Expanding Statistical Use of Administrative Data: A Research Proposal Focused on Privacy and Confidentiality[1]

Gerald W. Gates[2]
Privacy Consultant
8524 Wagon Wheel Road, Alexandria, Virginia 22309

Key Words: Data Access, Privacy, Confidentiality, Informed Consent

Abstract:

This paper outlines a research agenda necessary to help in understanding why barriers to the statistical use of administrative records exist and how they can be overcome. While there are legal issues that must be addressed by researchers in accessing these records, the most intransigent issues involve policy decisions related to confidentiality of personal information and privacy of individuals.[3] Addressing these policy concerns has been difficult and time consuming for statistical agencies and researchers, has led to missed opportunities, and has not necessarily facilitated privacy and confidentiality. To better address these concerns, it is critical that negotiations between administrative agencies and statistical agencies/researchers recognize real risks both to privacy and to data use. Regarding confidentiality, research should focus on identifying and limiting risks to confidentiality from security breaches or inadequate disclosure limitation measures. Privacy concerns are more subjective and are the most difficult to overcome. Research on privacy needs to focus on public awareness of these uses and how opinion may be swayed to support or oppose these uses. Additional research is proposed to determine the extent to which valuable statistical research is abandoned where agreements cannot be reached because guidance on addressing privacy and confidentiality is lacking. The findings will be helpful in establishing model agreements and forming more generalized legislation and policy support for the statistical use of administrative records.

1.0 Introduction

U.S. statistical agencies collect information directly from individuals and businesses to generate federal statistics. Also important is the information that is gathered from secondary sources that was originally obtained for administrative purposes.[4] Agencies have been obtaining and using these administrative records in their statistical activities for many

---

[1] This paper is an extension of a 2008 paper titled "Providing Researchers with Authorized, Safe, Useful Access to Administrative Records" that the author prepared for the Committee on National Statistics in support of a workshop on Protecting Student Records and Facilitating Education Research. The 2008 paper provides details on the legal and policy support for administrative records use and describes specific privacy and confidentiality issues and how they impact access and use of integrated survey and administrative data. The paper is available from the author upon request.

[2] Author was Chief Privacy Officer at the U.S. Census Bureau prior to retiring in July 2007. He 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 20 years.

[3] For the purposes of this discussion, confidentiality concerns the agreement reached with the individual when the information was collected about who can see the identifiable information. Privacy, on the other hand, pertains to the individual's right to control the use and disclosure of information about him.

[4] The Confidential Information Protection and Statistical Efficiency Act of 2002 defines administrative purpose 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.

decades. One of the earliest reported uses was for an evaluation of the 1950 Census income results using IRS and SSA data. (FSCM, 1980) Administrative records have also 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 in producing 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 the nature and impact of survey non-response;
- To aid survey methodologists in understanding the nature and extent of 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.

There have been many recent success stories about expanding these uses including the Census Bureau's Statistical Administrative Records System (StARS),[5] the Longitudinal Employer Household Dynamics Program (LEHD) (Census Bureau, 2006), and the Medicaid Undercount Project (Cox, et.al., 2006). There have also been some missed opportunities because legal and policy impasses could not be overcome.[6]

Without administrative records, agencies would have to spend considerably more taxpayer dollars to collect data or be forced to decrease greatly the geographic detail of published data. In addition, evaluations of survey and census data quality would be more difficult. There would also be increased burden on the public to report again information that they have already provided to the government. Fortunately, federal law and policy is supportive of these uses and recognizes that there is minimal privacy risk when administrative data are used to generate statistics.

Despite plenty of legal support, privacy and confidentiality play a significant role in the negotiations between the statistical and administrative agency and have delayed, and sometimes hampered, legitimate access and use. One of the key considerations revolves around how the public will view these uses and how these views may impact the agency's reputation and funding. Even when negotiations are successful and the data are shared, the statistical agency may not effectively use the data because of its own concerns about how the public views the use. But it does not end there. If the data are shared and effectively used in a federal statistical program, the resulting data may not be available to researchers because of difficulties in protecting confidentiality. This paper addresses the privacy and confidentiality concerns from each of these perspectives and lays out a research plan for better understanding the risks and how to mitigate them.

2.0 Legal and policy issues in accessing and using administrative records for statistics

The legal support for administrative records use is often overshadowed by the policy discussions that drive the decisions to share data. Often, the law allows significant discretion regarding what information can be shared and under what conditions. In the end, the record holder has the option to share or not depending on how the policy concerns are

---

[5] The StARS is a resource for much of the Census Bureau's administrative records program uses and is built using files from seven major federal agencies that are merged to develop the best possible measure of the population. StARS was used as an essential component of the Administrative Records Experiment of 2000 that was designed to assess the strengths and weaknesses of administrative data as a supplement to, or substitute for, decennial census population counts.

[6] It is difficult to find specific examples in the literature because these tend not to be well documented. Nevertheless, most agencies can point to examples where potential benefits have not been realized. Section 2.3 discusses the reasons for such impasses.

addressed. One issue that can complicate this process is how, and how much, to tell the public about these uses. The following describes the legal support for these uses, the expectations of the negotiators, and the public's knowledge of the circumstances for using their personal information in this way.

2.1 Legal Support

Title 13 of the United States Code (the Census Act) 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. There are also multiple examples of federal and state laws that permit the reuse of administrative data for research and statistics as long as confidentiality is assured and the information provided will not be used to take action against any individuals or businesses whose data are shared. (Gates, 2008)

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. [7] Agencies specify the recipient and the conditions for such disclosure 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 administrative agency recognizing and supporting 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 this does not 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 available to all statistical agencies covered under CIPSEA and its implementing regulations.

Title 26 (the Tax Code) is an example of a law pertaining to an administrative agency that specifically authorizes disclosure of identifiable records for research and statistics. For instance, Title 26 provides specifically for the Census Bureau to obtain tax return information "to the extent necessary in the structuring of censuses and national economic accounts and conducting related statistical activities authorized by law." Conditions for such disclosure are set out in regulations promulgated by the Secretary of Treasury.

Laws sometimes limit the types of statistical uses and users of administrative records. For instance, some laws provide that the research uses directly benefit the program for which the records were collected (for example, Food Stamp Records under 7 U.S.C. sec. 2026 b(1)(A)). Similarly, the recipients are sometimes limited to those specified in the legislation (for example, education records under 20 U.S.C. sec. 1232 g(b)(a)(F)). Where access is authorized and the 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.

---

[7] 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.

2.2 Public Awareness and Consent

Important in these legal exemptions is the notion that prior written consent of the individual is not required to permit sharing data because the information is to be protected from uses that can impact the individual. 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 an SSN from survey/census respondents to facilitate linkage. Refusal to provide one's SSN implies refusal to permit the linkage.[8] If SSNs are not collected but linkage is planned,[9] agencies may provide a notice of intent to link and an opportunity to opt out.[10] Such consents are frequently general in nature and may not identify each source file to be linked. Research that is covered under the Federal Policy for the Protection of Human Subjects, known as the Common Rule, is subject to IRB review. The IRB may require informed consent prior to linking survey and/or census data with the administrative data. Sometimes, the IRB may require that signed consent be obtained from the individual. IRBs may also exempt surveys from the informed consent requirements if they determine that there is no more than minimal risk to the individual. [11] Legal assurance that confidentiality is guaranteed without exception would be a basis for such a determination.

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 (required under the Privacy Act), reviewing privacy materials on agencies' Web sites, 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 occurred a little over 10 years 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) primarily on the grounds that it had been insufficiently publicized.[12]

2.3 Negotiations for statistical access and use

---

[8] There is some evidence that it is the growing concern over identity theft rather that record linkage, per se, that affects unwillingness to provide one's SSN.

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

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

[11] The IRB also considers if the waiver will adversely affect the rights and welfare of the subjects, why the research results depend on the waiver, and if subjects can be provided additional pertinent information after the fact.

[12] 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.

Where the law is supportive of the sharing of administrative records for statistical uses, there are still multiple hurdles to overcome before the data are transferred to the statistical agency. The negotiations revolve around various policy considerations pertaining to the costs and benefits for each party:

- 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.
- Incentives. Negotiations sometimes involve incentives for the administrative agency. This quid pro quo is usually some form of enhancement of the source data that may include the addition of metadata, geographic variables, or summary statistics. But under statistical agencies confidentiality laws in no case can identifiable survey information be provided back to the administrative agency.
- 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 be also be used for these purposes.
- 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.
- 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.
- 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.
- Opinion leaders. Related to the public's fears are concerns about the views of law makers, 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.
- 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.

These negotiations 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.

3.0 Assessing the real risks to privacy and confidentiality in the statistical use of administrative records

The administrative agency and the statistical agency rightly spend a great deal of attention in negotiations addressing privacy and confidentiality issues. The decisions reached have direct implications for the nature and quality of the statistical research that can be conducted. In assessing these tradeoffs, it is incumbent on the statistical agency to argue for the research use and to demonstrate that confidentiality and privacy are at minimal risk. You may note that I have not argued that the risk be zero. That is an impossible goal but one that all parties seem to accept in principle. The risk, or even the perceived risk, of a confidentiality or privacy violation is at the heart of any debate about whether and how to share administrative data for research and statistics.

Confidentiality breaches occur when security controls are inappropriate or when disclosure avoidance procedures are not adequate to protect the data. Security controls have improved recently as a result of new OMB and NIST requirements

that have been imposed on all agencies. There are well-documented best practices for disclosure avoidance in published statistical data, including those derived in part from administrative data. The success rate for these techniques is considered excellent but is mostly anecdotal. Where data are provided to researchers in a restricted environment, the record again is considered to be very good with few reported problems. A more systematic accounting of instances where violations do occur will help inform discussions about the real risks and possible trade offs. As far as the record indicates, we can assume that the confidentiality measures taken to date are adequate to protect the data. What we don't know definitively is whether changes in technology and tools available to intruders are weakening agencies' ability to protect the data in the future or if current techniques are excessive and perhaps unnecessarily hampering important research.

The real threats to privacy involve uses of personally identifiable information in ways that are inconsistent with the uses described to the individual at the time the information was collected. A data disclosure can certainly lead to a privacy violation but privacy concerns can arise even if confidentiality is protected. Combining records from different sources to obtain greater knowledge can violate privacy if the linkage is unknown to the individual whose records are linked. Any administrative use of information contained in files to be used only for statistical purposes would violate privacy. Inappropriate browsing of personal information would also create a privacy violation. Statistical agencies have every incentive to keep the data they collect or obtain from being used for non-statistical purposes since they depend on voluntary cooperation in most surveys. Barriers between them and any administrative functions of government are important deterrents to improper uses.[13]

4.0 Extending prior research on privacy and confidentiality

There has been considerable research in the fields of statistical disclosure avoidance, informed consent, and privacy attitudes as they relate to administrative records use. Disclosure research has drawn international support both in the national statistical offices and in academia. Privacy-related research has primarily been a focus in the U.S. and much of it has been funded by the Census Bureau. Recent research has begun to focus on how confidentiality and privacy protection measures impact statistical research.[14] The following discussion highlights what agencies have learned in prior research and what they still need to know.

It is important to remember that there are other factors (as mentioned above) that influence decisions in obtaining access to these records for statistical research. Additional research could be proposed that would benefit knowledge regarding each of these factors. My purpose here is to focus on privacy and confidentially as the key factors that overshadow most decisions to share and use administrative records for statistical research. Also, where privacy is concerned, the focus is on agencies that collect information on individuals rather than on businesses.

4.1 Privacy research

4.1.1 What we know

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. (NAS, 1979) The purpose of this study was to determine why 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

---

[13] The Privacy Protection Study Commission in its 1977 report *Personal Privacy in an Information Society* advised that in order to assure only statistical uses are made of information collected or obtained for statistical purposes, agencies should be "functionally separate" in that any administrative functions are organizationally separate from statistical functions.

[14] Early work by Duncan and Feinberg to map risk vs. utility for public use microdata have led to more recent efforts by Lane and Kennickell, among others.

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 focus shifted somewhat to include a series of studies directed toward the use of administrative records to "derive the census totals from some non-responding households, to assist coverage measurement activities, and to help provide missing content." (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 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-Bolton, 1998)

In 1997, plans to expand administrative records use in the 2000 census were postponed due to inadequate time to complete the necessary research and growing concerns from advisors about possible impacts on census participation. (Gates-Bolton, 1998) 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.[15] 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; 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-Tretham, 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

[15] 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. Census 2000 is the last census to include the long form questions.

about data sharing.[16] The research further shows an apparent trend toward increased concern over data sharing during the period of 1995-2000.

This research also provides insights into the impact of notification on acceptance of data sharing, how negative publicity affects privacy concerns, and how attitudes translate to behaviors. The notification experiment was associated with the request for SSN. The research findings reported that "notification of record linkage has a small but significant negative effect on the response rate but a positive effect on responding to the SSN item." 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 et al. "found that respondents who reported exposure to negative as well as positive publicity about the census had significantly higher scores on the privacy index 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)

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. Nevertheless, most prior research has not been designed to determine how attitudes carry over to behavior. The SSN experiment, involving a comparison between two samples drawn from the same population, did provide indirect measures of behavior. The study found 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 directed to do so." (Singer, 2003) This relationship between a belief (concern about privacy) and an action (refusal to comply) indicates that, at least in this context, behavior is strongly linked to attitude.

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 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.

4.1.2 What we need to know

Despite the considerable knowledge gained by past research, statistical agencies still do not feel comfortable that they fully understand how the public might react to their efforts to expand access to and use of their 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. They may think that they have considered everything from a legal, policy and ethical perspective but the public may still not be satisfied.

---

[16] Singer, Schaeffer, and Raghunathan (1997) have shown that opinions about data sharing are related in predictable ways to trust in government, confidence in the Census Bureau's promise of confidentiality, feelings of political effectiveness, and a more general inclination to share or withhold personal information.

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 has to be current and has to be able to adapt to unexpected events. 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 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 directly measure the impact of privacy attitudes on survey response. If agencies can better understand the "privacy hot buttons" that lead people to decide not to cooperate, they can develop ways to address those concerns.
- 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 Web sites. The results will also be helpful in convincing advisors that the agency is being proactive in gaining public support.

4.2 Confidentiality research

4.2.1 What we know

Confidentiality is specifically mandated in the statutes of several federal statistical agencies[17] and was extended to the principal statistical agencies by the Confidential Information Protection and Statistical Efficiency Act of 2002 (CIPSEA). Disclosure limitation in published data is a key component of confidentiality protection and has been the focus of considerable research over the years. Statistical Policy Working Paper #22 provides statistical agencies with mathematical techniques that are helpful in reducing the disclosure risk in data products. (FCSM, 2005) Such techniques include top- and bottom-coding, random noise, swapping, blurring, microaggregation, or subsampling. Each agency's Disclosure Review Board determines if the proposed microdata file is "safe" based on the inherent risks unique to the data set and the techniques applied. This should not be read to imply, however, that the risk is entirely known or even measurable. Rather,

[17] For example, section 9 of Title 13 for the U.S. Census Bureau and the Education Sciences Reform Act of 2002 and its predecessor the National Education Statistics Act of 1994 for the National Center for Education Statistics.

the assessment is a best judgment based on available research on population uniqueness, as well as the motivation, level of effort, and resources available to a potential intruder. Facilitating the assessment of population uniqueness is a growing body of knowledge within the federal statistical agencies about external data sources.

One of the greatest challenges in disclosure limitation comes from files containing administrative data. Where survey data are matched with administrative data, the risk includes the ability of someone holding the source administrative data using it 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 reidentify its program participants is not sufficient protection. It could also be argued that the administrative agency has an incentive to identify individual in the linked data if it determines that would be a cost effective method to assess whether individuals are getting benefits to which they are not entitled. Consequently, microdata derived, in whole or in part, from administrative data have not typically been made available to researchers in the form of public use microdata, although the demand for such data is great.

Because of this demand, techniques have recently been developed to produce a set of pseudo-data with the same specified statistical properties as the true microdata. These techniques rely on multiple imputation methodologies and in the case of "Inference-valid Synthetic Data," involve 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-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.

Where synthetic data do not meet researchers' needs, agencies have the option of providing controlled access to the data in a secure environment such as a research data center, at a licensed academic institution, or through computer-monitored remote access. See Gates 2008 for a detailed discussion of these options as they pertain to administrative records.

Disclosure avoidance is one aspect of confidentiality protection and security is the other. Increasingly, agencies are becoming aware of the risks associated with transferring, storing, and retrieving confidential information. Over the past five years, data breaches have been reported by most government agencies as a result of new federal reporting requirements or through Freedom of Information Act requests. Generally, such losses occur when unencrypted data are transmitted by 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. (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 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, NCES not only provided notification and free credit monitoring it also offered the opportunity to withdraw participation—both retrospectively and prospectively. Seastrom et al. 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.

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.[18] 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.

4.2.2 What we need to know

Federal statistical agencies' 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 jointly undertake 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 statistical file containing its data. Treating the administrative agency as a possible intruder results in greatly reducing the data available to everyone. Currently, the law provides no discretion here but perhaps the law could provide disincentives for administrative agencies to re-link to its own data. An assessment should be done to determine if this is an option worth pursuing.
- 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. An assessment of actual harm to individuals based on past breaches would be helpful in determining the real risks from breaches or disclosures.
- 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.
- The limitations and potential for synthetic data for various applications. Research, such as that promoted by Rubin, Abowd and Reiter, among others, should continue to assess

[18] The 1999 IRS Safeguard Review of the Census Bureau found deficiencies in controlling access and use of tax data but did not find any evidence of data breach.

the disclosure protection and analytic validity of synthetic data.[19] Applications for synthetic data, such as those currently supporting the Census Bureau's programs that are available through the Cornell Virtual RDC, should be promoted across all federal agencies that are seeking access mechanisms for linked data.

- 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.

4.3 Research on missed opportunities

In addition, and as a prelude to privacy and confidentiality research, there is a desperate need for research on how confidentiality and privacy are limiting the statistical use of administrative records. This includes missed opportunities because negotiations cannot be reached to obtain the records from the administrative agency as well as missed opportunities because of the statistical agency's inability to effectively use the data it does obtain. There are also lost opportunities from not allowing researchers to access integrated 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.

5.0 Why administrative records continue to play only a small role in the decennial census

Although the potential for expanding administrative records use encompasses all federal statistical agencies, the biggest program may offer the biggest payoff but also offers the biggest risk. For the past three censuses, the U.S. Census Bureau has planned and conducted research on various uses of administrative records in the decennial census to evaluate coverage and content, improve coverage of individuals and households, supplement or replace long form content, and even replace the direct enumeration. The outcome always seems to be the same: research demonstrates operational and policy issues that cannot be addressed in time for this census so a research program is planned to improve the chances for success in the next census. Granted, some limited uses have been adopted in past censuses, primarily for evaluations. And, it should be noted that for the 2010 census the Census Bureau plans to use the StARS to identify potentially undercounted cases, to improve race coding, and to evaluate agreements between MAF and StARS for future maintenance activities and to predict address validity. This recent progress offers promise and StARS offers enormous potential. But at the end of the day there seems to be hesitancy to make a significant commitment to using administrative records in a meaningful way to improve census coverage or fill in for missing content.

The measured progress in the decennial census environment can, at least in part, be attributed to the agency's concern that in an activity as visible as the census, a privacy protest has the potential to cause irreparable damage. This concern is justified. Over the past few years, the public has been exposed to media reports of privacy and confidentiality violations by government agencies, academic institutions, and corporations. Concerns have focused on overuse and abuse of Social

---

[19] 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.

Security Numbers, extensive data mining of personal information, and inadequate security controls that lead to data breaches. It is easy to imagine a scenario where one or more of these could become an issue engulfing a census with a major administrative records component.

In the case of the decennial census, the question comes down to whether the Census Bureau can manage the risk of a privacy protest on census participation or whether it should abandon any thought of further research toward a census that integrates administrative data in a significant way? Undertaking the privacy and confidentiality research proposals outlined above should better inform this decision and give the Census Bureau a course of action that maximizes the likelihood it can maintain the public's trust.

6.0 A critical need for formal leadership and an open dialog

In addition to the knowledge to be gained by an ongoing program of privacy and confidentiality research, there is a critical need for government leadership beyond what is already in place. Relying on piecemeal legislation permitting administrative agencies to share data for statistical research and on the 35-year-old Privacy Act as justification for access is not sufficient in today's environment where so many additional applications are possible. Although the Committee on National Statistics and other respected professional groups have recommended greater use of administrative records in such programs as the decennial census, there has been no formal statement by Congress or the Administration that this a specific goal. CIPSEA offered an opportunity to recognize the current data sharing environment and the conditions under which the data should be protected. Unfortunately, it did not go far enough. Nevertheless, there are things that can be done to improve the situation.

First, the Privacy Act needs to recognize that the exemption granted the Census Bureau to obtain and use data from other federal statistical agencies without consent is also applicable to those agencies covered by the confidentiality provisions of CIPSEA. Each of these 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 in granting the Census Bureau exemption. At the same time, the Privacy Act and/or legislative history, as revised, could address specifically the importance of this exemption in fostering the statistical uses of administrative records.

Second, an OMB order or directive focused on the statistical use of administrative records would help guide agency decisions on data sharing. Such a statement should make clear that functional separation in the statistical use of administrative records is sacrosanct and that in no case can the information shared for statistical research be put in a position that its permitted uses could be compromised. This would support the intent of the Privacy Act and CIPSEA, and the statement for the record would be an important argument to counter opinion leaders who choose to focus on historical arguments to undermine such uses. An important caveat to this recommendation is that that ongoing Administration discussions regarding government IT consolidation must reflect this commitment to functional separation if such a statement is to be trusted.

Third, a coordinated data stewardship effort like that currently in place in a few agencies should be put in place across the federal statistical agencies. The Census Bureau, following the 1999 IRS Safeguard Review, committed 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. This commitment recognized the importance of protecting and limiting 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.

Fourth, 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 needs to be carried forward with the development of model agreements. These model agreements should be disseminated by OMB as appropriate for government-wide use.

Finally, 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-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.

References:

Abowd, John and Julia Lane (2003). "Synthetic Data and Confidentiality Protection," LEHD Technical Paper TP2003-10, U.S. Bureau of the Census, 2003 http://lehd.did.census.gov/led/library/techpapers/tp-2003-10.pdf

Cox, Christine, Michael Berning and Rochelle Wilkie Martinez (2006). "Data Policy and Legal Issues in Creating and Managing Integrated Data Sets," Proceedings of the 2006 Federal Committee on Statistical Methodology Policy Conference, forthcoming.

Federal Committee on Statistical Methodology (FCSM) (1980). Statistical Policy Working Paper #6, Office of Management and Budget, p 15.

Federal Committee on Statistical Methodology (FCSM) (2005, revised), Statistical Policy Working Paper #22, Office of Management and Budget.

Gates, Gerald (2008). "Providing Researchers with Authorized, Safe, Useful Access to Administrative Data," prepared under contract to the Committee on National Statistics, DBASSE-P280884, National Academy of Sciences, Washington DC.

Gates, Gerald and Deborah Bolton (1999). "Privacy Research Involving Expanded Statistical Use of Administrative Records," 1998 Proceedings of the Section on Government Statistics and the Social Statistics Section of the American Statistical Association, Alexandria, VA, pp. 203-208.

Gerber, Eleanor (2003). "Privacy Schemas and Data Collection: An Ethnographic Account," U.S. Census Bureau, February 10, 2003

Kincannon, Louis, V. Barabba, S.W. Martinez, L. Blumerman, G. Gates, W. Alvey (2005). "Panel on Privacy and Data Use in the New Technological Environment," 2005 Proceedings of the Government Statistics Section, American Statistical Association, Alexandria, VA, pp. 1234-1241.

Larwood, Laurie and Susan Trentham (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, DC 20233.

Mayer, Thomas S. (2002). Research Report Series (Survey Methodology #202-01, Privacy and Confidentiality Research and the U.S. Census Bureau: Recommendations Based on a Review of the Literature, U.S. Census Bureau.

Minkel, JR (2007). "Confirmed: The U.S. Census Bureau gave up names of Japanese-Americans in WWII," Scientific American.com, March 30, 2007. http://www.scientificamerican.com/article.cfm?id=confirmed-the-us-census-b

National Academy of Sciences (1979), *Privacy and Confidentiality as Factors in Survey Response*, Committee on National Statistics, National Research Council, National Academy of Sciences, Washington DC.

Office of Management and Budget (OMB) (2007). M07-16, Safeguarding Against and Responding to the Breach of Personally Identifiable Information, May 22, 2007. http://www.whitehouse.gov/omb/assets/omb/memoranda/fy2007/m07-16.pdf

Seastrom, Marilyn, and C. Chapman, G Mulligan (2008), "The Impact of Privacy Breaches on Survey Participation in a National Longitudinal Survey," 2008 Proceedings of the Section on Survey Research Methods, American Statistical Association, Alexandria, VA, pp. 241-250.

Singer, E. (2003). *Census 2000 Testing, Experimentation, and Evaluation Program* Topic Report No.1, TR-1, Privacy Research in Census 2000, U.S. Census Bureau, Washington DC.

Singer, Eleanor (2004). "Risk, Benefit, and Informed Consent in Survey Research," Survey Research, University of Illinois at Chicago, Volume 25, Number 2-3.

Singer, E., J. Van Hoewyk, R. Tourangeau, D.M. Steiger, M. Montgomery, and R. Montgomery (2001). "Final Report on the 1999-2000 Surveys of Privacy Attitudes," Washington, DC, U.S. Bureau of the Census, Planning, Research and Evaluation Division, December 2001.

Singer, Schaeffer, and Raghunathan (1997). "Public Attitudes Toward Data Sharing by Federal Agencies." International Journal of Public Opinion Research 9:277-84.

U.S. Census Bureau (2006). LEHD Technical Working Paper #2006-1, http://lehd.did.census.gov/led/library/techpapers/tp-2006-01.pdf

U.S. Census Bureau (1996). "The Plan for Census 2000," U.S. Census Bureau, February 28, 1996

Vijayan, Jaikumar (2007), "One Year Later: Five Lessons Learned from the VA Data Breach." Computerworld.com. http://www.sfgate.com/cgi-bin/article.cgi?f=/c/a/2006/06/16/BUG77JER911.DTL