 (2-86)). Washington, D.C.: IRS Statistics of Income Division, 1985.

Korn, David. Letter to the Editor. The Washington Post. Aug. 1, 2000.

Kovar, Mary Grace, et al. "The Longitudinal Study of Aging: 1984-1990." Vital Health and Statistics 1:28 (Pub. 92-1304). Washington, D.C.: National Center for Health Statistics, 1992.

Lowrance, William W. Privacy and Health Research: A Report to the U.S. Secretary of Health and Human Services. Washington, D.C.: 1997.

Marsh, Catherine, et al. "Safe Data Versus Safe Settings: Access to Microdata From the British Census." International Statistical Review 62(1):35-53, 1994.

McGough, Helen. "Social Science Research: Privacy and Confidentiality Issues." Paper presented at conference on Privacy and Confidentiality in Clinical and Social Science Research: Myth or Reality? In Houston, TX, Feb. 10-11, 2000.

McMillen, Marilyn. "National Center for Education Statistics: Data Licensing Systems." In Workshop on Confidentiality of and Access to Research Data Files: Workshop Papers. Washington, D.C.: NAS, 1999.

Melton, L. Joseph. "The Threat to Medical-Records Research." New England Journal of Medicine 337(20):1466-70, Nov. 13, 1997.

Miller, Judith Droitcour. "The Nominative Technique: A New Method of Estimating Heroin Prevalence." In Beatrice Rouse et al., Self-Report Methods of Estimating Drug Use: Meeting Current Challenges to Validity (NIDA (National Institute on Drug Abuse) Research Monograph 57.) Washington, D.C.: U.S. Government Printing Office (GPO), 1985.

Mitchell, Olivia, et al. "Construction of the Earnings Benefits File (EBF) for Use With the Health and Retirement Survey." Working Paper W5707. Cambridge, MA: NBER, Aug. 1996.

National Bioethics Advisory Commission. "Ethical and Policy Issues in Research Involving Human Participants." Draft report, Dec. 2000 (2000a).

National Bioethics Advisory Commission. Federal Agency Survey on Policies and Procedures for the Protection of Human Subjects in Research. Draft distributed at Salt Lake City, UT Oct. 24-25, 2000 (2000b).

National Cancer Policy Board, Institute of Medicine, and National Research Council. Enhancing Data Systems to Improve the Quality of Cancer Care, Hewitt, Maria and Joseph V. Simone (eds.). Washington, D.C.: National Academy Press, 2000.

National Center for Health Statistics. NCHS Staff Manual on Confidentiality (rev. ed.). Hyattsville, MD: 1984.

National Center for Health Statistics. Programs and Activities. (Pub. No. 99-1200). Hyattsville, MD: Aug. 1999.

National Center for Health Statistics. Shaping a Vision for 21st Century Health Statistics: Interim Report. Hyattsville, MD: June 2000.

National Committee on Vital and Health Statistics, Subcommittee on Privacy and Confidentiality. Roundtable Discussion: Identifiability of Data. Washington, D.C.: Jan. 28, 1998.

National Research Council, Committee on National Statistics, Commission on Behavioral and Social Sciences and Education. Sharing Research Data, Stephen E. Fienberg et al. (eds.). Washington, D.C.: National Academy Press, 1985.

National Research Council. For the Record: Protecting Electronic Health Information. Washington, D.C.: National Academy Press, 1997.

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

NBAC. See National Bioethics Advisory Commission.

NRC. See National Research Council.

The New York Times. "Worried Swedes Questioning Wide Reach of Researchers," by Joseph Lelyveld. Mar. 11, 1986.

Newcombe, Howard B., et al., "Automatic Linkage of Vital Records." Science 130:954-9, 1959.

Newcombe, Howard B., et al. "The Use of Names for Linking Personal Records." Journal of the American Statistical Association 87:420, 1992. (Reprinted in Alvey and Jamerson (eds.), Record Linkage Techniques, 1997. Washington, D.C.: Federal Committee, 1997, pp. 335-46.)

OECD. See Organization for Economic Cooperation and Development.

Office of Management and Budget (OMB). "Privacy Act Implementation." 40 Fed. Reg., July 9, 1975, pp. 28948-78.

OMB. "Order Providing for the Confidentiality of Statistical Information and Extending the Coverage of Energy Statistical Programs Under the Federal Statistical Confidentiality Order." 62 Fed. Reg., June 27, 1997, pp. 35044-49.

Olson, Janice A. "The Health and Retirement Study: The New Retirement Survey." Social Security Bulletin 59(1), 1996.

Olson, Janice A. "Linkages with Data from Social Security Administrative Records in the Health and Retirement Study." (ORES Working Paper Series, No. 84.) Washington, D.C.: Social Security Administration, Office of Research, Evaluation and Statistics, Aug. 1999.

Organization for Economic Cooperation and Development. Guidelines on the Protection of Privacy and Transborder Flows of Personal Data, Sept. 1980. Available at: http://www.oecd.org/dsti/sti/it/secur/prod/PRIV-en.htm

Organization for Economic Cooperation and Development. Working Party on Information Security and Privacy. Inventory of Instruments and Mechanisms Contributing to the Implementation and Enforcement of the OECD Privacy Guidelines on Global Networks, 1999.

Ottawa Sun. "Data." Editorial, June 5, 2000a.

Ottawa Sun. "No More Big Brother: HRDC Bows to Public Pressure, Dismantles Massive Database," by Mark Dunn. May 30, 2000b.

Pandiani, John, et al. "Using Incarceration Rates to Measure Mental Health Program Performance." Journal of Behavioral Health Services and Research 25(3):300-11, 1998.

Panel on Civic Trust and Citizen Responsibility. A Government to Trust and Respect: Rebuilding Citizen-Government Relations for the 21st Century. Washington, D.C.: National Academy of Public Administration, June 1999.

Penslar, Robin Levin, and Joan P. Porter. Institutional Review Board Guidebook. HHS Office for Human Research Protections, 1993. Available at http://ohrp.osophs.dhhs.gov/irb/irb_guidebook.htm

Piccinino, Linda J., and William D. Mosher. "Contextual Data Files From the 1995 National Survey of Family Growth: Access and Analyses." In Federal Committee on Statistical Methodology Research Conference. 5 vols. (Volume for Wednesday All Sessions, pp. 89-97). Springfield, VA: NTIS (PB99-166795), Nov. 1999.

Potosky, Arnold L., et al. "Breast Cancer Survival and Treatment in Health Maintenance Organization and Fee-for-Service Settings." Journal of the National Cancer Institute 89(22):1683-91, 1997.

Potosky, Arnold L., et al. "Prostate Cancer Treatment and Ten-Year Survival Among Group/Staff HMO and Fee-for-Service Medicare Patients." Health Services Research 34(2):525-46, 1999.

Prevost, Ronald, and Charlene Leggieri. "Expansion of Administrative Records Uses at the Census Bureau: A Long-Range Research Plan (abr)." In Federal Committee on Statistical Methodology Research Conference, 5 vols. (Volume for Monday A Sessions, pp. 20-9). Springfield, VA: NTIS (PB99-166795), Nov. 1999.

Privacy Commissioner of Canada. Privacy Commissioner Annual Report, 1999-2000. Ottawa: May, 2000 (2000a).

Privacy Commissioner of Canada (Bruce Phillips). "Privacy Commissioner applauds dismantling of database." News release. Ottawa: May 29, 2000 (2000b).

Privacy Protection Study Commission. Personal Privacy in an Information Society. Washington, D.C.: GPO, 1977.

Purdy, Jedediah. "An Intimate Invasion." USA Weekend, June 30-July 2, 2000.[1]

Rasinski, Kenneth A., and Douglas Wright. "Practical Aspects of Disclosure Analysis." Of Significance, 2(1): 35-41, 2000.

Relyea, Harold C. The Privacy Act: Emerging Issues and Related Legislation. Washington, D.C.: Congressional Research Service, The Library of Congress, Jan. 2001.

Riley, Gerald F., et al. "Stage of Cancer at Diagnosis for Medicare HMO and Fee-for-Service Enrollees." American Journal of Public Health 84(10):1598-604, 1994.

Riley, Gerald F., et al. "Stage at Diagnosis and Treatment Patterns Among Older Women With Breast Cancer: An HMO and Fee-for-Service Comparison." Journal of the American Medical Association 281(8):720-6, 1999.

[1]This article and additional statistical information on a national privacy poll conducted by Opinion Research are on the Internet at www.usaweekend.com/00_issues/000702/000702privacy.html

Rittenhouse, Joan Dunne, and Judith Droitcour Miller. "Social Learning and Teenage Drug Use: An Analysis of Family Dyads." Health Psychology 3(4):329-45, 1984.

Robbin, Alice, et al. "A Survey of Statistical Disclosure Limitation (SDL) Practices of Organizations that Distribute Public Use Microdata." In Workshop on Confidentiality of and Access to Research Data Files: Workshop Papers. Washington, D.C.: NAS, 1999.

Rubin, Donald B. "Discussion: Statistical Disclosure Limitation." Journal of Official Statistics 9(2):461-8, 1993.

Ruggles, Steven. "Foreword—A Data User's Perspective on Confidentiality." Of Significance 2(1): 1-5, 2000.

Scheuren, Fritz. Book review, Private Lives and Public Policies: Confidentiality and Accessibility of Government Statistics, George T. Duncan, Thomas B. Jabine, and Virginia A. de Wolf (eds.). Journal of the American Statistical Association 90:386-7, 1995.

Scheuren, Fritz. "Linking Health Records: Human Rights Concerns." In Audrey Chapman (ed.), Health Care and Information Ethics: Protecting Fundamental Human Rights. Kansas City, MO: Sheed and Ward, 1997, pp. 252-78.

Scheuren, Fritz. "Administrative Records and Census Taking." Survey Methodology 25(2):151-60, 1999.

Scheuren, Fritz, and Jeri Mulrow. "The Confidentiality Beasties: A Fable About the Elephant, the Duck, and the Pig." In James Dalton and Beth Kilss (eds.), Turning Administrative Systems Into Information Systems, 1998-1999, pp. 137-41. Washington, D.C.: IRS (Pub. 1299), 1999.

Scheuren, Fritz, and Tom Petska. "Turning Administrative Systems Into Information Systems." Statistics of Income Bulletin 13(1):15-26. Washington, D.C.: IRS, Summer 1993.

Schwartz, Richard D., and Sonya Orleans. "On Legal Sanctions." University of Chicago Law Review, 34, pp. 274-300, 1967.

Seltzer, William. "Population Statistics, the Holocaust, and the Nuremberg Trials." Population and Development Review 24(3):511-52, 1998.

Seltzer, William. "Uniform Population Identifiers: The Case for Caution and Effective Safeguards." RSS News (Royal Statistical Society), Oct. 1999, p. 10.

Singer, Eleanor, et al. "Public Attitudes Toward Data Sharing by Federal Agencies." In Alvey and Jamerson (eds.), Record Linkage Techniques, 1997. Washington, D.C.: Federal Committee, 1997, pp. 237-47.

Skinner, C. J., and D.J. Holmes. "Estimating the Re-identification Risk Per Record in Microdata." Journal of Official Statistics 14(4):361-72, 1998.

Spruill, Nancy L., and Joseph L. Gastwirth. "On the Estimation of the Correlation Coefficient From Grouped Data." Journal of the American Statistical Association 77(379):614-20, 1982.

Stearns, Sally C., et al. "Estimates of National Hospital Use From Administrative Data and Personal Interviews." Journal of Official Statistics 12(1):47-61, 1996a.

Stearns, Sally C., et al. "Risk Indicators for Hospitalization During the Last Year of Life." Health Services Research 31(1):49-69, 1996b.

Steenland, Kyle, et al. "Cancer, Heart Disease, and Diabetes in Workers Exposed to 2,3,7,8-Tetrachlorodibenzo-p-dioxin." Journal of the National Cancer Institute 91(9):779-86, 1999.

Sweeney, Latanya. "Weaving Technology and Policy Together to Maintain Confidentiality." Journal of Law, Medicine, and Ethics (25):98-110, 1997.

Sweeney, Latanya. A Primer on Data Privacy Protection, (In press a.)

Sweeney, Latanya. The Identifiability of Data. (In press b.)

Sweeney, Latanya. Towards All the Data on All the People. (In press c.)

Thompson, Ross A. "Protecting the Health Services Research Data of Minors." In Protecting Data Privacy in Health Services Research, pp. 129-40. Washington, D.C.: National Academy Press, 2000.

Toronto Star. "Ottawa Pulls the Plug on Big Brother," by Valerie Lawton. May 30, 2000.

Toronto Sun. "Liberals Axe 'Big Brother': Stewart Nixes Fed Databank," by Mark Dunn. May 30, 2000.

U.S. Census Bureau. Survey of Income and Program Participation Information Booklet. 1992 Panel (Waves 1-9): 1993 Panel (Waves 1-9). (SIPP-9220.) Washington, D.C.: Jan. 2, 1995.

U.S. Census Bureau. Survey of Income and Program Participation (SIPP) Field Representative's Interviewing Manual. Sept. 1997.

U.S. Census Bureau. "ST-99-2, State Population Estimates and Demographic Components of Population Change: April 1, 1990 to July 1, 1999." Available at http://www.census.gov/population/www/estimates/state/st-99-2.txt

U.S. Census Bureau. "History of the 1997 Economic Census." (POL/00-HEC.) Washington, D.C.: July 2000 (2000a).

U.S. Census Bureau. "1990 to 1999 Annual Time Series of County Population Estimates by Age, Sex, Race, and Hispanic Origin." 2000b. Available at http://www.census.gov/population/www/estimates/co_casrh.html

U.S. Department of Commerce, Office of Federal Statistical Policy and Standards. A Framework for Planning U.S. Federal Statistics for the 1980's. Washington, D.C.: GPO, 1978.

U.S. Department of Energy, Creating an Ethnical Framework for Studies That Involve the Worker Community: Suggested Guidelines. 2000. Available at http://www.science.doe.gov/ober/humsubj/wsguidebk.html

U.S. Department of Health, Education, and Welfare. Secretary's Advisory Committee on Automated Personal Data Systems. Records, Computers and the Rights of Citizens. (DHEW Pub. No. (OS)73-94.) Washington, D.C.: U.S. Department of Health, Education and Welfare, 1973.

U.S. Department of Health and Human Services. Task Force on Privacy, Office of the Assistant Secretary for Planning and Evaluation and the Agency for Agency for Health Care Policy Research. Health Records: Social Needs and Personal Privacy. Conference Proceedings (Doc. No. PB94-168192). Washington, D.C.: NTIS, Feb. 11-12, 1993.

U.S. Department of Health and Human Services, Office of the Inspector General. Institutional Review Boards: Promising Approaches. Report No. OEI-01-91-00191, June 1998. Available on the Internet at http://www.dhhs.gov/progorg/oei.

U.S. Department of Health and Human Services. "Standards for Privacy of Individually Identifiable Health Information." 65 Fed. Reg., Dec. 28, 2000, pp. 82462-828. See also 66 Fed. Reg., Feb. 26, 2001, p. 12434 and 66 Fed. Reg., Feb. 28, 2001, pp. 12738-9.

U.S. Department of Labor, Office of Inspector General. BLS Information Technology, Survey Processing and Administrative Controls Must Be Improved. (Rept. No. 09-99-007-11-001.) July 1999.

U.S. Department of the Treasury, Office of Tax Policy. Scope and Use of Taxpayer Confidentiality Disclosure Provisions, Volume I: Study of General Provisions, Report to the Congress, October 2000.

U.S. General Accounting Office. Organizational Culture: Techniques Companies Use to Perpetuate or Change Beliefs and Values. GAO/NSIAD-92-105, Feb. 27, 1992.

U.S. General Accounting Office. Communications Privacy: Federal Policy and Actions. GAO/OSI-94-2, Nov. 4, 1993.

U.S. General Accounting Office. Information Superhighway: An Overview of Technology Challenges. GAO/AIMD-95-23, Jan. 23, 1995.

U.S. General Accounting Office. Job Training Partnership Act: Long Term Earnings and Employment Outcomes. GAO/HEHS-96-40, Mar. 4, 1996 (1996a).

U.S. General Accounting Office. Scientific Research: Continued Vigilance Critical to Protecting Human Subjects. GAO/HEHS-96-72, Mar. 8, 1996 (1996b).

U.S. General Accounting Office. Statistical Agencies: Statutory Requirements Affecting Government Policies and Programs. GAO/GGD-96-106, July 17, 1996 (1996c).

U.S. General Accounting Office. Social Security Administration: Internet Access to Personal Earnings and Benefits Information. GAO/T-AIMD/HEHS-97-123, May 6, 1997.

U.S. General Accounting Office. Decennial Census: Overview of Historical Census Issues. GAO/GGD-98-103, May 1998 (1998a).

U.S. General Accounting Office. Executive Guide: Information Security Management: Learning From Leading Organizations. GAO/AIMD-98-68, May 1,1998 (1998b).

U.S. General Accounting Office. Information Security: Serious Weaknesses Place Critical Federal Operations and Assets at Risk. GAO/AIMD-98-92, Sept. 23, 1998 (1998c).

U.S. General Accounting Office. Federal Information System Controls Audit Manual. GAO/AIMD-12.19.6, Jan. 1999 (1999a).

U.S. General Accounting Office. Information Security Risk Assessment: Practices of Leading Organizations. GAO/AIMD-00-33, Nov. 1, 1999 (1999b).

U.S. General Accounting Office. Medical Records Privacy: Access Needed for Health Research but Oversight of Privacy Protections Is Limited. GAO/HEHS-99-55, Feb. 24, 1999 (1999c).

U.S. General Accounting Office. Medicare: Improvements Needed to Enhance Protection of Confidential Health Information. GAO/HEHS-99-140, July 29, 1999 (1999d).

U.S. General Accounting Office. Social Security: Government and Commercial Use of the Social Security Number Is Widespread. GAO/HEHS-99-28, Feb. 1999 (1999e).

U.S. General Accounting Office. Survey Methodology: An Innovative Technique for Estimating Sensitive Survey Items. GAO/GGD-00-30, Nov. 1, 1999 (1999f).

U.S. General Accounting Office. Taxpayer Confidentiality: Federal, State, and Local Agencies Receiving Taxpayer Information. GAO/GGD-99-164, Aug. 30, 1999 (1999g).

U.S. General Accounting Office. Benefit and Loan Programs: Improved Data Sharing Could Enhance Program Integrity. GAO/HEHS-00-119, Sept. 13, 2000 (2000a).

U.S. General Accounting Office. The Challenge of Data Sharing: Results of a GAO-Sponsored Symposium on Benefit and Loan Programs. GAO-01-67, Oct. 20, 2000 (2000b).

U.S. General Accounting Office. Computer Security: Critical Federal Operations and Assets Remain at Risk. GAO/T-AIMD-00-314, Sept. 11, 2000 (2000c).

U.S. General Accounting Office. Information Security: Serious and Widespread Weaknesses Persist at Federal Agencies. GAO/AIMD-00-295, Sept. 6, 2000 (2000d).

U.S. General Accounting Office. Internet Privacy: Agencies' Efforts to Implement OMB's Privacy Policy. GAO/GGD-00-191, Sept. 5, 2000 (2000e).

U.S. General Accounting Office. Privacy Standards: Issues in HHS' Proposed Rule on Confidentiality of Personal Health Information. GAO/T-HEHS-00-106, Apr. 26, 2000 (2000f).

U.S. General Accounting Office. Health Privacy: Regulation Enhances Protection of Patient Records but Raises Practical Concerns. GAO-01-387T, Feb. 8, 2001 (2001a).

U.S. General Accounting Office. High Risk Series: An Update. GAO-01-263, Jan. 2001 (2001b).

U.S. General Accounting Office. Medical Privacy Regulation: Questions Remain About Implementing the New Consent Requirement. GAO-01-584, April 6, 2001 (2001c).

Wahl, Jenny B. "Linking Federal Estate Tax Records." In Alvey and Jamerson (eds.), Record Linkage Techniques, 1997. Washington, D.C.: Federal Committee, 1997, pp. 171-8.

Wahl, Jenny B. "Riches to Riches? The Importance of Intergenerational Transfers on Wealth Distribution." Unpublished article. Carleton College Department of Economics, Northfield, MN., Nov. 1998.

Wallman, Katherine K., and Jerry L. Coffey. "Sharing Statistical Information for Statistical Purposes." In Alvey and Jamerson (eds.), Record Linkage Techniques, 1997. Washington, D.C.: Federal Committee, 1997, pp. 268-75.

Ware, Willis. "Lessons for the Future: Privacy Dimensions of Medical Record Keeping." In Health Records—Social Needs and Personal Privacy. Conference proceedings, Feb. 11-12, 1993.[2]

Warner, Stanley L. "Randomized Response: A Survey Technique for Eliminating Evasive Answer Bias." Journal of the American Statistical Association 60:63-9, 1965.

[2]Sponsored by HHS Task Force on Privacy, Office of the Assistant Secretary for Planning and Evaluation and the Agency for Health Care Privacy and Research. Document No. PB94-168192. Washington, D.C.: NTIS, pp. 43-51.

The Washington Times. "Database Spurs Rage in Canadians," by Barry Brown. June 17, 2000.

Webb, Eugene J., et al. Unobtrusive Measures. Sage Classics 2, Rev. ed. Thousand Oaks, CA: Sage, 2000.

Winkler, William E. "Matching and Record Linkage." In Cox et al. (eds.), Business Survey Methods. New York: Wiley, 1995, pp. 355-84.

Ordering Copies of GAO Reports

The first copy of each GAO report and testimony is free. Additional copies are $2 each. Orders should be sent to the following address, accompanied by a check or money order made out to the Superintendent of Documents, when necessary. VISA and MasterCard credit cards are accepted, also. Orders for 100 or more copies to be mailed to a single address are discounted 25 percent.

Order by mail:

U.S. General Accounting Office
P.O. Box 37050
Washington, DC 20013

or visit:

Room 1100
700 4th St. NW (corner of 4th and G Sts. NW)
U.S. General Accounting Office
Washington, DC

Orders may also be placed by calling (202) 512-6000 or by using fax number (202) 512-6061, or TDD (202) 512-2537.

Each day, GAO issues a list of newly available reports and testimony. To receive facsimile copies of the daily list or any list from the past 30 days, please call (202) 512-6000 using a touch-tone phone. A recorded menu will provide information on how to obtain these lists.

Viewing GAO Reports on the Internet

For information on how to access GAO reports on the INTERNET, send e-mail message with "info" in the body to:

info@www.gao.gov

or visit GAO's World Wide Web Home Page at:

http://www.gao.gov

Reporting Fraud, Waste, and Abuse in Federal Programs

To contact GAO FraudNET use:

Web site: http://www.gao.gov/fraudnet/fraudnet.htm
E-Mail: fraudnet@gao.gov
Telephone: 1-800-424-5454 (automated answering system)

PRINTED ON RECYCLED PAPER

United States
General Accounting Office
Washington, D.C. 20548-0001

Official Business
Penalty for Private Use $300

Address Correction Requested

First Class
Postage & Fees Paid
GAO
Permit No. G100