Journal of Privacy and Confidentiality (2011) **3**, Number 2, 3–40

How Uncertainty about Privacy and Confidentiality Is Hampering Efforts to More Effectively Use Administrative Records in Producing U.S. National Statistics

Gerald W. Gates*

Abstract. U.S. federal statistical agencies continually face challenges in obtaining and using administrative records and in providing useful analytic products to support policy analysis and program planning. At each of three decision points—obtaining the administrative data, integrating the data into statistical programs, and releasing useful data products—concerns over privacy and confidentiality determine to a great extent how effectively these data are used. Although there is a long history of relevant research on privacy attitudes and methodologies to protect confidentiality in published data, agency decisions to share or publish data are not necessarily informed by known risks. Additional research is proposed to help identify and manage these risks. The paper also proposes government actions to ensure that U.S. federal statistical agencies are meeting the nation's data needs through the appropriate application of survey and administrative data.

Keywords: Data Sharing, Record Linkage, Informed Consent

1 Introduction

Administrative records represent profiles of individuals and businesses based on their past interaction with government for such things as determining eligibility for government programs, obtaining benefits and services, paying taxes, helping to improve general health and welfare, and preserving public rights. These records are also used by government to detect and prevent fraud and abuse, to assess program performance, and to set broad economic and social policy. When provided to statistical agencies, these records offer a cost-effective way to evaluate, enhance, and improve national statistics. Concerns about privacy and confidentiality play a key role in determining the extent to which these data are shared for these purposes.[1]

U.S. law is generally supportive of the sharing of identifiable administrative data for statistical uses as long as the data can be protected against non-statistical uses.[2] In practice, administrative agencies decide to share personally identifiable information with statistical agencies by first protecting their "business interests" with their clients

*Formerly Chief Privacy Officer, U.S. Census Bureau. Served as Chief of the Census Bureau's Policy Office from 1998–2005 where he led the establishment of the Census Bureau's Data Stewardship Program. He has worked on privacy, confidentiality, and data access issues and supported statistical uses of administrative records for over 25 years. mailto:gwgates@verizon.net

(program participants) and with funding sources (Administration, Congress).[3] Since violations of privacy or breaches of confidentiality can damage their mission, administrative agencies will assess the current climate and risk mitigation measures proposed by the statistical agency as key factors in deciding the conditions for sharing their data. Similarly, statistical agencies will decide how to use administrative data they receive based in part on how the uses fit within agreed upon privacy and confidentiality constraints and an assessment of their ability to recover should future access be denied. Finally, decisions by statistical agencies to release these data to researchers are dependent on the same assessments of confidentiality and privacy.

Unlike surveys, administrative records typically represent entire populations (e.g., all tax filers, all food stamp recipients). When used for statistics, these records are also frequently linked to other administrative records or to survey records for the same individuals. The extent of sharing and linkage is generally not apparent to the data subjects. For these reasons, there are unique privacy and confidentiality risks that must be addressed. Agencies are dealing with these risks through policies and statistical and procedural tools. Such tools include techniques for data masking, alternative data access procedures, and methods to reduce sensitivity. The Government Accountability Office's (GAO) 2001 report on *Record Linkage and Privacy* provides an excellent discussion of these tools and how they are applied.

To help inform decisions from the perspective of privacy and confidentiality, U.S. statistical agencies have sponsored research into disclosure risk and privacy attitudes. From the perspective of privacy, research suggests that although the public does not trust the government to protect their personal information, they are more likely to respond favorably to statistical studies involving their personal information when they understand the uses and potential benefits (Gerber, 2003) and (Guarino et al., 2001). Research has also determined that the public's knowledge and opinions about privacy in this context are fluid and quickly become out of date (Singer et al., 2001). Research on disclosure avoidance has led to new data products and alternative methods for researcher access, but demands for even greater access require creative solutions. Despite the progress made through privacy and confidentiality research, there remains considerable uncertainty about the degree of public understanding and acceptance of the statistical use of administrative records and whether current confidentiality protections are appropriate. As a result, we may not be realizing the full potential of these

[1] In the context of this paper, privacy refers to "information privacy" which I define as the individual's desire (claim) to control the terms under which information about him/her is acquired, used, or disclosed. Confidentiality is closely related to privacy and refers to the agreement reached with the individual/business when the information was collected about who can see the identifiable information. Changes to this agreement can be made only with the explicit consent of the individual. There are many variations of these definitions in the literature, but I believe these to be the most useful in the context of statistical activities.

[2] The Confidential Information Protection and Statistical Efficiency Act of 2002 defines non-statistical uses as the use of data in identifiable form for any purpose that is not a statistical purpose, including any administrative, regulatory, law enforcement, adjudicatory, or other purpose that affects the rights, privileges, or benefits of a particular identifiable respondent.

[3] Although government agencies do not operate for profit, they act to ensure that program funding is protected. Ensuring public participation is critical to program funding.

records.

In this paper I explore a broad range of policy, public opinion, and methodological issues surrounding the sharing and use of administrative records for federal statistics. My primary goal is to encourage new research on privacy and confidentiality in order to provide agency decision makers with relevant information to better understand the benefits and risks of sharing and using administrative data. It is also my goal to encourage a more open process that respects the interests of all parties, in particular the individuals whose records are to be shared.

In Section 2, I provide a summary of some important new uses of administrative records, and in Section 3 I discuss legal and policy support for such uses. The reader desiring more details on the relevant laws and policies is directed to the appendices. In Sections 4–6, I describe the role that privacy and confidentiality play in acquiring, using, and providing researcher access to administrative records. In Section 7, I discuss past research related to privacy and confidentiality in the context of record linkage, and I offer suggestions for new research. Section 8 provides some examples of how other countries' laws and policies on the statistical use of administrative records compare to the U.S. Finally, in Section 9 I suggest a government-wide approach to fostering administrative records use in U.S. federal statistics.

A few caveats are worth noting:

- First, many of the administrative records projects discussed in the paper involve the U.S. Census Bureau. There are three reasons for this: 1) the Census Bureau is the leading agency in conducting surveys and has a long history of linking administrative data from multiple sources; 2) much of the research on privacy and confidentiality has been conducted or funded by the Census Bureau; and 3) I have firsthand knowledge of negotiations involving the Census Bureau. I have included some examples in the text of work by the National Center for Health Statistics and I acknowledge that a great deal of important research with administrative records is being conducted in many other federal statistical agencies. The issues I discuss are similar for these agencies and proposals for additional research and policy steps should be coordinated across, and are applicable to, all statistical agencies.
- Second, although administrative records acquired for statistical uses can relate to either individuals or businesses, for the most part, this paper will focus on administrative records pertaining to individuals since privacy is a concept inherent only to individuals. Much of the discussion pertaining to confidentiality also applies to records of businesses.
- Third, the paper focuses on administrative data collected and held by government entities and does not specifically consider issues related to data collected and held by private sector organizations. Private sector records, such as those derived from credit reports and public records, are increasingly being used by U.S. statistical agencies, and privacy and confidentiality will play an important role in accessing and using these records as well.

- Finally, throughout the paper I refer both to data linkage and data integration as the process of integrating administrative records and statistical records at the individual level.

2 Recent Progress in Administrative Records Use in U.S. Federal Statistics

For many decades, administrative records have been an integral part of U.S. federal statistics. Tax records, Social Security records, unemployment insurance records, health records, education records, birth and death (vital) records, and many others have supported survey and census data in informing public policy decisions. Specifically, administrative records have served:

- As frames for economic surveys conducted by the Bureau of Economic Analysis, the Census Bureau, and the Bureau of Labor Statistics;
- To measure births, deaths, and migration within the U.S. to help produce estimates of the population between censuses;
- As a source of information about income, poverty, and health insurance at the sub-state level;
- To assess population coverage issues in surveys;
- To assess survey response accuracy;
- To assess the nature and impact of survey non-response;
- To aid survey methodologists in understanding the nature and extent of non-sampling error;
- To improve survey data editing and imputation;
- To improve questionnaire design;
- To provide improvements in survey sampling frames; and
- To improve simulation models for policy evaluation and review.

By judiciously using administrative records, agencies are saving tax dollars that would otherwise be required to collect this information directly from individuals and businesses. Reporting burden on the public is also reduced when administrative data are reused for statistics. Through administrative records, statistical agencies are also able to provide data at levels of geographic detail that they could not afford to produce otherwise. In addition, survey and census data quality are enhanced through evaluations with administrative data. Specific applications of administrative records in the U.S. are widely documented in professional papers of the American Statistical Association and

the Federal Committee on Statistical Methodology among others, and are published on statistical agencies' websites.

The following are a few of the recent advances in the statistical use of administrative records in the U.S. that involved negotiating new agreements and undertaking major record linkage efforts:

- The Census Bureau's Longitudinal Employer-Household Dynamics (LEHD) Program was initiated in 1999 to integrate census, survey, and administrative records data on workers and employers. (See `http://lehd.did.census.gov/led/`.) The program, "with the support of several national research agencies, has built a set of infrastructure files using administrative data provided by state agencies, enhanced with information from other administrative data sources, demographic and economic (business) surveys and censuses. The LEHD Infrastructure Files provide a detailed and comprehensive picture of workers, employers, and their interaction in the U.S. economy. Beginning in 2003 and building on this infrastructure, the Census Bureau has published the Quarterly Workforce Indicators (QWI), a new collection of data series that offers unprecedented detail on the local dynamics of labor markets. Despite the fine detail, confidentiality is maintained due to the application of state-of-the-art confidentiality protection methods" (Abowd et al., 2005, abstract). The backbone of the LEHD program is state-level employment data provided by partner states. The source data consist of Unemployment Insurance (UI) Wage Records, quarterly from 1990 to present, as available; Quarterly Census of Employment and Wages (QCEW and formerly ES-202) records, quarterly from 1990 to present, as available; and the latest geographical definitions of Workforce Investment Areas (WIA). While Internal Revenue Service (IRS) tax return information was originally intended to provide much of the source data for this project, LEHD managers could not get agreement with the IRS for adding such uses to the Tax Regulations that limit Census Bureau statistical use of tax data. This unique relationship is described in Section 4 of this paper.

- The Medicaid Undercount Project is a joint effort of the State Health Access Data Assistance Center at the University of Minnesota, the National Center for Health Statistics (NCHS), the Office of the Assistant Secretary For Planning and Evaluation in the Department of Health and Human Services, the Centers for Medicaid and Medicare Services, and the U.S. Census Bureau. This multiphase research project is designed to explain why discrepancies exist between survey estimates of enrollment in Medicaid and the number of enrollees reported in state and national administrative data. "Project results will benefit the Census Bureau and other participating agencies because they can be used to improve evaluation of the Medicaid programs (e.g., estimating the effects of proposed policy changes) and to improve survey methods used to collect health insurance coverage information." (See `http://www.census.gov/did/www/snacc/`). Source data for this project came from records of two surveys—the 2001–2002 Current Population Survey and the 2001–2002 National Health Interview Survey—and the Medicaid administrative data. Although this project has demonstrated that linking survey

and administrative data can lead to improvements to survey methodology and also expand the policy relevance of the data, it also made clear that issues surrounding data ownership and oversight add considerable complexity and administrative burden. According to Cox et al., "the fundamental basis for the policy and legal issues in linking the source files from the different organizations is that each organization has its own set of statutory and policy requirements to protect the confidentiality of its own data" (Cox et al., 2006, p. 2).

- The Census Bureau's Statistical Administrative Records System (StARS) was built as an essential component of the 2000 Administrative Records Experiment that was designed to assess the strengths and weaknesses of administrative data as a supplement to or substitute for decennial census population counts. It was originally designed to include records from seven major federal files held by six federal agencies. The linchpin to developing the StARS is the Social Security Administration's (SSA's) NUMIDENT file of Social Security Number (SSN) applicants which is essential to validating matches and providing demographic characteristics. For much of the 1990s the Census Bureau had sought permission to access the full NUMIDENT (rather than the 20% sample it had been receiving annually for its Population Estimates programs). SSA's concerns for privacy and confidentiality, heightened by negative public reaction to privacy abuses by its own employees, had contributed to prior failed negotiations. A breakthrough came in 2000 when the Census Bureau won approval for access on a trial basis.

  Ultimately, the Administrative Records Experiment did not lead to major changes for Census 2000, but the StARS was determined to be a valuable resource to support future censuses, ongoing demographic programs, and associated research. As a result, the Census Bureau has maintained and updated the StARS and has added additional data from state agencies as available. For the 2010 census the Census Bureau is using the StARS to identify potentially undercounted cases, to improve race coding, and to evaluate agreements between the Master Address File and StARS for future maintenance activities and to predict address validity. Moving forward, research plans for the 2020 census currently include an administrative records component that will rely heavily on an up-to-date StARS.

- The National Center for Health Statistics links its various health surveys with air monitoring data from the Environmental Protection Agency (EPA), death certificate records from the National Death Index (NDI), Medicare enrollment and claims data from the Centers for Medicare and Medicaid Services (CMS), and Retirement, Survivor, and Disability Insurance (RSDI) and Supplemental Security Income (SSI) benefit data from SSA. This work is designed to expand the uses of the Center's population-based surveys. Linked data files enable researchers to examine the factors that influence disability, chronic disease, health care utilization, morbidity, and mortality.[4] Recently, NCHS undertook a pilot study to link data from the National Health and Nutrition Examination Survey (NHANES) to Supplemental Nutrition Assistance Program (Food Stamp) data for the state of

[4]See `http://www.cdc.gov/nchs/data_access/data_linkage_activities.htm`.

Texas in order to assess the reason for differences in estimates of food stamp receipt between the NHANES and estimates from the Department of Agriculture. NCHS's linkage activities are supported by interagency agreements that provide mutual benefits to each of the parties. For example, the agreements often provide for free access to the restricted data by administrative agency employees and contractors at the NCHS Research Data Center.

These examples demonstrate the potential that lies in expanding the use of administrative records in U.S. federal statistical programs and highlight some of the issues that surface in acquiring administrative records, using them effectively, and making them available to researchers. These advances have not come overnight and they might not have happened at all had it not been for the interpersonal relationships that exist or were formed to overcome various legal, policy, and institutional barriers. To assess how the U.S. can advance the statistical use of administrative records we need to understand the motivations of the parties involved based on the environment in which they operate.

3 Legal and Policy Environment and the Role of the Individual

U.S. laws generally support the individual's right to decide who can see their personal information and how it can be used. In the case of administrative data, multiple laws permit the sharing of personally identifiable information without consent if the information will be protected from further disclosure and from being used for non-statistical purposes. In granting exceptions to the consent requirement in these cases, the law makers have recognized that the uses for statistics are compatible with the intended program uses and will not cause harm to the individual. In most cases, the laws permit, rather than mandate, sharing and, in some cases, specifically identify the recipient agencies and types of statistical uses.

Applicable laws can be categorized as 1) permitting limited sharing of administrative data for statistical purposes without consent; 2) requiring confidentiality and limiting uses when acquiring data for statistical purposes; and 3) encouraging administrative records use to enhance statistics and reduce reporting burden. The laws can be generic to all types of personal information collected by all agencies or specific to one agency or class of information (for example, health records or education records). Key laws that directly or indirectly impact the sharing, protection, and use of identifiable administrative data include the Privacy Act of 1974, the Confidential Information Protection and Statistical Efficiency Act of 2002, Title 13 of the United States Code (the Census Act), Title 26 of the United States Code (the Tax Code), and the Freedom of Information Act of 1996. Appendix 1 highlights how these and other laws impact sharing and use of administrative data for federal statistics.

Policy is also supportive of the statistical use of administrative records providing that confidentiality is preserved. The Privacy Protection Study Commission, the Office for Statistical Policy and Standards, and the Committee on National Statistics of the

National Academy of Sciences have issued guidance on this matter. The Federal Policy on Human Subjects Research, also known as the Common Rule, lays out informed consent requirements that come into play when data are shared for statistical research. Together, these documents highlight key aspects of the confidentiality/privacy challenges facing administrative and statistical agencies that propose to partner in a data sharing arrangement. In summary, the issues boil down to: 1) the role of the individual in determining the uses of personal information pertaining to them (informed consent), and 2) how to organizationally limit the risk of non-statistical uses of the shared data (functional separation). Appendix 2 provides brief descriptions of these policy statements.

Both law and policy recognize the individual's right to be informed of and consent to uses of their personal information. Laws permitting the sharing of administrative data for statistical uses recognize that these uses come with limited risks to the individual and provide for exceptions to the consent provision when the data are to be shared for solely statistical purposes and confidentiality is assured. The exception to the usual requirement that individual consent be obtained before using personal information is an important contribution to the effective use of records for research and statistics. Obtaining consent at the time of initial collection would complicate procedures for the administrative agency which would have to account for those who do not wish to allow their records used in this manner. Obtaining consent after the fact could be quite costly and time consuming, especially if some time has passed since the initial collection and the individuals are difficult to locate. Notice, however, is required by the Privacy Act and agencies accomplish this by publishing a System of Records Notice in the *Federal Register* describing the intended uses of the personally identifiable information, usually in a general way.

Where the administrative data are to be linked to survey or census data, rather than used alone or in combination with other administrative records, consent may come into play. Agencies may indirectly be obtaining consent for such uses by requesting a Social Security Number (SSN) from survey/census respondents to facilitate linkage. Under the Privacy Act (Title 5, U.S.C., Section 552a (note)), agencies are required to inform individuals whether providing their SSN is mandatory or voluntary, their authority to collect the SSN, and the uses to be made of it. Refusal to provide one's SSN implies refusal to permit the linkage.[5] If SSNs are not collected but linkage is planned,[6] agencies may provide a notice of intent to link and an opportunity to opt out.[7] Such consents are frequently general in nature and may not identify each source file to be linked.

Pascale (2011) describes how the Current Population Survey, conducted by the Census Bureau, uses implicit (passive) consent to seek approval for record linkage. Respondents receive a mailed letter informing them of their selection in the survey and

[5]There is some speculation that it is the growing concern over identity theft rather than record linkage, per se, that affects unwillingness to provide ones SSN.

[6]Because of increased concerns for privacy and data security, OMB issued guidance to agencies in 2007 to limit their collection and use of SSNs (Office of Management and Budget (OMB), 2007).

[7]The ability to opt out does not apply in programs like the decennial census that require mandatory reporting.

of the agency's intention to link records from other agencies. They are instructed to inform the interviewer during the interview if they object to this aspect of the survey. This approach depends upon the respondent reading the letter and understanding the implications of their decision not to object. The wording of the notice is also critical and has been the subject of privacy research as discussed in Section 7.1 of this paper.

Research that is covered under the Common Rule is subject to review by an Institutional Review Board (IRB). The Common Rule exempts research involving survey procedures where confidentiality is maintained without exception, so IRB approval is not required. However, several agencies, such as the National Center for Health Statistics and the Bureau of Justice Statistics, have established IRBs that assume responsibility for approving survey research protocols. These IRBs typically review informed consent procedures as well as procedures for ensuring confidentiality of individual records. In some cases, IRBs may require agencies to seek explicit consent prior to linking survey data with the administrative data. (See Appendix 2 for a discussion of implications of IRB review in social science research.)

Public knowledge of the statistical use of administrative records is then dependent upon an individual being informed at the time he/she responds to a survey or census or based on reading a System of Records Notice, reviewing privacy materials on agencies' websites, or finding a research report describing the methodology. There is no evidence that the public is generally knowledgeable about these uses.

An example of the potential implications of this lack of knowledge about record linkage uses occurred a decade ago when the Canadian Privacy Commissioner effectively shut down a major data linking project undertaken by the research arm of Human Resources Development Canada (HRDC), [8] primarily on the grounds that it had been insufficiently publicized. The Commissioner determined that the Longitudinal Labour Force File created by Human Resource Development Canada "is relatively invisible."[9] "HRDC is not trying to hide its existence. In fact, it describes the database in Info Source and on its Web site. Unfortunately, neither are widely read, nor easily understood, and the description of the database contains few details. Canadians don't know how much information is being collected about them or the extent to which it is being integrated and shared with others" (Privacy Commissioner of Canada, 2000, pp. 64-71). In response, the HRDC established a Privacy Management Framework that examines the operational, administrative, and research uses of personal information to ensure that all privacy issues are identified and mitigated through the use of guidelines, best practices, and tools. In 2007, the Commissioner recognized this effort as a "good practice" in privacy impact assessment.

[8] Human Resource Development Canada was renamed Human Resources and Social Development Canada in 2006.

[9] The Commissioner also determined that HRDC did not have a sufficient protective legal framework to fend off other government departments who might want to use the linked data for non-statistical uses.

4 Negotiating Access

Where both law and policy are supportive, record holders ultimately have the option to share depending on how their interests are protected. The negotiations revolve around various policy considerations pertaining to the costs and benefits for each party. There are eight factors that weigh heavily in the final outcome:

1. **Administrative costs**. Negotiations almost always involve provisions for reimbursing the administrative agency for the costs in terms of staff and computer time associated with providing the data in the formats required. During negotiations, administrative agencies must weigh the degree to which this work will detract from the primary functions of the agency.

2. **Incentives**. Negotiations sometimes involve incentives for the administrative agency. This quid pro quo may be some form of enhancement of the source data that includes the addition of metadata, geographic variables, or summary statistics. The administrative agency may also receive public use files or be given access to restricted data at a Research Data Center. Quid pro quos never involve providing identifiable linked data back to the administrative agency for non-statistical uses. Section 6 of this paper discusses mechanisms for researcher access that may accommodate some administrative agency needs.

3. **Self Interest**. Administrative agencies frequently have an interest in preserving their singular ability to analyze individual data for policy analysis, planning, and evaluation purposes. Sharing individual records with statistical agencies allows these agencies to produce data that can also be used for these purposes.

4. **Controls**. Negotiations usually stipulate the conditions for access and use of the administrative data for the stated statistical purpose. This frequently includes specific legal requirements, security requirements, employee training, disclosure avoidance measures to be taken prior to release of data products, and provisions for maintaining accountability and auditing compliance.

5. **Rights**. Negotiations usually define the roles of the parties in terms of custodianship of the identified data. For instance, signed agreements usually provide rights to the statistical agency to retain and use the identified data as well as any product that integrates the identified data with the agency's survey data. However, the agreements often impose limits and controls that imply ownership rights are jointly held.

6. **Public Support**. Attitudes of program participants and survey participants are always in the back of agency decision makers' minds when deciding to share information. Negative public reaction (frequently related to privacy and confidentiality) can have dramatic impacts on the agency's ability to function by reducing participation, increasing program complexity, and fostering greater oversight.

7. **Opinion leaders**. Related to the public's fears are concerns about the views of lawmakers, advocates, and the media, who have the power to alleviate or foster

the public's concerns. Although these groups do not work in unison, they will respond, or drive attention to, perceived or real privacy threats.

8. **Public good**. Negotiations for access often include either implicit or explicit assessment of the public good to be realized from the research use of the administrative data. A well understood appreciation for the research benefits can go a long way in moving discussions to a signed agreement.

In my experience, negotiations on these points tend to be very time consuming and can take months or even years. At various stages, the negotiations may involve lawyers, policy officials, program managers, technical staff, and eventually, senior management. The final decision to share or not ultimately rests with the administrative agency since there is no third party arbiter to reconcile differences.

Privacy and confidentiality play a direct role in four of the eight factors but can indirectly influence decisions pertaining to all eight. For instance, concerns for privacy and confidentiality (both real and perceived) will impact controls over the data, the extent to which custodianship is conveyed, and the support to be expected from the public and opinion leaders. While costs, incentives, public good, and self interest are not driven by privacy per se, they can overcome, or be overwhelmed by, privacy and confidentiality concerns expressed by the negotiators. So, regardless of the legal and policy support for data sharing, the negotiation may fall apart because the two sides cannot agree on the conditions for access and use.

Although I have firsthand knowledge of the role that privacy and confidentiality play in negotiations between statistical and administrative agencies, there is limited documentation regarding the successes and almost no documentation of the failures. This is not unexpected. Where an agreement is reached, the administrative agency's only obligation is to ensure that the Privacy Act System of Records Notice accommodates the sharing whereas the statistical agency is most interested in reporting on the methodology and statistical results. Where no agreement is reached, the administrative agency has little incentive to publicly document their reasons for denying access. On the other hand, statistical agencies who are rebuffed by administrative agencies may not want to publicly announce a failure to reach agreement since it may damage hopes for future negotiations.

Over the years, I have been personally involved in negotiations between the Census Bureau and the IRS for access to and use of federal tax information in Census Bureau programs. As required by IRS law, the Census Bureau can only receive tax return information that has been previously included in the IRS Regulations that have been published in the Federal Register. In addition, uses are limited to only those statistical uses previously approved for each item. Any changes to the Regulations are contingent upon the IRS's determination that the data are <u>necessary</u> to conducting authorized statistical activities. The IRS has further interpreted this to imply that access to, and uses of, the data must be to the <u>minimum</u> extent necessary to meet Census Bureau program needs. The Census Bureau, on the other hand, is required by its law to use administrative records to the <u>maximum</u> extent necessary to avoid over-burdening the

public. These opposing views add tension to the negotiations.

In pursuing these seemingly opposing legal mandates, the two agencies also apply different views of how privacy and confidentiality can be ensured. From the Census Bureau's perspective, it views its own confidentiality law to be as strong as, or stronger than, the IRS law when compared based on penalties for improper use and the ability to fend off demands from law enforcement and the courts. The IRS, on the other hand, views sharing as a potential confidentiality threat when additional tax return information from potentially millions of taxpayers is put in the hands of another agency's employees, contractors, and agents.[10] This concern is amplified in negotiations with the Census Bureau because of its expanded reliance on Research Data Centers as a means of providing access to researchers.

Further complicating these negotiations is the residual mistrust resulting from the 1999–2000 IRS Safeguard Review that identified accounting irregularities and uses that were not properly approved by the IRS (Potok, 2009). As a result, the agencies adopted the "criteria document" that stipulates more stringent requirements for access to and use of data derived from federal tax information. (See `http://webserver03.ces.census.gov/index.php/ces/researchguidelines`). Although the control issues have been corrected by the agency's commitment to data stewardship and subsequent IRS Safeguard Reviews have been favorable, the new review and approval process has not been viewed with favor by most Census Bureau managers.

5 Integrating Administrative Records into Statistical Programs

Statistical agencies determine the appropriate uses of administrative data they obtain by considering their fitness for a particular program use that meets quality standards consisting of utility, objectivity, and integrity.[11] Further, agencies must be able to demonstrate that these data provide an acceptable alternative to data collected using survey or census methods. Once the decision is made to use records that have been obtained, the agency is committing to its Congressional funders and its data users that the source data will be available through the life of the program. Agencies are also committing to abide by the terms of agreements made with administrative agencies when accessing the records as well as commitments made to respondents in their surveys and censuses regarding uses and protections of these data when combined with administrative data. Finally, they are committing to data users that resulting data products will be available for research.

In terms of privacy and confidentiality, there are two major factors agencies con-

[10] The IRS has been under pressure from Congressional oversight bodies such as the Government Accountability Office and the Joint Committee on Taxation to address taxpayer privacy issues that have surfaced over the years.

[11] The Census Bureau further defines data quality in terms of relevance, accuracy, timeliness, accessibility, interpretability, and transparency. For example, see `http://www.census.gov/quality/standards/Quality_Standards.pdf`

sider in deciding to integrate previously-obtained administrative records in a statistical program. The first consideration is how access and use controls that have been agreed upon with the administrative agency supplier will impact program operations. Layering on additional access and use controls may be manageable for some programs but not for others. A second consideration is how the public might perceive linking data from multiple sources to create a new data set that has potential uses beyond that of each individual source. Efforts in the late 1960s to centralize U.S. federal statistics were met with outcries by civil libertarians over the proposal to establish "central databanks" which they viewed as an "Orwellian threat to personal privacy" (Robertson, 1968). Extensive linkages, even where privacy is protected, could serve to rekindle these longstanding fears.

The U.S. decennial census offers a good example of how these factors come into play. A redesigned census that incorporates administrative records could benefit from the use of federal tax information. If an agreement could be reached with the IRS that permitted this use, there is still the matter of controlling access and use of the commingled data. The IRS has ruled that any data that are commingled with tax information must follow all IRS access and use requirements. This implies that the entire decennial census and its resulting data would have to follow IRS rules, and uses would be subject to IRS review and approval similar to what is currently in place for the Census Bureau's Business Register.[12] Based on the experiences of their colleagues, those responsible for the decennial census may be reluctant to commit to integrating IRS data in the decennial census because of the additional controls and oversight that would result.

Another deterrent to the use of administrative records in the census is uncertainty about how the public might react to these uses. Although there would be cost savings and reduced reporting burden, the public might view these secondary uses of their personal information as a privacy threat and opinion leaders could champion that cause. If SSNs were requested in the census or became part of a census-administrative records database, privacy concerns would be magnified.[13] Should a privacy protest be successful, the agency would need to retrofit previous census procedures at great additional costs and suffer damage to its reputation that would certainly spill over to its survey work. This concern about how the public might view these uses has been the subject of extensive privacy research as described in Section 7 of this paper.

6 Providing Researchers with Access to Administrative Data

Data products derived from administrative data present unique challenges in ensuring that specific individuals or businesses are not identifiable. For surveys with matched

[12]The Business Register, which is the frame for most economic surveys conducted by the Census Bureau, is generated in part from IRS records.

[13]In order to limit risk of misuse, the U.S. Census Bureau has begun replacing SSNs in integrated data files with Protected Identification Keys created by the agency.

administrative data, the primary risk involves the ability of someone holding the source administrative data using record linkage techniques to identify the individual. Since the confidentiality requirements (both in Title 13 and CIPSEA) apply to both the survey and any administrative data, the fact that only the administrative agency has the ability to re-identify its program participants is not sufficient protection. Consequently, administrative data linked to survey data have not typically been made available to researchers in the form of public use microdata, although the demand for such data is great.

New techniques, such as Latin Hypercube Sampling and Inference Valid Synthetic Data, have been used to create simulated microdata that reduces the risk of disclosure (Federal Committee on Statistical Methodology (FCSM), 2005). These techniques are particularly useful for linked datasets. Latin Hypercube Sampling involves creating a replacement file with simulated values for the sensitive variables which retains the same specified statistical properties as the true microdata. Inference-Valid Synthetic Data involves replacing confidential variables using a controlled data adjustment constraint algorithm. Using this method, multiple public use files can be created from the same underlying data with each customized to different groups of users. The inference valid synthetic data methodology was applied to the Survey of Income and Program Participation (SIPP) data after the SIPP data were linked to earnings data from the Social Security Administration (Abowd and Lane, 2003). This work has considerable promise but, as Abowd and Lane acknowledge, a body of knowledge is needed about the quality of the synthetic data in relation to the confidential data. The Workshop on Synthetic Data and Confidentiality Protection held at the Census Bureau on July 31, 2009, demonstrated the advances that have been made in these techniques as well as areas where further research is needed. See `http://www.vrdc.cornell.edu/news/` for papers presented at the workshop.

Where synthetic data do not meet researchers' needs, users sometimes have the option of accessing integrated data remotely through a computer-based data analysis system that monitors queries to ensure that confidentiality is not breached. The major limitations of these data analysis systems are that users are limited in their ability to use specialized software, and heavy-duty processing will be limited by the capability of the agency's dedicated servers. Researchers who want to understand the data by looking at the outliers will be disappointed by data analysis systems as these will be cases that are suppressed to protect confidentiality.

Where access to the confidential data is required for the analysis, researchers have two additional options depending upon the agency's statutory authority. A few agencies, including the National Center for Education Statistics and the National Center for Science and Engineering Statistics of the National Science Foundation, have the authority to license universities to hold non-public data for use by authorized researchers. Individual researchers apply for access through their organization. Violations of confidentiality are subject to fines and denial of future access by the researcher. For the organization, researcher violations may mean cancellation of the license and removal of the data.

The other option, for those agencies with the authority to do so, is to designate researchers as agents to use the non-public data in a secure environment known as a research data center (RDC). Agents are sworn to protect confidentiality and are subject to the same legal penalties as agency employees. Access is monitored by agency employees or designates and products are reviewed prior to removal from the center. In some cases, researchers can bring their own data into the center to be linked to the survey and administrative data held by the agency. Some agencies house in their RDCs data from other statistical agencies in addition to their own. The primary disadvantage of this option is that researchers may have to relocate temporarily to the city where the center is housed.

A final option that is used infrequently as a last resort for providing researcher access to confidential data involves obtaining the consent of the respondent for the disclosure of identifiable information. CIPSEA provides that: "Data or information acquired by an agency under a pledge of confidentiality for exclusively statistical purposes shall not be disclosed by an agency in identifiable form, for any use other than an exclusively statistical purpose, except with the informed consent of the respondent." This envisions situations where individuals may be asked to waive confidentiality to the extent needed to permit some important secondary use of their personal information.

In the context of linking administrative records with survey and census information, one might decide to ask upfront (if part of the survey planning), or during a re-contact (if not planned), for permission to provide non-public data to researchers. Such informed consent would acknowledge the small risk that researchers could re-identify them if they have access to their administrative records. Major limitations with consent are the additional costs, the ability to locate people long after a survey is completed, and the proportion and characteristics of those who do not agree to waive their confidentiality. In addition, the wording of the waiver must adequately convey any additional risk to the individual. Prior testing is critical to assessing the nature of the waiver and the potential bias created by under-representing certain populations.

Regardless of the mechanism, when data derived in part from administrative data are provided to researchers, the statistical agency must assure the administrative agency that confidentiality is protected. In general, administrative agencies do not insist that they participate in the decision making process for release of each data file as long as professionally accepted disclosure avoidance practices are employed.[14] When data are provided to agents or to researchers under licenses, the administrative agency may require that the projects undergo a review and approval process, that security controls meet specified guidelines, and that the statistical agency permits random inspections of the facility to ensure that there are no security violations.

Researcher access to linked data has been the subject of considerable discussion over the years. The Committee on National Statistics has held multiple workshops on options for providing researchers with safe and convenient access to data derived in whole or in part from administrative records. For example, the 2008 Workshop on Protecting

[14]The National Center for Health Statistics will not release a public use file with linked survey and administrative data without the prior approval of the administrative agency.

Student Records and Facilitating Education Research looked at how to reconcile privacy protections with current educational needs and goals. One of the workshop findings was a call that researchers assume responsibility for protecting confidentiality in order to avoid harm to both research and agency missions (National Research Council, 2009).

7 Research Focused on Privacy and Confidentiality Risks that Impact the Statistical Use of Administrative Records

To understand and overcome the various privacy and confidentiality risks that deter greater statistical use of administrative records in national statistics, agencies have funded research that sheds light on public attitudes about privacy and about risks of disclosure in published data. Privacy-related research has primarily been a focus in the U.S. and much of it has been funded by the Census Bureau. Disclosure research has drawn international support both in the national statistical offices and in academia. More recently, research has begun to focus on how confidentiality and privacy protection measures for tabular data and public use microdata impact statistical research.[15] The following summarizes what we know from past research and what I believe we can learn from additional research.

7.1 Privacy Attitudes and Informed Consent

What Research Tells Us

One of the first major quantitative research studies on privacy attitudes was undertaken by the Committee on National Statistics in its 1979 report Privacy and Confidentiality as Factors in Survey Response (National Research Council, 1979). The purpose of this study was to determine whether individuals, based on their concerns about individual privacy and confidentiality, might choose not to respond to questions posed in household surveys as well as the upcoming 1980 census and what might be done to assuage those concerns. Subsequent privacy studies sponsored by the Census Bureau also focused on better understanding and improving participation in household surveys and the decennial censuses.

In the 1990s, the Census Bureau undertook research focused on the use of administrative records to supplement the count in non-responding households, fill in missing information for responding households, and assist in coverage measurement. Recognizing the potential privacy concerns, the plan also called for continuing efforts to study public attitudes about obtaining and using other agencies' data and exploring the best ways to inform the public about these uses (U.S. Census Bureau, 1996). The research associated with this effort consisted of several public opinion surveys focused on administrative records use, focus group discussions, cognitive interviews, and a facilitated discussion with privacy experts. The research was designed to address four key issues: 1) what new notices should be provided to census respondents to inform them about

[15] See, for example, Fienberg (1994), Duncan et al. (2003), and Kennickell and Lane (2006).

the use of administrative records and how that would affect their response; 2) does the public currently believe the confidentiality promise, and how will obtaining and using other agencies' data affect that belief; 3) if Social Security Numbers were requested of census respondents, would it be perceived as a privacy violation; and 4) would combining records of individuals at a national level be perceived as a privacy threat despite reassurances to the contrary. Based on the research findings, the Census Bureau concluded that the public: 1) believes that the Census Bureau already shares its data with others; 2) believes that federal computers are all connected; 3) feels that individuals have lost control over how their personal information is used; 4) thinks there is no law prohibiting the Census Bureau from sharing its information; and 5) worries that the federal government cannot be trusted and does not care about individuals (Gates and Bolton, 1998).

In 1997, plans to expand administrative records use in Census 2000 were postponed due to inadequate time to complete the necessary research and growing concerns from members of the Census Advisory Committees about possible impacts on census participation. In anticipation of renewed efforts in 2010, the Census 2000 Testing, Experimentation, and Evaluation Program included various studies to better understand how privacy concerns impact the mail back of census forms, as well as how increased data sharing among agencies as a result of greater administrative records use might increase the public's concerns about privacy.[16] The studies, both quantitative and qualitative, that comprised this research included the Surveys of Privacy Attitudes; the Social Security Number, Privacy Attitudes, and Notification Experiment (SPAN); a survey of partners participating on outreach for the census; the report of focus groups held in Puerto Rico on why households do not mail back their questionnaire; an ethnographic investigation focused on privacy; and an Internet survey of privacy attitudes conducted during Census 2000. See Larwood and Trentham (2004) and Singer (2003). For a comprehensive literature review of this and other privacy research impacting federal statistics see Mayer (2002).

The Census 2000 privacy research provides some helpful insights into how the public views the sharing of data within the government and with the Census Bureau specifically. A key finding suggests that even as more people become knowledgeable about the law protecting their census data, they continue to believe that government does not keep personal information confidential. This is especially true among members of minority groups. This suggests that trust in the government and in the Census Bureau to protect information plays a significant role in attitudes about data sharing.[17] The research further shows an apparent trend toward increased concern over data sharing during the period of 1995–2000 (Singer et al., 2001).

This research also provides insights into the impact of notification on acceptance of

[16]It should be noted that the research highlighted potential uses of administrative records that would substitute in part for questions obtained on the Census Long Form questionnaire. With the full implementation of the American Community Survey in 2005, Census 2000 is the last census to include the long form questions.

[17]Singer et al. (1997) have shown that opinions about data sharing are closely related to trust in government, confidence in the promise of confidentiality, and sense of political effectiveness.

data sharing, how negative publicity affects privacy concerns, and how attitudes translate to behaviors. The Notification Experiment, which was associated with a request for SSN, showed that "notification of record linkage has a small but significant negative effect on the response rate[18] but a positive effect on responding to the SSN item" (Singer, 2003, p. 26). This result is consistent with ethnographic research by Gerber which shows respondents attach legitimacy to questions based on their understanding of the nature and purpose of the survey, including why the data are needed and how they will be used (Gerber, 2003).

The 2000 Privacy Research included an analysis of how negative publicity affects privacy concerns. Singer and her colleagues "found that respondents who reported exposure to negative as well as positive publicity about the census had significantly higher scores on the privacy index[19] and were significantly more likely to regard the census as an invasion of privacy, and less likely to be willing to provide their Social Security Number, than those reporting no exposure to publicity about the census" (Singer et al., 2001, p. 11).

One aspect of the research led to conclusions about how attitudes impact response. This has been a subject of some interest at the Census Bureau. Although the agency is aware that the public is concerned about privacy and these concerns have been growing over time, it is not clear that response to surveys is being affected proportionately. Most prior privacy research has not been designed to determine how attitudes carry over to behavior. The SPAN experiment, in conjunction with the Survey of Privacy Attitudes did, however, provide an opportunity to indirectly assess whether expressed unwillingness to provide one's SSN in the census context results in a failure to do so. The study suggests that "approximately one half of those saying they would be unwilling to provide their SSN to the Census Bureau would actually fail to provide an accurate number if they were directly asked to do so" (Singer, 2003, p. 21). Although this comparison of attitudes and behavior was based on two samples of different individuals, the context (request for SSN for the purpose of obtaining government records) was virtually identical and both studies were conducted around the same time frame. Thus, at least in this context, there appears to be a substantial relationship between an attitude (expressed willingness to provide one's SSN) and an action (compliance with a request for one's SSN).

Although not part of the formal research on privacy attitudes, further evidence of the public's reaction to administrative records use can be found in the reactions from stakeholder groups. At a meeting of the Census Bureau's Advisory Committee on Racial and Ethnic Populations in April 2006, strong concerns were voiced by some members about the Census Bureau's research of administrative records to develop improved imputation methods for the 2010 decennial census. The discussion centered on perceived privacy concerns about record linkages by racial and ethnic populations who were growing more and more distrustful of government. As a result, the Hispanic Committee recommended the Census Bureau not use administrative records for imputing missing

[18]Response rate refers to completing the census questionnaire.
[19]The Privacy Index consists of answers to five general privacy questions.

data (U.S. Census Bureau, 2006).

From the household survey perspective, a recent field study designed to test various consent-to-link questions as part of the Survey of Health Insurance and Program Participation (SHIPP) offers some additional insights. The 2010 Computer Assisted Telephone Interview (CATI) study was designed to test three different motivational rationales for gaining consent for data sharing for the purpose of record linkage: 1) improved accuracy, 2) reduced costs, and 3) reduced burden on respondents. Contrary to expectations, the research showed that none of these rationales was better than the other in gaining cooperation for record linkage. The research also affirmed prior findings that older respondents and those with less than a high school education are less likely to agree to data sharing or to provide information needed to link data. This implies that targeted efforts may be needed for these demographic groups or statistical adjustments will be needed to account for the resulting bias. Finally, and most surprisingly, whereas a similar 2004 study had reported 63% expressing no objection to data sharing, that number increased to 84% in 2010 (Pascale, 2011). This finding is likely attributable to differences in question wording and/or changes in attitudes over the period between surveys.[20] Additional research is needed to determine the extent to which these are contributing. If its mainly the former, cognitive research would be helpful in determining whether such general notices are meaningful given the potential implications for increased participation.

Proposed Research

Despite the considerable knowledge gained by past research, statistical agencies still may not feel comfortable that they fully understand how the public might react to their efforts to expand access to and use of individuals' personal information. This unease arises from the fact that privacy opinions shift over time and are influenced by people and events over which the agency has little control. Agencies may think that they have considered everything from a legal, policy, and ethical perspective, but the public may still not be satisfied.

Since this issue impacts all federal statistical agencies that collect or obtain information on individuals, a statistical system-wide approach is needed. To assure agencies that they have made the right decision to commit to administrative records, privacy research should be current and should be able to adapt to unexpected events. Most importantly, privacy research should be input to a program of outreach and education that promotes awareness and fosters discussion. A coordinated research effort should consider the following components:

- Conduct ongoing surveys to monitor changes in public opinion pertaining to privacy and confidentiality. Assuming a consistent set of questions is replicated over

[20]The 2010 study asked to produce additional statistical data "by combining your survey responses with data from other government agencies" whereas the 2004 study asked for "permission to obtain the information that you have given to other government agencies on topics such as Social Security and Medicare benefits."

time, such surveys could alert agencies to reduced levels of trust in government, increased concerns about data sharing, and false impressions about the confidentiality of personal information.

- Cognitively test and disseminate messages to broadly convey concepts of confidentiality, statistical use, and functional separation. These are difficult concepts to communicate and understand and are at the heart of any debate over whether administrative records should be shared for statistical purposes.

- Conduct studies on how trust is influenced by those in leadership positions and how negative messages can be counteracted. Despite legal protections, sound research protocols, and all the proper policies and procedures, our historical failures (such as the reports of the Census Bureau's involvement in the government internment of Japanese Americans in WWII) or the failures of other agencies (such as the loss of millions of personal records on a VA laptop in 2006) have and will continue be used to question our motives (Minkel, 2007) (Vijayan, 2007).

- Prepare a public outreach effort beyond the statistical profession to include privacy advocates and advocates for minority populations to discuss the conditions under which administrative data are being used for statistical research. It is clearly to the agency's advantage to discover "show stoppers" before plans are set in stone.

- Design studies to cognitively test informed consent notices related to data linkage. Research has shown that people respond more favorably to general notices that emphasize benefits to them. In crafting new notices it is important to determine whether they are conveying meaningful information to respondents without being unduly alarming.

- Conduct focus groups and cognitive interviews to assess the public's current knowledge of the statistical use of administrative records and the factors that make the public agreeable to such uses. The results should be used to craft messages to include on survey brochures and agency websites. The results will also be helpful in convincing stakeholders that the agency is being proactive in gaining public support.

7.2 Confidentiality and Security

What Research Tells Us

Confidentiality is maintained through the application of security controls and disclosure avoidance techniques. Confidentiality is put at risk when security procedures to limit access and use of personally identifiable information are inadequate or not followed. Confidentiality can also be breached when intruders are able to defeat the disclosure protections applied to published data.

Disclosure avoidance research for integrated data involves disguising the administrative data in such a way that anyone holding the source data cannot, with certainty,

match data items to identify individual persons. The residual risk of re-identification in published microdata has been the subject of research by Lambert, Sweeney, Truta, and Winker, among others. The goal is to limit the risk of disclosure by disguising characteristics that are unique to one person in the entire population and still preserve the analytic validity of the original data. As described previously, imputation methods are proving to be effective in disguising administrative data and preserving data utility. What is unclear, however, is whether future research needs can be satisfied by these techniques and whether synthetic data can withstand future attacks as technology continues to provide intruders with more sophisticated matching tools.

From the perspective of security, agencies are becoming increasingly aware of the risks associated with transferring, storing, and retrieving confidential information. Over the past five years, data breaches have been reported by many government agencies as a result of new federal reporting requirements or through requests under the Freedom of Information Act. Generally, such losses occur when unencrypted data are transmitted through the internet or are present on lost or stolen laptops or flash drives. The federal government has issued requirements for agencies with regard to storing and transmitting personally identifiable information (PII) residing in electronic form (Office of Management and Budget (OMB), 2007). Requirements include encrypting PII on mobile computers/devices, transmitting PII only with two-factor authentication; using password controls and timeouts for remote access; logging all computer readable data extracts; and ensuring accountability of employees. Federal statistical agencies are subject to these requirements.

When data breaches occur, agencies are required to report them to the U.S. Computer Emergency Readiness Team (US-CERT). This process is designed to protect the U.S. cyber infrastructure by identifying willful attacks. If PII is breached, the OMB guidance provides requirements for determining if individuals should be notified and whether free credit monitoring[21] is warranted. This assessment is based on the likely risk of harm to the individual when considering: 1) the nature of the data elements breached; 2) number of individuals affected; 3) likelihood the information is accessible and usable; 4) likelihood the breach may lead to harm; and 5) the ability of the agency to mitigate the risk of harm.

A recent, first of its kind, assessment by the National Center for Education Statistics took an interesting look at the effect on survey participation of data breaches in the Early Childhood Longitudinal Study. For this study, the NCES not only provided notification and free credit monitoring, it also offered the opportunity to withdraw participation—both retrospectively and prospectively. Seastrom and her colleagues found that providing respondents who suffered a data breach the option to withdraw previous responses and/or decline future participation results in a differential loss that can bias results (Seastrom et al., 2008). What is yet to be studied is the degree to which harm to the individual is mitigated by notification, credit monitoring, or the withdrawal of participation.

[21]This service is provided to individuals who suspect their identity has been or may be stolen by others to commit fraud. Companies providing this service monitor the individuals credit and alert them to changes so improper activity can be identified and stopped.

Data breaches involving administrative data used for statistical research would most often occur when employees process and analyze the data or the data are transferred to research data centers, placed on remote servers, or provided to licensees. There is no evidence that such breaches are occurring. Should administrative data be breached, agencies would be required to report to US-CERT and assess whether notification is warranted. Most likely, they would also be required to report the breach to the administrative agency under the terms of the agreement.

Proposed Research

Disclosure avoidance research in the U.S. and internationally has benefited from the involvement of renowned statisticians and computer scientists both in government and academia. In addition, federal agencies have been open to exploring innovative methods of providing researchers with access to unpublished data in secure settings. Nevertheless, access and use of administrative records would benefit from ongoing, extended, and coordinated research on aspects of disclosure avoidance, security, and data access, as well as a review of current legal confidentiality requirements. Specifically, federal statistical agencies should consider jointly undertaking research to help them better understand:

- The pool of potential intruders. Currently, data are not published if the disclosure review boards determine that the administrative agency can use its source data to find someone on a public use file containing its data. Treating administrative agencies' employees and contractors as possible intruders results in greatly reducing the data available to everyone. Currently, the law provides no discretion here but perhaps the law could provide disincentives for others trying to identify individuals on a public use file.[22] An assessment should be done to determine if this is an option worth pursuing.

- The limitations and potential of synthetic data for various applications. Research, such as that promoted by Rubin, Abowd, and Reiter, among others, should continue to assess the disclosure protection and analytic validity of synthetic data. Applications for synthetic data, such as those currently supporting the Census Bureau's programs that are available through the Cornell Virtual RDC (see `http://www.vrdc.cornell.edu/news/`), should be promoted across all federal agencies that are seeking access mechanisms for linked data.

- Effectiveness of security controls on limiting administrative data breaches. Currently there is no public record of PII breaches since US-CERT incidents are not published. Public reporting of data breaches in such a way that national cyber security is not compromised would provide evidence of whether security controls are working and would facilitate transparency.

[22]The law that governs the National Center for Education Statistics provides for legal penalties to "any person who uses any data provided by the Center, in conjunction with any other information or technique, to identify any individual student, teacher, administrator, or other individual and who knowingly discloses, publishes, or uses such data for a purpose other than a statistical purpose..." This law is unique within the Federal Statistical System.

- The potential and realized impacts on individuals of disclosures/breaches and notification. PII breaches/disclosures are not all equal and OMB guidelines recognize this by requiring an assessment of risk based on likelihood and magnitude of harm to the individual. This assessment is mainly subjective. Agencies should share information on breaches/disclosures and any known impacts on individuals.[23] Agencies should also consider following up with individuals affected by breaches to assess if and how individuals have been harmed.

- The costs and benefits of various access mechanisms from the perspective of individual privacy and research utility. Despite the variety of mechanisms available, some researchers find that the choices available to meet their unique requirements are not workable and agencies are not willing to accept the additional risk created from options that, to the researcher, are workable. A risk assessment should look at this issue from both perspectives.

- The impacts of disclosure protections on data utility. Coordinated research should focus on determining the degree to which various disclosure protection techniques are limiting the usefulness of data for policy analysis. Research could provide insight into the best data/access options for different types of users.

7.3 Proposed Research on Missed Opportunities

In addition to privacy and confidentiality research, there is a pressing need for research on the degree to which concerns for confidentiality and privacy have limited the statistical use of administrative records. Missed opportunities may result when agreements cannot be reached to obtain the records from the administrative agency as well as when the statistical agency does not effectively use the data it does obtain. There are also lost opportunities from not allowing researchers to access linked survey and administrative datasets. An analysis of such missed opportunities would be useful to inform debates over the tradeoffs between the public good and individual privacy and whether the proper attention is being focused on both.

8 An International Perspective

There has been a great deal of international collaboration over the years on research in support of the statistical use of administrative records. From the perspective of privacy and confidentiality, research has been the focus of seminars organized by Statistics Canada, Eurostat, the Conference of European Statisticians, UNESCO Chair in Data Privacy, and the United Nations Economic Commission for Europe. These collaborative efforts have greatly enhanced our knowledge of data confidentiality, data access, and privacy. The research proposals listed in Section 7 above would benefit from continued international participation.

From a policy perspective, it is helpful to examine the experiences of other coun-

[23] This should be done in a way to ensure individuals' privacy is protected.

tries in terms of the evolution of laws permitting/requiring sharing and the public support for such sharing. In Canada, for example, the 1985 Statistics Act mandates that Canadian departments, municipal offices, businesses, or organizations provide Statistics Canada with documents or records for the purpose of completing or correcting information collected under the Act. Since Canada has a centralized statistical system, this facilitates access to administrative records for all statistical programs. (See `http://www.infosource.gc.ca/emp/emp06-eng.asp` for a description of files currently accessed by Statistics Canada.) To protect individual privacy, Canada has established the Office of the Privacy Commissioner that administers the Privacy Act and handles complaints from individuals about the handling of their personal information by government institutions. Statistics Canada is keenly aware of the public's concern for privacy and ensures that record linkage activities meet its policy guidelines (`http://www.statcan.gc.ca/record-enregistrement/policy4-1-politique4-1-eng.htm`).

Like the U.S., the UK has no single law permitting data sharing and, instead, depends upon various statutory provisions and common law rules. "Despite, or more likely, because of the broad range of provisions, the legal basis for setting out whether and how information can be shared in every situation is far from clear-cut" (Thomas and Walport, 2008, p. 22). Where the legal basis is clear, "barriers" to sharing "are most often cultural or institutional—an aversion to risk, a lack of funds or proper IT, poor legal advice, an unwillingness to put the required safeguards in place or to seek people's consent." The public concerns over data security and government intrusiveness seem to also parallel those in the U.S. According to Dibben et al. (2009), p. 5, "[Recent] events and public discourse have generated an environment that is not especially conducive to arguments for the extended use of administrative data for research purposes. On the whole they have tended to lead to an environment where the risks associated with the extension of these types of uses are very salient but the potential benefits are not."

Despite the cultural differences, we can also learn a great deal from the experiences of the Nordic countries. In 2007, the United Nations Economic Commission on Europe (UNECE) issued a report on "Register-based Statistics in Nordic Countries" to highlight best practices in the use of administrative records for population and social statistics (United Nations, 2007). The report notes that Nordic Countries (Finland, Norway, Denmark, and Sweden) have a long history of successfully accessing and using administrative records. In fact, in 1981 Denmark was the first country to move to a totally register-based population census.

The legal basis for administrative records use in the Nordic countries is the national statistics act that grants a right of access to the National Statistics Institutes (NSIs) and stipulates obligations for data protection. Some countries obligate the NSIs to first examine available administrative data before attempting to collect information directly from individuals. National legislation on processing of personal data as well as EU regulations on community statistics support these uses in the Nordic countries. Public approval of these uses has generally been positive. In Finland and Denmark, there has been little controversy. In Norway, there was a public debate over these uses in the 1970s that has seemed to lessen over time. In Sweden, the discussion has been ongoing since 1970. The UN report notes that in all Nordic countries, a key principle in the statistical

use of administrative records has been an open discussion and debate explaining the rationale and benefits of register use. The authors note that it is important to be vigilant so as not to lose the public's confidence.

The experience of the Nordic countries is clearly having an impact as more countries are moving toward register-based population censuses. In fact, in 2011 Germany, which has not conducted a population census since 1986, is using administrative records together with its population register to obtain basic census information. It will supplement this with information not available in registers by surveying 10% of the population. In 2011, India is integrating the preparation of its National Population Register with its 2011 census enumeration with the goal of providing real-time population data. Laws and public opinion will likely determine the speed at which other countries follow suit.

Also important from an international perspective, in 2009 the UNECE published "Principles and Guidelines on Confidentiality Aspects of Data Integration Undertaken for Statistical or Related Research Purposes" to address the fact that different countries have different degrees of experience in integrating administrative data in their statistical programs. Because data integration is relatively new in many countries and there is no supporting legal and policy framework, the UNECE proposes a common framework to guide such uses in these countries. The principles highlight the need to balance the public benefits from data integration with the public's concerns for privacy as well as potential risks to the other statistical operations of the organization. Focusing on uses and protections once the data are acquired, these principles stress the importance of controls and limits on uses, the rights of respondents regarding the use of their personal information, openness and transparency, and the protection of confidentiality (United Nations, 2009).

9 Moving Forward

In this paper I have attempted to demonstrate the complex environment in which administrative records are accessed and used for U.S. national statistics. Additional research on privacy and confidentiality will help address some of the uncertainty that surrounds administrative records sharing but that alone will not suffice.

In the context of an administrative records census, Scheuren (1999) recommended the Census Bureau seek legislation to ensure the cooperation of administrative agencies. He noted that seeking such legislation would provide an opportunity for needed public debate. He went on to suggest that the Census Bureau should consider establishing an advisory body to represent the public's interest in linking individual records in this context.

Like Scheuren, I fully support the need for new legislation and more public participation. I would not, however, limit this to the Census Bureau or the decennial census. I believe the law should facilitate the sharing of administrative records with all statistical agencies where individual privacy is protected through law, policy, and procedures. I would further propose the establishment of an official arbiter who can decide if the

proper conditions for sharing and using administrative records are being met and who can encourage cooperation. Finally, I would encourage openness and public debate with regard to the benefits and risks. Accordingly, the following actions are recommended.

Action 1: Revise the Privacy Act

The Privacy Act needs to be revised to recognize that the routine use exemption that permits agencies to share information with the Census Bureau without individual consent is also applicable to those agencies covered by the confidentiality provisions of CIPSEA. Each of these statistical agencies now has the legal requirement to ensure confidentiality, even to the extent of refusing to comply with compulsory legal process such as subpoena or court order and to limit use of this information. These were the conditions that lawmakers considered when granting the Census Bureau exemption.

The Privacy Act also needs to be updated to reflect a key recommendation of the Privacy Protection Study Commission and the enactment of CIPSEA. The functional separation principle, as outlined by the Privacy Protection Study Commission in its 1977 Report, noted that personal information collected for an administrative purpose can be shared for a statistical purpose, but in order to ensure that no personal information obtained or collected for a statistical purpose can be used for an administrative purpose, organizational barriers must be in place to separate administrative and statistical functions in agencies. Functional separation is essential to promoting the statistical use of administrative records.

Action 2: Expanded Role of OMB

The role of the Statistical and Science Policy (SSP) Office in the Office of Information and Regulatory Affairs (OIRA) at the Office of Management and Budget (OMB) should be strengthened to promote agency cooperation in the statistical use of administrative records. The SSP has responsibility for reviewing and approving statistical data collections on the basis of adherence to sound statistical practice and compliance with legal requirements to limit reporting burden. The office plays a coordinating role in ensuring that statistical agencies effectively use administrative data in surveys and censuses, and through OIRA, can influence participation by administrative agencies. However, in my experience, negotiations to obtain administrative data are rarely mediated with the active participation of SSP. This is likely due to the small size of the office and the complex nature of these negotiations.

In December 2010, OMB/OIRA issued a Memorandum to all agency heads emphasizing the benefits of data sharing and the importance of protecting privacy.[24] The Memorandum cites the value of data sets held by program, administrative, and regulatory offices and agencies in support of the statistics initiative. OMB offers assistance to agencies and indicates that OMB may ask specific agencies to evaluate options to share data. While signaling the Administration's attention to this issue generally, it is too

[24] http://www.whitehouse.gov/sites/default/files/omb/memoranda/2011/m11-02.pdf

early to determine if this will lead to meaningful changes. The proposed amendments to the Privacy Act (Action 1 above) should specifically recognize a role of the Statistical and Science Policy Office in promoting the statistical use of administrative records and as serving as final arbiter in resolving disputes between statistical and administrative agencies over data access and use. This role would be accomplished primarily through the issuance of memoranda laying out general principles for sharing data for statistical uses while protecting privacy. Arbitration should be a last resort.

Action 3: Model Agreements

Privacy Act amendments will take time and any efforts to amend the law should not delay the ongoing efforts of the FCSM's Subcommittee on the Statistical Uses of Administrative Records to assess commonalities and differences in agreements between/among statistical and administrative agencies. This work is intended to support the development and dissemination by OMB of model agreements for use by statistical and administrative agencies. I recommend that OMB issue a memorandum to agencies disseminating these model agreements and requesting that agencies use them to promote the statistical use of administrative records.

Action 4: Data Stewardship Programs in Statistical Offices

A coordinated data stewardship effort like what is currently in place in a few agencies should be put in place across the federal statistical agencies. The Census Bureau, for example, committed in 2001 to data stewardship through the establishment of a senior-level committee and the necessary support staff to develop and implement wide-ranging policies focused on privacy, confidentiality, and data access and use. The program encompasses all personal and business information collected or acquired by the Census Bureau. This commitment recognizes the importance of protecting and controlling the use of valuable administrative records. A statistical system-wide approach would bolster the government's claim that administrative records can be safely used for statistical programs.

Action 5: Public Debate

A more public conversation needs to take place with privacy advocates, representatives for minority groups, and the media about the current uses of administrative records and the conditions for such use. Small targeted efforts were led by the Census Bureau in workshops conducted in 1997 (Gates and Bolton, 1998) and again in 2005 (Kincannon et al., 2005). Also, the issues have been addressed in various public meetings of the Census Bureau's advisory committees. These discussions identified some important issues and concerns but lacked the size and scope needed to determine what conditions would make sharing data for statistical purposes workable or unworkable. This conversation needs to be led by OMB on behalf of all federal statistical agencies since it is really a government-wide issue. Significant issues that surface should be published for public

comment and any conclusions factored into new Administration and/or Congressional actions.

As a final note, at the time I prepared this paper, the Obama Administration's proposed fiscal year (FY) 2011 budget initiative to stimulate new uses of administrative data in the production of U.S. statistics was not funded by Congress. However, these projects started modestly in FY 2010 and are continuing, albeit on a smaller scale than if the agencies had gotten the requested resources. The projects remain a priority for the Interagency Council on Statistical Policy and it hopes to expand to additional pilots over time. This initiative includes three initial pilot projects. One project, led by the Census Bureau, is designed to replicate the 2010 census coverage. Another project would link National Center for Health Statistics (NCHS) health surveys to administrative data at the Census Bureau then return them to NCHS. Both projects depend upon an infrastructure at the Census Bureau. The third project, led by the Economic Research Service in the Department of Agriculture, will provide data on nutrition and food assistance by acquiring, linking, and studying the quality of state administrative files. This commitment to administrative records is significant, but achieving the full potential of administrative data will require leadership to address the policy and legal issues discussed here. I am confident that those involved are committed to this goal.

Acknowledgments

I would like to thank Shelly Martinez, John Eltinge, and Bill Iwig of the U.S. Federal Committee on Statistical Methodology's (FCSM) Subcommittee on the Statistical Uses of Administrative Records who encouraged me to outline a research agenda focused on understanding and overcoming barriers to accessing and using administrative records in statistical programs. They, along with Jeffrey Rodamar, John Fanning, Kathleen Styles, Dan Weinberg, Cynthia Nickerson, and Patricia Melvin provided comments on various versions of this paper. I greatly appreciate their support and helpful suggestions but take full responsibility for any errors or omissions as well as the opinions expressed.

Appendix 1

Selected Laws Supporting the Acquisition and Use of Administrative Records in U.S. Federal Statistics

Laws permitting limited sharing of administrative data for statistical purposes without consent.

The Privacy Act of 1974 provides that agencies may establish a "routine use" in their System of Records Notice (SORN) that would allow the disclosure of personally identifiable information for research and statistics.[25] Agencies specify the categories of users and purposes for the uses in the SORN that is published in the Federal Register for public comment. An example of such a routine use provision is the United States Renal Data System (see `http://oma.od.nih.gov/ms/privacy/pa-files/0160.htm`). Although helpful in fostering statistical uses of administrative data, this approach depends upon the conditions for disclosure in the administrative agency's statute as well as the agency's willingness to recognize and support the research and statistical uses in advance of creating the data system.

The Privacy Act also allows for the disclosure, without prior written consent, of a record "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." Since such records are not identifiable and most uses require exact matching with survey and census data, this does not generally facilitate the sharing of administrative records for statistical purposes. However, the Privacy Act does explicitly permit the disclosure of personal information "to the Census Bureau for the purpose of planning or carrying out a census or survey or related activity pursuant to the provisions of Title 13." This special provision recognizes that the Census Bureau's statute limits the uses which may be made of the records and makes them immune from legal process. With the enactment of the Confidential Information Protection and Statistical Efficiency Act (CIPSEA) in 2002, it could be argued that the same Privacy Act exemption granted the Census Bureau should be made available to all statistical agencies covered under CIPSEA and its implementing regulations.

The Computer Matching and Privacy Protection Act of 1988 amended Title 5 of the United States Code (U.S.C.) to specifically address computer matching agreements between federal agencies. This law requires that agencies proposing to initiate or amend a matching program must provide prior notice to Congress and the OMB so that the privacy impacts can be evaluated. Further, each agency is required to establish a Data Integrity Board to review and approve proposed matching agreements. Similar to the Privacy Act's approach in recognizing the low privacy risk from systems that are intended for statistical uses only, this law exempts matches "performed to produce aggregate statistical data without any personal identifiers (and) matches performed to support any research or statistical project, the specific data of which may not be used

[25] A routine use is defined as the use of a record for a purpose which is compatible with the purpose for which it was collected.

to make decisions concerning the rights, benefits, or privileges of specific individuals."

Agency-specific laws permit sharing by agencies for statistical purposes as long as confidentiality is maintained by the receiving party. Frequently, these laws limit the types of statistical uses and/or users of the records. For instance, Food Stamp Records under Title 7 U.S.C. sec. 2026 b(1)(A) can be shared with other (unnamed) agencies for statistical uses provided that the research uses "improve the administration and effectiveness" of the Food Stamp Program. On the other hand, education records under Title 20 U.S.C. sec. 1232 g (b) may be shared for statistical research only with the federal and state education agencies mentioned in the Family Educational Rights and Privacy Act of 1996. Where access is authorized by law and the receiving statistical agency has the authority to designate agents to work on behalf of the agency, these agents may also be authorized access to the administrative data under the same conditions as agency employees. Where the law does not permit agents to access the identifiable records, or where the arrangement is not agreeable to the researcher, obtaining written consent for such access is sometimes an option.

An important example of an agency-specific law that authorizes limited sharing and use of administrative records is Title 26 U.S.C. (the Tax Code). The strict limits on access and use of tax return information originated with the Tax Reform Act of 1976 that came on the heels of privacy abuses surfacing in the aftermath of the Watergate Scandal. In recognition of the longstanding uses by the U.S. Census Bureau, Congress provided in Section 6103(j) of Title 26 that the Secretary of Treasury shall provide, upon request in writing by the Secretary of Commerce, "such returns, or return information reflected thereon, to officers and employees of the Bureau of the Census, as the Secretary may prescribe by regulation for the purpose of, but only to the extent necessary in, the structuring of censuses and national economic accounts and conducting related statistical activities authorized by law."[26] In practice, the Internal Revenue Service, on behalf of the Secretary of Treasury, has issued regulations that have restricted Census Bureau access to specific tax return items for specific uses. Title 26 also requires the Census Bureau to provide safeguards determined by the Secretary of Treasury to be necessary or appropriate to protect the confidentiality of the returns or return information. Regular Safeguard Reviews, including on-site inspections, are conducted by the IRS to ensure that security and use limitation requirements are met.

Laws requiring confidentiality and limiting uses when data are collected or acquired for statistical purposes

The Privacy Act of 1974 prohibits federal agencies from disclosing personal information they obtain from individuals unless the individual provides written permission for such disclosure or the disclosure meets one of twelve categories of permitted disclosures stipulated in the act.[27] Intended disclosures must be described to the individual in a written notice provided at the time of collection and must be consistent with the published System of Records Notice. Notices must describe the categories of users and purpose of each use. Using the information in ways not specified or sharing with unau-

[26]The Bureau of Economic Analysis was also provided limited access to tax data at this time.
[27]Businesses are not covered by the Privacy Act.

thorized persons can lead to criminal penalties including individual fines up to $5000.

Title 13 U.S.C. authorizes the activities of the Census Bureau. Section 9 of Title 13 stipulates that the Census Bureau may not: 1) use the information furnished for any purpose other than the statistical purposes for which it is supplied; 2) make any publication whereby the data furnished by any particular establishment or individual can be identified; or 3) permit anyone other than the sworn officers and employees of the Department or bureau or agency thereof to examine the individual reports.[28] The courts have also determined that Title 13 exempts the individual records from legal process so they are not available, even under subpoena. Any administrative records obtained by the Census Bureau under this statute are also protected by this confidentiality requirement. Violations of the confidentiality requirement are subject to fines of up to $250,000 and up to five years in prison or both. Section 23c of Title 13 permits the Census Bureau to designate agents to perform work on a temporary basis and holds these "special sworn status" persons to the same legal obligations as regular employees to protect any personal information they may see on the basis of their appointment.

Title 20, U.S.C, Section 9573, provides for confidentiality in the collection, maintenance, use, and dissemination of personal information by the National Center for Health Statistics. This law provides confidentiality assurances similar to those in Title 13. Unlike the Census law, it also provides for legal penalties to any individual who uses any data provided by the agency to identify an individual student, teacher, or administrator and knowingly discloses, publishes, and uses this information for a non-statistical purpose. This provides additional protections should disclosure avoidance or security measures prove to be inadequate.

The Confidential Information Protection and Statistical Efficiency Act (CIPSEA) of 2002 provides uniform confidentiality protections for the 70+ federal agencies collecting information for statistical purposes under a pledge of confidentiality. Prior to CIPSEA, agencies used a variety of authorities to protect this information—some more ironclad than others. In addition to the legal protection on par with Title 13, CIPSEA permits statistical agencies to designate agents to use confidential information. Such agents have the same legal requirement to protect the information as do agency employees—similar to the Census Bureau's special sworn status authority in Title 13. In June 2007, the OMB issued implementation guidance for CIPSEA spelling out the requirements for agencies collecting or acquiring information protected under CIPSEA; minimum standards for safeguarding confidential information; requirements when designating agents to access and use confidential information; and requirements when acquiring information that may be used for non-statistical purposes.

The Freedom of Information Act (FOIA) of 1996 requires agencies to release, upon request, government information to the public. The FOIA exempts from release certain categories of information. Exemption B-3 exempts from release "information specifically exempted by statute provided that such statute requires that the matters be withheld from the public in such a manner as to leave no discretion on the issue, or establishes particular criteria for withholding or refers to particular types of matters to be with-

[28] Limited exceptions apply to Section 8, Section 16, and Chapter 10 activities.

held." Agencies can cite their own statute, the Privacy Act, or CIPSEA in withholding personal information requested under the FOIA. In addition, Exemption B-6 permits the government to withhold all information about individuals in "personnel and medical files and similar file" when the disclosure of such information "would constitute a clearly unwarranted invasion of personal privacy."

Laws that encourage use of administrative records for statistics

The Paperwork Reduction Act (PRA) of 1995 is supportive of statistical uses of administrative data as one way to reduce reporting burden. Title 44, Section 3506 includes instructions for agencies to, among other things, certify that each information collection is not unnecessarily duplicative of information otherwise reasonably accessible to the agency; make data available to statistical agencies and readily accessible to the public; and implement and enforce applicable policies, procedures, standards, and guidelines on privacy, confidentiality, security, disclosure, and sharing of information collected or maintained by or for the agency. Each information request must be approved by the Office of Management and Budget and display the OMB assigned number signifying that the collection meets the PRA requirements.

Title 13 of the United States Code explicitly acknowledges the importance of administrative records in the creation of federal statistics. Section 6 of Title 13 requires that the Census Bureau use administrative data from other agencies, state and local governments and other instrumentalities, and private organizations instead of conducting direct inquiries if such data meet the quality and timeliness standards of the Census Bureau.

Appendix 2

Selected Policy Pertaining to the Acquisition and Use of Administrative Records in U.S. Federal Statistics

1. Privacy Protection Study Commission's 1977 report *Personal Privacy in an Information Society.*

   This report first recognized the role of "functional separation" as an important determinant in allowing administrative data to be transferred and used for federal statistics. Functional separation is defined as "separating the use of information about an individual for a research or statistical purpose from its use in arriving at an administrative or other decision about that individual." (Privacy Protection Study Commission, 1977, p. 574) The commission recommended creating standards and guidelines for agency information practice to limit exposure and the risk that statistical information may be used for an administrative purpose. It went on to recommend the creation of legal protections to prevent information collected and maintained for statistical purposes from being used to take action against the individual. CIPSEA, in part, addresses the Commission's recommendation.

2. The Office of Federal Statistical Policy and Standards' 1978 report *A Framework for Planning U.S. Federal Statistics for the 1980s.*

   This report addressed the issues surrounding an individual's control over statistical and research uses of data in administrative records systems. The report dismissed the notion that individuals have a right to consent to interagency transfers for statistical uses of their administrative information and instead recommended that blanket notices should be used to inform of such uses. The report notes that the price of individual consents "is great, leading to biased data, increased public expenditure, and the failure or impossibility of some valuable statistical and research studies" (Office of Federal Statistical Policy and Standards, 1978, p. 259).

3. Committee on National Statistics' Panel on Confidentiality and Data Access 1993 report, *Private Lives and Public Policies.*

   One focus of this study highlighted the individual's ability to control information about themselves when provided to a government agency on a mandatory basis. The committee admitted this was "one of the more difficult questions [they] faced"(National Research Council, 1993, p. 71). The ethical problem centered on the balance between the data needs of society and the individual's control over information they are required to provide to government agencies. The committee did not make a specific recommendation to address the control issue but noted: "Whatever general principles may be developed for statistical and research uses of mandatory data sets, their application in specific instances will require the establishment of an orderly and fair process that takes into account the interests of data subjects, users, and custodians" (National Research Council, 1993, p. 73). In a related comment, the committee further noted that "in keeping with the objective of giving individuals control over their own information whenever societal

needs do not clearly take precedence, data subjects or data providers should be allowed to waive certain aspects of confidentiality protection that would usually be accorded to the information they provide" (National Research Council, 1993, p. 75).

4. The 1991 Federal Policy for the Protection of Human Subjects, known as the Common Rule.

   This Federal Regulation acknowledges that individuals have the right to consent to participation in federally funded research, including research involving administrative data. Such consent must be "informed" in that the participant must be provided eight basic elements and up to six additional elements appropriate to the research (Title 45, Part 46, Subtitle A, Section 46.116). These elements can be found at: `http://www.hhs.gov/ohrp/humansubjects/commonrule/index.html`. The Common Rule provides that, in cases where informed consent may adversely affect the research, an Institutional Review Board (IRB) may waive the requirement if the research involves no more than minimal[29] risk to the individual and the waiver will not adversely affect their rights and welfare. Concerns have been raised over the years about the application to social science research of procedures designed to protect subjects of clinical studies, including the unevenness with which IRBs apply requirements on such research. In a 2003 report, the Panel on Institutional Review Boards, Surveys, and Social Science Research of the Committee on National Statistics recommended guidelines for IRBs in making decisions for social science research related to obtaining consent, guaranteeing confidentiality, and using appropriate review procedures for minimal-risk research (National Research Council, 2003). To help IRBs assess minimal risk for social science research, the National Science Foundation has posted Frequently Asked Questions and Vignettes: Interpreting the Common Rule for the Protection of Human Subjects for Behavioral and Social Science Research to its website at: `http://www.nsf.gov/bfa/dias/policy/hsfaqs.jsp`. There have also been efforts to improve training for IRBs on addressing different types of research.

[29]The Common Rule refers to "minimal" risk as a condition in which the probability and magnitude of harm or discomfort anticipated in the research are not, in and of themselves, greater than those ordinarily encountered in daily life or during the performance of routine physical or psychological examinations or tests.

References

Abowd, J. and Lane, J. (2003). Synthetic data and confidentiality protection. LEHD Technical Paper TP2003-10, U.S. Census Bureau, Washington, D.C. http://lehd.did.census.gov/led/library/techpapers/tp-2003-10.pdf.

Abowd, J., Stephens, B., Vilhuber, L., Andersson, F., McKinney, K., Roemer, M., and Woodcock, S. (2005). The LEHD infrastructure file and the creation of the quarterly workforce indicators (Abstract). Technical Report #TP-2006-01, U.S. Census Bureau, Washington, D.C. http://lehd.did.census.gov/led/library/techpapers/tp-2006-01.pdf.

Cox, C., Berning, M., and Martinez, S. W. (2006). Data policy and legal issues in creating and managing integrated data sets. In *Proceedings of the 2006 Federal Committee on Statistical Methodology Policy Conference*, p. 2. Statistical Policy Office, Office of Management and Budget, Washington, D.C.

Dibben, C., Gowans, H., Elliot, M., Anttila, C., Boyle, P., Kaye, J., McLennan, D., Noble, M., Smith, G., and Wilkinson, K. (eds.) (2009). Encouraging the wider and more creative use of administrative data in the UK–The 'Administrative Data Liason Service'. Paper presented at 2009 Conference on New Techniques and Technologies for Statistics, Eurostat. http://epp.eurostat.ec.europa.eu/portal/page/portal/research_methodology/documents/POSTER_5A_ENCOURAGING_THE_WIDER_AND_MORE_CREATIVE_USE_OF.pdf.

Domingo-Ferrer, J. (2003). Disclosure risk assessment in statistical microdata protection via advanced record linkage. *Statistics and Computing*, 13(4). Springer.

Duncan, G., Keller-McNulty, S., and Stokes, L. (2004). Database security and confidentiality: Examining disclosure risk vs. data utility through the R -U confidentiality map. Technical Report #142, National Institute of Statistical Sciences, Research Triangle Park, NC. http://nisla05.niss.org/technicalreports/tr142.pdf.

Elliot, M. and Skinner, C. (2002). A measure of disclosure risk for microdata. *Journal of the Royal Statistical Society B*, 64(4): 855–867.

Federal Committee on Statistical Methodology (FCSM) (2005). Statistical policy working paper #22 (revised), report on statistical disclosure limitation methodology. Statistical Policy Office, Office of Management and Budget. http://www.fcsm.gov/working-papers/SPWP22_rev_ch5.pdf.

Fienberg, S. E. (1994). Conflicts between the needs for access to statistical information and demands for confidentiality. *Journal of Official Statistics*, 10: 115–132. http://www.jos.nu/Articles/abstract.asp?article=102115.

Gates, G. and Bolton, D. (1998). Privacy research involving expanded statistical use of administrative records. In *1998 Proceedings of the Section on Government Statistics and the Social Statistics Section,* American Statistical Association, Alexandria, VA. 203–208.

Gerber, E. (2003). Privacy schemas and data collection: An ethnographic account. U.S. Census Bureau, Washington, D.C. https://www.census.gov/pred/www/rpts/Privacy\%20Schemas\%20Final\%20Report.pdf

Government Accountability Office (GAO) (2001). Record linkage and privacy: Issues in creating new federal research and statistical information. U.S. Government Accountability Office, GAO-01-126SP. http://www.gao.gov/new.items/d01126sp.pdf.

Guarino, J. A., Hill, J. M., and Woltman, H. F. (2001). Analysis of the social security number notification component of the social security number, privacy attitudes, and notification experiment. U.S. Census Bureau, Washington, D.C. http://www.census.gov/pred/www/rpts/SPAN_Notification.pdf.

Kennickell, A. and Lane, J. (2006). Measuring the impact of data protection techniques on data utility: Evidence from the survey of consumer finances. In *Privacy in Statistical Databases,* vol. 4302 of *LNCS,* J. Domingo-Ferrer and L. Franconi, eds., 291–303.

Kincannon, C. L., Barabba, V., Martinez, S. W., Blumerman, L., Gates, G., and Alvey, W. (2005). Panel on privacy and data use in the new technological environment. In *2005 Proceedings of the Government Statistics Section,* American Statistical Association, Alexandria, VA. 1234–1241.

Larwood, L. and Trentham, S. (2004). Census 2000 testing, experimentation, and evaluation program synthesis report No. 19, TR-19, Results from the Social Security Number, Privacy Attitudes, and Notification Experiment in Census 2000. U.S. Census Bureau, Washington, D.C. http://www.census.gov/pred/www/rpts/TR19.pdf.

Mayer, T. (2002). Privacy and confidentiality research and the U.S. Census Bureau: Recommendations based on a review of the literature. Research Report Series (Survey Methodology #202-01). U.S. Census Bureau, Washington, D.C. http://www.census.gov/srd/papers/pdf/rsm2002-01.pdf.

Minkel, J. (2007). Confirmed: The U.S. Census Bureau gave up names of Japanese-Americans in WWII. *Scientific American*. http://www.scientificamerican.com/article.cfm?id=confirmed-the-us-census-b.

National Research Council (1979). Privacy and confidentiality as factors in survey response. Committee on National Statistics, National Research Council, National Academies Press, Washington, D.C.

— (1993). Private lives and public policies: Confidentiality and accessibility of government statistics. Panel on Confidentiality and Data Access, Committee on National Statistics, Division of Behavioral and Social Sciences and Education, National Academies Press, Washington, D.C.

— (2003). Protecting participants and facilitating social science and behavioral sciences research. Committee on National Statistics and the Board on Behavioral, Cognitive, and Sensory Sciences and Education, National Academies Press, Washington, D.C.

— (2009). Protecting student records and facilitating education research: A workshop summary. M. Hilton, Rapporteur, Committee on National Statistics and Center for Education, Division of Behavioral and Social Sciences and Education, Washington, D.C., National Academies Press.

Office of Federal Statistical Policy and Standards (1978). A framework for planning U.S. federal statistics for the 1980s. Office of Federal Statistical Policy and Standards, U.S. Department of Commerce, Washington, D.C.

Office of Management and Budget (OMB) (2007). Memorandum M07-16, Safeguarding against and responding to the breach of personally identifiable information. May 27, 2007. http://www.whitehouse.gov/sites/default/files/omb/memoranda/fy2007/m07-16.pdf.

Pascale, J. (2011). Requesting consent to link survey data to administrative records: Results from a split-ballot experiment in the survey of health insurance and program participation (SHIPP). Center for Survey Measurement, Research and Methodology Directorate (Survey Methodology #2011-03), U.S. Census Bureau, Washington, D.C. http://www.census.gov/srd/papers/pdf/ssm2011-03.pdf.

Potok, N. (2009). Creating useful integrated data sets to inform public policy. Doctoral Dissertation, Columbian College of Arts and Sciences, George Washington University.

Privacy Commissioner of Canada (2000). Annual report to parliament, 1999–2000, 64–70. http://www.priv.gc.ca/information/ar/02_04_08_e.cfm.

Privacy Protection Study Commission (1977). The relationship between citizen and govenment: The citizen as participant in research and statistical studies. Chapter 15 in Personal Privacy in an Information Society: The Report of the Privacy Protection Study Commission. Washington, D.C. http://aspe.hhs.gov/datacncl/1977privacy/c15.htm.

Reiter, J. (2002). Satisfying disclosure restrictions with synthetic data sets. *Journal of Official Statistics*, 18(4): 531–543.

Robertson, N. (1968). Data Bank: Peril or Aid? *New York Times.* Page 1, Col. 7, January 6.

Rubin, D. (1993). Comments on confidentiality: A proposal for satisfying all confidentiality constraints through the use of multiply-imputed synthetic microdata. *Journal of Official Statistics*, 9(2): 461–468.

Scheuren, F. (1999). Administrative records and census taking. *Survey Methodology, Statistics Canada*, 25(2): 151–160. http://www.statcan.gc.ca/pub/12-001-x/1999002/article/4878-eng.pdf.

Seastrom, M., Chapman, C., and Mulligan, G. (2008). The impact of privacy breaches on survey participation in a national longitudinal survey. In *Proceedings of the Section on Survey Research Methods,* American Statistical Association, Alexandria, VA. 241–249.

Singer, E. (2003). Privacy Research in Census 2000, Census 2000 Topic Report No. 1. Census 2000 Testing, Experimentation, and Evaluation Program, U.S. Census Bureau, Washington, D.C. http://www.census.gov/pred/www/rpts/TR-1.pdf.

— (2004). Risk, benefit, and informed consent in survey research. *Survey Research,* University of Illinois at Chicago, 35(2–3). http://www.srl.uic.edu/Publist/Newsletter/2004/04v35n2-3.pdf.

Singer, E., Schaeffer, N. C., and Raghunathan, T. (1997). Public attitudes toward data sharing by federal agencies. *International Journal of Public Opinion Research*, 9: 277–287. http://ijpor.oxfordjournals.org/content/9/3/277.short.

Singer, E., Van Hoewyk, J., Tourangeau, R., Steiger, D. M., Montgomery, M., and Montgomery, R. (2001). Final report on the 1999–2000 surveys of privacy attitudes. Planning, Research and Evaluation Division, U.S. Census Bureau, Washington, D.C. http://www.census.gov/pred/www/rpts/SPAN_SPA.pdf.

Sweeney, L. (1997). Weaving technology and policy together to maintain confidentiality. *Journal of Law, Medicine and Ethics*, 25(2–3): 98–110. http://dataprivacylab.org/dataprivacy/projects/law/law1.html.

Thomas, R. and Walport, M. (2008). Data sharing review report. Report to the Prime Minister and Secretary of State for Justice, Ministry of Justice, United Kingdom. http://webarchive.nationalarchives.gov.uk/+/http://www.justice.gov.uk/docs/data-sharing-review.pdf.

United Nations (2007). Register-based statistics in the nordic countries: Review of best practices with focus on population and social statistics. United Nations Economic Commission for Europe, New York and Geneva. http://unstats.un.org/unsd/censuskb20/Attachments/Register_based_statistics_in_Nordic_countries-GUID6c4bc07a61994d2fa92e7c1dd79438fa.pdf.

— (2009). Principles and guidelines on confidentiality aspects of data integration undertaken for statistical or related research purposes. United Nations Economic Commission for Europe, Geneva. http://www.unece.org/stats/publications/Confidentiality_aspects_data_integration.pdf.

U.S. Census Bureau (1996). The plan for Census 2000. U.S. Census Bureau, Washington, D.C. 147–153.

— (2004). Census 2000 Synthesis Report #16: Results from Administrative Records Experiment in 2000. U.S. Census Bureau, Washington, D.C. http://www.census.gov/pred/www/rpts/TR16.pdf.

— (2006). Transcript of race and ethnic advisory committee meeting of April 2006. U.S. Census Bureau, Washington, D.C. 147–153.

Vijayan, J. (2007). One year later: Five lessons learned from the VA data breach. *Computerworld.com*. http://www.computerworld.com/s/article/9022678/One_year_later_Five_lessons_learned_from_the_VA_data_breach.