MATHEMATICA Policy Research

REPORT

FINAL REPORT

Minimizing Disclosure Risk in HHS Open Data Initiatives

September 29, 2014

John Czajka
Craig Schneider
Amang Sukasih
Kevin Collins

Submitted to:
Office of the Secretary
Department of Health and Human Services
200 Independence Avenue, SW
Washington, DC 20201

Project Officer: Joan Turek
Contract Number: HHSP23320095642WC, Task Order HHSP23337049T

Submitted by:
Mathematica Policy Research
1100 1st Street, NE
12th Floor
Washington, DC 20002-4221
Telephone: (202) 484-9220
Facsimile: (202) 863-1763

Project Director: John L. Czajka
Reference Number: 40287.301

This page has been left blank for double-sided copying.

ACKNOWLEDGMENTS

The authors would like to acknowledge the contributions of numerous individuals to this report. We especially want to thank the panelists who shared their expertise and the moderators who guided the discussions, and we want to express our appreciation to all who attended the meeting. All of these persons are named in Appendix B. Their thoughtful discussion of the issues underlies the most important results of this project. We are grateful as well to our colleagues Bonnie Harvey and Melissa Medeiros, who contributed to the first background paper; Myles Maxfield, who reviewed the earlier drafts of this report; and Kimberly Ruffin, LaTia Downing, and Lisa Walls, who prepared the final manuscript. We also want to thank William Winkler, U.S. Census Bureau, who provided a bibliography of materials on statistical disclosure limitation. Lastly, we want to thank our project officer, Joan Turek, and her colleagues Susan Queen and Jim Scanlon in the Office of the Assistant Secretary for Planning and Evaluation, Department of Health and Human Services, the sponsoring agency, for their experienced and valuable guidance throughout the project and their helpful comments on the draft report.

The findings, conclusions, and recommendations presented in this report are those of the authors and do not necessarily represent the views of ASPE.

This page has been left blank for double-sided copying.

CONTENTS

EXECUTIVE SUMMARY

Federal agencies have a long history of releasing data to the public, and they also have a legal obligation to protect the confidentiality of the individuals and organizations from which the data were collected. Federal agencies have successfully balanced these two objectives for decades. With the new emphasis on expanding public access to federal data, coupled with the increasing availability of data from other sources, federal agencies are continuing to ensure that the combination of data already available and the data they are preparing to release does not enable the identification of individuals or other entities through what has been termed the "mosaic effect." The concept of a mosaic effect is derived from the mosaic theory of intelligence gathering, in which disparate pieces of information become significant when combined with other types of information (Pozen 2005).

To gain more insight into the mosaic effect and its implications for the continued release of data to the public while minimizing the risk of disclosing personal information, the Office of the Assistant Secretary for Planning and Evaluation (ASPE) in the U.S. Department of Health and Human Services (HHS) contracted with Mathematica Policy Research to convene a technical expert panel (TEP), prepare background materials, and summarize what was learned from the panel discussion and the background research in a final report. The goals of the project were (1) a balanced and scientifically sound assessment of the mosaic effect, (2) identification of any unique increased risk associated with the mosaic effect, and (3) identification of data release policies and best practices that can prevent or reduce disclosure due to the mosaic effect.

In assessing the increased risk that open data initiatives may create or incur through the mosaic effect, we and our expert panelists reviewed what is known about the sources of disclosure risk and the effectiveness of various ways to control such risk. From the discussion at the TEP meeting and the materials we reviewed in preparing the two background papers, we prepared a synthesis, which is followed by concluding observations.

Synthesis

Data collected by the federal government are covered by a substantial array of regulations intended to protect the confidentiality of the data and the privacy of those from whom such data are obtained. While remaining attentive to these requirements, federal agencies have been able and willing to provide researchers across a wide range of disciplines with open access to public use versions of their data. While some users can accept forms of restricted access in return for more extensive content or an ability to link records across datasets, many applications require a level of access that can only be supplied by public use files, for which federal agencies have gone to great lengths to protect respondent confidentiality.

There have been no documented breaches of federal public use data—survey or administrative. And while occasional, documented breaches involving non-federal data have received a lot of attention, these breaches have been confined almost entirely to data that were not protected in ways consistent with current standards.

Despite this strong track record, there was general agreement among the panelists that disclosure risk cannot be driven to zero without seriously weakening the analytical value of a dataset.

Sources and Evidence of Disclosure Risk. In general, the removal of direct identifiers is not sufficient to de-identify a dataset, as some of the remaining variables can serve as indirect or quasi-identifiers. In combination, they can uniquely identify individuals within the population or well-defined subpopulations. Voter registration files, which typically contain name, address, gender, and date of birth, have figured prominently in a number of the past breaches reported in the literature because a substantial proportion of the U.S. population bears a unique combination of gender, date of birth, and ZIP code.

The Health Insurance Portability and Accountability Act (HIPAA) requires that covered entities de-identify the data they share using methods that would protect against a wide range of potential threats, yet only 3 of the 33 states that sell inpatient discharge data are applying HIPAA standards. Recently, Sweeney was able to re-identify a substantial proportion of a small subsample of individuals in hospital discharge data released by the State of Washington, based on information compiled from newspaper accounts. It must be noted, however, that the Washington State data did not comply with widely used standards developed by the National Association of Health Data Organizations. This episode underscores the potential vulnerability of public use data to the kinds of information becoming more widely accessible through Internet searches and web scraping techniques.

While the Washington State data may represent an exception among data distributed by states, many sales or transfers of personal data—particularly health data and financial data—are neither regulated by government nor held to professional standards designed to minimize disclosure risk. Consequently, such data may be especially vulnerable to re-identification.

The Mosaic Effect and Related Threats. There was a consensus among the TEP participants that the public use files released by HHS and other federal agencies bear limited risk of disclosure, even in combination with other publicly available data. If de-identified properly, the public use data released by federal agencies are not useful to the intruder. In general, the confidentiality of federal datasets is threatened less by the release of other federal data than by data from two other sources: (1) datasets compiled by other organizations and released with weak or no de-identification, and (2) personal data revealed through social media. While agencies can control what they release, they cannot control what other organizations and private individuals release. Federal agencies recognize the growing threat from external data and are actively engaged in assessing and responding to the disclosure risks that they pose.

The explosion of personal information posted to the Internet through social media and other avenues means that the availability of data that might be of use to a potential intruder has grown as well. Such information does not cover the entire population, however. Its coverage is far less extensive than voter registration records, for example, and some investigators have noted the impact of incomplete coverage.

If the confidentiality of a dataset has been breached to the extent that many of the records have been correctly re-identified, then all of the variables contained in that dataset for these

named individuals become available as potential indirect identifiers that could be used to break into another dataset containing some of the same individuals and some of the same variables. In actuality, the threat posed by the re-identification of records in a single database covering a narrow subset of the population is likely to be very small, given the limited number of records involved and their minimal overlap with records released by the federal government—particularly from sample surveys.

The release of well-protected federal files does not appear to increase the re-identification risk for other federal files; however, a number of agencies are conducting informed risk assessments. The sheer volume of data files made accessible through the Open Data Initiative is striking. Are there large numbers of individuals who appear repeatedly because of the way that file universes and samples are defined? This may be difficult if not impossible to determine, given restrictions on sharing or linking personal data across agencies, but risk assessment would be enhanced with such information.

Protecting Public Use Data. To maximize their effectiveness, statistical disclosure limitation methods must be tailored to the data they are being used to protect. Each dataset faces unique risks, depending on the type of data, the population covered, the sample design, the variables that require the most protection, and the distribution of values on these variables.

Statistical Policy Working Paper 22 has provided valuable documentation of federal agency practice in protecting the confidentiality of federal data. However, the last update of this important resource was nearly 10 years ago, and agencies have upgraded their statistical disclosure limitation methods since then. TEP panelists asserted that regular updates—perhaps as often as every five years—would help to ensure that the document remains current.

A useful way to represent disclosure risk is that the probability of a re-identification is equal to the product of the probability of an attack, and the probability of a re-identification conditional on an attack. Disclosure risk can be lowered by strengthening the protections applied to public use data *or* by taking steps that reduce the likelihood of an attempted re-identification. That no breaches of federal data have occurred to date may be due in part to the fact that incentives were not high enough to inspire serious efforts to challenge the disclosure protections, although the protections themselves may have contributed to reducing these incentives.

Working in the opposite direction, however, the penalties to which data users are subject for attempting to re-identify records in public use files are light to non-existent. For most agencies the onus falls entirely upon the data producers in the event of a re-identification. This creates a tension for agencies releasing public use files in order to comply with Open Data policies and initiatives.

The confidentiality of public use data is reinforced in ways besides de-identification. Sampling is an important tool for reducing disclosure risk. This speaks to the security of the federal government's many public use files of sample survey data but underscores the inherent risks in creating public use files from administrative records, which may not be sampled at all or are sampled at much higher rates than is typically found in surveys.

Reporting error, which can be particularly high in sample surveys, also provides protection against disclosure. Sometimes the application of additional protection in the form of masking may not be needed and may only reduce the quality of the data.

The effects of disclosure limitation methods on the quality of public use data are an important concern. Over-zealous confidentiality protection can weaken data quality with little improvement in data security. It is possible to apply such strong protections to public use data that they become useless—and unused. To guard against this, some agencies consult with subject matter experts and major data users to better assess the trade-offs between analytic utility and effective disclosure limitation. Secure remote access is growing as a solution to the problem of maintaining quality while protecting confidentiality, but it cannot serve all data needs.

False re-identification is generally not addressed in regulations, yet it may present a more serious problem, potentially, than positive re-identifications. From a technical standpoint, an agency can protect against a positive (or true) re-identification but not a false re-identification. Furthermore, agencies cannot deny alleged re-identifications except to reiterate that a re-identification is not possible, as an explicit denial may reveal information that assists an intruder with a correct re-identification.

Evaluation of the effectiveness of disclosure limitation methods by attempting to re-identify records internally remains the most powerful approach to establishing that a public use file is secure. A number of agencies obtain external, identified data—both public and commercial—to use in their evaluations, which may enable more realistic assessments of risk.

Frontiers of Research. Research is providing important enhancements to risk assessment and the ability to assign probabilities to disclosure risk. Disclosure risk and intrusion can be modeled; this has been done in a variety of ways by different investigators. Much of the recent research on protecting microdata has addressed the impact of statistical disclosure limitation on the analytic utility of the data. In particular, research has focused on measuring the information loss due to the application of disclosure limitation measures. More recent research is exploring the problem of maximizing utility while minimizing risk. Lastly, a prominent topic of recent research on statistical disclosure limitation is improving the quality of synthetic data.

Concluding Observations

Because federal agencies have worked hard to develop, maintain, and update their procedures for protecting the confidentiality of public use data, releasing multiple, well-protected files does not appear to produce a significant increase in disclosure risk. A greater threat comes from the personal information that individuals reveal about themselves and others through social media, as this information is identified. Incomplete population coverage reduces the threat, however. Files released with inadequate protection by states, local areas, and commercial organizations pose a threat as well if large numbers of records can be re-identified. The few examples discussed in this report indicate that such public files are not common, however, and their rarity of such files and their generally small size limit the threat they present.

Skilled hackers present more of a concern because of their potential ability to break into nonpublic databases and obtain access to data that has not been well protected. Whether they would also have interest in uncovering identities in public use files is not clear. The highly

publicized thefts of credit card numbers and other personal identifiers suggest that the threat from hackers breaking into internal, fully-identified databases may be greater than the risk of their re-identifying records in public use files protected with the most effective methods.

Federal agencies have demonstrated that they remain vigilant and forward-looking in their evaluation and application of disclosure limitation techniques to the data that they release to the public. The track record for federal public use files is unblemished, but agencies have not rested on these accomplishments. Well aware that once a dataset is released to the public it cannot be recalled, federal agencies have devoted resources to anticipating future threats. Such active engagement promises continued security for the data that federal agencies release.

This page has been left blank for double-sided copying.

I. INTRODUCTION AND BACKGROUND

On May 9, 2013, President Obama issued the executive order, "Making Open and Machine Readable the New Default for Government Information," in which he directed the Office of Management and Budget (OMB) to issue an Open Data Policy throughout the federal government. The objectives of this executive order were to advance the management of government information as an asset throughout its life cycle; to promote interoperability and openness; and, whenever possible and legally permissible, to ensure that data are released to the public in ways that make the data easy to find, accessible, and usable. Federal agencies have a long history of releasing data to the public, and they also have a legal obligation to protect the confidentiality of the individuals and organizations from which the data were collected. Federal agencies have successfully balanced these two objectives for decades. With the new emphasis on expanding public access to federal data, coupled with the increasing availability of data from other sources, federal agencies are continuing to ensure that the combination of data already available and the data they are preparing to release does not enable the identification of individuals or other entities through what has been termed the "mosaic effect." [1]

To gain more insight into the mosaic effect and its implications for the continued release of data to the public while minimizing the risk of disclosing personal information, the Office of the Assistant Secretary for Planning and Evaluation (ASPE) in the U.S. Department of Health and Human Services (HHS) contracted with Mathematica Policy Research to convene a technical expert panel (TEP), prepare background materials, and summarize what was learned from the panel discussion and the background research in a final report. [2] The goals of the project were (1) a balanced and scientifically sound assessment of the mosaic effect, (2) identification of any unique increased risk associated with the mosaic effect, and (3) identification of data release policies and best practices that can prevent or reduce disclosure due to the mosaic effect.

A. Background

The Open Data Initiatives launched by the White House have as their goal making government information resources more accessible to the public in machine-readable form, and encouraging the use of such data by entrepreneurs to aid in the creation of new products, services, and jobs. However, there is concern that certain users of the datasets made publicly available via these open data initiatives will be able to re-identify individuals or firms whose information is contained in these datasets. The challenge faced by federal agencies is to achieve

[1] The concept of a mosaic effect is derived from the mosaic theory of intelligence gathering, in which disparate pieces of information—though individually of limited utility—become significant when combined with other types of information (Pozen 2005).

[2] More specifically, the project's components are: (1) A pair of background papers, one reviewing federal policies and procedures regarding the use and protection of personal data and the other an environmental scan of literature relevant to releasing federal microdata in light of the risks presented by the mosaic effect; (2) a TEP tasked with addressing the mosaic effect through a discussion of best practices in protecting confidentiality in open data initiatives; and (3) this report, which synthesizes the findings from the background papers and the proceedings of the TEP meeting. The TEP meeting was held on June 27, 2014. The meeting agenda is reproduced in Appendix A, and a list of attendees is included in Appendix B. Minutes from the TEP meeting are presented in Appendix C. The two background papers are included in Appendices D and E.

the appropriate balance between (1) providing the public with useful datasets, and (2) protecting the privacy and confidentiality of individuals whose information is contained in this data.

The mosaic effect refers to the concept that the availability of increasing numbers of micro datasets, including de-identified datasets, increases the risk of disclosure beyond the risk associated with any particular dataset because of the totality of the information in the other datasets. Although there is great interest in the mosaic effect, ASPE had not yet been able to identify analyses of the issue, or evidence of risk quantification, or principles and best practices for addressing the issue.

The federal and HHS-specific Open Data Initiatives are releasing more data for public use. At the same time, there is a proliferation of data from other, non-federal sources, including social media. A key question is whether the risk of re-identification grows as more datasets become available. If it does, what principles and practices can agencies use to offset this increased risk of disclosure? Changing technology has also increased the potential threat of re-identification. Faster and less expensive data processing capabilities and sophisticated software make it much more feasible for nefarious actors to combine the information released in numerous datasets, and then use these data to try to determine individuals' identities.

To counter these threats, there is a large body of statistical disclosure avoidance techniques that have been developed by statisticians and computer scientists to minimize the risk of disclosure, and researchers continue to advance the state of the art. Federal agencies have implemented a variety of policies and procedures to protect the confidentiality of the data they release, and these have been highly effective in protecting against disclosures in individual datasets. [3] The TEP addressed the topic of whether the current techniques and data release procedures are sufficient to protect confidentiality in light of the mosaic effect and growing threats elsewhere, or if new techniques and data release mechanisms are needed. [4] The meeting provided a self-assessment regarding potential new threats to federal data privacy protection and the agencies' capacity to address them.

B. Organization of the Report

The concerns discussed above—new technologies, increasing amounts of data being made available to the public, growing numbers of other data sources, and the tools available to determined adversaries—provide a compelling motivation for federal agencies to continuously re-assess the risks of disclosure due to the mosaic effect. ASPE established this project to promote greater sharing of information about methods, data sources, and how to minimize disclosure risk among federal agencies in order to benefit the government and the public. The communication of best practices, lessons learned, and the state of the art in de-identification and re-identification methodologies should be useful to federal officials and others who make data publicly available and are simultaneously responsible for ensuring the privacy of respondents and the confidentiality of these data files.

[3] Many of these policies are described in Statistical Policy Working Paper 22 produced by the Federal Committee on Statistical Methodology (FCSM) in 2005.

[4] Multiple disclosure avoidance techniques are discussed in Chapter III. The data release mechanisms include public use data files, de-identified data, data use agreements, and research data centers (RDCs).

This report summarizes the principal findings from the project. Chapter II presents a summary of federal legislation and regulations regarding the release and protection of personal data along with recent policy statements with respect to open data. The chapter is based on material presented in the first background paper (Appendix D). Chapter III summarizes federal procedures for providing the public with access to government data while preserving the confidentiality of the individuals and businesses from whom the data was collected. This chapter draws on material presented in both background papers and one of the TEP sessions. Chapter IV reviews key issues in protecting public use microdata from disclosure. The chapter draws on the second background paper (Appendix E). This is followed in Chapter V by a summary of the experts' views expressed during two panel discussions at the TEP meeting. Chapter VI synthesizes the key findings from the project.

This page has been left blank for double-sided copying.

II. RELEVANT FEDERAL POLICIES

Federal policies covering the use and protection of personal data and the Open Data Initiative focus on data collected or obtained by the federal government. This chapter discusses the key legislation that has helped to shape federal policy on the use and protection of personal data; additional laws governing data use; illustrative examples of agency regulations and guidelines; and the major documents defining federal open data policy.

A. Key Legislation

Several key pieces of federal legislation govern the types of personal information that government and other organizations, such as health providers and educational institutions, can disclose about individual citizens or consumers. Most privacy laws focus on an individual's rights over the privacy of personal information—including ability to access and correct information—and the circumstances under which an organization may disclose information, with or without consent from the individual. This summary provides an overview of the acts that created the foundation for U.S. privacy law as it relates to data held by the federal government. We discuss the Privacy Act of 1974, the Computer Matching and Privacy Protection Act of 1988, the Health Insurance Portability and Accountability Act (HIPAA) of 1996, the Confidential Information Protection and Statistical Efficiency Act (CIPSEA) of 2002, and the Health Information Technology for Economic and Clinical Health Act (HITECH Act) of 2009.

1. Privacy Act of 1974

The Privacy Act of 1974 was one of the first pieces of legislation to recognize the rights of individuals to privacy and the government's responsibility to safeguard information that citizens provide to it. This law was based in part on a report prepared by an advisory committee under what was then the Department of Health, Education, and Welfare (HEW), which recommended a Code of Fair Information Practice that was intended to prevent information collected for one use to be made available for other purposes (without the consent of the individual), and would require agencies to have mechanisms in place that allow individuals to learn what information is being kept on them and to correct or amend a record (HEW 1973). The Privacy Act requires federal agencies to provide citizens with access and correction rights to personal information and limits how agencies share information. An agency can only disclose a person's record with the individual's written consent or under special circumstances. Under these exceptions, information may be shared within the agency or for uses for which it was intended (defined as routine use), for purposes of the Census, to the National Archives and Records Administration if the information is deemed worthy of preservation, to another agency for civil or criminal law enforcement activities that are authorized by law, and to individuals who have provided agencies with advance written notice that information will be used only for statistical research or reporting records. Records shared for statistical research or reporting must be "transferred in a form that is not individually identifiable." [5]

[5] "The Privacy Act of 1974," Title 5 U.S. Code, Sec. 552a. Available at [http://www.gpo.gov/fdsys/pkg/USCODE-2012-title5/pdf/USCODE-2012-title5-partI-chap5-subchapII-sec552a.pdf]. Accessed May 30, 2014.

2. Computer Matching and Privacy Protection Act of 1988

The Computer Matching and Privacy Act of 1988 updated the language of the Privacy Act to address concerns about how agencies share and match data across agencies. Agencies must notify individuals at the time of data collection that the information provided could be used for matching purposes, and give individuals 30-days advance notice before taking adverse action based on the matched data. In addition, the law requires that agencies create internal review boards to approve matching activities, publish matching agreements between agencies, and report to OMB and Congress about matching. The law does not apply to two types of matches: (1) matches that aggregate data stripped of personal identifiers, and (2) matches made to support research or statistical purposes. [6]

3. Health Insurance Portability and Accountability Act of 1996

HIPAA applies to health plans, clearinghouses, and health care providers. This legislation is often considered to represent the "high water mark" for how entities "balance risks to privacy against valuable uses of information" (Ohm 2010). There are two key regulations that emerge from HIPAA: the Privacy Rule and the Security Rule.

a. Standards for Privacy of Individually Identifiable Health Information (Privacy Rule)

Under the Privacy Rule, HIPAA-covered entities cannot disclose individually identifiable health information—known as protected health information (PHI)—unless the individual has authorized the release in writing or the disclosure or use is permitted under the Privacy Rule's exceptions. These exceptions allow for the information to be shared within the covered entity for treatment, payment, or health care operations or for public interest and benefit activities—for example, law enforcement purposes, or public health activities (HHS 2003). De-identified PHI can be disclosed if the data no longer identifies the individual or provides a reasonable basis to identify the individual. HIPAA-covered entities must de-identify data using one of two methods: (1) by receiving a formal determination of de-identification by a qualified statistician, or (2) by removing 18 specific identifiers (the "Safe Harbor" method), such as names, addresses, and account number. The full list of 18 identifiers may be found on p. D-3, Appendix D.

b. The Security Standards for the Protection of Electronic PHI (The Security Rule)

The Security Rule established a national security standard to safeguard health information and addresses the technical and non-technical safeguards that entities must put in place to uphold the Privacy Rule standards. Under the Security Rule, entities must "ensure the confidentiality, integrity, and availability" of all PHI that are created, received, maintained, or transmitted electronically, identify and protect against "reasonably anticipated threats" to security or integrity of data and uses or disclosures, and ensure workforce compliance. The rule includes physical and technical safeguards and other organizational and policy requirements that entities must implement.[7]

[6] "Computer Matching and Privacy Protection Act of 1988," Public Law 100-503. Available at [http://www.whitehouse.gov/sites/default/files/omb/inforeg/final_guidance_pl100-503.pdf]. Accessed May 30, 2014.

[7] "Summary of the HIPAA Security Rule." Available at [http://www.hhs.gov/ocr/privacy/hipaa/understanding/srsummary.html]. Accessed June 3, 2014.

4. Confidential Information Protection and Statistical Efficiency Act of 2002

CIPSEA limits federal agencies that collect data for statistical purposes from using the data for any other purpose – agencies must clearly distinguish between data collected for statistical and non-statistical reasons, and inform individuals at the start of data collection if the information will be used for other purposes. Additionally, identifiable information cannot be disclosed for any use other than statistical analysis or research without the consent of the respondent, unless the purpose is authorized by the head of the agency and the disclosure is not prohibited by any other law. CIPSEA also authorizes the Census Bureau, the Bureau of Economic Analysis, and the Bureau of Labor Statistics to share business data for the sole purpose of statistical analysis. [8] Each bureau must come up with systems and security protocols to protect the confidentiality of shared information and must remove any identifying information when publishing data and findings.

5. Health Information Technology for Economic and Clinical Health ACT

The HITECH Act is part of the American Recovery and Reinvestment Act of 2009. This law strengthened several of the privacy and security protections under HIPAA. For example, business associates of HIPAA-covered entities, such as contractors, must comply with HIPAA privacy and security requirements. The Act also strengthened rules related to disclosure of PHI for marketing and fundraising and prohibits the sale of PHI without an individual's authorization. The Act also requires HIPAA-covered entities to notify individuals and HHS of any breach of unsecured PHI and to report breaches affecting more than 500 residents to media outlets in the affected area.[9]

B. Additional Laws, Regulations, and Agency-wide Guidance

Appendix D contains background and details on several other laws related to privacy that are pertinent to federal agencies: the Family Educational Rights and Privacy Act of 1974, federal alcohol and drug confidentiality regulations, the Driver's Privacy Protection Act, and proposed legislation regarding the resale of consumer information. The Appendix also addresses regulations for the Census Bureau, the Internal Revenue Service (IRS), and the Department of Education and summarizes two manuals of internal rules maintained by the National Center for Health Statistics (NCHS).

To provide agency-wide guidance, in 1999 an FCSM interest group, the predecessor to the current Confidentiality and Data Access Committee (CDAC), prepared a "Checklist on Disclosure Potential of Proposed Data Releases." The Checklist, which has been updated a number of times, asks a number of questions about the proposed release of data—both public use microdata and tabular data. For microdata, for example, the Checklist asks about geographic

[8] "Confidential Information Protection and Statistical Efficiency," Public Law 107-347, December 17, 2012. Available at [http://www.gpo.gov/fdsys/pkg/PLAW-107publ347/pdf/PLAW-107publ347.pdf]. Accessed June 17, 2014.

[9] "Modifications to the HIPAA Privacy, Security, Enforcement and Breach Notification Rules Under the Health Information Technology for Economic and Clinical Health Act and the Genetic Information Nondiscrimination Act," 45 CFR Parts 160 and 164. Available at [http://www.gpo.gov/fdsys/pkg/FR-2013-01-25/pdf/2013-01073.pdf]. Accessed June 3, 2014.

detail reported on the file, top coding of continuous variables, and a number of other factors associated with disclosure risk. The Checklist also provides guidance on many of these topics. One purpose of the checklist was to provide a standardized document that agencies could prepare and submit to their disclosure review boards. Some agencies reference the Checklist as a resource in the preparation of public use data within their agencies. Other agencies have updated or developed their own versions of the Checklist to use in much the same way.

Another resource available to all agencies is Statistical Policy Working Paper 22, "Report on Statistical Disclosure Limitation Methodology" (FCSM 2005). The report describes numerous techniques for protecting public use data from disclosure, provides guidance in their use, and summarizes the practices of more than a dozen federal statistical agencies in preparing data for public release.

C. HHS Dissemination Activity

HHS is actively engaged in disseminating information relating to the privacy, confidentiality, and protection of health data through conferences and workshops covering the topics of legal issues, restricted data and restricted access procedures, disclosure risk analysis, and statistical disclosure limitation methods and techniques. For example, HHS staff were instructors in the workshop Privacy, Confidentiality, and the Protection of Health Data—A Statistical Perspective during the 1999 Joint Statistical Meetings in Baltimore, Maryland. There have been several workshops with topics in health care privacy, confidentiality, and data security, as well as statistical disclosure avoidance techniques, at which HHS staff presented. In 2004, the Confidentiality and Data Access Committee and the Washington Statistical Society organized the workshop Privacy, Confidentiality, and the Protection of Health Data: A Statistical Perspective on the HIPAA Privacy Rule. The Workshop on the HIPAA Privacy Rule's De-Identification Standard sponsored by the Office for Civil Rights was held in 2010 and featured speakers on five different panels, including statistical disclosure control and HIPAA privacy rule protections. The panelists provided guidance regarding the statistical and/or scientific methods that can be used or applied in practice to protect health data in accordance with the Privacy Rule. The slides from this panel can be accessed at http://hhshipaaprivacy.com/assets/5/resources/.

D. Open Data Documents

In January 2009, the President issued a Memorandum for the heads of Executive Departments and Agencies on Transparency and Open Government (White House 2009). [10] The Memorandum required the development of recommendations for an Open Government Directive to establish a system of transparency, public participation, and collaboration. In response, OMB issued an Open Government Directive in December 2009.

Four documents issued by the Executive Office of the President over a six-month period in 2013 define the scope and provide guidance on implementation of the new open data policy. These four documents were:

1. Increasing Access to the Results of Federally Funded Scientific Research (Office of Science and Technology Policy 2013)

[10] Reprinted in *Federal Register*, vol. 74, no. 15, pp. 4685-4686.

2. Making Open and Machine Readable the New Default for Government Information (White House 2013c)
3. Open Data Policy—Managing Information as an Asset (OMB 2013a)
4. Supplemental Guidance on the Implementation of M-13-13 "Open Data Policy—Managing Information as an Asset" (OMB 2013b)

Summaries of these documents are presented below.

1. Increasing Access to the Results of Federally Funded Scientific Research

This memorandum was issued on February 22, 2013 by the Office of Science and Technology Policy (OSTP) and calls for all federal agencies that are engaged in research and development to outline plans to provide public access to all results of scientific projects that are receiving federal funds. With respect to scientific data, OSTP requires that agencies maximize free public access to digitally-formatted data created with federal funds, while protecting privacy and proprietary information. This involves requesting that grant recipients outline data management plans and detail any reasons why their data cannot be made publicly accessible. OSTP also requests that agencies allow for the inclusion of appropriate costs associated with data management and access in federal grant proposals.

2. Making Open and Machine Readable the New Default for Government Information

Executive Order 13642 was issued on May 9, 2013, and calls for a shift in the default policy in federal agencies toward that of free public access to information. The order describes government information as an asset, the dissemination of which is likely to create new jobs, provide inspiration for entrepreneurship, and stimulate the American economy. The order calls for the adoption of an Open Data Policy, as outlined in the OMB memorandum issued on the same day (and discussed next). The Chief Information Officer and Chief Technology Officer were instructed to publish an online resource to assist agencies in their efforts to implement open data policies; Open Data Policy requirements were to be integrated into federal acquisition and grant-making; and agencies were to report quarterly on their progress, including a full analysis of risks to individual privacy and confidentiality.

3. Open Data Policy—Managing Information as an Asset

Memorandum M-13-13 was issued by OMB in conjunction with the Executive Order, establishes a framework to support effective information management strategies that will promote open data. Agencies are directed to adopt the following policies: use machine-readable and open formats; use data standards; remove all restrictions on distribution of public data (open license); and describe data using common core metadata (for example, origin, linked data, geographic location, time period/interval, and data quality). Agencies were requested to take the following actions:

- Create and maintain an enterprise data inventory
- Create and maintain a public data listing

- Create a process to engage with customers to help facilitate and prioritize data release
- Clarify roles and responsibilities for promoting efficient and effective data release practices

The memorandum gives particular attention to the protection of privacy and confidentiality, and the Mosaic Effect is noted as an issue of particular concern to this goal. Risk minimization guidelines include: collect or create only necessary and useful information; limit collection of identifying information; limit sharing identifying or proprietary information; take into account the levels of risk and potential harm that are associated with the dissemination of particular datasets; and consider information that is already public when releasing de-identified data (that is, be aware of the mosaic effect).

4. Supplemental Guidance on the Implementation of M-13-13 "Open Data Policy – Managing Information as an Asset"

This document, issued in August 2013, provides additional information and establishes minimum requirements for the objectives of the Executive Order and OMB Memorandum M-13-13. Minimum requirements for an Enterprise Data Inventory include submitting a schedule to OMB of how the agency plans to identify all its data; posting all datasets in machine-readable format to Data.gov; and updating the inventory schedule on a quarterly basis. Minimum requirements for a Public Data Listing include publishing all data that are described in the inventory metadata as public and publishing the data listing at www.[agency].gov/data.json. The minimum requirements to engage with customers are establishing a mechanism for receiving and reviewing customer feedback.

III. PROVIDING ACCESS TO GOVERNMENT DATA

Since the re-identification risk that might result from the mosaic effect more often involves microdata—that is, individual records representing persons or organizations—than tabular data, this chapter focuses on disclosure avoidance procedures for microdata. [11] In this chapter we discuss procedures, methods and techniques that federal agencies and other data holders may use to avoid disclosure of confidential information in microdata. The main reference for federal practice in this area is Statistical Policy Working Paper 22 (FCSM 2005), particularly Chapters III ("Current Federal Statistical Agency Practices") and V ("Methods for Public-Use Microdata Files"). In Chapter III of the working paper, fourteen federal agencies from across governmental departments reported their data disclosure avoidance practices. This information was collected in 2004, and during the past 10 years the agencies may have modified some of their practices. Mathematica contacted these fourteen agencies to collect information regarding any updates to these practices. Of the 11 agencies that release public use microdata, 6 reported at least some modification of their procedures, although these were generally minor. For the 5 that reported no updates, the descriptions provided in 2004 remain accurate, but in some cases the earlier practices were described in broad terms that could encompass at least some level of revision (for example, citing procedures documented in a separate manual). It is possible, too, that the responses in some cases reflected a reluctance to make public the specifics of the disclosure techniques that are applied to the agency data, as this information could be of use in an attempted re-identification. Table D.1 in Appendix D provides a summary of the earlier practices as well as the reported updates.

We begin by discussing the principal forms of access that federal agencies provide to users of their microdata and then provide an overview of the methods used to protect public use files. We conclude with a brief summary of recent advances in protecting microdata.

A. Forms of Access

There are two general approaches that are used to release microdata in a way that protects the data from disclosure. One is by restricting access to the data, and the other is by restricting the data that are released for public use (National Research Council 2005). The latter approach encompasses a wide range of techniques that include suppressing variables and changing their values.

1. Restricted Access

There are three basic mechanisms that federal agencies use to provide researchers with restricted access to data that are not released to the public. These include licensing, research data centers (RDCs), and secure remote access.

Under licensing arrangements, prospective users request restricted data files through a formal application process. To obtain such data, users must demonstrate that the data will be stored and used in a secure environment that meets the issuing agency's standards. As part of the proposal the user will generally have to explain why the data are needed and how they will be

[11] Strategies for tabular data are discussed briefly in Chapter V and in Appendices C and D.

used, and access may be limited to variables and records for which the user can demonstrate a critical need. To receive the data, the user typically has to sign a nondisclosure agreement.

Several federal agencies maintain RDCs, in which approved users can access agency data that are not released to the public. The data never leave the site, and output produced from data held in the RDC cannot be removed without a disclosure review, which can take different forms. For example, RDC staff may be authorized to review output, or the output may have to be screened by an agency disclosure review board. The types of data manipulations allowed to RDC users are limited. Linkages between databases may be prohibited or restricted. Users may not be allowed to attach portable storage devices to the computers or terminals that they use, and even printing of output may not be permitted (one RDC emails output to users after it has been reviewed). Obtaining access to an RDC requires submission of a proposal, and acceptable uses may be restricted to applications that carry potential benefits to the agency. Some agencies require its RDC users to undergo a background check and obtain employee-like status. The entire approval process may require several months.

A number of federal agencies allow users remote access to agency data that are not released on public use files. This can take a number of different forms. For example, the Census Bureau allows users to request tabulations from decennial census files that include more detail than the numerous tabulations that can be obtained from the bureau website (FCSM 2005). The requests are reviewed to ensure that the tabulations do not present a disclosure risk. The National Center for Health Statistics allows approved RDC users to submit programs remotely, although the software that can be used for this purpose is more limited than what is available in the RDC, and certain functions are not accessible. The advantage to the user lies in not having to travel to an RDC. This may be important when the research involves submission of a series of programs that take little time to run but require extensive review of the results before the next program can be prepared. Some RDCs charge a daily fee for in-person visits, not to mention long-distance travel costs and overnight accommodations, which can make a series of brief visits to the RDC very costly.

Additional information on modes of restricted access is provided in Appendix E. Open data initiatives imply public use data, for the most part, so we direct the rest of this chapter to the discussion of procedures used to prepare data for public release.

2. Restricted Data for Public Access

Public use microdata play a critical role in research and policy analysis. Exploratory research and many types of policy analysis do not lend themselves well to the conditions that govern restricted access as described above. The creation of public use data that protect the confidentiality of the subjects begins with de-identification, but depending on the contents of the data and the characteristics of the subjects, it may require the application of a number of additional techniques to reduce the risk of re-identification to a satisfactory level

a. De-identification

HIPAA codified a de-identification process for health records that includes the removal of 18 specific direct and indirect identifiers, which are listed in Appendix D. HIPAA requirements apply to a narrow range of datasets, but most of these identifiers have relevance outside of the

health data that fall under the HIPAA regulations. The protections mandated by HIPAA go well beyond the simpler de-identification practices that were common in unregulated health data prior to HIPAA.

b. The Concept of *k*-Anonymity

To protect the individuals in a dataset from re-identification, one must be certain that the characteristics reported on the file do not define unique individuals in a separate, identified database that is accessible to potential intruders. In theory, the way to achieve this level of protection is to ensure that no combination of characteristics is shared by fewer than some minimum number of persons in the population. This concept is called "*k*-anonymity," where *k* is the chosen minimum number (Sweeney 2002). This is a fundamental concept in protecting public use data from disclosure (Ciriani et al. 2007, El Emam and Dankar 2008). [12] If the data producer has access to a population database containing characteristics that will be reported on the public use file, the application of *k*-anonymity as a principle of disclosure limitation is straightforward and rigorous. Characteristics that in combination define unique individuals can be altered so that, when combined, they point to no fewer than *k* people. Typically, however, the data producer is not able to access population data for this purpose and applies *k*-anonymity to the file that is to be released. This is a conservative approach in that it yields more protection than is necessary to achieve *k*-anonymity at the population level. However, when the files to which it is applied contain a non-trivial proportion of the population, it may not be excessively conservative.

c. Statistical Disclosure Limitation

More generally, statistical disclosure limitation encompasses a wide range of techniques for reducing detail, modifying data values, or creating alternative data values to minimize the likelihood of a successful re-identification and, secondarily, lessen the information that would be gained if a re-identification were actually accomplished. The techniques that federal agencies apply tend to vary across the agencies, and within an agency they are likely to vary with the dataset, as different datasets present different challenges, depending on the type of information collected, the depth of the information recorded, and the design of the sample. The next section provides an overview of the methods commonly employed for statistical disclosure limitation.

B. Methods of Statistical Disclosure Limitation

Statistical disclosure avoidance techniques for microdata have been well developed and widely published in journals, textbooks, and workshop and conference proceedings. The following two sources provide comprehensive accounts of these techniques: (1) Statistical Policy Working Paper 22 (FCSM 2005); and (2) Handbook on Statistical Disclosure Control (Hundepool et al. 2010). Techniques to protect microdata for public release include those approaches pertaining to the file in general and those related to variables within the file. For example, Statistical Policy Working Paper 22 identified the following approaches:

1. Include data from only a sample of the population

[12] The Office for Civil Rights (OCR) guidance on methods for de-identification of data under the HIPAA Privacy Rule discusses k-anonymity as a principle that can be applied to protect health data under HIPAA (OCR 2012).

2. Do not include obvious identifiers
3. Limit geographic detail
4. Limit the number and detailed breakdown of categories within variables on the file
5. Truncate extreme codes for certain variables (top or bottom coding)
6. Recode into intervals or round continuous variables
7. Add or multiply by random numbers (adding noise)
8. Swap or rank swap the values on otherwise similar records (also called switching)
9. Select records at random and blank out selected variables and impute the missing values (also called blank and impute)
10. Aggregate across small groups of respondents and replace each individual's reported value with the average

A more complete list of statistical disclosure limitation methods is presented below, divided between nonperturbative and perturbative methods. Some of these techniques are suitable only for categorical variables, or only for continuous variables, whereas others can be applied to both types of variables.

Nonperturbative methods. These methods do not alter data values; rather, they implement partial suppressions or reductions of detail in the original dataset. These techniques include the following:

- **Sampling:** releasing a subsample of the original microdata
- **Global recoding:** combining several categories to form new, less specific categories
- **Top and bottom coding:** combining values in the upper (or lower) tail of a distribution (a special case of global recoding)
- **Local suppression:** suppressing the values of individual variables for selected records so that no information about these variables is conveyed for these records

Sampling introduces or increases the uncertainty that a particular individual is included in a microdata file, and in doing so it provides a strong disincentive for a would-be intruder to attempt to re-identify records on the file. Sampling can produce a very strong disincentive if the intruder has access to identified records for only a small subset of the population, as there may be no overlap between the two files—that is, no records included in both files. On the other hand, sampling will provide less of a disincentive for attempted re-identification if the intruder has data on the entire population, as the intruder can be nearly certain that every record in the public use microdata file is represented in the population data.

- **Perturbative methods.** With these methods, values in the microdata are distorted, but this is done in such a way that key statistical properties or relationships in the original data are preserved. These techniques include the following:
- **Noise addition:** random noise technique is to add or multiply the original value by random numbers

- **Data swapping:** selecting a sample of records, finding a match in the database on a set of predetermined variables, and swapping all other variables
- **Rank swapping:** unlike regular swapping, in which the match/pair is defined based on exact match, in rank swapping the pair can be defined to be close based on their proximity to each other on a list sorted by the continuous variable; frequently the variable used in the sort is the one that will be swapped
- **Shuffling:** like shuffling a deck of cards, the values of a confidential variable are reordered in a way that preserves the correlation between the confidential variable and a non-confidential variable while also preserving the correlation between the rank order of the confidential variable and that of a non-confidential variable in the original data
- **Rounding:** replace the original values of variables with rounded values
- **Resampling:** for a variable in the original data, a new variable for released data is created in which the values of this new variable are calculated as the average of a set of resampled values from the original variable
- **Blurring:** replacing a reported value (or values) by the aggregate values (for example, the mean) across small sets of respondents for selected variables
- **Microaggregation:** a form of data blurring in which records are grouped based on a proximity measure of all variables of interest, and the same groups of records are used in calculating aggregates for those variables; Domingo-Ferrer and Mateo-Sanz (2002) note that microaggregation provides a way to achieve *k*-anonymity with respect to one or more quantitative attributes
- **Post-randomization method or PRAM (Gouweleeuw et al. 1997):** a probabilistic, perturbative method for a categorical variable; in the masked file, the scores on some categorical variables for certain records in the original file are changed to a different score according to a prescribed probability mechanism, namely a Markov matrix
- **Micro agglomeration, substitution, subsampling and calibration, or MASSC (Singh et al. 2004):** this creates sets of identifying variables (called strata) to find records that might be at risk of disclosure (that is, unique records) and calculates a disclosure risk measure for each stratum (unique records are also assigned a disclosure risk associated with that stratum); an overall measure of disclosure risk can be calculated for an entire database by collapsing over the strata
- **Synthetic microdata (Rubin 1993):** Some or all of the variables in the original dataset are replaced with imputed (synthetic) variables developed from models based on the original data; while certain statistics or internal relationships in the original dataset are preserved, the synthetic variables do not characterize actual individuals

We note that because of swapping and other techniques, the Census Bureau has been willing to publish tabulations from the decennial long form that, for small geographic areas, included frequency counts as low as 1, even for sensitive variables such as income class.

Even when the microdata have been protected using one or more of these statistical disclosure limitation methods and are perceived to be safe for release, the risk of re-identification, in all likelihood, is still not zero. [13] Chapter IV discusses re-identification risk and its potential sources.

C. Recent Advances in Protecting Microdata

Much of the recent research on protecting microdata has focused on how the usefulness of the data is affected when methods of statistical disclosure limitation are applied. This topic is addressed in the next chapter. Research on ways to improve the protection afforded to public use microdata has addressed ways to enhance existing approaches rather than the development of entirely new approaches.

Singh (2009) proposes an enhanced version of MASSC that generalizes the risk measures used in altering the data to encompass cases with "partial risk," defined as having risk scores between 0 and 1. All records with nonzero risk are subject to treatment (that is, alteration of data values), but only a random subset is actually treated. Both disclosure risk and information loss are assessed in developing the final dataset.

Machanavajjhala et al. (2005) show limitations of *k*-anonymity in two situations: (1) one in which the *k* individuals are homogeneous with respect to particular characteristics, resulting in attribute disclosure; and (2) one in which the intruder possesses background knowledge that makes it possible to differentiate between the target individual and the *k*-1 other individuals. To overcome these limitations, the authors propose the concept of *l*-diversity, which requires that the values of sensitive attributes be well-represented in each group. Further work will focus on extending the concept of *l*-diversity to multiple sensitive attributes and to continuous sensitive attributes.

Efforts to improve the quality of synthetic data have received attention as well. Zayatz (2008) notes that this is one of three areas of current research on disclosure avoidance at the Census Bureau (the other two being the use of noise addition for tabular magnitude data and the development of a system for remote microdata analysis). The Census Bureau uses the synthetic method to produce two databases that incorporate data from administrative records and is also applying synthetic methods to produce group quarters microdata from the American Community Survey.

[13] Theoretically, a fully synthetic file has no risk of disclosure because none of the records corresponds to an actual person. Concerns about synthetic data focus almost exclusively on their usefulness for analysis, a concept discussed in Chapter IV. Nevertheless, if a synthetic file mimics the original data sufficiently closely, it can still reveal information about the individuals in the original data. In other words, if a synthetic file captures relationships in the original data so well that it is highly useful analytically, it may also carry some disclosure risk.

IV. ISSUES IN PROTECTING MICRODATA FROM DISCLOSURE

In the previous chapter we reviewed methods that federal agencies and other organizations use to protect the confidentiality of the data they release. In this chapter we examine what the literature tells us about disclosure risk and about an important side-effect of protecting data from disclosure: a reduction in the data's usefulness for research. Section A discusses the concept of disclosure, reviews some instances in which individuals have been re-identified in data released to the public (although not by the federal government), explores the potential sources of disclosure risk, discusses the legal environment, and outlines approaches to assessing disclosure risk. Section B discusses ways in which the utility of data is reduced by strategies to limit disclosure and, related to this, how the loss of information can be measured. A fuller discussion of these issues is presented in Appendix E.

A. Disclosure Risk

In a seminal paper on the protecting the confidentiality of data released to the public, Dalenius (1977) described the problem in the following terms: "access to a statistical database should not enable one to learn anything about an individual that could not be learned without access" (cited in Dwork and Naor 2010). The literature on disclosure distinguishes between identity disclosure and attribute disclosure (Duncan and Lambert 1989). An identity disclosure assigns a name to a record in a database while an attribute disclosure assigns a characteristic to an individual or small group of individuals. Identity disclosure implies attribute disclosure, but attribute disclosure can occur without identity disclosure. A database may reveal that all of the members of a particular subpopulation share a specific characteristic. If an individual is known to belong to this subpopulation, something is learned about the individual even though no record in the database can be assigned unambiguously to that individual. Research on protecting confidentiality in tabular data has recognized the risk of attribute disclosure and devoted considerable attention to it, but research on protecting confidentiality in microdata has focused on identity disclosure. For federal agencies in particular, the goal in protecting the confidentiality of microdata is to prevent the re-identification of records that have been released as anonymous.

Some confidentiality provisions in federal legislation interpret any disclosure of an individual identity—regardless of how it is accomplished—as a violation of the law. Other confidentiality provisions—for example, those in HIPAA—acknowledge that it is impossible to release data for which the risk of disclosure is zero. If the effort to prevent disclosure produced a very low risk of re-identification, that effort would satisfy the legal requirement for protecting the data, even if a breach of confidentiality occurred. The United Kingdom's National Office of Statistics does not consider a re-identification to be a disclosure by the agency if the breach required more than a reasonable amount of effort (Duncan et al. 2011, p. 28).

1. Re-identification of Individuals in Data Released to the Public

There are exceedingly few documented instances of the re-identification of individual persons in datasets that have been released to the public. None has involved a sample survey or a federal government database, and few have involved data that were protected by methods that would be considered rigorous by today's standards.

The most famous re-identification, which predated the HIPAA Privacy Rule and influenced its development, was Latanya Sweeney's 1996 re-identification of a substantial fraction of the records in a database of Massachusetts state employees discharged from hospitals (Cavoukian and Castro 2014). The records had been de-identified by removal of names, Social Security numbers, health insurance IDs, hospital names, doctors' names, and other obvious identifiers, but ZIP codes, sex, and date of birth had been retained because of their analytic value. Sweeney's re-identification used the values of these three variables obtained from a city voter registration list that was purchased for a nominal fee. The employees who were re-identified included the state's governor. With this work Sweeney (1997) demonstrated that the removal of explicit identifiers does not guarantee that records are anonymous—that is, unable to be associated with individual persons. Under-scoring the latter point, Sweeney (2000) estimated from 1990 census data that 87 percent of the U.S. population could be uniquely identified by the combination of 5-digit ZIP code, date of birth, and gender.[14]

El Emam et al. (2011) conducted a systematic review of known re-identification attacks on health data and other types of data. The review uncovered 14 re-identification attacks in which at least one individual was accurately re-identified. Of the 14 examples, 11 were conducted by researchers solely to demonstrate or evaluate the risk of re-identification. Notably, only 2 of the 14 involved databases that were protected in accordance with current standards. One of the two was a health database, consisting of records from a regional hospital that were protected with the HIPAA Safe Harbor Privacy Rules, and the rate of re-identification was found to be very low—just 0.022 percent, representing two persons—despite strong assumptions about what an intruder might know (see Kwok and Lafky 2011). Overall these results confirm the value of current best practices for de-identification but also indicate that there is merit in complementary legal protection, where possible. The study also highlights a need for better information on disclosure risk, which could be obtained from re-identification attacks on large databases protected with the best current methods.

Another example of re-identification, which received considerable attention in the media, was the re-identification of published Netflix rental histories from the movie reviews submitted by (identified) Netflix customers (see Narayanan and Shmatikov 2008). Although this example does not bear directly on the risks associated with federal data in general or health data in particular, it demonstrates what can be possible with data that are publicly available.

2. Sources of Disclosure Risk

To understand the potential sources of disclosure risk requires an awareness of who might attempt to re-identify records in federal microdata, what are their capabilities, and what are their resources, including what data they might use in their re-identification attempts, and what tools are available to assist them in doing so.

[14] A replication of Sweeney's calculations with 2000 census data found that the percentage of the population that could be uniquely identified with these same variables had fallen to 63 percent (Golle 2006).

a. Potential Intruders

Terms such as intruder, adversary, attacker, and snooper have been applied to describe the individuals who might attempt to re-identify entities in public use data and apply that information in some, possibly malicious way. We will stick with the term intruder.

Nearly all of the documented instances of records being re-identified in public use data have been accomplished by researchers—and for the purpose of demonstrating data vulnerabilities. Researchers have incentives in the form of publication and possible career enhancement aside from contributing to a public good—namely, better protection of public data in the future. Other potential intruders include hackers, persons with access to proprietary information, neighbors, family members, and former spouses. Protecting data against re-identification by family members is a particular concern when data were collected confidentially from other members. The example discussed in Appendix E involves drug use by children. A different kind of threat is presented by former spouses, who may be able to realize a financial benefit by learning of a former partner's finances and may have extensive information on which to base a re-identification.

b. Auxiliary Data

Following Dwork and Naor (2010), the data that a potential intruder would use to re-identify records on a public use file may be described as auxiliary data. The challenge in protecting a public use file from any possibility of re-identification is the inability to guarantee that there are no auxiliary data in anyone's possession that would enable re-identification of even one record.

Voter registration lists—such as the one used by Sweeney in Massachusetts—are often cited as a key source of auxiliary data for potential intruders. Such data include identifiers along with demographic and residential characteristics that uniquely identify a lot of people. Benitez and Malin (2009) estimated the risk of re-identification from voter registration lists by state for datasets protected by the HIPAA Safe Harbor and Limited Dataset policies.[15] Because voter registration lists vary in cost, the study addresses both the probability of successful re-identification, given the attempt, and cost factors that may affect the likelihood of an attempt. In their study, the Safe Harbor dataset included year of birth, gender, and race while the Limited Dataset policy added county and date of birth. The results showed wide variation in estimated risk and in the unit price of each potential re-identification by state, with substantially greater risk under the Limited Dataset versus the Safe Harbor policy. They concluded that blanket protection policies expose different organizations (in different states) to differential disclosure risk.

Duncan et al. (2011) observe that "the Achilles' heel for data stewardship organizations in dealing with the practicalities of risk assessment is their lack of knowledge about what data snoopers know." While much is known or can readily be determined about the contents and coverage of certain public databases maintained by the states, the same cannot be said about the data that are compiled, maintained, and resold by commercial entities. A recent study by the U.S. Government Accountability Office (2013) concluded that "the advent of new and more advanced

[15] Limited datasets are not intended to be released as public use files but through licensing or other restricted arrangements.

technologies and changes in the marketplace for consumer information have vastly increased the amount and nature of personal information collected and the number of parties that use or share this information." The report provides examples of the types of information collected, some of which are discussed in Appendix E.

Although the data that are "out there" in the public domain or held by private sources may be considerable, the threat that it presents is mitigated to at least some degree by discrepancies between the values recorded in these data sources and the values reported by survey respondents or collected by administrative agencies. Such "data divergence," as described by Duncan et al. (2011), includes not only measurement error but conceptual differences in the way that data elements are defined in different sources. Timing is another factor in data divergence. The data accessible to a would-be intruder and the information reported in federal datasets may be separated by years, which can matter a great deal for health conditions, income, and even geographic location.

c. Record Linkage

One of the most important tools available to the more sophisticated potential intruders is record linkage methodology. When two files contain some of the same individuals, the records common to the two files can be linked if the two files also contain some of the same variables. When the two files contain the same unique and valid numeric identifiers, the records can be linked using "exact matching" on those fields—as is commonly done when files contain Social Security numbers. When the conditions for exact matching are absent but other, non-unique or imperfect identifiers are present, either "probabilistic record linkage" or distance-based matching can be used as an alternative. Probabilistic record linkage separates all possible combinations of records into likely matches, likely non-matches, and a group that cannot be confidently assigned to either and would require a manual "clerical" review to determine in which category they belong. Probabilistic record linkage is often applied to link records based on names and addresses, where duplicate names and spelling errors are possible. Distance-based matching may be used instead of probabilistic record linkage when one or both files contain no explicit identifiers but the two files contain quantitative variables—such as income.

While record linkage is typically applied using variables that appear in both files being linked, it is also possible to use record linkage methods to link files that have no variables in common. For example, one approach relies on correlations between variables. The effectiveness of such matching increases with the strength of the correlation between variables in the two files and the degree of overlap between the two files—that is, the percentage of records in each file that appear in the other file.

3. The Legal Environment

As we reported in Chapter II, there are extensive federal regulations designed to protect the confidentiality of the individuals and organizations whose private information is reported in federal databases. The laws regulating the sharing of federal data place much more responsibility upon the data producer than the user. Many of these laws specify severe penalties for agency employees in the event that disclosures occur, but these penalties rarely extend to the individuals outside the agency who are actually responsible for the disclosures. A 2005 National Academy of Sciences (NAS) panel report on expanding access to research data notes that "at present, the

obligation to protect individual respondents falls primarily on those who collect the data, thereby creating a disincentive for providing access to other researchers" (National Research Council 2005). The NAS panel addressed two recommendations to this problem:

> Recommendation 7. All releases of public-use data should include a warning that the data are provided for statistical purposes only and that any attempt to identify an individual respondent is a violation of the ethical understandings under which the data are provided. Users should be required to attest to having read this warning and instructed to include it with any data they redistribute.
>
> Recommendation 8. Access to public-use data should be restricted to those who agree to abide by the confidentiality protections governing such data, and meaningful penalties should be enforced for willful misuse of public-use data.

Achieving these objectives—particularly the second—would require new legislation authorizing agencies to impose penalties.

4. Assessing Disclosure Risk

A useful way to view disclosure risk was expressed by Marsh et al. (1991): the probability of disclosure is the product of two terms: (1) the probability of a successful re-identification conditional on someone trying to re-identify a record and (2) the probability that someone will try to re-identify a record. A data producer can lower the risk of disclosure by reducing either of these probabilities. For example, charging a high fee for a public use file reduces the probability that a potential intruder will even acquire the file. Sampling reduces the certainty that someone of interest is included on the file, which will also discourage potential intruders. Altering the data values in various ways reduces the likelihood of a re-identification, and publicizing the fact that such measures were applied may also discourage attempts at re-identification. Altering the data also reduces the potential value of the information gained by re-identification, which may further reduce the likelihood that a would-be intruder will attempt a re-identification.

When microdata protection is based on *k*-anonymity, the assessment of disclosure risk involves determining if *k*-anonymity is satisfied. Ideally, this is done with population data, but strategies applicable to sample data exist, as noted in Chapter III.

When public use files contain numerous variables or include continuous variables, sample uniqueness across the range of variables is almost assured. Under these circumstances a different approach to assessing disclosure risk is required. Commonly, this involves using one or more alternative files with identifiers and attempting to match records on the public use file. The accuracy of unique matches can be measured and, depending on the results, the data producer may decide to exclude high-risk records from the public use file or increase the level of masking on these records to prevent matches. If the accuracy of unique matches is sufficiently low, and there is no indication that correct matches can be differentiated from the vastly greater number of incorrect matches, the data producer may conclude that deleting or further masking the records that were matched correctly is not necessary. However, correct matches from publicly-available data do provide direct evidence of vulnerability.

The most rigorous way to assess disclosure risk is to attempt to identify records in the public use file from the source records in the original or internal file. The rigor in this approach comes from two factors. First, the only data divergence between the public use file and the internal file is that which was created deliberately to reduce disclosure risk. Second, unless the public use file was subsampled from the internal file, the two files will contain the same records, which is analogous to an intruder knowing with certainty who is included in the public use file. Because these factors can make re-identification rather easy, data producers using this approach must introduce some limitations on the match attempt to produce a realistic assessment of risk. Typically, this involves first determining what variables from the internal file might be available to a potential intruder, as it is likely that all or nearly all of the records in the public use file could be re-identified if all of the variables appearing in both files were used in the attempt. Data producers also need to account for the impact of sampling if the internal file is itself a sample from a larger population.

B. Maintaining the Utility of Public Use Data

Steps taken to preserve the confidentiality of public use data have an adverse effect on the quality of the data and its general usefulness for research. Purdam and Elliot (2007) classify the impact of statistical disclosure limitation on data utility into two categories: (1) reduction of analytical completeness and (2) loss of analytical validity. The former implies that some analyses cannot be conducted because critical information has been removed from the file. The latter implies that some analyses will yield different conclusions than if they had been conducted on the original data. For example, suppressing state geography, as is done for some national databases, precludes analysis of characteristics by state. Adding noise to variables reduces the degree of fit in predictive models and lowers simple measures of association. Swapping, if not monitored carefully, can distort distributions, as was demonstrated recently with multiple Census Bureau household surveys (Alexander et al. 2010).

Preserving the utility of the data is a prominent topic in the statistical literature on disclosure limitation but much less so in the public discussion of data security. In the statistical literature, research has focused on measuring the information loss due to the application of the protective measures. As a general principle, statistics computed from the protected dataset should not differ significantly from the statistics obtained from the original dataset. An approach to measuring information loss is to compare statistics—totals, means, medians, and covariances—between the public use data and the source data. Some of the methods of statistical disclosure limitation discussed in the previous chapter have been shown to protect certain statistics—in particular, totals—or to introduce less distortion into covariances than other methods. Consider, for example, top coding. One can assign top codes in such a way that the original totals are preserved (by assigning the mean of the top coded values as the top code), but this benefit will not extend to variances, which will be reduced.

Shlomo (2010) reviews several approaches to measuring information loss. These include distance metrics, impacts on measures of association, and impacts on regression analyses. Because statistical disclosure limitation introduces error into the data, measures of goodness of fit may capture information loss particularly well. Depending on how disclosure limitation affects the data, the error added may reduce between-group variance and increase within-group variance in regression analysis or ANOVA. Alternatively, it is possible that disclosure limitation

may artificially increase between-group variance, creating more association than was present in the original data. Calculating a range of information loss measures will enhance the data producer's understanding of the impact of disclosure limitation on the analytic utility of the data. Shlomo (2010) argues for a coordinated analysis of disclosure risk and information loss by data producers in order to maximize the analytic utility of the public use data consistent with providing the desired level of protection.

This page has been left blank for double-sided copying.

V. EXPERT VIEWS

The technical expert panel that was held on June 27, 2014 included two panel discussions: What Are the Re-identification Threats to Releasing Federal Data to the Public? and Good Practices for Protecting Public Use Data (see Appendix C for a detailed summary of the TEP presentations and discussions). In this chapter we summarize highlights of the experts' presentations and the ensuing discussion.

A. What Are the Re-identification Threats to Releasing Federal Data to the Public?

Panelists in this third session on the day's agenda included:

Khaled El Emam, University of Ottawa and Privacy Analytics
Brad Malin, Vanderbilt University
Latanya Sweeney, Federal Trade Commission and Harvard University
Denise Love, National Association of Health Data Organizations (NAHDO)
Daniel Barth-Jones, Columbia University
Steve Cohen, AHRQ (moderator)

The session included both presentations by panel members and discussion.

1. Panel Presentations

Khaled El Emam noted that de-identification has been simplified through automation. The process of de-identification in practice involves assessing risk, classifying the variables in the file, and mapping the data. These contribute to specifications in an automated anonymization engine through which the original data are run to produce the anonymized data for release.

Adversaries (that is, those who might re-identify the data) may include academia, the media, a person's acquaintances, the data recipient, and malicious actors. Interestingly, there is no apparent economic case for malicious re-identification of health data; the bigger concern is the media.

There are direct identifiers and quasi-identifiers. Examples of direct identifiers include name, address, telephone number, fax number, medical record number, health care number, health plan beneficiary number, voter identification number, license plate number, email address, photograph, biometrics, Social Security number, device number, and clinical trial record number. Examples of quasi-identifiers include sex, date of birth or age, geographic locations (such as postal codes, census geography, information about proximity to known or unique landmarks), language spoken at home, ethnic origin, total years of schooling, marital status, criminal history, total income, visible minority status, profession, event dates, number of children, high-level diagnoses and procedures.

An identifier must satisfy three general criteria: it must be replicable, distinguishable, and knowable. Replicable means that the identifier is sufficiently stable over time and has the same values for the data subject in different data sources (for example, blood glucose level is not replicable, but date of birth is replicable). A potential identifier is distinguishable if there is

sufficient variation in the values of the field that it can distinguish among data subjects (example: a diagnosis field will have low distinguishability in a database of only breast cancer patients but high distinguishability in a claims database). An identifier must be knowable by an adversary, and how much an adversary knows will depend on whether the adversary is an acquaintance of the data subject or not. If an adversary is not an acquaintance, the types of information that are available include inferences from existing identifiers, such as date of hospital discharge at birth and public data such as voter registration lists.

Determining risk is a solvable computational problem. Make assumptions about the knowledge of the adversary and how many quasi-identifiers it has, consider all combinations of these, and then manage the risk for every combination.

Some special types of data require specialized techniques. There are good techniques to de-identify geo-spatial information (including movement trajectories), dates and long sequences of dates (for example, transactional data), and streaming data—that is, data that is continuously being updated.

If de-identified properly, open data is not particularly useful for further attacks because it has no identifiable information, and the success rate of linking these data to other data should be small. Decent data can be created for public release, and we can add terms of use or conditions in order to release higher quality data.

Brad Malin described the de-identification system for DNA sequence data that his team constructed. The database contains 2 million patients and biospecimens for 200,000 patients, and the data is being used by 200 researchers (subject to a DUA with the National Institutes of Health).

His team published a paper in June 2014 on a probabilistic model for patient disclosure based on estimating population uniqueness across datasets (Sattar et al. 2014). One needs to be cognizant of data over time: if a data holder anonymizes someone in different ways at different points in time, this may actually make that person easier to identify.

Research has shown the variety of characteristics and behaviors that can distinguish an individual. These characteristics and behaviors include demographics, diagnosis codes, lab tests, DNA, health survey responses, location visits, pedigree structure, movie review, social network structure, search queries, Internet browsing, and smart utility meter usage. A study he conducted found that re-identification risk was substantially greater for a HIPAA limited dataset than a dataset protected with HIPAA Safe Harbor methods.

A simplified view of risk is that the probability of re-identification is approximately equal to the product of the probability of an attack, and the probability of a re-identification conditional on an attack. Deterrents to attack include DUAs, access gateways, unique login IDs and passwords, and audits. Data characteristics that affect the conditional probability of a re-identification include uniqueness, replicability, availability, and cost.

Latanya Sweeney began her remarks by noting that this conversation is not much different than it was in 1997, but the world has changed a lot since then. The Data Privacy Lab at Harvard University initiated the DataMap project (thedatamap.org) to document where personal health

data goes outside of the doctor-patient relationship. Maps show the flow of data from the patient to various entities and from the physician and hospital back to the patient. Flows that do not directly involve the patient are numerous, and less than half of the documented data flows are covered by HIPAA, including inpatient discharge data transmitted without explicit identifiers.

A study she led found that only three of the 33 states that sell or share de-identified versions of their hospital inpatient discharge data are using HIPAA standards to protect the data. In a separate study, her team purchased a public use version of patient-level hospital discharge data from Washington State, and using accounts of accidents published in newspapers in 2011, they was able to re-identify 43 percent of a sample of 81 accident victims in the hospital discharge data based on characteristics reported in both sources.

With colleagues she submitted a FOIA request to determine who are the buyers of publicly available health data, and found that predictive analytic companies are the big buyers. They are producing data products that exploit publicly available health data.

There are four ways to add transparency to the system: (1) public notice of privacy breaches should be required; (2) data holders should be required to list publicly those with whom they share data; (3) each person should be able to acquire copies of their personal data from any entity holding their data; and (4) each person should also be able to acquire an audit trail of the of the organizations with which the data was shared.

Re-identification is a key part of the cycle of improving the protection of data. We improve protective techniques only after protections fail. For example, encryption techniques have improved because they were used, problems were identified, and better techniques were developed. We now have strong encryption, and we need the prevention of re-identification to advance to that stage as well.

Denise Love explained that NAHDO has been involved for years in discussions regarding these issues with states. The state data agencies have come up solutions to balance transparency and confidentiality.

The state inpatient discharge and all-payer claims data systems are essential to public health and multiple other purposes, including public safety, injury and disease surveillance, health planning, market share analyses, quality assessments and improvement, and identification of overuse/underuse/misuse of health care services.

There is a critical “iron triangle” to public data, representing three principles of data policy: transparency, data utility, and data safety. There must be a balance among all three. Over-emphasis on any one of the three does not serve the public good.

DUAs can mitigate the risk of inappropriate use. The Washington state story is the first breach that we’ve ever heard about. NAHDO spent a year developing guidelines for data release by states, which was published in January 2012, but Washington State was not following these guidelines.

Daniel Barth-Jones discussed his recent work using uncertainty analysis through a flow chart that lays out several components including intrusion scenarios and information on what variables are needed by an intruder for re-identification. Adding an uncertainty distribution at each step of the flowchart gives a sense of how the data protection and disclosure avoidance techniques can reduce re-identification risk.

Intrusion scenarios include a "nosy neighbor" attack, a mass marketing-type attack to re-identify as many individuals as possible for marketing purposes, and a demonstration attack by a researcher in academia or a journalist. There could be as many as 3,000 potential variables/data elements. However, since most often the data is not necessarily accurate and the intruder cannot build a complete population register, there are often false positives. Each step in the flow chart has a probabilistic distribution—then you can sample across the scenario with a hyper-grid multiple times, which gives us a robust idea of the re-identification risk. There are dependencies at each step in the chain to determine the economic motivation or benefit to the entity.

It is important to consider the impact of de-identification on statistical analysis. Poorly implemented de-identification can distort multivariate relationships and hide heterogeneities. Data reduction through sampling and other means can destroy the ability to identify heterogeneity among the races, or by educational level, for example.

A forthcoming paper by T.S. Gal et al. evaluates the impact of four different anonymization methods on the results obtained from three different types of regression models estimated with colon cancer and lung cancer data. For each combination the authors calculated the percentage of coefficients that changed significance between the original data and the anonymized data.

HIPAA lacks a penalty if data is re-identified by the user, even if these are false positives; currently there is no cost for false positive identification. We need to change the cost for false positive identification to change the economic incentives for efforts at re-identification.

2. Discussion

Moderator **Steve Cohen** identified the following themes during these presentations: game theory, data resources that allow for a breach, and how to simulate the threat of disclosure. He asked the panelists to address where we are heading in the next five years to address these threats.

Malin responded that social media is a serious threat: people self-disclose and disclose about others. His team is doing research on how Twitter is used, and preliminary findings are that people talk more about others than about themselves, and it is a minefield of potential disclosures (for example, "pray for my mom, she has breast cancer"). Another challenge is that electronic health records are becoming commercialized, and start-ups are using data without regulation, which is a big loophole.

Sweeney added that no one is really studying the predictive analytics industry, so we don't know how big an industry it is. Re-identification is a way of illustrating risk—it's big although unquantified – we don't know how much really goes on, because DUAs don't stop it, they just hide it because the penalties are so draconian. Federal agencies should try to figure out how to link data in a secure way in the cloud to produce aggregated data for the public.

Barth-Jones stated that the future concern is harm from bad de-identification practice—from bad science and inefficiency. We should focus on reducing bad de-identification practices.

Love is concerned that data will be too protected, and opt-in/opt-out will be disastrous for public health and for population health (an example is when parents do not vaccinate their children).

El Emam noted that techniques are becoming more sophisticated, including protection of data. Risks can be managed with appropriate de-identification practices.

Malin recommended that data holders have a dialogue with the community regarding use of data for research purposes. They should create an advisory board and keep them in place, and make them partners. This will reduce the risk of research being shutdown in the event of a breach.

B. Good Practices for Protecting Public Use Data

Panelists in this fourth session included:

Mark Asiala, Census Bureau
Barry Johnson, Statistics of Income Division, Internal Revenue Service
Allison Oelschlaeger, Centers for Medicare & Medicaid Services
Eve Powell-Griner, National Center for Health Statistics
Fritz Scheuren, NORC at the University of Chicago
Connie Citro, Committee on National Statistics (Moderator)

1. Panel Presentations

Mark Asiala explained that public use files that include microdata are only one part of a "suite" of data types released by the Census Bureau. Other types include tables produced from aggregated data for low levels of geography, special tabulations, and research papers.

The potential threats that the Census Bureau faces include an ability to identify individuals by using the tables directly, matching external data to public use files, or using data products in combination.

The Bureau's strategies for protecting data from disclosure vary with the type of data. To reduce disclosure risk for tables, they alter the table design and use combinations of data swapping and partially synthetic variables on the source files. For public use files, they apply size thresholds for geographic and category detail; noise addition for some variables; and additional data swapping and/or partial synthesis of data. Rounding is the primary strategy in special tabulations and research papers. The Bureau prefers to minimize the use of suppression techniques because they harm the utility of the data, and he recommends that data holders think about whether they can mask a particular characteristic rather than an entire record.

For tables, the granularity of data cells raises the risk of re-identification—too much detail leads to a "pseudo-microdata" file. A good rule of thumb is not to publish tables with more than 100 cells. Treating the records at risk before producing tabulations is preferable to having to suppress cells.

Strategies used for public use files include sub-sampling, thresholds for identification of geographic areas and categories, additional data swapping for "special uniques," noise infusion, and synthetic data. The threshold for identification of geographic areas is 100,000 population size, and the threshold for categories is 10,000 nationally. A "special unique" case will stand out even with a large sample size, so additional swapping is done for such cases.

For special tabulations and research papers, the Bureau rounds the data to protect small cells and coarsens the detail. In some cases they impose a hard rule such as publishing no detail of a given characteristic below the county or state.

The Census Bureau is working on a microdata analysis system that allows tabulations off the entire data file, but with certain restrictions/protections, as an alternative to public use files. They are also considering creating a bridge between public use files and research data centers to find a middle ground between these approaches.

Barry Johnson discussed the role of the statistics arm of the IRS, which has data from tax returns but no survey data. IRS public use data has been the core of tax and economic modeling for the Congressional Budget Office, the Urban Institute, and the National Bureau of Economic Research. The IRS works with the Federal Reserve Board to plan disclosure protection of the data collected in the Survey of Consumer Finances.

Tax data is releasable because there are not many demographic pieces of data on the 1040 form, and this makes the intruder's job more difficult. Their data is constrained by accounting rules, so it is difficult to perturb, and because of the alternative minimum tax rules and other complexities, it is important to preserve these accounting relationships in the data. IRS removes obvious identifiers, and relies on a portfolio of techniques to protect especially vulnerable records and variables.

IRS works with experts in disclosure limitation and with fellow agencies to protect data/variables, and based on these reviews updates the individual tax public use file regularly, and evaluates how effective the changes have been. Having access to the full population dataset makes evaluation/simulation effective (because the public use file can be matched to the population data to assess risk).

Allison Oelschlaeger commented that CMS has mostly administrative data. CMS formed the Office of Information Products and Data Analytics a few years ago to maximize data for internal and external users. CMS produces two types of de-identified data products: (1) stand-alone public use files of basic Medicare claims data, with direct identifiers removed and careful review of indirect identifiers; and (2) a synthetic file, which is a good way for researchers to develop expertise before doing research with the actual data. CMS has launched a "virtual RDC." Researchers can submit a research protocol, conduct the research, and then any outputs are reviewed/cleared by CMS; researchers do not have to satisfy security requirements at their own facilities this way.

Eve Powell-Griner pointed out that most of NCHS's data is survey data, but they also offer vital statistics records and physical exam data. NCHS relies on Statistical Policy Working Paper 22 and standard limitation techniques, and in addition thinks about disclosure limitation from the beginning of the process and discusses any data issues with their review board.

NCHS is becoming somewhat more conservative in what is being released in public use files (for example, geography fields). None of NCHS data is inaccessible except for personally identifiable information. They focus on rare characteristics that would be identifiable and are sensitive to rare information fields. NCHS has deployed new software to extend risk assessment and assign a probability of disclosure, and a priority is to keep the genetic data collected in the National Health and Nutrition Examination Survey under tight control.

Fritz Scheuren stated that one advantage data holders have is the variety of disclosure prevention techniques available, but on the other hand the disadvantage is the extent of the variety available. He said there is a "civil war" going on between the data quality people and the information quality people. The user cannot rely on tables; they want to use it in a microdata simulation model. Regarding disclosure prevention, he is concerned that data holders are not keeping up with the prey-predator problem: whatever federal agencies do to protect the data will eventually be defeated.

2. Discussion

Johnson noted that even revealing that a person filed a tax return is considered a disclosure by the IRS, so they set a high bar to prevent disclosure. IRS balances transparency and confidentiality by working in cooperation with the users. They formed a user group, and ask them to help make choices. Two outside users helped develop the updated version of the public use file, which increased utility and strengthened protection, and helped justify removing the geographic variable.

Oelschlaeger commented that, recently, CMS has focused on aggregated files rather than de-identified files. The stand-alone de-identified public use files are so focused on removing all variables that could lead to re-identification that the files might be considered as useless. CMS has a number of ways to share data, only one of which is for researchers. Commercial researchers receive HIPAA Limited Datasets. The Qualified Entity program in the Affordable Care Act gives CMS the authority to release data for quality improvement/performance measures. Any disclosure of identifiable data requires a DUA.

Scheuren observed that the mosaic effect comes into play when someone extracts data and tries to match it to another dataset. Billions of records in the insurance world are used for data mining. This is largely a good thing, but there are downsides, such as hackers. He noted that we have a trust system, but we need a trust-but-verify system. There are three things to do: penalize people, enforcement, and scale back overzealous confidentiality.

Johnson commented that the IRS needs a legislative change to allow a DUA that would put the responsibility on users. Right now, the data the IRS releases must be completely safe, or it cannot be made available.

This page has been left blank for double-sided copying.

VI. SYNTHESIS AND CONCLUSIONS

The Open Data Initiative launched by the Executive Office of the President and OMB has encouraged the release of increasing numbers of datasets containing individual records (microdata) collected or sponsored by federal agencies from survey respondents, doctor and hospital visits, and medical claims. At the same time, federal agencies that release data collected from individuals and establishments have an obligation under the law to protect the confidentiality of those supplying the data as well as any information provided that could disclose the identity of individuals. The challenge faced by HHS and other federal agencies is to achieve an appropriate balance between providing the public with useful datasets and protecting the confidentiality of the individuals and establishments whose information is contained in the data.

Information in an individual dataset, in isolation, may not pose a risk of identifying an individual, but when combined with other available information, may pose such a risk. As stated by OMB in M-13-13, before disclosing potential identifiable information, "agencies must consider other publicly available data—in any medium and from any source—to determine whether some combination of existing data and the data intended to be publicly released could allow for the identification of an individual or pose another security concern" (OMB 2013a). The concern is that the datasets being released in large numbers—by mid-2014 more than 1,000 datasets had been released by HHS alone (HHS 2014)—may provide the pieces of information that, in combination with other publicly available data may disclose information that the federal government is required to maintain as confidential.

ASPE established this project in order to better understand how federal agencies are meeting the challenge of releasing more and more data while simultaneously maintaining the confidentiality of those who provided the data. The goals of this project were: (1) to obtain a balanced, scientifically sound assessment of the mosaic effect, (2) to identify any unique increased risk associated with the mosaic effect, and (3) to identify data release policies and best practices that can prevent or reduce the risk of disclosure through the mosaic effect.

In assessing whether there is an increased risk that the Open Data Initiative may create or incur through the mosaic effect, we and our expert panelists reviewed what is known about the sources of disclosure risk and the effectiveness of various ways to control such risk. From the discussion at the TEP meeting and the materials we reviewed in producing the two background papers, we have prepared a synthesis, which is presented here. We close this final chapter with some concluding observations that will assist federal agencies in going forward as they comply with open data policies while maintaining the confidentiality of the data they release.

A. Synthesis

Data collected by the federal government are covered by a substantial array of regulations intended to protect the confidentiality of the data and the privacy of those from whom such data are obtained. While remaining attentive to these requirements, federal agencies have been able and willing to provide researchers across a wide range of disciplines with open access to public use versions of their data. While some users can accept forms of restricted access in return for more extensive content or an ability to link records across datasets, microsimulation modeling

and quick turn-around policy analyses are among the many applications that require unrestricted access to their data, which only public use files can provide.

There have been no documented breaches of federal public use data—survey or administrative. And while occasional, documented breaches involving non-federal data have received a lot of attention, these breaches have been confined almost entirely to data that were not protected in ways consistent with current standards.

Despite this strong track record, views on the adequacy of de-identification or anonymization strategies cover the spectrum. From panelists we heard statements to the effect that, on the one hand, disclosure risks can be managed, and properly de-identified data can be released with little risk; while on the other hand, whatever we do to protect the data will eventually be defeated, and given enough time, effort, incentive, and money, some records may be re-identified. Such statements should not be taken out of context, as the assumptions that underlie them are important. There was general agreement; however, that the disclosure risk cannot be driven to zero without seriously weakening the analytical value of a dataset—an outcome that we will revisit below.

Sources and Evidence of Disclosure Risk. In general, the removal of direct identifiers is not sufficient to de-identify a dataset, as some of the remaining variables can serve as indirect or quasi-identifiers. In combination, they can uniquely identify individuals within the population or well-defined subpopulations. Voter registration files, which typically contain name, address, gender, and date of birth, have figured prominently in a number of the past breaches reported in the literature because a substantial proportion of the U.S. population bears a unique combination of gender, date of birth, and ZIP code.

In the wake of earlier, well-publicized re-identifications and federal legislation (specifically HIPAA), even non-federal files rarely report the combinations of characteristics that would permit large-scale re-identifications from voter registration records. Nevertheless, only 3 of the 33 states that sell inpatient discharge data are applying HIPAA standards. Recently, Sweeney was able to re-identify a substantial proportion of a small subsample of individuals in hospital discharge data released by the State of Washington, based on information compiled from newspaper accounts. It must be noted, however, that the Washington State data did not comply with widely used NAHDO standards, which were developed to improve confidentiality protections in the high volume of claims data released by states and other organizations. This episode underscores the potential vulnerability of public use data to the kinds of information becoming more widely accessible through Internet searches and web scraping techniques.

While the Washington State data may represent an exception among data distributed by states, many sales or transfers of personal data—particularly health data and financial data—are neither regulated by government nor held to professional standards designed to minimize disclosure risk. Consequently, such data may be especially vulnerable to re-identification, and such breaches can tarnish the movement toward greater transparency.

The Mosaic Effect and Related Threats. With regard to the mosaic effect and its implications, there was general agreement among the TEP participants that the public use files released by HHS and other federal agencies bear limited risk of disclosure, even in combination

with other publicly available data. If de-identified properly, the public use data released by federal agencies are not useful to the intruder. On the whole, the confidentiality of federal datasets is threatened less by the release of other federal data than by data from two other sources: (1) datasets compiled by other organizations and released with weak or no de-identification, and (2) personal data revealed through social media. While agencies can control what they release, they cannot control what other organizations and private individuals release. Federal agencies recognize the growing threat from external data and are actively engaged in assessing and responding to the disclosure risks that they pose.

The explosion of personal information posted to the Internet through social media and other avenues means that the availability of data that might be of use to a potential intruder has grown as well. Such information does not cover the entire population, however. Its coverage is far less extensive than voter registration records, for example, and some investigators have noted the impact of incomplete coverage.

If the confidentiality of a dataset has been breached to the extent that many of the records have been correctly re-identified, then all of the variables contained in that dataset for these named individuals become available as potential indirect identifiers that could be used to break into another dataset containing some of the same individuals and some of the same variables. This is an unlikely scenario for a federal database; in actuality, the threat posed by the re-identification of records in a single database covering a narrow subset of the population is likely to be very small, given the limited number of records involved and their minimal overlap with records released by the federal government—particularly from sample surveys.

The release of well-protected federal files does not appear to increase the re-identification risk for other federal files; however, a number of agencies are conducting informed risk assessments. The sheer volume of data files made accessible through the Open Data Initiative is striking. Are there large numbers of individuals who appear repeatedly because of the way that file universes and samples are defined? This may be difficult if not impossible to determine, given restrictions on sharing or linking personal data across agencies, but there is no question that risk assessment would be enhanced with such information.

Protecting Public Use Data. To maximize their effectiveness, statistical disclosure limitation methods must be tailored to the data they are being used to protect. Each dataset faces unique risks, depending on the type of data, the population covered, the sample design, the variables that require the most protection, and the distribution of values on these variables.

Statistical Policy Working Paper 22 has provided valuable documentation of federal agency practice in protecting the confidentiality of federal data. However, the last update of this important resource was nearly 10 years ago, and agencies have upgraded their statistical disclosure limitation methods since then. TEP panelists asserted that regular updates—perhaps as often as every five years—would help to ensure that the document remains current.

A useful way to represent disclosure risk is that the probability of a re-identification is equal to the product of the probability of an attack, and the probability of a re-identification conditional on an attack. This implies that disclosure risk can be lowered by strengthening the protections applied to public use data or by taking steps that reduce the likelihood of an attempted re-

identification. An important element in the security of data protected with rigorous methods may have been the absence of serious re-identification attacks. That no breaches of federal data have occurred to date may also be due in part to the fact that incentives were not high enough to inspire serious efforts to challenge the disclosure protections, although the protections themselves may have contributed to reducing these incentives.

Working in the opposite direction, however, the penalties to which data users are subject for attempting to re-identify records in public use files are light to non-existent. For most agencies the onus falls entirely upon the data producers in the event of a re-identification. This creates a tension for agencies releasing public use files in order to comply with Open Data policies and initiatives.

While de-identification has occupied much of our attention, the confidentiality of public use data is reinforced in other ways. Sampling is an important tool for reducing disclosure risk. This speaks to the security of the federal government's many public use files of sample survey data. At the same time, however, the added protection afforded by sampling underscores the inherent risks in creating public use files from administrative records, which may not be sampled at all or are sampled at much higher rates than is typically found in surveys.

Reporting error, which can be particularly high in sample surveys, also provides protection against disclosure. For some variables in some databases, the application of additional protection in the form of masking may not be needed and may only reduce the quality of the data even further.

More generally, protection versus utility is the key trade-off. The effects of disclosure limitation methods on the quality and general analytic usefulness of public use data are an important concern. Over-zealous confidentiality protection can hamper data quality with no material improvement in data security. Moreover, it is possible to apply such strong protections to public use data that they become useless—and unused. To guard against this outcome, some agencies consult with subject matter experts and their major data users to better assess the trade-offs between analytic utility and effective disclosure limitation.

Secure remote access to restricted (not public use) data is growing as a solution to the problem of maintaining quality while protecting the confidentiality of the most sensitive data, and permitting access to data for applications that require specific indirect or even direct identifiers—for example, for analyses requiring linkage of files. Secure remote access is popular because it provides more convenient access than an RDC. In the end, however, neither of these options will replace public use data, as critical applications of public use data cannot be served by these alternatives, and the rapid expansion of the Open Data Initiative and related efforts only underscores the importance of public use data.

False re-identification is generally not addressed in regulations, yet it may present a more serious problem, potentially, than positive re-identifications. From a technical standpoint, an agency can protect against a positive (or true) re-identification but not a false re-identification. Furthermore, agencies cannot deny alleged re-identifications except to reiterate that a re-identification is not possible, as an explicit denial may reveal information that assists an intruder with a correct re-identification.

Evaluation of the effectiveness of disclosure limitation methods by attempting to re-identify records internally remains the most powerful approach to establishing that a public use file is secure. A number of agencies obtain external, identified data—both public and commercial—to use in their evaluations. The use of external data in attempts to re-identify records in a preliminary public use file directly addresses the potential threat that such data poses. Such efforts can produce even stronger tests when performed by experienced external consultants, who can provide a fresh approach that is not influenced—or encumbered—by detailed knowledge of how the public use data were created.

Frontiers of Research. Research is providing important enhancements to risk assessment and the ability to assign probabilities to disclosure risk. Disclosure risk and intrusion can be modeled; this has been done in a variety of ways by different investigators. Much of the recent research on protecting microdata has addressed the impact of statistical disclosure limitation on the analytic utility of the data. In particular, research has focused on measuring the information loss due to the application of disclosure limitation measures. More recent research is exploring the problem of maximizing utility while minimizing risk. Lastly, a prominent topic of recent research on statistical disclosure limitation is improving the quality of synthetic data.

B. Concluding Observations

Because federal agencies have worked hard to develop, maintain, and update their procedures for protecting the confidentiality of their public use data, releasing multiple, well-protected files under the Open Data Initiative and related programs does not appear to produce a significant increase in disclosure risk. A greater threat comes from the personal information that individuals reveal about themselves and others through social media, as this information is identified. An intruder cannot know when a set of characteristics uniquely identifies that individual, however. In general, incomplete population coverage reduces the threat—just as it does with voter registration records. Files released with inadequate protection by states, local areas, and commercial organizations pose a threat as well if large numbers of records can be re-identified. The few examples discussed earlier in this report indicate that such public files are not common, however, and their rarity and generally small size limit the threat they present.

Skilled hackers present more of a concern because of their potential ability to break into nonpublic databases and obtain access to data that has not been protected by the removal of identifiers and the application of more sophisticated disclosure limitation techniques. Whether they would also have interest in uncovering identities in public use files is not clear. In fact, the highly publicized thefts of credit card numbers and other personal identifiers suggest that the threat from hackers breaking into internal, fully-identified databases may be greater than the risk of their re-identifying records in public use files protected with the most effective methods.

There is little question that the threat from these sources is growing. But at the same time, federal agencies have demonstrated that they remain vigilant and forward-looking in their evaluation and application of disclosure limitation techniques to the data that they release to the public. The track record for federal public use files is unblemished, but agencies have not rested on these accomplishments. Protective measures that have worked well in the past can become less effective over time. To guard against this possibility, many agencies regularly assess whether their procedures have become vulnerable to new data sources, new software, or

expanded computational capacity. Furthermore, being well aware that once a dataset is released to the public it cannot be recalled, federal agencies have devoted resources to anticipating future threats. Such active engagement promises continued security for the data that federal agencies release.

REFERENCES

Alexander, J. Trent, Michael Davern, and Betsey Stevenson. "Inaccurate Age and Sex Data in the Census PUMS Files: Evidence and Implications." *Public Opinion Quarterly*, vol. 74, no. 3 (Fall 2010): 551-569.

Benitez K. and B. Malin. "Evaluating Deidentification Risks with Respect to the HIPAA Privacy Rule." *Journal of the American Medical Informatics Association*, vol. 17, 2010, pp. 169–177.

Cavoukian, Ann, and Daniel Castro. "Big Data and Innovation, Setting the Record Straight: De-identification Does Work." Toronto, Ontario, Canada: Office of the Information and Privacy Commissioner, June 16, 2014.

Ciriani, V., S. De Capitani di Vimercati, S. Foresti, and P. Samarati. "*k*-Anonymity." In *Secure Data Management in Decentralized Systems*, edited by Ting Yu and Sushil Jajodia. New York: Springer, 2007.

Dalenius, Tore. "Towards a Methodology for Statistical Disclosure Control." *Statistik Tidskrift*, vol. 15, 1977, pp. 429-444.

Domingo-Ferrer, J., and J.M. Mateo-Sanz. "Practical Data-oriented Microaggregation for Statistical Disclosure Control." *IEEE Transactions on Knowledge and Data Engineering*, vol. 14, no. 1 (2002), pp. 189-201.

Duncan, George T., Mark Elliot, and Juan-Jose Salazar-Gonzalez. *Statistical Confidentiality: Principles and Practice*. New York: Springer, 2011.

Duncan, George T., and Diane Lambert. "The Risk of Disclosure for Microdata." *Journal of Business and Economic Statistics*, vol. 7, no. 2, April 1989, pp. 207-217.

Dwork, Cynthia, and Moni Naor. "On the Difficulties of Disclosure Prevention in Statistical Databases or the Case for Differential Privacy." *Journal of Privacy and Confidentiality*, vol. 2, no. 1, 2010, pp. 93-107.

El Emam, Khaled, and Fida Kamal Dankar. "Protecting Privacy Using *k*-Anonymity." *Journal of the American Medical Informatics Association*, vol. 15, no. 5, September/October 2008, pp. 627-637.

El Emam, Khaled, Elizabeth Jonker, Luk Arbuckle, and Bradley Malin. "A Systematic Review of Re-Identification Attacks on Health Data." *PLoS ONE*, vol. 6, no. 12, December 2011, pp. 1-12.

Federal Committee on Statistical Methodology. "Report on Statistical Disclosure Limitation Methodology." Statistical Policy Working Paper 22 (second version). Washington, DC: Office of Information and Regulatory Affairs, Office of Management and Budget, 2005.

Golle, Phillippe. "Revising the Uniqueness of Simple Demographics in the U.S. Population." Palo Alto, CA: Palo Alto Research Center, 2006.

Gouweleeuw, J.M., P. Kooiman, L.C. Willenborg, and P.P. DeWolf. "Post Randomisation for Statistical Disclosure Control: Theory and Implementation." Research paper no. 9731. Voorburg, Netherlands: Statistics Netherlands, 1997.

Hundepool, A., J. Domingo-Ferrer, L. Franconi, S. Giessing, R. Lenz, J. Naylor, E.S. Nordholt, G. Seri, and P.P. de Wolf. *Handbook on Statistical Disclosure Control*. A Network Excellence in the European Statistical System in the Field of Statistical Disclosure Control (ESSNet SDC). Hoboken, NJ: Wiley, January 2010.

Kwok, Peter, and Deborah Lafky. "Harder Than You Think: A Case Study of Re-identification Risk of HIPAA-compliant Records." *Proceedings of the Joint Statistical Meetings*. Alexandria, VA: American Statistical Association, 2011.

Machanavajjhala, Ashwin, Johannes Gehrke, Daniel Kifer, and Muthuramakrishnan Venkitasubramaniam. "*l*-Diversity: Privacy Beyond *k*-Anonymity." Cornell Computer Science Department Technical Report. Ithaca, NY: Cornell University, 2005.

Marsh, C., C. Skinner, S. Arber, B. Penhale, S. Openshaw, J. Hobcraft, D. Lievesley, and N. Walford. "The Case for Samples of Anonymized Records from the 1991 Census." *Journal of the Royal Statistical Society*, Series A, vol. 154, 1991, pp. 305-340.

Narayanan, A. and V. Shmatikov. "Robust De-anonymization of Large Spare Datasets. *Proceedings of the IEEE Symposium on Security and Privacy*, 2008, pp. 111-125.

National Research Council. *Expanding Access to Research Data: Reconciling Risks and Opportunities*. Panel on Data Access for Research Purposes, Committee on National Statistics, Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press, 2005.

Office of Management and Budget, Executive Office of the President. "Memorandum Number M-13-13: Open Data Policy—Managing Information as an Asset." Washington, DC: OMB, May 9, 2013a. Available at [http://www.whitehouse.gov/omb/memoranda_default/]. Accessed July 15, 2013a.

Office of Management and Budget, Executive Office of the President. "Supplemental Guidance on the Implementation of M-13-13 'Open Data Policy—Managing Information as an Asset.'" Washington, DC: OMB, May 9, 2013b. Available at [http://www.whitehouse.gov/sites/default/files/omb/memoranda/2013/m-13-13.pdf]. Accessed July 15, 2013.

Office of Science and Technology Policy, Executive Office of the President. "Increasing Access to the Results of Federally Funded Scientific Research." Memorandum. Washington, DC: OSTP, February 22, 2013. Available at [http://www.whitehouse.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_2013.pdf].

Ohm, Paul. "Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization." *UCLA Law Review*, vol. 57, 2010, pp. 1701-1777.

Pozen, D.E. "The Mosaic Theory, National Security, and the Freedom of Information Act." *The Yale Law Journal*, December 2005, pp. 628–679.

Purdam, K. and M.J. Elliot. "A Case Study of the Impact of Statistical Disclosure Control on Data Quality in the Individual UK Samples of Anonymised Records." *Environmental Planning*, Vol. A 39, 2007, pp. 1101-1118.

Rubin, Donald B. "Discussion of Statistical Disclosure Limitation." *Journal of Official Statistics*, vol. 9, no. 2, 1993, pp. 461-468.

Sattar, A.H.M. Sarowar, Jiuyong Li, Jixue Liu, Raymond Heatherly, and Bradley Malin. "A Probabilistic Approach to Mitigate Composition Attacks on Privacy in Non-coordinated Environments." *Knowledge-Based Systems*, vol. 67, September 2014, pp. 361-372.

Shlomo, Natalie. "Releasing Microdata: Disclosure Risk Estimation, Data Masking, and Assessing Utility." *Journal of Privacy and Confidentiality*, vol. 2, no. 1, 2010, pp. 73-91.

Singh, Avinash C. "Maintaining Analytic Utility while Protecting Confidentiality of Survey and Nonsurvey Data." *Journal of Privacy and Confidentiality*, vol. 1, no. 2, 2009, pp. 155-182.

Singh, A.C., F. Yu, and G.H. Dunteman. "MASSC: A New Data Mask for Limiting Statistical Information Loss and Disclosure." In *Work Session on Statistical Data Confidentiality 2003*, *Monographs in Official Statistics*, edited by H. Linden, J. Riecan, and L. Belsby, pages 373-394. Luxemburg, Belgium: Eurostat, 2004.

Sweeney, Latanya. "Weaving Technology and Policy Together to Maintain Confidentiality." *Journal of Law, Medicine & Ethics*, vol. 25 (1997), pp. 98-110.

Sweeney, Latanya. "Uniqueness of Simple Demographics in the U.S. Population." Technical report. Pittsburgh, PA: Carnegie Mellon University, 2000.

Sweeney, Latanya. "*k*-Anonymity: A Model for Protecting Privacy." *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems*, vol. 10, no. 5, 2002, pp. 557-570.

U.S. Department of Health and Human Services. "Summary of the HIPAA Privacy Rule." Washington, DC: U.S. Department of Health and Human Services. Last revised May 2003. Available at [http://www.hhs.gov/ocr/privacy/hipaa/understanding/summary/privacysummary.pdf]. Accessed June 3, 2014.

U.S. Department of Health and Human Services. "Health Data Initiative." Washington, DC: HHS, 2013a. Available at [http://www.hhs.gov/open/initiatives/hdi/index.html]. Accessed July 17, 2013.

U.S. Department of Health and Human Services. "Heritage Health Prize Announcement at 2013 Health Datapalooza." Washington, DC: HHS, 2013b. Available at [http://www.slideshare.net/HealthDataConsortium/health-datapalooza-2013-heritage-health-prize]. Accessed July 26, 2013.

U.S. Department of Health and Human Services. Healthdata.gov, Washington, DC. Available at [http://www.healthdata.gov/dataset/search]. Accessed July 24, 2014.

U.S. Department of Health, Education and Welfare, "Records, Computers, and the Rights of Citizens," Report of the Secretary's Advisory Committee on Automated Personal Data Systems. Washington, DC: U.S. Department of Health, Education and Welfare. Available at [http://www.justice.gov/opcl/docs/rec-com-rights.pdf]. July 1973.

U.S. Government Accountability Office. "Information Resellers: Consumer Privacy Framework Needs to Reflect Changes in Technology and the Marketplace." Document number GAO-13-663. Washington, DC: GAO, September 2013.

White House. "Transparency and Open Government." Memorandum for the Heads of Executive Departments and Agencies. Washington, DC: White House, January 2009. Available at [http://www.whitehouse.gov/the-press-office/TransparencyandOpenGovernment/]. Accessed September 5, 2014.

White House. "President's Remarks Presenting New Management Agenda." Washington, DC: White House, July 8, 2013a. Available at [http://www.whitehouse.gov/the-press-office/2013/07/08/remarks-president-presenting-new-management-agenda]. Accessed July 20, 2013.

White House. "First Look at Next.Data.gov." Washington, DC: White House, 2013b. Available at [http://www.whitehouse.gov/blog/2013/07/16/first-look-nextdatagov]. Accessed July 20, 2013.

White House. "Executive Order 13642—Making Open and Machine Readable the New Default for Government Information." Washington, DC: White House, May 9, 2013c. Available at [http://www.whitehouse.gov/the-press-office/2013/05/09/executive-order-making-open-and-machine-readable-new-default-government-]. Accessed December 26, 2013.

Zayatz, Laura. "New Ways to Provide More and Better Data to the Public While Still Protecting Confidentiality." *Proceedings of the Joint Statistical Meetings, Section on Survey Research Methods*. Alexandria, VA: American Statistical Association, 2008.

APPENDIX A

MINUTES

This page has been left blank for double-sided copying.

AGENDA

Addressing the Mosaic Effect: Best Practices in Protecting Confidentiality in Open Data Initiatives

Friday, June 27, 2014
Hubert H. Humphrey Building, Room 705A
9:30 AM to 4:00 PM

9:30-9:45 Welcome and Overview of the Day

John Czajka (Mathematica) and Joan Turek (ASPE)

9:45-10:15 Session 1: Purpose of the Meeting: The Mosaic Effect and Open Data

Jim Scanlon (ASPE)

10:15-10:45 Session 2: Current Policies and Practices for Disclosure Avoidance with Federal Survey and Administrative Data

Jacob Bournazian (Energy Information Administration)

John Czajka (Mathematica)

10:45-11:00 Break

11:00-12:30 Session 3: What Are the Re-identification Threats in Releasing Federal Data to the Public?

Moderator: Steve Cohen (Agency for Healthcare Research and Quality)

Daniel Barth-Jones (Columbia University)

Khaled El Emam (University of Ottawa and Privacy Analytics)

Denise Love (National Association of Health Data Organizations)

Brad Malin (Vanderbilt University)

Latanya Sweeney (Federal Trade Commission and Harvard University)

12:30-1:30 Lunch in the Cafeteria

1:30-3:00 Session 4: Good Practices for Protecting Public Use Data

Moderator: Connie Citro (Committee on National Statistics)

Mark Asiala (U.S. Census Bureau)

Barry Johnson (Statistics of Income, Internal Revenue Service)

Allison Oelschlaeger (Centers for Medicare & Medicaid Services)

Eve Powell-Griner (National Center for Health Statistics)

Fritz Scheuren (NORC at the University of Chicago)

3:00-3:15 Break

3:15-4:00 Wrap-up

John Czajka (Mathematica)

Jim Scanlon (ASPE)

APPENDIX B

ATTENDEES

This page has been left blank for double-sided copying.

ATTENDEES

Panelists and Moderators

Mark Asiala	Census Bureau
Daniel Barth-Jones	Columbia University
Jacob Bournazian	Energy Information Administration
Connie Citro	Committee on National Statistics
Steve Cohen	AHRQ
Khaled El Emam	Privacy Analytics
Barry Johnson	Internal Revenue Service
Denise Love	NAHDO
Brad Malin	Vanderbilt University
Allison Oelschlaeger	CMS
Eve Powell-Griner	CDC/NCHS
Fritz Scheuren	NORC at the University of Chicago
Latanya Sweeney	Federal Trade Commission

ASPE Project Staff

Susan Queen	ASPE
Jim Scanlon	ASPE
Joan Turek	ASPE

Mathematica Project Staff

Kevin Collins	Mathematica
John Czajka	Mathematica
Craig Schneider	Mathematica
Amang Sukasih	Mathematica

Invited Guests

Don Asmonga	Privacy Analytics
Maya Bernstein	ASPE
Lily Bradley	ASPE
Victoria Bryant	Internal Revenue Service
Ben Busby	NIH
Jeff DeGrasse	FDA
Nancy Donovan	GAO
John Eltinge	BLS
Brian Harris-Kojetin	OMB
Nathalie Holmes	Privacy Analytics
Rachel Hornstein	ASPE
Hyon Kim	GSA
Elizabeth Kittrie	Office of the Secretary, HHS
Doris Lefkowitz	AHRQ
Brooklyn Lupari	SAMHSA
Jennifer Madans	CDC/NCHS
Charles Pineles-Mark	CBO
Jenny Schnaier	AHRQ
Margo Schwab	OMB
Harnam Singh	HRSA
Mike Simpson	CBO
James Sorace	ASPE
Phil Steel	Census Bureau
Ann Waldo	Privacy Analytics
Chris Zogby	CBO

Telephone Participants

Jodi Duckhorn	HRSA
Rashida Dorsey	Office of Minority Health, HHS
Rachel Kaufman	CDC

APPENDIX C

MINUTES OF THE TECHNICAL EXPERT PANEL MEETING

This page has been left blank for double-sided copying.

Addressing the Mosaic Effect: Best Practices in Protecting Confidentiality in Open Data Initiatives
Technical Expert Panel
June 27, 2014

MINUTES[1]

Welcome and Overview and Introduction of All Attendees

Joan Turek, U.S. Department of Health and Human Services (HHS), Office of the Assistant Secretary for Planning and Evaluation (ASPE)

John Czajka, Mathematica Policy Research

Session 1: Purpose of the Meeting—The Mosaic Effect and Open Data

Jim Scanlon, ASPE

Thank you for participating today and giving us the benefit of your expertise and experience. Let me set the stage for today's discussion. We have two objectives for today's meeting. First, we will discuss and assess the extent to which *additional* risks of re-identification in federal statistical data access initiatives may arise because of the increasing availability of data and datasets from other sources, the so called "mosaic effect." The concept of the mosaic effect derives from issues in the security community when different types and sources of data are brought to bear on a problem to yield new insights when the individual pieces are combined. Does the risk of re-identification increase as more datasets become available?

Second, we will discuss the current and emerging body of statistical disclosure avoidance policies and techniques used by federal statistical programs and activities today, as well as promising areas of new research. Our focus is whether the current set of statistical disclosure avoidance techniques and data release mechanisms are adequate to protect confidentiality in the light of more data becoming available from all sources, or are new techniques and access mechanisms needed. Again, our interest today is on federal statistical data access and release activities, and primarily on microdata releases.

Federal statistical agencies have a long tradition of making data available through a broad continuum of access policies and mechanisms. Agencies constantly balance access to the data with protecting the confidentiality of the individuals who provide the information, weighing increasing demands for data with confidentiality protection, advancing technology and other publicly available data. Several initiatives have been launched in recent years to make federal

[1] This is a detailed summary that is presented in transcript form for ease of identification of comments with individual speakers, but the contents of this summary do not represent an actual transcription of the speakers' verbal remarks.

data more available and accessible—the Federal Open Data Initiative, President Obama's Executive Order 13642, and other health data initiatives.

As many of you are aware, a large body of effective and well developed statistical disclosure avoidance policies, techniques and practices has been developed, and virtually all federal agencies use them in their data release policies. We will discuss those techniques here today, and they are described in OMB Statistical Policy Working Paper 22. Similarly, there is a continuum of mechanisms through which federal agencies release data, including public use data files, de-identified data, data use agreements and restricted access mechanism such as research data centers (RDCs).

So the goal of today's meeting is a self-assessment regarding 1) potential new threats to data privacy protection in federal agency statistical practice and 2) our current capacity to address them. Our preliminary sense is that the current portfolio of disclosure avoidance techniques are close to the state of the art and effective in protecting against disclosure, but we are eager to learn of any new issues, risks or new approaches, or new areas of research that we might want to pursue.

This morning we will hear about the current state of the art in federal statistical disclosure avoidance techniques as well as some new developments. This afternoon we'll hear about specific statistical agency practices. And then we will discuss potential new threats to data protection.

Question from Brad Malin (Vanderbilt): Can you please define "open data?"

Scanlon: We view the concept of open data as a policy of making federal data available in whatever purpose/form would be helpful, with appropriate confidentiality protections. For example we have a long tradition of making survey data, research data and administrative data available to a variety of data uses. The data is either de-identified or made available through a data use agreement or a restricted access data center.

Session 2: Current Policies and Practices for Disclosure Avoidance with Federal Survey and Administrative Data

Jacob Bournazian, Energy Information Administration

John Czajka, Mathematica

Jacob Bournazian:

Much of this is what I wrote in Statistical Policy Working Paper 22, chapters 3 to 5.

Microdata is highly sensitive to identity disclosure if:

- Data are from administrative records
  - Categorize so there is no unique combination of variables
  - Or add some type of disturbance to the data
- The microdata contains a "population unique"
  - Relate the distribution of sample uniques to the distribution of population uniques
- The file is linkable to administrative record files or other exogenous data files

The bottom line is this. A file is adequately protected if the disturbed microdata cannot be successfully matched to the original data file or to another file with comparable variables.

EIA regularly combs through social media and newspapers to identify sources of microdata disclosure risk. All government agencies are affected by the impact of "sensational events." Catastrophic events and sensational news stories create re-identification risks that can exceed the capabilities of the methodologies applied to a file. For example, in April 2010 there was a mine explosion in West Virginia. Workers compensation files were eventually released, and although they were run through software that anonymizes microdata, it was easy to figure out who the 29 deceased miners were in the data.

The first step in protecting microdata files is to remove direct identifiers. Direct identifiers are usually personally identifiable information (PII). Examples include:

- Names and addresses
- Small geographic subdivisions
- Telephone or fax numbers
- E-mail addresses
- Social Security numbers
- Medical record numbers
- Health plan numbers

- Patient record numbers
- Account numbers
- Certificate or license numbers
- Vehicle identifiers and license plate numbers
- Medical device and serial numbers
- Personal URLs or websites
- Internet Protocol address numbers
- Biometric identifiers (finger and voice prints)
- Full face photographic images
- Any other unique identifying number, characteristic, or code

The second step is to assess and modify indirect identifying information. See pages 12 to 15 of the Mathematica background paper that was provided to the attendees for a summary of techniques to avoid disclosure. Three broad types of microdata disclosure limitation methods for protecting data are: (1) data reduction, (2) data modification, and (3) data creation.

Considering the overall file, three data reduction options are:

- Do not release microdata (that is, only release tabular data)
- Only release a sample of a data file. A census is more likely to disclose a respondent than a sample of respondents
- Only release a selection of variables. Remove sensitive variables (especially direct identifiers and certain indirect identifiers)

Considering the respondent record, options for data reduction include:

- Delete highly unique respondents from the file
- Identify outliers or sensitive values within a respondent's record and set them to missing (local suppression)

Considering individual data fields or variables, choices include:

- Truncate variable distributions—that is, top or bottom code
  - Do not rely on top coding or bottom coding a fixed percentage of records, however. Check the frequency distribution of variables identified for top or bottom coding. Since policies are built around quartiles, this can bring up some issues.
  - Recode threshold values with measure of central tendency—for example, using the mean or median

- Recode or collapse a number of categories in a variable distribution
- Round variable values

Data modification encompasses a number of techniques. Perturbation or noise addition involves, for a particular sensitive variable or subset of respondents, adding "error" to the information. In general, the error added is randomly assigned and should have a mean of zero and a known variance. It is useful to check the file for the percentage of records where more noise is added than the threshold level or less noise is added than the threshold level.

Data swapping and data shuffling—a similar technique—are commonly used with categorical variables, but they can also be used with continuous variables. One approach is to first sort the values of continuous variables; values close in rank are designated as pairs and are swapped between the two records. With an alternative approach, some percentage of records is matched with other records in the same file on a set of predetermined categorical variables, and the values of the variables are swapped between the two records. In either case, pay attention to the frequency distribution in deciding how many values to swap. This is more important than the percentage of data swapped.

Data creation implies replacing the actual data with imputed (or synthetic) data. This is done by first constructing an imputation model, then running the model using the original data, then creating new distributions of the variables that you are trying to protect. Only the created variable is released. The quality of the inferences from synthetic data is only as good as the specificity of the imputation model and the validity of the data used. There are no really good measures to assess the quality of the imputation of synthetic data. One can apply this approach to missing values, select variables, or the entire file.

Federal statistical agencies do a good job of securing their data files. A critical question is whether your microdata file can be matched to an external data base. Link Plus software is available from the Centers for Disease Control and Prevention (CDC) at: www.cdc.gov/cancer/npcr/tools/registryplus/lp_tech_info.htm.

Record linkage using non-unique identifiers can be rule-based, distance-based, or involve the use of string comparators. Examples of potential matching variables include gender, month, year of an event or treatment, ZIP code, education, or medical condition. As an aside, while it is true that because of increased computational power, it may not be necessary to block on particular variables (that is, restrict potential matches to records with common values on key fields), blocking remains important as a way of minimizing false positives. The priority is to modify blocking rather than to eliminate it.

If records on the file can be matched to an external data base, then the file should be modified. That is, the variables that contribute to the match would need to be deleted or changed in order to prevent matches or disrupt the blocking strategy.

Examples of popular administrative data sources for linking include health billing records, Social Security records, cancer registries, voter registration records, birth and death records, real property tax records, and health insurer claims data.

Not all agencies follow the same protocols and checks, but a best practice for checking data quality after the application of protection is to conduct statistical analysis comparing the original data to the protected data to ascertain quality, usefulness, and the ability to make unbiased inferences. For example, one can run "before and after" analyses for:

- Univariate statistics (means, skewness)
- Bivariate statistics (correlation, association)
- Multivariate statistics (model parameters)

The univariate and bivariate statistics are an obvious thing to check, but the use of multivariate tests needs to increase.

Finally, there are a number of limitations of Statistical Policy Working Paper 22:

- Breaking the link between the person and identification in a file is not enough in the digital age environment.
- Anonymity cannot be sustained even if anonymity can be preserved in a data file.
- Re-identification versus reachability; the role of predictive analytics keeps increasing.
- Risk assessment needs to change to include classes of persons. We need to include the kind and intensity of harm that people are exposed to by governments basing their decisions on algorithms that identify correlations in files.

False positives are a new kind of harm. If you modify the data, make sure that people are not re-identified inaccurately.

There are "3 C's"—coincidence, causation, and correlation. Correlation is dominating these days. This is a new age in data confidentiality – we need to go beyond merely preventing re-identification.

John Czajka:

To test the effectiveness of the disclosure protection by matching the public use file back to the original data, the matching is typically done with just a subset of variables—those that might be available to a potential intruder. This involves some judgment about what variables are "out there." It would be too easy to re-identify records with the original data if all fields were used, but it would not provide a realistic measure of risk.

Sensitive variables may not be highly correlated. Perturbing them independently may not be very effective in preventing re-identification. There are procedures for multivariate masking. Unfortunately, one may have to do a lot of damage to protect the data.

When introducing additive noise, it may be useful to assign the noise with a "donut" distribution. The Census Bureau explored the use of additive noise in the context of tabular data—work described by Laura (then Zayatz) McKenna. First you determine at random whether to make a positive or a negative adjustment to a value and then assign random noise with a

nonzero mean and specified variance. This assures that most records are altered at least a small amount, but the mean noise is still zero.

How much data swapping do you need to do? One view is that much of the effectiveness of swapping comes from introducing uncertainty about whether a value was swapped or not. It may not be necessary to swap a very high percentage of values for swapping to be effective.

A strategy in releasing synthetic data is to give users the opportunity to have their final estimates run on the real data. This provides a way of validating the users' findings. This is not always viable, however. A number of years ago the Statistics of Income division of the Internal Revenue Service (IRS) set up a data users group. An important application of public use tax data is in microsimulation models. The thinking was that the IRS could test the impact of different masking strategies on the quality of the data by bringing one of these models into the IRS and running alternative public use datasets through the model in order to compare results. The users explained that a dataset requires months of work before it can be used in a model. It wasn't feasible to test alternative datasets by simply running them through the model.

Agencies need to be concerned about not just current threats but future threats. Once you release a dataset, it's out there. You cannot take it back.

Discussion:

Joan Turek: You don't want to manipulate the data so much that you take away its usefulness to the user. Synthetic data have not provided acceptable estimates of the characteristics of small groups who are critical to policy making—for example, groups such as unwed mothers, SSI recipients, the disabled. Good estimates of small characteristics that permit detailed cross-tabs are needed. Groups such as these account for large amounts of federal expenditures and need to be accurately portrayed. People using data often have turnaround deadlines that are too short to allow them to wait for runs on the fully accurate data.

Scanlon: We found that the more we use techniques that modify data, the more disturbed the researchers become, and they lose confidence in the data. There is an anecdote of trying to produce estimates for a regulation. We got lots of pushback from the public because they didn't understand the statistical adjustments we had made.

Bournazian: With non-disturbance methods, the danger is high of re-identifying certain persons. Or of making false positives through predictive analytics. We have additional responsibility to protect people from these erroneous re-identifications.

Turek: We don't publish individual data, but there is a need for the data to be valuable/useful. It's a two-sided coin.

Steve Cohen (AHRQ): We can use research data centers to balance the two principles. One limitation is that geographical size reduces the ability to protect individuals. Some say there is "no zero risk of disclosure." We hear that, but we try to minimize risk. We should be more concerned about false positive identification by faulty re-identification. We are legally obligated to protect against true positive identification but technically not responsible for false positive identification. What liability do federal agencies have?

Bournazian: The main limitation of Statistical Policy Working Paper 22 in the section on risk assessment is its focus entirely on re-identification. We can't be blind to the era we're living in. And talking about Open Data, there is a software package that takes audio information and converts it into digital. It not only digitizes the data; it takes digitized data and converts it into a computational format, which is more powerful. It's happening all around in multiple subject matters. Someone can be included in a loan application, and predictive analytics say that this person is a high risk, which may result in a loan denial even though the consumer had paid on time to date. Someone's kid can be identified as not making it into an IB program because predictive analytics says he is not a performer. We can get a lot of these events in health data. Someone could be suffering from a medical condition—say a bone fracture—and had to stay in the hospital and was prescribed medicine, and an insurance company makes its decision on coverage based on the file.

Brian Harris-Kojetin (OMB): See Hermann Habermann's article on distinguishing confidentiality versus group harm. For example, survey responses may result in effects (such as where to site a hospital); we cannot guarantee that there will be no such harm. That's a totally different issue than protecting confidentiality.

John Eltinge (BLS): Take the RDC as a gold standard. There is a tradeoff between disclosure risk and data quality. What about the cost/burden on the analyst? Has anyone studied the incremental cost/burden of going to an RDC?

Scanlon: There are burdens involved with going to an RDC and complying with data use agreement stipulations. Since risk rises with public use data, we need to increase restrictions on the data content. There is a spectrum of availability.

Margo Schwab (OMB): Data quality is an issue of marginally lower data quality. Minor perturbations may have very little effect on certain analyses. Data quality usually refers to major problems, rather than precision with very specific analyses. CMS is experimenting with virtual data centers—this is another part of the spectrum. I hope today's meeting helps us figure out what are the right places on the spectrum in matching datasets to uses. We should push boundaries farther in coming up with creative ways to make data both available and protected.

Session 3: What Are the Re-identification Threats to Releasing Federal Data to the Public?

Steve Cohen, AHRQ (moderator)

Daniel Barth-Jones, Columbia University

Khaled El Emam, University of Ottawa and Privacy Analytics

Denise Love, National Association of Health Data Organizations (NAHDO)

Brad Malin, Vanderbilt University

Latanya Sweeney, Federal Trade Commission and Harvard University

Cohen: We need to be forward looking—let's think about what's coming during the next 3, 5, or 10 years to address potential threats.

Khaled El Emam:

De-identification has been simplified through automation. In a graphic representation, the process of de-identification in practice involves assessing risk, classifying the variables in the file, and mapping the data. These contribute to specifications in an automated anonymization engine through which the original data are run to produce the anonymized data for release. Our organization has 10 years' experience in helping clients share health data. We published three books on this topic last year.

Who is an adversary? This can include academia, the media, acquaintances (neighbor, ex-spouse, employer, relative, co-worker), the data recipient, malicious actors. There is no apparent economic case for malicious re-identification of health data. The bigger concern is the media.

There are direct and quasi-identifiers. Examples of direct identifiers include name, address, telephone number, fax number, medical record number, health care number, health plan beneficiary number, voter identification number, license plate number, email address, photograph, biometrics, Social Security number, social insurance number, device number, clinical trial record number. Examples of quasi-identifiers include sex, date of birth or age, geographic locations (such as postal codes, census geography, information about proximity to known or unique landmarks), language spoken at home, ethnic origin, total years of schooling, marital status, criminal history, total income, visible minority status, profession, event dates, number of children, high-level diagnoses and procedures.

An identifier must satisfy three general criteria. It must be replicable, distinguishable (that is, variable), and knowable. Replicable means that the identifier is sufficiently stable over time and has the same values for the data subject in different data sources. For example, blood glucose level is not replicable, but date of birth is replicable. A potential identifier is distinguishable if there is sufficient variation in the values of the field that it can distinguish among data subjects. A diagnosis field will have low distinguishability in a database of only breast cancer patients but high distinguishability in a claims database. An identifier must be knowable by an adversary.

The likelihood of its being known has to be high. How much an adversary knows will depend on whether the adversary is an acquaintance of the data subject or not. It may also depend on the expected resources of the adversary.

If an adversary is not an acquaintance, the types of information that are available include inferences from existing identifiers—for example, determining birth date from the date of the hospital discharge at birth; public data such as voter registration lists (available for free in some states), white pages, and whatever the subject reveals in public forums; semi/quasi-public data having a nominal fee or terms of use, such as voter registration lists in other states; and non-public sources such as commercial databases or provider databases that have significant costs to acquire.

How do we protect confidentiality when there are multiple quasi-identifiers? How much will an adversary know? If there are 10 quasi-identifiers, what assumptions do we make about the knowledge of the adversary? These are assumptions about adversary power. Assume, for example, that an adversary will know only 5 of the 10 quasi-identifiers. We can consider all combinations of 5 things and manage the risk for every combination. This becomes a solvable computational problem.

Some special types of data require specialized techniques. There are good techniques to de-identify geo-spatial information (including movement trajectories), dates and long sequences of dates (for example, transactional data), and streaming data—that is, data that is continuously being updated.

What is the impact of the mosaic effect? If de-identified properly, open data is not particularly useful for further attacks because it has no identifiable information, and the success rate of linking these data to other data should be small. Will the risks increase over time? Probably. But we have the same case with encryption algorithms, yet we still encrypt data. Should we release data publicly? Decent data can be created for public release. We can add terms of use or conditions in order to release higher quality data.

I also wonder about the cost-effectiveness of RDCs. They are not used very much. I recently conducted research at a Statistics Canada RDC, and the only other person I ever saw was the center librarian.

Malin:

We have constructed a de-identification system for DNA sequence data. Our database uses de-identification techniques for 2 million patients in the Vanderbilt system. The data is being used by 200 researchers, and we have biospecimens for 200,000 patients. Researchers may use the data, subject to a DUA with the National Institutes of Health (NIH).

We published a paper two weeks ago on a probabilistic model for patient disclosure based on estimating population uniqueness across datasets (Sattar et al. 2014 in *Knowledge-Based Systems*). One needs to be cognizant of data over time. If you anonymize someone in different ways at different points in time, this may actually make that person easier to identify.

Research has shown the variety of characteristics and behaviors that can distinguish an individual. This includes demographics, diagnosis codes, lab tests, DNA, health survey responses, location visits, pedigree structure, movie review, social network structure, search queries, internet browsing, and smart utility meter usage.

A colleague and I showed that the potential number of individuals who could be identified with demographic data from voter registration lists and the cost per person identified varied dramatically across a subset of states. The risk was substantially greater for a HIPAA Limited Dataset than a dataset protected with HIPAA Safe Harbor methods.

We are working on research to understand the incentives behind re-identification. A simplified view of risk is that the probability of re-identification is approximately equal to the product of the probability of an attack and the probability of a re-identification conditional on an attack. Our incentive system is broken. Incentives exist for researchers to re-identify and publish the results. This, in turn, may allow private industry to learn re-identification techniques from published literature. Deterrents to attack include DUAs, access gateways, unique login IDs and passwords, and audits. Data characteristics that affect the conditional probability of a re-identification include uniqueness, replicability, availability, and cost.

Latanya Sweeney:

This conversation is not much different than it was in 1997, but the world has changed a lot since then. Technology is constantly clashing with society—it's a matter of trust.

Under my direction, the Data Privacy Lab at Harvard University initiated the DataMap project (thedatamap.org) to document where personal health data goes outside of the doctor-patient relationship. Maps show the flow of data from the patient to various entities and from the physician and hospital back to the patient. These data maps also show the flow of the patient's personal health data from the immediate recipients to numerous other entities. The maps indicate that much of the data is transmitted with personal identifiers although some is not. Flows that do not directly involve the patient are numerous.

Less than half of the documented data flows are covered by HIPAA, including inpatient discharge data transmitted without explicit identifiers. Almost all states collect inpatient discharge data, and 33 states sell or share de-identified versions of their discharge data. HIPAA does not cover these data, and only 3 of the 33 states are using HIPAA standards to protect the data, according to a 2013 survey we conducted.

Recently, I purchased a public use version of patient-level hospital discharge data from Washington State. Using accounts of accidents published in newspapers in 2011, I was able to re-identify 43 percent of a sample of 81 accident victims in the hospital discharge data based on characteristics reported in both sources. The kinds of information reported in news stories is often known by others, including family, friends, neighbors, employers, and creditors.

Data brokers make data available very cheaply compared to the states.

We did a FOIA request to determine who were the buyers of publicly available health data. Predictive analytic companies are the big buyers. They are producing data products that exploit publicly available health data.

There are four ways to add transparency to the system. First, public notice of privacy breaches should be required. Second, data holders should be required to list publicly those with whom they share data. Third, each person should be able to acquire copies of their personal data from any entity holding their data. Fourth, each person should also be able to acquire an audit trail of the of the organizations with which the data was shared.

Re-identification is a key part of the cycle of improving the protection of data. We improve protective techniques only after protections fail. Encryption techniques have improved because they were used, problems were identified, and better techniques were developed. We now have strong encryption. We need the prevention of re-identification to get there as well.

A written summary is available at http://www.ftc.gov/news-events/blogs/techftc/2014/04/transparency-establishes-trust.

Denise Love:

We have been involved for years in discussions regarding these issues with states. We are proud of the solutions that states have come up with to balance transparency and confidentiality. These data systems are essential to public health and multiple other purposes. Documented uses of state health care databases include:

- Public safety, injury surveillance, and prevention
- Disease surveillance, public health registries
- Health planning, such as community needs assessments, hospital conversions and closures
- Market share analyses, hospital strategic planning
- Quality assessments and improvement, patient safety, outcomes studies
- Public reporting, informed purchasing (outcomes and charges)
- Transparency
- Health systems performance
- Identification of overuse, underuse, and misuse of health care services

There is a critical “iron triangle” to public data, representing three principles of data policy: transparency (public availability and data information), data utility, and data safety. There must be a balance among all three. Over-emphasis on any one of the three does not serve the public good.

Public data resources are valuable assets. A one-size-fits-all approach to data governance is not feasible. Each data system has a set of stakeholders, laws, and history. Useful data can be

shared while controlling for risks using statistical methods, management controls, and web query systems. We hear more complaints about the lack of data sharing than about the risks.

DUAs can mitigate the risk of inappropriate use. The Washington state story is the first breach that we've ever heard about. NAHDO spent a year developing guidelines for data release by states, which were published in January 2012, but Washington state was not following these guidelines.

Daniel Barth-Jones:

My recent work is using uncertainty analysis through a flow chart that lays out several components including intrusion scenarios and information on what variables are needed by an intruder for re-identification. I add an uncertainty distribution at each step of the flowchart to give a sense of how the data protection and disclosure avoidance techniques can reduce re-identification risk. I have included intrusion scenarios such as a "nosy neighbor" attack; a mass marketing type attack to re-identify as many individuals as possible—for marketing purposes; or a demonstration attack by a researcher in academia or a journalist, to try to identify individual or random people just to show vulnerability or seek attention in order to influence public policy. The flow charts that I have developed include different data elements—variables that pose a risk of de-identification—needed by the intruder for different intrusion scenarios. There could be as many as 3,000 potential variables. However, since most often the data is not necessarily accurate and the intruder cannot build a complete population register, there are often false positives. Each step in the flow chart has a probabilistic distribution—then you can sample across the scenario with a hyper-grid multiple times. This gives us a robust idea of the re-identification risk. The chart may include a model of trade-offs between the cost to protect the data versus the volume of disclosure. Playing up single re-identifications may convey the wrong message to policymakers. There are dependencies at each step in the chain to determine the economic motivation or benefit to the entity.

It is important to consider the impact of de-identification on statistical analysis. Poorly implemented de-identification can distort multivariate relationships and hide heterogeneities. This can be illustrated using plots of data from census public use microdata samples, where each dot is a combination of three quasi-identifiers: age, income, and education in years. Each color represents a different race. Data reduction through sampling and other means can destroy the ability to identify heterogeneity among the races. Starting with two percent sample data, I show the percentage of records that are population unique (3.5 percent), sample unique but not population unique (40.6 percent), and not unique (56 percent). When education in years is replaced with six education categories, the population unique are reduced to zero percent, the sample uniques that are not population unique are reduced to 8 percent, and the fraction that is not unique is increased to 92 percent. If education is dichotomized as greater than high school graduation versus less than or equal to high school graduation, the sample unique that are not population unique are further reduced to just 0.6 percent while those that are not unique are increased to 99.4 percent. If education is removed, no records are unique.

A forthcoming paper by T.S. Gal et al. evaluates the impact of four different anonymization methods on the results obtained from three different types of regression models estimated with colon cancer and lung cancer data. For each combination the authors calculated the percentage of

coefficients that changed significance between the original data and the anonymized data. Depending on the de-identification technique that was used, the health dataset, and the type of regression model that was evaluated, these percentages varied but were mostly non-trivial, ranging between 40 and 80 percent for the one de-identification technique that fared worst. The best method, one proposed by the authors, was consistently below 20 percent.

HIPAA lacks a penalty if data is re-identified by the user, even if these are false positives. Robert Gelman has proposed a personal data de-identification act. Currently, there is no cost for false positive identification. We need to change the cost for false positive identification to change the economic incentives for efforts at re-identification.

Discussion:

Cohen: I identified the following themes during these presentations: Brad discussed game theory; Latanya noted data sources that could result in a breach and observed that even one breach is too many; and Daniel discussed how to simulate the threat. Where are we heading in the next five years to address these threats?

Malin: Social media are a serious threat. People self-disclose and disclose about others. We are doing research on how Twitter is used (it turn out that people talk more about others than about themselves), and it is a minefield of potential disclosures ("pray for my mom; she has breast cancer"). Another challenge is that electronic health records are becoming commercialized, and start-ups are using data without regulation—this is a big loophole.

Sweeney: No one is really studying the predictive analytics industry, so we don't know how big an industry it is. Re-identification is a way of illustrating risk—it's big although unquantified. We don't know how much really goes on. DUAs don't stop it; they just hide it, because the penalties are so draconian. The focus of our conversation should not be on re-identification, but rather on disclosure risk. The future of releasing data to the public should not be from large private-sector organizations such as Google or Facebook. Instead, federal agencies should try to figure out how to link data in a secure way in the cloud to produce aggregated data for the public.

Barth-Jones: The future concern is harm from bad de-identification practice—from bad science and inefficiency. We should focus on reducing bad de-identification practices.

Love: Washington State withdrew from a best practices process. In the future, discharge and all payer data will be derived from claims, but then it becomes a statistical abstract rather than identifiable data. I'm worried that data will be too protected, and opt-in/opt-out will be disastrous for public health and for population health (for example, when parents do not vaccinate their children).

El Emam: Techniques are becoming more sophisticated, including protection of data. Adoption of good de-identification practices has been less than ideal. Risks can be managed with appropriate practices.

Malin: We have been having a four-year dialogue with the community regarding use of data for research purposes. A catastrophic breach will shut down research efforts. You need to

involve the community in these discussions—create an advisory board and keep them in place, and make them partners. This reduces the risk of research being shut down in the event of a breach.

Session 4: Good Practices for Protecting Public Use Data

Connie Citro, Committee on National Statistics (Moderator)

Mark Asiala, Census Bureau

Barry Johnson, Statistics of Income, Internal Revenue Service

Allison Oelschlaeger, Centers for Medicare & Medicaid Services

Eve Powell-Griner, National Center for Health Statistics

Fritz Scheuren, NORC at the University of Chicago

Connie Citro:

People voluntarily offer information to health researchers (for example, participants in a clinical trial) with the understanding that the researchers will keep the information confidential. A breach would be a violation of this trust. What are the implications if this data is made public? How much thinking have agencies done regarding the probability of disclosure? What are the likely risks versus the potential risks? It is probably time to update Statistical Policy Working Paper 22.

Mark Asiala:

Public use files that include microdata are only one part of a "suite" of data types released by the Census Bureau. Other types include tables produced from aggregated data for low levels of geography, special tabulations, and research papers.

Potential threats include an ability to identify individuals by using the tables directly, matching external data to public use files, or using data products in combination.

Strategies for protecting data from disclosure vary with the type of data. For tables, the table design and combinations of data swapping and partially synthetic variables on the source files are used to reduce disclosure risk. For public use files, size thresholds for geographic and category detail; noise addition for some variables; and additional data swapping and/or partial synthesis of data are used. Rounding is the primary strategy in special tabulations and research papers. We want to minimize the use of suppression techniques because they harm the utility of the data. We would prefer to mask a particular characteristic rather than an entire record.

For tables, the granularity of data cells raises the risk of re-identification. Too much detail leads to a "pseudo-microdata" file. A good rule of thumb is to not publish tables with more than 100 cells. Skewed distributions are another concern, even for less detailed tables and/or larger

geographies. Treating the records at risk before producing tabulations is preferable to having to suppress cells.

Strategies used for public use files include subsampling, setting thresholds for identification of geographic areas and categories, additional data swapping for "special uniques," noise infusion, and synthetic data. The threshold for identification of geographic areas is 100,000 (population size). The threshold for categories is 10,000 nationally. A special unique case will stand out even with a large sample size, so additional swapping is done for such cases. Noise infusion is used for age, with some constraints.

For special tabulations and research papers, we evaluate the detail of the tables and the underlying data to avoid inadvertent disclosure. We round the data to protect small cells and coarsen the detail. In some cases we impose a hard rule such as publishing no detail for a given characteristic below the county or state.

A group has been meeting to anticipate new disclosure risks. The Census Bureau is working on a microdata analysis system that allows tabulations off the entire data file, but with certain restrictions and protections, as an alternative to public use files. We are also asking if we can create a bridge between public use files and RDCs. Is there a middle ground?

Citro: We need to first understand the probabilities of data disclosure and what the actual effects of a disclosure may be. Why develop hypothetical scenarios where a neighbor knows so much about a sample member that they could pick that person out from a public use file?

Barry Johnson:

I represent the statistics arm of the IRS. We have data from tax returns and other documents filed with the IRS (we do not have survey data). We produce tabulations and analyses for the general public, and a couple of public use files. The individual tax public use file has existed since 1962. Our public use data has been the core of tax and economic modeling for the Congressional Budget Office, the Urban Institute, and the National Bureau of Economic Research.

The IRS works with the Federal Reserve Board to plan disclosure protection of the data collected in the Survey of Consumer Finances. Tax data is releasable because there are not many demographic pieces of data on the 1040 form, and this makes the intruder's job more difficult.

Data is constrained by accounting rules, so it is difficult to perturb. Because of the alternative minimum tax rules and other complexities, it is important to preserve these relationships in the data. Non-linear relationships also make the data hard to adjust. We remove obvious identifiers and rely on a portfolio of techniques to protect especially vulnerable variables.

We have partnered with experts in disclosure limitation and with fellow agencies to protect data and variables. Based on these reviews, the individual tax public use file is updated regularly. We then evaluate how effective the changes have been. Having access to the full population dataset makes evaluation or simulation effective (you can match the public use file to the population data to assess risk).

Allison Oelschlaeger:

CMS has mostly administrative data and less voluntary or survey data. Our office was formed a few years ago to maximize data for internal and external users. CMS policy is inclined toward aggregated data; we don't publish cell sizes of 10 or less.

CMS produces two types of de-identified data products, and it was a struggle to create useful de-identified products. (1) Stand-alone public use files of basic Medicare claims data—we remove direct identifiers and look carefully at indirect identifiers. (2) A synthetic file. This is a good way for researchers to develop expertise before doing research with the actual data.

Regarding access to researchers for files with identifiable information: historically this data is encrypted and sent via hard drive, with a DUA. Last year we launched a "virtual RDC"—you can submit a research protocol, conduct the research, and then any outputs are reviewed/cleared by CMS. Researchers don't have to satisfy security requirements at their own facilities this way.

Eve Powell-Griner:

Most of NCHS's data is survey data, but we also have vital statistics records and physical exam data. We rely on Working Paper 22 and standard limitation techniques. What is different: we think about disclosure limitation from the get-go. We identify potential problems with data each year and discuss them with our review board.

We have been developing on-line resources during the past few years. About 95 percent of our data is released as public use files. For the remaining 5 percent, we require a DUA and designated-agent agreements (for other federal agencies and researchers supported by organizations with security protections), and use of the data in RDCs, which can be used either in person or remotely.

NCHS is becoming somewhat more conservative in what is being released in public use files (for example, geography fields). There is a trade-off between accessibility and control. None of our data is inaccessible except for personally identifiable information.

We focus on rare characteristics that would be identifiable, and we are sensitive to rare information fields. NCHS has deployed new software to extend risk assessment and assign a probability of disclosure. In addition, we need to keep the genetic data collected in the National Health and Nutrition Examination Survey under tight control.

Fritz Scheuren:

Advantage: the variety of disclosure prevention techniques available. Disadvantage: the extent of the variety available.

I agree that it is time to update Statistical Policy Working Paper 22; it should be updated every five years (at least that frequently).

There is a "civil war" going on: the data quality people are at war with the information quality people. The user can't rely on tables; they want to use it in a microdata simulation model.

To date we have only done a level-one fix, which is not really enough. We haven't gone beyond this, due to resource constraints. We are not keeping up with the prey-predator problem. Whatever we do to protect the data will eventually be defeated. And there need to be penalties for intruders. It is important to note that with public use files, there is no contract, no DUA.

Discussion:

Citro: Mosaic effect is a term like "big data." The more multivariate you are, the greater the risk of disclosure. We need to be careful not to be too restrictive for public use files, especially for variables that hold little re-identification value, as some of these variables hold great research value.

Johnson: Revealing that a person filed a tax return is considered a disclosure by the IRS; we set a high bar to prevent disclosure. So how do we balance transparency and confidentiality? By working in cooperation with the users. We formed a user group, and we ask them to help us make choices. Two outside users helped develop the updated version of the public use file, which increased utility and strengthened protection. Their participation helped justify removing the geographic variable; "there is no zealot like a convert." Users worked through the process to make a better balance among the trade-offs.

Oelschlaeger: Recently, CMS has focused on aggregated files rather than de-identified files. The stand-alone de-identified public use files are so focused on removing all variables that could lead to re-identification that the files are useless. My boss calls them public useless files. Earlier this year CMS released aggregated data at the physician level. Historically, a 1979 injunction prevented Medicare physician payment disclosure. Dow Jones filed suit to overturn this injunction, and a judge agreed. After the injunction was overturned, CMS had to determine what to release at the physician level, and we have since published a file with National Provider Identifier-specific data. We give more weight to beneficiary privacy than to physician privacy in this data release.

Asiala: There is transparency within the agency as well as outside of it. We need a better solution for transparency plus protection. It is becoming more important that subject-matter experts work with the statistical experts to improve the effectiveness and reasonableness of de-identification practices.

Powell-Griner: Federal statistical agencies are trying to be more responsive to users. We get feedback from data users who report what data fields they want, and we try to make appropriate trade-offs. Maybe they will get access, but not as conveniently as perhaps they would like.

Scanlon: CMS—what if a data aggregator asks for data? Do you only allow it for research?

Oelschlaeger: CMS has a number of ways to share data, only one of which is for researchers. HIPAA has a concept of Limited Datasets (this is what commercial researchers get). The Qualified Entity program in the Affordable Care Act gives CMS the authority to release data for quality improvement/performance measures, provided the entity that is receiving the CMS data is combining it with other payer data. Any disclosure of identifiable data requires a DUA.

Scheuren: The mosaic effect comes into play when someone extracts data and tries to match it to another dataset. Billions of records in the insurance world are used for data mining. This is largely a good thing, but there are downsides (such as hackers). We have a trust system, but you need a trust-but-verify system. There are three things to do: penalize people, enforcement, and scale back overzealous confidentiality.

The upper tail of the CPS income distribution is so bad that the disclosure protection isn't really necessary—wasting resources.

Johnson: IRS needs a legislative change to allow a DUA that would put the responsibility on users. Right now, our data must be completely safe, or it cannot be made available.

Turek: How do you find hackers so you can punish them?

Barth-Jones: You could use computer forensics, and ensure that re-identification has legal consequences (more whistle blowing), and impose a penalty for false positives.

The postal code presents big problems—it is too detailed.

Connie: Communities need to be involved in the process, as they may help to identify variables that pose a high re-identification risk.

El Emam: DUAs only solve the issue of local adversaries. Children are the most easily re-identifiable because they reveal a lot of personal information. But their parents also reveal a lot of information about them (through social media).

Love: The Washington State problem is a good, real world example of how multiple layers of data privacy management are necessary.

El Emam: I have never seen any evidence that predictive analytic firms are trying to re-identify individuals.

Wrap-Up

Jim Scanlon, ASPE

John Czajka, Mathematica

Scanlon:

This has been the progression of the day: we began with policies, proceeded to discuss disclosure risk assessment, and then addressed agency practices to protect data and ideas for further research. Our discussion of risk assessment went beyond current procedures. We discussed broad issues of privacy and transparency. Those issues will be referred to other parts of HHS and to the FTC to address. The HHS recommendations to Congress for HIPAA medical privacy were not implemented, so we did the best we could with our regulatory authority.

The portfolio of disclosure prevention practices are where we thought they are, but we heard some interesting ideas for going forward for disclosure avoidance techniques. We also heard about the continuum of data release concept: for example, quasi-public use (a public use file with terms of use), and the CMS virtual RDC (although users have to go through ResDAC and pay for access to the data).

Discussion:

Asiala: Could property rights for government agency data be used as a means to hold users accountable?

El Emam: This approach worked at the Louisiana code fest. Privacy Analytics produced a file for CMS for Louisiana with terms of agreement.

Love: DUAs are used by the states, but enforcement has been uneven. A DUA is not only a means of restricting the use of the data; it can also be a tool to communicate and educate data users of the data's importance, of ethics and proper handling of data, and of the agency's values in protecting and securing the data. Washington State did not have a DUA. This was an example of why we need multiple layers of protection.

Brooklyn Lupari (SAMHSA): We have aggregated data, public use files, restricted use files, and a virtual RDC. We have confidence in our data protection technology—but the concern is about human behavior, and how users might avoid monitoring protocols. This can only be mitigated through training and education.

Jenny Schnaier (AHRQ): We communicate our rules to data users, and we can write other things into the DUA—including that they are personally responsible (not just their organizations). In fact, AHRQ will not accept requests from organizations—we insist on a person taking responsibility. The data we distribute is considered limited use.

Nancy Donovan (GAO): We are tracking data methodology issues. There are a variety of technologies for pooling data. There are also barriers to linkage across agencies—how to combine datasets and deal with unstructured data. Are there examples of data pooling?

Scanlon: The underlying statutes make it difficult to share identifiable information with another federal agency.

Craig Schneider (Mathematica): Two types of users have been discussed: researchers and commercial organizations. Clinical users were not discussed. These may start to do more analysis.

Ben Busby (NIH): The purpose of the data commons at NIH is to get large datasets such as genomes in and out for research. We are trying to make it easier for investigators to upload and download data and starting to put public use files in the cloud—but no PII.

Malin: NIH doesn't want to host the data for everyone. There will need to be a host for all this data, and it may not even be a federal agency. Amazon? Google? Trust will be a big issue. BD2K (Big Data to Knowledge) will fund national centers for biomedical computing.

El Emam: Data brokers haven't re-identified data. Some have actually participated in re-identification studies using their own data. Researchers push back against changes in data access and the ability to do work on their own computers.

Ann Waldo (Privacy Analytics): What about consumer-generated health data? This is unregulated by HIPAA. The ONC (Office of the National Coordinator for Health Information Technology) will need to consider the impact of this.

Scanlon and Czajka: Thank you for your participation today. This has been a very informative discussion.

This page has been left blank for double-sided copying.

APPENDIX D

BACKGROUND PAPER:
REVIEW OF FEDERAL POLICIES AND PROCEDURES REGARDING THE USE AND PROTECTION OF PERSONAL DATA

This page has been left blank for double-sided copying.

REVIEW OF FEDERAL POLICIES AND PROCEDURES REGARDING THE USE AND PROTECTION OF PERSONAL DATA[1]

Federal policies covering the use and protection of personal data focus on data collected or obtained by the federal government, but legislation extends to data collected at lower levels of government and by non-governmental entities. This review covers the key legislation that has helped to shape federal policy on the use and protection of personal data; additional laws governing data use; illustrative examples of agency regulations and guidelines; the major documents defining federal open data policy; an overview of datasets released by the Department of Health and Human Services (HHS); and methods of disclosure limitation used by federal agencies.

A. Key Legislation

Several key pieces of federal legislation govern the types of personal information that government and other organizations, such as health providers and educational institutions, can disclose about individual citizens or consumers. Most privacy laws focus on an individual's rights over the privacy of personal information—including the ability to access and correct information—and the circumstances under which an entity may be allowed to disclose information, with or without consent from the individual. This summary provides an overview of the acts that created the foundation for U.S. privacy law as it relates to data held by the federal government. These include the Privacy Act of 1974, the Computer Matching and Privacy Protection Act of 1988, the Health Insurance Portability and Accountability Act (HIPAA) of 1996, the Confidential Information Protection and Statistical Efficiency Act (CIPSEA) of 2002, and the Health Information Technology for Economic and Clinical Health Act (HITECH Act).

1. Privacy Act of 1974

The Privacy Act of 1974 was one of the first pieces of legislation to recognize the rights of individuals to privacy and the government's responsibility to safeguard information that citizens provide to it. With the computerization of public records during the 1960s and growing concern over how the government might use information about private citizens, an advisory committee under what was then the Department of Health, Education, and Welfare (HEW) was appointed to advise the government on the potential consequences of automated data systems and suggest possible safeguards to protect information. The committee's report, entitled "Records, Computers and the Rights of Citizens," became the cornerstone of the subsequent privacy legislation and policies moving forward. In particular, the report recommended a "Code of Fair Information Practice," which would prevent information collected for one use to be made available for other purposes without the consent of the individual and would require agencies to have mechanisms in place that allow individuals to learn what information is being kept on them and to correct or amend a record (HEW 1973). The report also highlighted the dangers of using Social Security numbers as universal identifiers.

[1] This background paper was prepared by Kevin Collins, John L. Czajka, Bonnie Harvey, Melissa Medeiros, Craig Schneider, and Amang Sukasih.

The Privacy Act of 1974 requires federal agencies to provide citizens with access and correction rights to personal information and limits how agencies share information. According to the Act, an agency can disclose a person's record only with the individual's written consent or under special circumstances. Under these exceptions, information may be shared within the agency or for uses for which it was intended (defined as routine use), for purposes of the Census, to the National Archives and Records Administration if the information is deemed worthy of preservation, to another agency for civil or criminal law enforcement activities that are authorized by law, and to individuals who have provided agencies with advance written notice that information will be used only for statistical research or reporting. Records shared for statistical research or reporting must be "transferred in a form that is not individually identifiable."[2]

2. Computer Matching and Privacy Protection Act of 1988

The Computer Matching and Privacy Act of 1988 updated the language of the Privacy Act to address concerns about how agencies share and match data across agencies. Under this law, agencies must notify individuals at the time of data collection that the information provided could be used for matching purposes. Agencies must also give individuals 30-days advance notice before taking adverse action based on the matched data. Finally, the law provides some guidance on oversight, requiring that agencies create internal review boards to approve matching activities, publish matching agreements between agencies, and report to the Office of Management and Budget (OMB) and Congress about matching. The law does not apply to two types of matches: (1) matches that aggregate data stripped of personal identifiers, and (2) matches made to support research or statistical purposes. For matches that will be used for research purposes, information collected through the matching process cannot be used to make decisions that "affect the rights, benefits, or privileges of specific individuals." However, data can be used to make decisions about the program in general.[3]

3. Health Insurance Portability and Accountability Act (HIPAA) of 1996

HIPAA applies to health plans, clearinghouses, and health care providers. This legislation is often considered the "high water mark" for how entities "balance risks to privacy against valuable uses of information" (Ohm 2010). There are two key regulations that emerged from HIPAA: the Privacy Rule and the Security Rule.

a. Standards for Privacy of Individually Identifiable Health Information (Privacy Rule)

Under the Privacy Rule, HIPAA-covered entities cannot disclose individually identifiable health information—known as protected health information (PHI)—unless the individual has authorized the release in writing, or the disclosure or use is permitted under the Privacy Rule's exceptions. These exceptions allow for the information to be shared within the covered entity for treatment, payment, or health care operations or for public interest and benefit activities (for

[2] "The Privacy Act of 1974," Title 5 U.S. Code, Sec. 552a. Available at [http://www.gpo.gov/fdsys/pkg/USCODE-2012-title5/pdf/USCODE-2012-title5-partI-chap5-subchapII-sec552a.pdf]. Accessed May 30, 2014.

[3] "Computer Matching and Privacy Protection Act of 1988," Public Law 100-503. Available at [http://www.whitehouse.gov/sites/default/files/omb/inforeg/final_guidance_pl100-503.pdf]. Accessed May 30, 2014.

example, law enforcement purposes or public health activities.) PHI includes information and demographic data related to an individual's past, present, or future physical or mental health and the provision or payment of health care services (HHS 2003).

De-identified PHI can be disclosed if the data no longer identifies the individual or provides a reasonable basis to identify the individual. HIPAA-covered entities must de-identify data using one of two methods: receive a formal determination of de-identification by a qualified statistician or by removing 18 specific identifiers (the "Safe Harbor" method), such as names, addresses, and account number (HHS 2012).

Under the Safe Harbor method, the following 18 identifiers of the individual or his or her relatives, employers, and/or household members must be removed:

1. Names
2. All geographic subdivisions smaller than a state. (The first three digits of the ZIP code may be included if the geographic region formed by combining all areas with the same first three digit ZIP code has more than 20,000 residents. Geographic regions with 20,000 or fewer residents will have a ZIP code of 000.)
3. All dates (except year) directly related to an individual, such as birth date, admission or discharge date, date of death, and all ages over 89 and dates (including year) that indicate such age. Ages and information related to ages over 89 can be aggregated into a single category of 90 years or older
4. Telephone Numbers
5. Fax Numbers
6. E-mail addresses
7. Social Security numbers
8. Medical record numbers
9. Health plan beneficiary numbers
10. Account numbers
11. Certificate/license numbers
12. Vehicle identifiers and serial numbers, including license plate numbers
13. Device identifiers and serial numbers
14. URLs
15. IP addresses
16. Biometric identifiers, including finger and voice prints
17. Full face photographic images and any comparable images
18. Any other unique identifying number, characteristic, or code

Following removal of the 18 identifiers, the HIPAA-covered entity cannot "have actual knowledge that the information could be used alone or in combination with other information to identify an individual" (Office for Civil Rights 2012).

b. The Security Standards for the Protection of Electronic Protected Health Information (The Security Rule)

The Security Rule established a national security standard to safeguard health information and addressed the technical and non-technical safeguards that entities must put in place to uphold the Privacy Rule standards. Under the Security Rule, entities must "ensure the confidentiality, integrity, and availability" of all PHI that are created, received, maintained, or transmitted electronically; identify and protect against "reasonably anticipated threats" to security or integrity of data and uses or disclosures; and ensure workforce compliance. The rule includes physical and technical safeguards and other organizational and policy requirements that entities must implement (HHS no date).

4. Confidential Information Protection and Statistical Efficiency Act of 2002

CIPSEA, which is Title V of the E-Government Act of 2002, limits federals agencies that collect data for statistical purposes from using the data for any other purpose.[4] Under this law, agencies must clearly distinguish between data collected for statistical and non-statistical reasons and inform individuals at the start of data collection if the information will be used for other purposes. Additionally, identifiable information cannot be disclosed for any use other than statistical analysis or research without the consent of the respondent, unless the purpose is authorized by the head of the agency and the disclosure is not prohibited by any other law.

CIPSEA also authorizes the Census Bureau, the Bureau of Economic Analysis, and the Bureau of Labor Statistics to share business data for the sole purpose of statistical analysis. The law aims to improve the efficiency and accuracy of the data that the three bureaus collect and reduce the burden of reporting data for businesses. Under the law, each bureau must come up with systems and security protocols to protect the confidentiality of shared information and must remove any identifying information when publishing data and findings.

5. Health Information Technology for Economic and Clinical Health ACT (HITECH Act)

The HITECH Act, which was passed in 2009 under the American Recovery and Reinvestment Act (ARRA), strengthened several of the privacy and security protections under HIPAA. Under HITECH, business associates of HIPAA-covered entities, such as contractors, must comply with HIPAA privacy and security requirements. The Act strengthened rules related to disclosure of PHI for marketing and fundraising and prohibits the sale of PHI without an individual's authorization. The Act also requires HIPAA-covered entities to notify individuals

[4] "Confidential Information Protection and Statistical Efficiency," Public Law 107-347, December 17, 2012. Available at [http://www.gpo.gov/fdsys/pkg/PLAW-107publ347/pdf/PLAW-107publ347.pdf]. Accessed June 17, 2014.

and HHS of any breach of unsecured PHI and to report breaches affecting more than 500 residents of a state or jurisdiction to media outlets in the affected area.[5]

B. Additional Laws and Proposed Legislation

A number of additional laws govern the use and protection of data from a variety of specific sources. Several examples are presented below, followed by a summary of a recent federal report that proposes new legislation to improve the protection afforded to data provided by or collected from consumers.

1. Family Educational Rights and Privacy Act (FERPA) of 1974

Under FERPA, federally-funded educational institutions and agencies are prohibited from disclosing protected educational information to any entity other than a student or a student's parents. Personally identifiable information (PII) may be disclosed only if the student or parent signs a document identifying the information to be released, to whom it should be released, and the reason for the disclosure. Educational institutions can release directory information, such as students' names, addresses, and telephone numbers, but must notify students and parents that the information will be released and give them time to opt out of the disclosure. Students and parents have the right to review and make changes to the students' records and must be notified annually of this right and the institution's policies for disclosing records.

Educational institutions can disclose PII without consent under special circumstances. These include, but are not limited to, the following:

- To other school officials within the same institution who have a legitimate educational interest in receiving the information
- To another institution where the student intends to enroll
- In relation to financial aid for which the student has applied or has received, for the purpose of determining eligibility, amount of aid, and conditions of the aid and/or to enforce terms and conditions of the financial aid
- To the juvenile justice system with the intention of improving the system's ability to serve the student

Institutions can also disclose PII to organizations conducting research on behalf of the institution and that have a "legitimate interest" in the information. The researchers must enter into a written agreement detailing the purpose, scope, and duration of the study and plans to destroy all PII once it is no longer needed for the study.[6]

[5] "Modifications to the HIPAA Privacy, Security, Enforcement and Breach Notification Rules Under the Health Information Technology for Economic and Clinical Health Act and the Genetic Information Nondiscrimination Act," 45 CFR Parts 160 and 164. Available at [http://www.gpo.gov/fdsys/pkg/FR-2013-01-25/pdf/2013-01073.pdf]. Accessed June 3, 2014.

[6] "Family Educational and Privacy Rights," Title 20 U.S. Code, Sec. 1232g. Available at [http://www.gpo.gov/fdsys/pkg/USCODE-2011-title20/pdf/USCODE-2011-title20-chap31-subchapIII-part4-sec1232g.pdf]. Accessed May 30, 2014

2. Federal Alcohol and Drug Confidentiality Regulations

Two laws protect the identity and medical records of patients receiving treatment for alcoholism or drug abuse and restrict the types of information that federally-assisted drug or alcohol abuse programs can release about current and former patients: (1) the Comprehensive Alcohol Abuse and Alcoholism Prevention, Treatment, and Rehabilitation Act of 1970 and (2) the Drug Abuse Prevention, Treatment, and Rehabilitation Act (1972).[7] Under these regulations, information may be released without a patient's written consent if the information is shared within the program or to qualified-service organizations that provide services to the program; is needed to meet a "bona fide medical emergency;" or is authorized by a court order for good cause. Information may also be released for purposes of research, management or financial audits, or program evaluations. An independent body must review the research proposal to determine that patients' rights are adequately protected and that the benefits outweigh the risks of disclosing personal information. The researchers cannot identify individual patients, either directly or indirectly, in the final report.

3. Protection of Financial Data

In addition to the above legislation, a couple of other acts restrict how private companies, particularly financial institutions, can disclose private information. Passed in 1970, the Fair Credit Reporting Act (FCRA) was one of the first federal laws to regulate how the private sector uses and discloses personal information. Under the act, consumer reports can be used only for specific purposes, such as determining eligibility for credit or background checks for employment. Consumer reporting agencies must provide individuals access to their records and investigate and address any mistakes that individuals find in their reports. Additionally, the Act requires that organizations contact an individual before taking adverse action based on information in his or her credit report (Solove and Hoofnagle 2006). Under the Gramm-Leach-Bliley Act (1999), financial institutions are required to initially and annually provide consumers with a privacy notice detailing the types of information they collect about the consumer; how the information is shared, used, and protected; and what rights the consumer has to opt out.[8]

4. Driver's Privacy Protection Act

Under the Driver's Privacy Protection Act, which was passed in 1996 and amended in 1999, a state department of motor vehicles can release personal information from an individual's motor vehicle record, such as name, address, or other demographic information, only with the individual's permission or for selected purposes. Under disclosure exceptions, federal, state, or local agencies can obtain drivers' personal information to carry out its functions or proceedings. Information may also be disclosed for automobile and driver safety purposes, such as vehicle recalls; for "use in the normal course of business by legitimate businesses" to verify accuracy of or correct personal information submitted by individuals to the business; for research activities, so long as personal information is not published or used to contact individuals; for use by

[7] "Confidentiality of Alcohol and Drug Abuse Patient Records," Title 42 U.S. Code, Sec. 1.A.2. Available at [http://www.ecfr.gov/cgi-bin/text-idx?tpl=/ecfrbrowse/Title42/42cfr2_main_02.tpl]. Accessed June 11 2014.

[8] "Gramm-Leach-Bliley Act," Public Law 106-102, Nov. 12, 1999. Available at [http://www.gpo.gov/fdsys/pkg/PLAW-106publ102/pdf/PLAW-106publ102.pdf]. Accessed June 3, 2014.

insurance companies; and for private investigative agencies to conduct authorized activities. An authorized recipient of the information may resell or redisclose information for permitted uses.

5. Resale of Consumer Information—Proposed Legislation

In late May 2014, the Federal Trade Commission (FTC) released a report recommending that Congress enact legislation that would create greater transparency around the activities of companies that collect and sell consumers' personal information (FTC 2014). These companies, commonly known as data brokers, typically collect data from a variety of public and non-public online and offline sources and sell information to other entities for marketing, risk mitigation, and people search purposes. In addition to collecting raw data, data brokers often create and sell derived data, or data that is inferred from a person's actions or choices. For example, a data broker may assume an individual's hobbies or interests based on the individual's magazine subscriptions. Some companies may also use an individual's locations or activities to infer more sensitive information, such as race and ethnicity or income level.

In its report, the FTC analyzed the work of nine data brokers and determined that most of the data that data brokers collect is done largely without consumers' knowledge and can have negative implications for the consumers' every day transactions, such as how an insurance company may classify consumers' risk profiles or the types of advertisements that they receive. The FTC report proposed legislation that would require data brokers to provide consumers access to their personal information and allow consumers to opt out of having their information shared for marketing purposes. The proposed legislation would also require data brokers to disclose that they draw inferences from raw data and provide the names of their data sources so that individuals can correct their information.

C. Agency Regulations and Guidelines

Federal regulations governing individual agencies sometimes include specific provisions regarding the collection and use of personal data. Prominent examples include Title 13 (The Census Act), Title 26 (The Internal Revenue Code), and a section of Title 20 that applies to the National Center for Education Statistics (NCES) and is notable for how it assigns legal liability for disclosure. Some agencies, like the National Center for Health Statistics (NCHS), have documented their internal rules in manuals, two of which are discussed here. To provide agency-wide guidance, the Federal Committee on Statistical Methodology (FCSM) developed a "Checklist on Disclosure Potential of Proposed Data Releases," which was intended primarily for public use data products. Recently, the Statistical and Science Policy Office in OMB issued a proposed statistical policy directive that reaffirms the importance of protecting the confidentiality of the information that statistical agencies collect from the public. All of these regulations and guidelines are discussed below.

1. Title 13 (The Census Act)

Section 9 of Title 13 specifies 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" (Gates 2011). Title 13 protections extend to any administrative records that the Census Bureau obtains in performance of its census

functions. It is notable that employees who violate the confidentiality requirement are subject to significant fines or imprisonment.

2. Title 26 (The Internal Revenue Code)

Section 6103 of the Internal Revenue Code (IRC) outlines the legal requirements for safeguarding the data received on federal income tax returns (Johnson 2014). In so doing it defines Federal Tax Information (FTI) and spells out the conditions governing the use of such data. As a general rule, "no officer or employee of the United States, state government, or local law enforcement agency is permitted to disclose FTI except as authorized by law." Section 1603 specifies a number of uses allowed by other agencies. These include the states (federal tax data on their residents, for state revenue functions), the Congressional Joint Committee on Taxation, the Treasury Department's Office of Tax Analysis (for tax administration and economic studies), the Department of Agriculture (to support the Census of Agriculture), the Bureau of Economic Analysis (to support the development of National Accounts), and the Census Bureau. In additional to specific census uses listed in Section 6103, the Internal Revenue Service (IRS) and the Census Bureau have an agreement listing nine types of statistical applications for which the Census Bureau may receive FTI.

3. Title 20, Section 9573

The Institute of Education Sciences within the Department of Education compiles statistics and conducts research and evaluations to help further knowledge of education from early childhood through postsecondary schooling. Under federal law, the Institute cannot use any individually identifiable information for purposes other than research or evaluation or publish any information that may identify an individual. For example, individually identifiable information that is collected by the Institute cannot be admitted as evidence in any judicial or administrative proceeding without the consent of the individual. Section 9573 of Title 20 defines the requirements for preserving the confidentiality of data collected, maintained, used, and disseminated by the Institute.[9] While the assurances of confidentiality are similar to those found in Title 13, however, Section 9573 goes a significant step farther in specifying legal penalties to *any* individual—not just an employee—who uses data from the Institute to identify individuals and knowingly discloses or uses the information for purposes other than research and evaluation. Such persons may be found guilty of a class E felony and subject to no more than five years in prison and/or a fine of up to $250,000.

4. NCHS Confidentiality Manual

Legislative and Regulatory Background. There are several pieces of legislation that regulate how NCHS must maintain the confidentiality of the data it collects. According to Section 308(d) of the Public Health Service Act, NCHS may use its information only as specified by the supplier and may never release identifiable information without the approval of the supplier or the person or establishment described in the information. The Privacy Act of 1974 provides additional standards on the treatment of records, although NCHS has a "K-4

[9] "Education Sciences Reform: Confidentiality" Title 20 U.S. Code, Sec. 9573. Available at [http://www.gpo.gov/fdsys/pkg/USCODE-2010-title20/pdf/USCODE-2010-title20-chap76-subchapI-partF-sec9573.pdf]. Accessed June 17, 2014.

exemption" for its statistical systems, meaning that the agency does not have to allow subjects of its data files to have access to the records about themselves in those files. CIPSEA requires that all data collected for statistical purposes be used only for statistical purposes, and it provides strong criminal penalties for unauthorized disclosure of data. The Freedom of Information Act (FOIA) requires that Federal agencies make their records available to persons who request them, although several kinds of records may be exempted. Finally, the Federal Law Governing Federal Employees' Behavior provides additional information about the consequences and penalties of unauthorized data use.

Definitions. There are some important terms that require clear understanding with regard to confidentiality. *Identifiable information* refers to information that can be used to establish individual or establishment identity, whether directly or indirectly (that is, by linking data with external information). *Confidential information* is identifiable information or information associated with identifiable information that was collected under an assurance that restricts the degree to which the information can be shared with others. *Disclosure* of identifiable information occurs when the information is made known to a third party. Disclosure may be classified as authorized, unauthorized, or inadvertent. An *agent* is a person designated by an agency to perform activities authorized by law and specified in a written document. A *collaborator* is one with whom NCHS has a formal working relationship at the inception of a survey or project. *Consent* is written, oral, or inferred approval by an NCHS respondent to provide the requested information.

Employee Responsibilities. The *Confidentiality Officer* will assist the *Center Director* and staff in a variety of ways, while s*upervisors* are responsible for informing all employees about NCHS policies and procedures relating to confidentiality. *Individual employees*, as well as *contractors*, *agents,* and *collaborators*, are expected to follow the rules and regulations at all times. Each new employee or contractor is required to view a confidentiality video, sign a confidentiality pledge, and receive documents and materials describing their responsibilities with respect to confidential information while working at NCHS. The *Administrative Officer* is responsible for making sure new staff comply with all requirements.

Policies on Consent and Assurances of Confidentiality. Consent from an individual may be obtained by signature or by construction, which means that permission has been indicated in writing or verbally. To obtain consent from an establishment, if the request for information is made in person by an NCHS staff member or agent, he/she must inquire as to who is authorized to provide the requested data on behalf of the establishment. When the authorized person is informed of the uses of the data and he/she supplies the data, NCHS construes that the establishment has given consent. When data are sought from an establishment by mail, the request may be addressed to the establishment itself, to the manager of the establishment, or another authorized person. When NCHS receives the data, it is construed that the establishment has consented.

Whenever NCHS requests data concerning an individual or an establishment, it is obligated to provide certain information and assurances to the supplier of information, such as the authority (i.e. a statute or by executive order of the President) that approves the solicitation of the information, the principal purpose for the information, intended disclosure of identifiable information, and the effects of not providing all or part of the requested information. If the

release of any identifiable information is to be made, then the law requires that consent be obtained in advance. The set of information given to an individual or establishment must contain a statement of reassurances. When data is collected directly from individuals or establishments, a "Confidential Information" notice (found in Section 5.4, page 7 of the manual) must be included on the data collection instrument, and some additional information about the data usage must be provided either on the instrument itself or in a separate letter/document. When data is collected by telephone, the respondents must be given information about the survey, and the telephone interviewer must sign a statement saying that the information was given orally to each respondent. In computer-assisted telephone interviewing, the respondent must acknowledge having read all of the statements by checking a box or symbol.

Treatment of Requests for Information under FOIA. Whenever a request is received for a specified record concerning a named individual, that request is subject to the requirements of FOIA. However, two important exceptions can apply to NCHS: (1) "personal and medical files and similar files, the disclosure of which would constitute a clearly unwarranted invasion of personal privacy" and (2) matters specifically exempted from disclosure by statute.

The Protection of Records and Data Systems. Employees of NCHS are responsible for protecting all confidential records from prying eyes, unauthorized access, theft, and from accidental loss or misplacement due to carelessness. Confidential records must be kept locked up at all times when they are not being used, and copies of confidential records are not to be made except as needed for operational purposes. Records containing PII should be held to the minimum number deemed essential to perform the necessary functions, kept in a highly secure manner, and kept only so long as needed to carry out those functions. No record containing personal identifiers may be sent to or accessed from an alternate work site or removed from NCHS offices except as required in the process of data collection activities. When records are transferred to the National Archives and Records Administration or record centers for storage, their containers must be sealed, and when records are transmitted between NCHS offices or between NCHS and its contractors, they must be packaged securely and sent by the most secure and trackable means available. Finally, the DHHS released a directive, called the DHHS Automated Information Systems Security Program Manual, which provides practices and procedures intended to carry out OMB Circular A-130, "Management of Federal Information Resources." All automatic data processing system users must familiarize themselves with the contents of this manual.

Authorized Disclosures. No information about a person or establishment may be disclosed to anyone without the informed consent of the person or establishment providing it, with one exception—to the Parent Locator Service. If such a request is ever received, it is to be referred immediately to the Confidentiality Officer. Under Section 308(d), NCHS is permitted to publicly release data for identifiable individual persons or establishments if 1) such release is included in the purpose for which the data were supplied, and 2) the particular person or establishment supplying the information or described in it has consented to such release. Division-level approval is required for the use of confidential data by other NCHS programs. Although the Privacy Act of 1974 considers DHHS in its entirety as one agency, NCHS is not required to disclose confidential records to other parties. Similarly, although information may be supplied to other departments of the federal government, transfers are rarely made, and they must conform to all the rules and regulations, as well as relevant federal laws. In the case that NCHS is one of two

or more organizations involved in a cooperative agreement, certification must be included indicating that the party or parties receiving NCHS data understand their obligation to abide by all NCHS rules and regulations.

Avoiding Inadvertent Disclosures through Release of Microdata. It is Center policy to make its files on individual elementary data units widely available to the scientific community. These microdata files consist of individual records each containing values of variables for a single person or establishment. However, even when all personal identifiers are removed, a large amount of information remains, and this information may identify NCHS respondents to a person who has access to that information from another source. Therefore, there are some rules that apply to all files released by NCHS. Before files are published, they must be approved by the Confidentiality Officer, and the file must not contain any detailed information about the subject that could facilitate identification and that is not essential for research purposes. Geographic places that have fewer than 100,000 people are not to be identified on the file, as well as characteristics of an area if they would uniquely identify an area of less than 100,000 people. Finally, information on the drawing of the sample that might assist in identifying a respondent must not be released outside the Center.

Avoiding Inadvertent Disclosures in Published Tabular Data. Any tabulations or calculations based upon approved public use microdata can be released to the public without additional disclosure protection measures. Tabulations based upon data that are not approved for public release must conform to the following special guidelines:

- In no table should all cases of any line or column be found in a single cell.
- In no case should the total figure for a line or column of a cross-tabulation be less than five unweighted cases.
- In no case should a quantity figure be based upon fewer than five unweighted cases.
- In no case should a quantity figure be published if one case contributes a disproportionate amount to the total.
- In no case should data on an identifiable case be derivable through subtraction or other calculations.
- Data published by NCHS should never permit disclosure when used in combination with other known data.

There are two methods that are customarily used in the Center to prevent disclosure through tabulations: (1) the table is reduced in size when rows or columns are combined into larger categories and (2) unacceptable data in cells are suppressed.

5. NCHS Disclosure Manual

The NCHS disclosure manual focuses on the use of the agency's Research Data Center (RDC), which provides restricted access to data elements not included on public use files and affords users the opportunity to work with datasets with linked survey and administrative records.

Part 1: Confidentiality and the RDC. There are several RDC procedures to prevent disclosure. Restricted data cannot leave the secure access modes, and output, programs, and files cannot be saved to transportable electronic media. A research proposal is required, and the Review Committee will carefully examine the variables requested, the plan of analysis, and the desired output. The RDC provides confidentiality training and requires researchers to complete confidentiality paperwork. Each mode of access has specific policies, procedures, and rules designed to prevent disclosure. Analytic datasets will be created for researchers, and all output must be reviewed by the remote access system or an RDC Analyst before it can be released to the researcher.

There are two laws that govern the NCHS RDC: Section 308(d) of the Public Health Service Act and CIPSEA. The Public Health Service Act asserts the importance of protecting confidentiality and that the only people who can access confidential data must become Designated Agents, while CIPSEA stipulates the penalties for violating confidentiality as up to five years in prison and/or a $250,000 fine.

Part 2: The RDC Research Process. Researchers are required to follow the proposal process instructions and use the proposal format provided on the RDC website. A list of variables to be used in the analysis must be provided, including a description of how they will be used. If restricted merge variables can be removed, coarsened, or substituted with randomized versions, this must be stated in the proposal. Although analysis plans may change, the RDC Analyst must be made aware of these changes throughout the process.

Part 3: Approved Proposals: Next Steps. Approval of a proposal does not explicitly or implicitly guarantee that all output generated by the analysis will be released. Once a proposal is approved, the principal investigator and all research team members who come in contact with the data must take the confidentiality orientation and complete the confidentiality forms. The forms are specific to the proposal, so they need to be completed each time a different proposal is approved by the RDC. The analytic dataset will be created by NCHS staff, who will follow certain policies to protect geographic, temporal, and perturbed and masked information. Researchers are responsible for providing the RDC with an extract from the NCHS public dataset as well as any non-NCHS data. The extract from the public dataset may include only variables that were specified in the proposal; original NCHS variables must retain the same name in the public dataset; and any derived variables should be clearly defined. The data files along with a list of variables must be emailed to the RDC Analyst, and any other questions should be discussed with the RDC Analyst as well.

Part 4: Working with Restricted Data. When working at an NCHS RDC, the researcher must abide by a number of rules developed to decrease the likelihood of a disclosure. Only one research project may be worked on at a time, and no individual-level data can leave the RDC facilities. No communication devices are allowed, and any items that may enable the identification of individuals and/or establishments are prohibited. Researchers cannot introduce new data using their computer code, and they are not allowed to put any content in code that would facilitate re-identification of a subject/establishment. Output must be submitted in a human-readable plain text file. While working at a Census Bureau RDC, all of the NCHS rules and restrictions apply. When working through the Remote Access System, only statistical code that is related to the analysis plan outlined in the research proposal may be submitted. Remote

access rights are granted to only one person, and any output results that pose a disclosure risk will be suppressed.

Part 5: Disclosure Review Policies and Procedures. There are some general output policies that exist to protect the confidentiality of NCHS study participants. Datasets, including output in the form of datasets, will not be released, and no output will leave the RDC facilities without first being reviewed by an RDC Analyst or the Remote Access System. Before submitting output for review, it must be in a form that can be released by the RDC, and any individual-level data or extreme values representing an individual must be removed. Furthermore, all cells with a frequency less than 5 should be asterisked. Approved output is returned via email and must match the research questions/output suggested in the proposal.

Part 6: Publishing Research. When publishing, all additional requirements specified in the approval email must be adhered to. Information that could identify individuals, establishments, or geographic areas must not be revealed, as well as information about specific dates from external sources of data that have been merged to NCHS data based on temporal or geographic components. Citations for all publications, presentations, and reports that refer to research conducted using the RDC must be emailed to rdca@cdc.gov and the RDC Analyst as soon as possible. In the publication, the methods section should specify which restricted variables were accessed through the RDC and why they were essential to the research questions. Finally, the following disclaimer must be added to the conclusion of the publication: "The findings and conclusions in this paper are those of the author(s) and do not necessarily represent the views of the Research Data Center, the National Center for Health Statistics, or the Centers for Disease Control and Prevention."

6. FCSM/CDAC Checklist

In 1999, an FCSM interest group, the predecessor to the current Confidentiality and Data Access Committee (CDAC), prepared a "Checklist on Disclosure Potential of Proposed Data Releases." The Checklist, a 48-page document, asked a number of questions about the proposed release. For microdata, for example, the Checklist asked about geographic detail reported on the file, top coding of continuous variables, and a number of other factors associated with disclosure risk. The Checklist also provided guidance on many of these topics. One purpose of the checklist was to provide a standardized document that agencies could prepare and submit to their disclosure review boards. Some agencies have updated or developed their own versions of the Checklist to use in much the same way.

7. Proposed Statistical Policy Directive from OMB

On May 21, 2014, the Office of Information and Regulatory Affairs in OMB published a Proposed Statistical Policy Directive that "provides a unified articulation of federal statistical agency responsibilities" (OMB 2014). The directive lists four responsibilities of federal statistical agencies:

1. Produce and disseminate relevant and timely information
2. Conduct credible and accurate statistical activities

3. Conduct objective statistical activities
4. Protect the trust of information providers by ensuring the confidentiality of their responses

On this last item the directive underscores the importance of maintaining a consistent level of protection as a way of reducing respondents' confusion, uncertainty, and concern about the use of the information they report. Statistical agencies are reminded that they must "follow best practices for protecting the confidentiality of (their) data."

D. Open Data Documents

Four documents issued by the Executive Office of the President over a six-month period in 2013 define the scope and provide guidance on implementation of the new open data policy. These four documents were:

1. Increasing Access to the Results of Federally Funded Scientific Research (Office of Science and Technology Policy 2013)
2. Making Open and Machine Readable the New Default for Government Information (White House 2013)
3. Open Data Policy—Managing Information as an Asset (OMB 2013a)
4. Supplemental Guidance on the Implementation of M-13-13 "Open Data Policy—Managing Information as an Asset" (OMB 2013b)

Summaries of the four documents are presented below.

1. Increasing Access to the Results of Federally Funded Scientific Research

This memorandum, issued on February 22, 2013 by the Office of Science and Technology Policy (OSTP) and directed to the heads of executive departments and agencies, calls for all federal agencies that are engaged in research and development to outline plans to provide public access to all results of scientific projects that are receiving federal funds. These plans were to be submitted to OSTP for review within six months of the publication of this memorandum. Plans must contain strategies addressing the following concerns: fostering public-private partnerships with scientific journals; improving access to digital data; optimizing archival and search features; plans for notifying federally funded researchers of their obligations; measurement and potential enforcement of compliance; identification of resources for implementation of the plan; and identification of circumstances that may exempt agency participation. OSTP further outlines specific objectives in increasing access to scientific publications as well as scientific data.

Regarding scientific publications, OSTP requests that agencies ensure that the public can access final manuscripts in digital form, following a 12-month publication embargo as a guideline. OSTP also asks agencies to facilitate public search and analysis of peer-reviewed scholarly work and to ensure public access to publications' metadata. In addition, agencies must encourage public-private collaboration, maintain accurate attribution to original authors, and provide archival solutions to storing metadata and publications that are publicly accessible.

With respect to scientific data, OSTP requires that agencies maximize free public access to digitally formatted data created with federal funds, while protecting privacy and proprietary

information. This involves requesting that grant recipients outline data management plans and detail any reasons why their data cannot be made publicly accessible. OSTP also requests that agencies allow for the inclusion of appropriate costs associated with data management and access in Federal grant proposals.

2. Making Open and Machine Readable the New Default for Government Information

Executive Order 13642, issued by President Obama on May 9, 2013, calls for a shift in the default policy in federal agencies toward that of free public access to information. The order describes government information as an asset, the dissemination of which is likely to create new jobs, provide inspiration for entrepreneurship, and stimulate the American economy. The order calls for the adoption of an Open Data Policy, as outlined in the OMB memorandum issued on the same day.

To ensure government-wide implementation of the Open Data Policy, the order directs all executive departments and government agencies to complete four actions in accordance with specified deadlines. First, within 30 days of the issuance of the Open Data Policy, the Chief Information Officer and the Chief Technology Officer must publish an online resource to assist agencies in their efforts to implement open data policies. Second, within 90 days, top government councils in governmental information (for example, the Chief Information Officers Council, the Federal Records Council) shall initiate implementation of measures to integrate the Open Data Policy requirements into federal acquisition and grant-making. Third, also within the 90 day time frame, the Chief Performance Officer must establish a Cross-Agency Priority (CAP) Goal to set metrics and milestones for agencies to follow. Fourth, within 180 days the agencies must report their progress in implementing the CAP Goal, with quarterly reporting thereafter.

The order acknowledges multiple times that agencies must continue to safeguard individual privacy, confidentiality, and national security, incorporating a full analysis of risks to each of these at each stage of the information life cycle so that information that should not be released can be identified.

3. Open Data Policy—Managing Information as an Asset

Memorandum M-13-13, issued by OMB in conjunction with the Executive Order, and also directed to the heads of executive departments and agencies, establishes a framework to support effective information management strategies that will promote open data. An attachment to the memorandum includes four sections: a list of legal definitions relevant to the Open Data Policy, a synopsis of the scope of open data policies, a brief description of the policy requirements, and guidelines for implementation.

This memo applies to "all new information collection, creation, and system development efforts as well as major modernization projects that update or re-design existing information systems." National Security Systems are noted to be exempt from these policies.

The policy requirements described by OMB are intended to support downstream information processing in the most efficient, cost-effective, and safest manner. To this end OMB will require agencies to begin planning management of information resources at the earliest possible stage, in order to minimize costly future maintenance. Agencies are directed to adopt the following

policies: use machine-readable and open formats; use data standards; remove all restrictions on distribution of public data (open license); and describe data using common core metadata (i.e. origin, linked data, geographic location, time period/interval, and data quality). Agencies are also directed to build systems supporting interoperability, such as those outlined in OMB's Common Approach to Federal Enterprise Architecture. These policies will likely support development of further requirements (detailed in the memo), which are to create an enterprise data inventory and maintain a public listing on Data.gov.

Within six months of the release of this memorandum, agencies were requested to take the following actions:

- Create and maintain an enterprise data inventory
- Create and maintain a public data listing
- Create a process to engage with customers to help facilitate and prioritize data release
- Clarify roles and responsibilities for promoting efficient and effective data release practices

Agencies were also asked to document their decisions that particular information should not be released as public datasets.

The memorandum gives particular attention to the protection of privacy and confidentiality, and the mosaic effect is noted as an issue of particular concern to this goal. To counteract potential breaches of privacy, guidelines for risk-minimization are detailed. These guidelines include the following: collect or create only necessary and useful information; limit collection of identifying information; limit sharing identifying or proprietary information; take into account the levels of risk and potential harm that are associated with the dissemination of particular datasets; and consider information that is already public when releasing de-identified data (that is, be aware of the mosaic effect). OMB also requires that a Senior Agency Official of Privacy or the equivalent assume a central role in the implementation process.

4. Supplemental Guidance on the Implementation of M-13-13 "Open Data Policy – Managing Information as an Asset"

This document, which was issued in August 2013, provides additional, in-depth information to agencies on how to carry out the objectives of the Executive Order and OMB Memorandum M-13-13. Most importantly, this document provides a list of minimum goals that must be met by agencies to fulfill these policy requirements. Minimum requirements for an Enterprise Data Inventory (goal 1) include: submitting a schedule to OMB of how the agency plans to identify all their data; posting all datasets in machine-readable format to Data.gov; and updating the inventory schedule on a quarterly basis. To assist agencies with their inventories, OMB and the General Services Administration have provided a data inventory tool called CKAN. Minimum requirements for a Public Data Listing (goal 2) include publishing all data that are described in the inventory metadata as public and publishing the data listing at www.[agency].gov/data.json. The memorandum also highlights the availability of two additional tools. The minimum requirements to engage with customers (goal 3) are establishing a mechanism for customer feedback and describing the process for review of customer feedback. Tools are provided for this purpose as well. The policy requirements for documenting non-releasable data (goal 4) include

simply describing and publishing the process by which the safety of releasing the data is determined. Finally, the only requirement for clarifying the roles and responsibilities for promoting efficient and effective data use (goal 5) is to report the point of contact for each these roles via the E-Gov website.

E. Datasets Released by HHS

HHS and the Institute of Medicine launched the Health Data Initiative (also known as the Open Data Initiative) in 2010. The purpose of the Health Data Initiative is to encourage "innovators to utilize health data to develop applications to raise awareness of health and health system performance and spark community action to improve health."[10]

The HHS Open Data Initiative is part of a broader federal government movement to make datasets available to the public. President Obama issued an executive order on May 9, 2013 that directed OMB to issue the Open Data Policy throughout the federal government. The objectives of the executive order are to advance the management of government information as an asset throughout its life cycle, to promote interoperability and openness, and to make the data accessible to and usable for the public.

These data are made accessible in machine-readable form, and innovators are encouraged to use the data to create new products and services, and thereby to create jobs. In addition to health care, open data efforts have been implemented in the fields of energy, education, finance, public safety, and global development. The executive order was refined by OMB memorandum M-13-13 (discussed above), which gave federal agencies guidance on definitions, scope, policy requirements, and implementation guidelines.

HHS alone has made an enormous amount of data available. As of June 2014 there were 1,530 datasets accessible at www.healthdata.gov, up from 428 one year earlier. The Centers for Medicare and Medicaid Services offers the most data, with 625 datasets, 515 of which are related to Medicare. The HHS agency with the next largest number of datasets is the Centers for Disease Control and Prevention, with 118. One reason for the large growth in the offerings of healthdata.gov since 2013 is the addition of state-specific databases, which totaled 425, including 111 from New York.

F. Methods of Disclosure Limitation Used by Federal Agencies

Table D.1 summarizes the methods of disclosure limitation used by federal statistical agencies at the time of the CDAC 2005 update of Statistical Policy Working Paper 22. Mathematica surveyed representatives of 14 agencies to ask if the agency has made any changes to its procedures. Results are reported in the final column of the table.

[10] U.S. Department of Health and Human Services, Healthdata.gov, Washington, DC. Available at [http://www.hhs.gov/open/initiatives/hdi/index.html].

This page has been left blank for double-sided copying.

Table D.1. Summary of Agency Practices for Protecting Public use Microdata as Reported in Statistical Policy Working Paper 22 (2005), with Updates

Agency	Public use microdata and who reviews	Restricted access allowed for researchers	Statistical disclosure limitation methods for public use microdata	Any update?
Energy Information Administration (EIA)	Yes – office review	No	EIA does not have a standard for statistical disclosure limitation techniques for microdata files. The only microdata files for confidential data released by EIA are for the Residential Energy Consumption Survey (RECS) and the Commercial Buildings Energy Consumption Survey (CBECS). In these files, various standard statistical disclosure limitation procedures are used to protect the confidentiality of the data for individual households and buildings. These procedures include: eliminating identifiers, limiting geographic detail, omitting or collapsing data items, top-coding, bottom-coding, interval-coding, rounding, substituting weighted average numbers (blurring), and introducing noise through a data adjustment method that randomly adjusts respondent level data within a controlled maximum percentage level around the actual published estimate. After applying the randomized adjustment method to the data, the mean values for broad population groups based on the adjusted data are the same as the mean values generated from the unadjusted data.	No updates. EIA still applies the same methodologies for protecting the CBECS and RECS public use files.

Table D.1. (*continued*)

Agency	Public use microdata and who reviews	Restricted access allowed for researchers	Statistical disclosure limitation methods for public use microdata	Any update?
National Science Foundation (NSF)	Yes – Meet or exceed Census public use products that are merged	Yes	When releasing public-use microdata files, individual identifiers are removed from all records and other high risk variables that contain distinguishing characteristics are modified to prevent identification of survey respondents and their responses. Top codes and bottom codes are employed for numeric fields to avoid showing extreme field values on a data record. Values beyond the top code or bottom code are replaced either by the average of the values in excess of the respective top code or bottom code or through the application of various imputation methodologies.	No updates; 2005 description remains accurate.

Table D.1. (*continued*)

Agency	Public use microdata and who reviews	Restricted access allowed for researchers	Statistical disclosure limitation methods for public use microdata	Any update?
U.S. Census Bureau	Yes -- Disclosure Review Board	Yes	"Microdata cannot show geography below a population of 100,000. For the most detailed microdata, that threshold is raised to 250,000 or higher." "For small populations or rare characteristics noise may be added to identifying variables, data may be swapped, or an imputation applied to the characteristic. Census data, which lacks the component of protection provided by sampling, employs targeted swapping in addition to the combination of table design and thresholds described above." Hawala, Zayatz, and Rowland (2004): "To insure that any data tabulation requested by external users will not disclose respondents' identities, the U.S. Census Bureau uses data recoding and data swapping (Zayatz 2003)." Zayatz (2005): "There are several disclosure avoidance techniques that we are currently using for our microdata files including geographic thresholds, rounding, noise addition, categorical thresholds, topcoding, and data swapping."	Subsampling (only a fraction of the full microdata file from the survey/census is released) was used for the Decennial long form prior to ACS and is now used for ACS releases. Synthetic data use is limited to the production of partially synthetic estimates for certain, small, specialized sub-populations. These subpopulations comprise only a small subset of the microdata files released. Synthetic data, in its various forms, is not widely used to protect Census microdata files.

Table D.1. (*continued*)

Agency	Public use microdata and who reviews	Restricted access allowed for researchers	Statistical disclosure limitation methods for public use microdata	Any update?
Bureau of Labor Statistics (BLS)	BOC Collects Title 13	Yes	BLS releases very few public use microdata files. Most of these microdata files contain data collected by the Census Bureau under an interagency agreement and Census's Title 13 authority. For these surveys (Current Population Survey, Consumer Expenditure Survey) the Census Bureau determines the statistical disclosure limitation procedures that are used. BLS releases public-use data files from three surveys in the family of the National Longitudinal surveys. Disclosure limitation methods used for the public use microdata files containing data from the National Longitudinal Survey of Youth, collected under contract by Ohio State University and Research Center at the University of Chicago, are similar to those used by the Census Bureau.	No update, but 2005 description has been edited.

Table D.1. (*continued*)

Agency	Public use microdata and who reviews	Restricted access allowed for researchers	Statistical disclosure limitation methods for public use microdata	Any update?
National Center for Education Statistics (NCES)	Yes -- Disclosure Review Board	Yes	All direct individually identifiable information (for example, school name, individual name, addresses) is stripped from the public use file. Continuous variables are top and bottom coded to protect against identification of outliers. After this has been done, a casual data intruder might identify an individual respondent by first identifying the sampled institution for the individual. To prevent identification of the sampled institution, all known publicly available lists of education institutions that contain institutions' names and addresses are gathered. Each list is matched with the sample file using all common variables between the two files. If an institution can be identified to within 2 other institutions, using an appropriate distance measure, then that is a disclosure risk and must be resolved before releasing the data. If too many disclosure risks are obtained then a common variable(s) may be dropped from the public-use file, or the variable(s) may be coarsened. If there are only a few identified disclosure risks found, the appropriate action is to selectively perturb a set of the common variables until all disclosure risks are resolved. This analysis is repeated sequentially for each list file until it can be applied to each list file without identifying any disclosure risks. Whenever institution head, teacher, student, or parent data are clustered, a subsampling of respondents is required. Data from respondents selected into this subsample are reviewed using an additional disclosure edit. The edit is either: (1) a blanking and imputing, or data swapping of a sample of sensitive items collected; or (2) a data swapping of the key identification variable of the respondent or institution. The amount of editing is set at a level sufficient to protect the confidentiality of the respondent, while not compromising the analytic usefulness of the data file.	The basic procedures are still the same. NCES has added additional measures as diagnostics to determine which of several trial data perturbations to select to meet the requirement to protect the confidentiality of the respondent, while not compromising the analytic usefulness of the data file.

Table D.1. (*continued*)

Agency	Public use microdata and who reviews	Restricted access allowed for researchers	Statistical disclosure limitation methods for public use microdata	Any update?
National Center for Health Statistics (NCHS)	Yes – Con-fidentiality Officer and Disclosure Review Board	Yes	It is NCHS policy to make microdata files available to the scientific community so that additional analyses can be made for the country's benefit. Such files follow guidance and principles contained in the NCHS Staff Manual on Confidentiality (September, 2004), Section 9 "Avoiding Inadvertent Disclosures Through Release of Microdata," and the NCHS Checklist for the Release of Micro Data Files. These guidelines require that detailed information that could be used to identify individuals (for example, date of birth) should not be included in microdata files. The identities of geographic places and characteristics of areas with fewer than 100,000 people are never to be identified, and it may be necessary to set this minimum at a higher number if research or other considerations so indicate. Information on the drawing of the sample that could identify data subjects should not be included.	The techniques, methods, and guidance used to protect NCHS's public use microdata have largely remained unchanged since 2004, although there are a few exceptions. The changes detailed below have been in response to changes in technology and proliferation of external data, and were made to reduce disclosure risk to individuals in NCHS data systems. 1. Vital record (birth, death, fetal death and linked birth/infant death) public use microdata files beginning with the 2005 data year contain individual-level vital event data at the national level only. The files for births, deaths, fetal deaths and linked birth/infant death generally include most other items from the vital record with the exception of exact dates. 2. Some NCHS surveys collect information on observable health conditions/limitations or rare conditions. This information is often excluded from public use microdata files because the information, in combination with the extensive information for other characteristics, is considered to pose too great a risk of respondent re-identification by knowledgeable insiders or from media coverage. 3. The level of detail for some variables has been reduced on public use microdata files. This includes geographic information for almost all files, but also includes items such as household relationships, race/ethnic categories and other observable characteristics that could increase risk of identification when combined with other indirect identifying information. Compared to 2004, NCHS staff responsible for developing public use microdata files spend more time identifying and researching external files available via the Internet to assess whether external sources can be used to re-identify NCHS survey respondents. Advances in computer technology, the introduction of Big Data and Open Data initiatives pose new challenges for preparing public use microdata files that were not present 10 years ago.

Table D.1. (*continued*)

Agency	Public use microdata and who reviews	Restricted access allowed for researchers	Statistical disclosure limitation methods for public use microdata	Any update?
National Center for Health Statistics (NCHS) continued	Yes – Confidentiality Officer and Disclosure Review Board	Yes	Refer back to previous row.	Although NCHS has reduced the level of detail available on public use microdata files since 2004, we have attempted to balance this by making non-public use microdata files more available through expansion of RDC sites, use of special agreements permitting access under controlled conditions (e.g., Designated Agent Agreements or DUAs), and development of new access tools. a. The NCHS RDC now offers researchers four access modes to access restricted use NCHS microdata including: (1) on-site at the NCHS RDC, (2) on-site at a Census RDC, (3) remote access, and (4) staff assisted research option. Additional information about each access mode can be found at the following location: http://www.cdc.gov/rdc/B2AccessMod/ACs200.htm. b. NCHS is developing new tools for data access. For example, NCHS is developing a National Health Interview Survey Online Analytic Real-Time System (OARS) to help meet the need for state-level estimates. This tool will allow health experts, policymakers, journalists, and others to search and compare health statistics by county, region, and state nationwide for grant proposals, needs assessments, research, news reporting, and policymaking. Additional information on OARS can be found at: http://www.cdc.gov/nchs/data/bsc/nhis_online_analytic_realtime_system.pdf NCHS remains committed to making data as widely available as possible while protecting the confidentiality of respondents. Approximately 95 percent of NCHS collected data are released in public use microdata files and most of the remaining data are available under controlled conditions that meet our legislative mandates to protect respondent identity.
Agency for Healthcare Research and Quality (AHRQ)	Yes – Disclosure Review Board	Yes	The disclosure limitation procedures used by AHRQ are similar to those of NCHS.	No updates; AHRQ continues to use procedures similar to NCHS but without the NCHS-specific revisions detailed above.

Table D.1. (*continued*)

Agency	Public use microdata and who reviews	Restricted access allowed for researchers	Statistical disclosure limitation methods for public use microdata	Any update?
National Agricultural Statistics Service (NASS)	No	Yes	NA	No updates
Economic Research Service (ERS)	No	Yes	NA	No updates
Bureau of Economic Analysis (BEA)	No	Yes	NA	No updates
Social Security Administration (SSA)	Yes - 2 Disclosure Review Boards. One handles Title 13 data; the other does not.	Yes	When releasing public use microdata files, individual identifiers are removed from all records, and other distinguishing characteristics are modified to prevent identification of persons to whom a record pertains. Records are sequenced in random order to avoid revealing information due to the ordering of records on the file. Top codes and bottom codes are employed for numeric fields to avoid showing extreme field values on a data record. Values beyond the top code or bottom code are replaced by the average of the values in excess of the respective top code or bottom code. Top code and bottom code values are derived at the national level and the replacement values are derived and applied at the state level when appropriate. Values shown for some categorical fields are combined into broader groupings than those present on the internal file, and dollar amounts are rounded. Top code and bottom code values, replacement values, and related information are provided to users as part of the file documentation.	Since 2010, the DRB has built a working relationship with the Office of Open Government in part to prevent the mosaic effect. Based on White House Open Government initiatives, SSA has enhanced their procedures for releasing data on the Agency website and onto Data.gov. The Data.gov National/Homeland Security and Privacy/Confidentiality Checklist and Guidance (referred to as the NHSP Checklist) is part of the guidance from the White House and is to be used by departments and agencies submitting datasets for publication on Data.gov. This Checklist augments the processes SSA is using to meet its existing statutory, regulatory or policy requirements for protecting national/homeland security and privacy/confidentiality interests. Since 2012, the DRB includes an external voting board member from the U.S. Census Bureau. This provides an avenue for the DRB to ensure that agency staff are informed of the latest disclosure avoidance techniques utilized and recommended by the Bureau's DRB.

Table D.1. (*continued*)

Agency	Public use microdata and who reviews	Restricted access allowed for researchers	Statistical disclosure limitation methods for public use microdata	Any update?
Internal Revenue Service (IRS)	Yes - Legislatively Controlled	No	SOI produces one annual public- se microdata file, known as the SOI "tax model", containing a sample of data based on the Form 1040 series of individual tax returns. The disclosure protection procedures applied to this file include: (1) subsampling certainty records at a 33 percent rate; (2) removing certain records having extreme values; (3) suppressing certain fields from all records and geographical fields from high income records; (4) top coding and modifying some fields; (5) blurring some fields of high income records by locally averaging across records; and (6) rounding amount fields to four significant digits. To help ensure that taxpayer privacy is protected in the SOI tax model file, SOI has periodically contracted with experts who employ "professional intruder" techniques to both verify that confidentiality is protected and to inform the techniques to be applied to future releases of the SOI tax model file.	SOI reviews its statistical disclosure limitation procedures for its public use microdata file and introduces enhancements on an ongoing basis. For example, the maximum sampling rate was changed to 10 percent several years ago, and multivariate blurring replaced univariate bluring for key fields on high-income returns. SOI is currently redesigning its public use file.

Table D.1. (*continued*)

Agency	Public use microdata and who reviews	Restricted access allowed for researchers	Statistical disclosure limitation methods for public use microdata	Any update?
Bureau of Transportation Statistics (BTS)	Yes – Disclosure Review Board	No	The BTS Confidentiality Procedures Manual documents the confidentiality procedures for the agency. For most microdata and tabular data products, BTS program managers are required to complete a checklist identifying potential disclosure risks and outline any steps taken to mitigate such risks. The BTS's DRB reviews the data product and checklist and makes a final determination on disclosure risk. The DRB can recommend application of SDL methods prior to public dissemination. BTS uses various microdata SDL methods based on the disclosure review findings and the unique characteristics of the data files. Some SDL procedures used include data suppression and modification. Data modification includes recoding continuous variables into categorical variables, collapsing categories, top and bottom coding, introduction of noise, and data swapping. BTS program managers must also identify any external data that could be matched to BTS datasets and take steps to minimize the ability to match.	No updates; 2005 description remains accurate.

Table D.1. (*continued*)

Agency	Public use microdata and who reviews	Restricted access allowed for researchers	Statistical disclosure limitation methods for public use microdata	Any update?
Bureau of Justice Statistics (BJS)	Yes - legislatively controlled agency review	No	The same requirements under Title 13 of the U.S.C. that cover the Census Bureau are followed by BJS for those data collected for BJS by the Census Bureau. Standards for microdata protection are incorporated in BJS enabling legislation. Individual identifiers are routinely stripped from all microdata files before they are released for public use.	BJS has allowed access to restricted files since at least the year 2000, if not before. Direct identifiers are routinely removed from all microdata files prior to release. Indirect identifiers—for example, geographic identifiers, dates of unique events, or age—undergo disclosure avoidance measures commensurate with the level of release (public, restricted, or enclaved). Measures commonly used include categorization of continuous variables, top- or bottom-coding, rounding, addition of noise, and data swapping. Most restricted microdata files are available by application from the National Archive of Criminal Justice Data (NACJD). The Archive is in the process of implementing and expanding technology that allows remote access to restricted microdata files. With this technology, the user does not receive or download the microdata, but rather logs into and analyzes the data on a secure NACJD server. In 2011, BJS began making extremely sensitive microdata files available onsite at the University of Michigan in the Interuniversity Consortium for Political and Social Research (ICPSR) Data Enclave in Ann Arbor, MI (also by application).

This page has been left blank for double-sided copying.

REFERENCES

Federal Trade Commission. "Data Brokers: A Call for Transparency and Accountability." Washington, DC: Federal Trade Commission. Available at [http://www.ftc.gov/system/files/documents/reports/data-brokers-call-transparency-accountability-report-federal-trade-commission-may-2014/140527databrokerreport.pdf]. May 2014.

Gerald Gates, "How Uncertainty about Privacy and Confidentiality Is Hampering Efforts to More Effectively Use Administrative Records in Producing U.S. National Statistics," *Journal of Privacy and Confidentiality*, vol. 3, no. 2, 2011, pp. 3 to 40.

Hawala, Sam, Laura Zayatz, and Sandra Rowland. "American FactFinder: Disclosure Limitation for the Advanced Query System." *Journal of Official Statistics*, vol. 20, no. 1, March 2004, pp. 115-124.

Johnson, Barry W. "Presentation to the Council of Professional Associations on Federal Statistics." Presented at the COPAFS Quarterly Meeting, Washington, DC, June 6, 2014.

Office of Management and Budget, Executive Office of the President. "Open Data Policy—Managing Information as an Asset." Memorandum Number M-13-13. Washington, DC: OMB, May 9, 2013a. Available at [http://www.whitehouse.gov/sites/default/files/omb/memoranda/2013/m-13-13.pdf]. Accessed July 15, 2013.

Office of Management and Budget, Executive Office of the President. "Supplemental Guidance on the Implementation of M-13-13 'Open Data Policy—Managing Information as an Asset.'" Washington, DC: OMB, May 9, 2013b. Available at [http://www.whitehouse.gov/sites/default/files/omb/memoranda/2013/m-13-13.pdf]. Accessed July 15, 2013.

Office of Management and Budget, Executive Office of the President. "Statistical Policy Directive: Fundamental Responsibilities of Federal Statistical Agencies and Recognized Statistical Units." *Federal Register*, vol. 79, no. 98, May 21, 2014, pp. 29308-29312.

Office of Science and Technology Policy, Executive Office of the President. "Increasing Access to the Results of Federally Funded Scientific Research." Memorandum. Washington, DC: OSTP, February 22, 2013. Available at [http://www.whitehouse.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_2013.pdf].

Ohm, Paul. "Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization." *UCLA Law Review*, vol. 57, 2010, pp. 1701-1777.

Solove, Daniel J., and Chris Jay Hoofnagle. "A Model Regime of Privacy Protection," *University of Illinois Law Review*, vol. 2006, no. 2, 2006, pp. 357-404.

U.S. Department of Health and Human Services. "Summary of the HIPAA Privacy Rule." Washington, DC: U.S. Department of Health and Human Services. Last revised May 2003. Available at [http://www.hhs.gov/ocr/privacy/hipaa/understanding/summary/privacysummary.pdf]. Accessed June 3, 2014.

U.S. Department of Health and Human Services, Office for Civil Rights. "Guidance Regarding Methods for De-identification of Protected Health Information in Accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule." Washington, DC: U.S. Department of Health and Human Services, November 26, 2012. Available at [http://www.hhs.gov/ocr/privacy/hipaa/understanding/coveredentities/De-identification/hhs_deid_guidance.pdf]. Accessed June 3, 2014.

U.S. Department of Health and Human Services. "Summary of the HIPAA Security Rule." Available at [http://www.hhs.gov/ocr/privacy/hipaa/understanding/srsummary.html]. Accessed June 3, 2014.

U.S. Department of Health, Education and Welfare, "Records, Computers, and the Rights of Citizens," Report of the Secretary's Advisory Committee on Automated Personal Data Systems. Washington, DC: U.S. Department of Health, Education and Welfare. Available at [http://www.justice.gov/opcl/docs/rec-com-rights.pdf]. July 1973.

White House. "Executive Order 13642—Making Open and Machine Readable the New Default for Government Information." Washington, DC: May 9, 2013 (White House 2013). Available at [http://www.whitehouse.gov/the-press-office/2013/05/09/executive-order-making-open-and-machine-readable-new-default-government-]. Accessed December 26, 2013.

Zayatz, Laura. "Disclosure Limitation for Census 2000 Tabular Data." Paper presented at the Joint European Commission for Europe and EUROSTAT Work Session on Statistical Data Confidentiality. Working Paper No. 15. Available at [http://www.unece.org/stats/documents/2003/04/confidentiality/wp.15.e.pdf]. 2003.

Zayatz, Laura. "Disclosure Avoidance Practices and Research at the U.S. Census Bureau: An Update." Research Report Series, Statistics #2005-06. Washington, DC: U.S. Census Bureau, August 31, 2005.

APPENDIX E

BACKGROUND PAPER:
RELEASING FEDERAL MICRODATA:
STRATEGIES, ISSUES, AND DEVELOPMENTS

This page has been left blank for double-sided copying.

I. INTRODUCTION[1]

Recent open data initiatives by the Department of Health and Human Services (HHS) and the White House have encouraged the release of increasing numbers of datasets containing individual records (microdata) collected from survey respondents, doctor and hospital visits, and medical claims. At the same time, federal agencies that release data collected from individuals and establishments have an obligation under the law to protect the confidentiality of those supplying the data as well as the information provided.2 The challenge faced by HHS and other federal agencies is to achieve an appropriate balance between providing the public with useful datasets and protecting the confidentiality of the individuals and establishments whose information is contained in the data. Addressing this challenge is made more difficult by what has been termed in other circles the "mosaic effect," the idea that disparate pieces of information—though individually of limited utility—become significant when combined with other types of information (Pozen 2005). The concern is that the datasets being released in large numbers—more than 1,000 by HHS alone—provide the pieces of intelligence that when assembled correctly disclose information that the federal government is required to maintain as confidential.

This paper reviews the issues that arise in protecting the confidentiality of data collected by the federal government. The literature discussed includes standard references on the topic from the past 10 years as well as additional articles and unpublished papers identified through searches of relevant journals and conference proceedings and bibliographic sources. Key words searched included de-identification, re-identification, disclosure avoidance, public use files, confidentiality, disclosure risk, and mosaic effect.

Chapter II discusses the concept of disclosure, reviews some instances in which individuals have been re-identified in data released to the public (although not by the federal government), explores the sources of disclosure risk, and concludes with observations on the mosaic effect. Chapter III reviews approaches to protecting data against disclosure, comments on the legal environment, and discusses methods commonly used to assess disclosure risk in public use datasets.

There is an inherent contradiction between protecting data from disclosure and maximizing its value to users. Restricting access to the data or altering its contents so as to reduce the risk of disclosure also diminishes the data's usefulness for research. Chapter IV discusses ways in which the utility of data is reduced by strategies to limit disclosure and, related to this, how the loss of information can be measured.

[1] This background paper was prepared by John L. Czajka, Amang Sukasih, and Craig Schneider.

[2] For a summary of recent documents explaining the open data policy and a review of relevant laws establishing the government's object to project the confidentiality of the data it collects, see Appendix D.

II. DISCLOSURE RISK

In a seminal paper on the protecting the confidentiality of data released to the public, Dalenius (1977) described the problem in the following terms: "access to a statistical database should not enable one to learn anything about an individual that could not be learned without access" (cited in Dwork and Naor 2010). The literature on disclosure distinguishes between identity disclosure and attribute disclosure (Duncan and Lambert 1989). An identity disclosure assigns a name to a record in a database while an attribute disclosure assigns a characteristic to an individual or small group of individuals. Identity disclosure implies attribute disclosure, but attribute disclosure can occur without identity disclosure. A database may reveal that all of the members of a particular subpopulation share a specific characteristic. In this instance something is learned about individual X even though no record in the database can be assigned unambiguously to that individual. Research on protecting confidentiality in tabular data has recognized the risk of attribute disclosure and devoted considerable attention to it, but research on protecting confidentiality in microdata has focused on identity disclosure. For federal agencies in particular, the goal in protecting the confidentiality of microdata is to prevent the re-identification of records that have been released as anonymous.

Some confidentiality provisions in federal legislation interpret any disclosure of an individual identity—regardless of how it is accomplished—as a violation of the law. Other confidentiality provisions—for example, those in the Health Insurance Portability and Accountability Act (HIPAA)—acknowledge that it is impossible to release data for which the risk of disclosure is zero. If the effort to prevent disclosure produced a very low risk of re-identification, that effort would satisfy the legal requirement for protecting the data, even if a breach of confidentiality occurred. An even more liberal definition of protection is used by the United Kingdom's National Office of Statistics, which doesn't consider a re-identification to be a disclosure by the agency if the breach required more than a reasonable amount of effort (Duncan et al. 2011, p. 28).

A. Re-identification of Individuals in Data Released to the Public

There are exceedingly few documented instances of the re-identification of individual persons in datasets that have been released to the public. None has involved a sample survey or a federal government database, and few have involved data that were protected by methods that would be considered rigorous by today's standards.

The most famous re-identification, which predated the HIPAA Privacy Rule and influenced its development, was Latanya Sweeney's 1996 re-identification of a substantial fraction of the records in a database of Massachusetts state employees discharged from hospitals (Cavoukian and Castro 2014). The records had been de-identified by removal of names, Social Security numbers, health insurance IDs, hospital names, doctors' names, and other obvious identifiers, but ZIP codes, sex, and date of birth had been retained because of their analytic value. Sweeney's re-identification used the values of these three variables obtained from a city voter registration list that was purchased for a nominal fee. The employees who were re-identified included the state's governor. With this work Sweeney (1997) demonstrated that the removal of explicit identifiers does not guarantee that records are anonymous—that is, unable to be associated with individual persons.

Underscoring the latter point, Sweeney et al. (no date) used voter registration and on-line public records to re-identify a subset of personal records in a public database of medical and genomic information. Participants in the Personal Genome Project could choose to make their data public, and while the profiles were not explicitly identified, the authors report that the consent forms did not guarantee privacy. Out of 1,130 profiles that were made public as of September 1, 2011, 579 or 51 percent included the participant's full date of birth, gender and five-digit ZIP code. Using a sample of voter registration records acquired from a commercial source for the ZIP codes represented among the 579 profiles, the authors were able to match 130 of the profiles uniquely to individual voter records and obtain names. The authors were also able to match 156 of the profiles to the on-line public records, although nearly half of these duplicated matches to the voter records. Genome project staff confirmed that 93 percent of the names obtained from the voter records and 87 percent of the names obtained from the on-line public records agreed with the names recorded in the profiles, and allowance for nicknames would have raised these rates even higher. These results provide further evidence of the uniqueness of many combinations of ZIP code, date of birth, and gender. Indeed, Sweeney (2000) estimated that 87 percent of the U.S. population could be uniquely identified by the combination of 5-digit ZIP code, date of birth, and gender.[3]

El Emam et al. (2011) conducted a systematic review of known re-identification attacks on health data and other types of data. The review uncovered 14 re-identification attacks in which at least one individual was accurately re-identified. Of the 14 examples, 11 were conducted by researchers solely to demonstrate or evaluate the risk of re-identification. Notably, only 2 of the 14 involved databases that were protected in accordance with current standards. One of the two was a health database, consisting of records from a regional hospital that were protected with the HIPAA Safe Harbor Privacy Rules, and the rate of re-identification was found to be very low—just 0.022 percent, representing two persons—despite strong assumptions about what an intruder might know (see Kwok and Lafky 2011). Overall these results confirm the value of current best practices for de-identification but also indicate that there is merit in complementary legal protection, where possible. The study also highlights a need for better information on disclosure risk, which could be obtained from re-identification attacks on large databases protected with the best current methods.

Another example of re-identification, which received considerable attention in the media, was the re-identification of published Netflix rental histories from the movie reviews submitted by (identified) Netflix customers (see Narayanan and Shmatikov 2008). Although this example does not bear directly on the risks associated with federal data in general or health data in particular, it demonstrates what can be possible with data that are publicly available.

B. Sources of Disclosure Risk

To understand the potential sources of disclosure risk, we need to be aware of who are the potential intruders—that is, those who might attempt to re-identify records in federal microdata—and their capabilities, what data they might use in their re-identification attempts, and what tools are available to assist them in doing so.

[3] A replication of Sweeney's calculations with 2000 census data found that the percentage of the population that could be uniquely identified with these same variables had fallen to 63 percent (Golle 2006).

1. Potential Intruders

Potential intruders—those who might attempt to re-identify entities in the data and use the information in some way—encompass a wide range of possible users. Among these the greatest threat is posed by those with exceptional computer skills or with access to information on a large number of identified individuals or with exceptionally detailed information on a particular individual. These attributes afford them an advantage in circumventing the protections that have been applied to the data that the government releases to the public. Hackers have demonstrated their ability to defeat high-level security measures, and they pose a constant threat, but attention must also be focused on individuals in organizations that collect and maintain extensive personal information for business purposes—such as credit bureaus. The challenge presented by those with access to proprietary databases is significant. First, the records in these databases are identified. Second, the records include basic demographic information that is repeated in many public databases of individuals. Third, because the data are proprietary, they are not readily available to government researchers to use in testing the adequacy of the protections applied to the data the government releases. Fourth, linking to information in the government databases could provide important enhancements to the propriety data, providing a significant incentive for individuals and organizations to attempt to link their records to the government data.

Family members of respondents or subjects in administrative datasets are also potential intruders. This is particularly true for data sources that contain sensitive information that other family members know was collected. For example, the National Survey on Drug Use and Health (NSDUH), which is conducted by the Substance Abuse and Mental Health Services Administration (SAMHSA), collects information from teenage children through a questionnaire administered privately to the respondents. Parents will know that their children were respondents but are not privy to the information that their children provide. Before releasing data from the survey, SAMHSA contracted with RTI International to develop a disclosure limitation methodology that would be capable of addressing the exceptional challenges presented by the NSDUH data (National Research Council 2005).[4] Another example of potential intruders from the family is former spouses, who may be in a position to realize a financial benefit by gaining knowledge of the finances of their ex-partners following the divorce and may have access to extensive financial and other information prior to the divorce.

2. Auxiliary Data

Following Dwork and Naor (2010), the data that a potential intruder would use to re-identify records on a public use file may be described as auxiliary data. The challenge in protecting a public use file from any possibility of re-identification is the inability to guarantee that there are no auxiliary data in anyone's possession that would enable re-identification of even one record.5 To assess the level of risk, however, some understanding of the nature and amount of

[4] The MASSC methodology (Singh et al. 2003) developed for this purpose is discussed in Chapter III.

[5] Combining this understanding of auxiliary data with Dalenius's assertion that access to a statistical database should not enable one to learn anything about an individual that could not be learned without access, Dwork and Smith (2009) explain the concept of differential privacy, which views disclosure risk in a different light. They argue that with the right auxiliary information, an intruder could learn something about an individual whether or not that individual was included in a particular database. Differential privacy compares the risk that an individual encounters by being included versus not included in that database.

information that a potential intruder could access to re-identify individuals in a database is critical.

Purdam and Elliot (2002) conducted a review of public data sources in Europe to determine what data could be used, potentially, to re-identify records in data released by government agencies. A comparable assessment for the U.S. does not exist although numerous papers have been written that discuss various sources of data.

Benitez and Malin (2009) estimated the risk of re-identification from voter registration lists by state for datasets protected by the HIPAA Safe Harbor and Limited Dataset policies. Because voter registration lists vary in cost, the study addresses both the probability of successful re-identification, given the attempt, and cost factors that may affect the likelihood of an attempt. In their study, the Safe Harbor dataset included year of birth, gender, and race while the Limited Dataset policy added county and date of birth. The results showed wide variation in estimated risk and in the unit price of each potential re-identification by state, with substantially greater risk under the Limited Dataset versus the Safe Harbor policy. They concluded that blanket protection policies expose different organizations (in different states) to differential disclosure risk.

Barth-Jones (2012) cautions that voter registration lists may exclude a significant proportion of the population. In Cambridge, Massachusetts, when Governor Weld was re-identified, voter registration records covered about half of the adult population. The implication is that a set of characteristics that is unique in an area's voter registration records may not be unique within the entire population. One cannot know that from just the voter records, however. To re-identify individuals within a small geographic area using simple demographic characteristics, one would need, in effect, a population register containing such characteristics for all individuals in the population.

Duncan et al. (2011) observe that "the Achilles' heel for data stewardship organizations in dealing with the practicalities of risk assessment is their lack of knowledge about what data snoopers know." While much is known or can readily be determined about the contents and coverage of certain public databases maintained by the states, the same cannot be said about the data that are compiled, maintained, and resold by commercial entities. A recent study by the U.S. Government Accountability Office (2013) concluded that "the advent of new and more advanced technologies and changes in the marketplace for consumer information have vastly increased the amount and nature of personal information collected and the number of parties that use or share this information." The report provides examples of the types of information collected. For instance, in addition to individual and household demographic information, Acxiom collects household wealth indicators such as estimated ranges of household income, indicators of income-producing assets, and estimated ranges of net worth. More precise information includes the year, make, and model of household vehicles and household life event indicators. Experian's data include a variety of physical ailments such as diabetes, high blood pressure, high cholesterol, and visual impairments; types of financial investments; and consumption tastes. The detailed contents are known by their internal users and, perhaps to a lesser degree, those who have bought data extracts, but this information—and, especially, the quality of the data—is not readily accessible to researchers seeking ways to protect federal data from re-identification. Some government agencies have purchased sets of data from these sources for their own research

purposes—not only to learn how to better protect the confidentiality of their data from such sources but as an additional data source that might be useful for nonresponse adjustment and imputation, and for reduction in nonsampling error generally.

Although the data that are "out there" in the public domain or held by private sources may be considerable, the threat that it presents is mitigated to at least some degree by discrepancies between the values recorded in these data sources and the values reported by survey respondents or collected by administrative agencies. Such "data divergence," as described by Duncan et al. (2011), includes not only measurement error but conceptual differences in the way that data elements are defined in different sources. For example, surveys that collect income data usually ask for gross earnings, but the earnings data collected by the Internal Revenue Service—some of which may end up in commercial databases by way of loan applications—are taxable earnings, which can be considerably less than gross earnings. Timing is another factor in data divergence. The data accessible to a would-be intruder and the information reported in federal datasets may be separated by years, which can matter a great deal for health conditions, income, and even geographic location.

3. Record Linkage

When two files contain some of the same individuals, the records common to the two files can be linked if the two files also contain some of the same variables. When the two files contain unique and valid numeric identifiers, the records can be linked using "exact matching" on those fields—as is commonly done when files contain Social Security numbers. When the conditions for exact matching are absent but other, non-unique or imperfect identifiers are present, either "probabilistic record linkage" or distance-based matching can be used as an alternative.

Probabilistic record linkage separates all possible combinations of records into likely matches, likely non-matches, and a group that cannot be confidently assigned to either and would require a manual "clerical" review to determine in which category they belong. Probabilistic record linkage is often applied to link records based on names and addresses, where duplicate names and spelling errors are possible and people may have been living at different addresses when the two files were created. The Census Bureau uses probabilistic record linkage to unduplicate the records collected in the decennial census, as some people may have been enumerated multiple times.

Distance-based matching may be used instead of probabilistic record linkage when one or both files contain no explicit identifiers but the two files contain quantitative variables—such as income. Different distance functions may be used. The Euclidean distance is commonly used because of its simplicity and general effectiveness. It can also be used to match a single observation to a dataset, simulating an intruder who is trying to find a single, target individual. Torra et al. (2006) explored an alternative metric, the Mahalanobis distance, with notable success, although this approach is not nearly as straightforward to implement and is designed for matches between two datasets.

While record linkage is typically applied to variables that are common between two files, it is also possible to apply record linkage methods to files with no variables in common. One approach relies on correlations between variables (see, for example, Domingo-Ferrer and Torra 2003, which employs clustering methods). The effectiveness of such matching increases with the

strength of the correlation between variables and the degree of overlap between the two files—that is, the percentage of records appearing in both files. An alternative approach applied by Torra (2000) uses a method called ordered weighted aggregation.

Record linkage techniques other than exact matching are computationally intensive because they entail "looking at" all possible pairs of records to determine a most likely match in one file for each record in the other file. When probabilistic record linkage in its present form was introduced (see Fellegi and Sunter 1969) and for many years afterwards, it was common to subset or "block" the files being linked to reduce the computational time. Only those potential links within the same block were evaluated. This precluded the possibility of identifying links across blocks. As processing capacity and computational efficiency have increased, this constraint has largely disappeared. For example, whereas the Census Bureau had to employ blocking when searching for duplicates in the decennial census file in earlier censuses, the bureau routinely matches entire population files when performing record linkage for census unduplication and other purposes.[6] This capability is accessible to potential intruders as well and represents perhaps the most important dimension in which the risk of re-identification in public use files has increased.

C. The Mosaic Effect

The "mosaic effect" is a new term in the literature on confidentiality. It received prominent mention in Memorandum M-13-13 from the Office of Management and Budget (OMB), "Open Data Policy—Managing Information as an Asset" (OMB 2013), but a search for the term in the database Google Scholar produced no relevant hits.

The notion of a mosaic effect is derived from the mosaic theory of intelligence gathering, in which disparate pieces of information—although individually of limited utility—become significant when combined with other types of information (Pozen 2005). Applied to public use data, the concept of a mosaic effect suggests that even anonymized data, which may seem innocuous in isolation, may become vulnerable to re-identification if enough datasets containing similar or complementary information are released. Even though personal identifiers are removed from these datasets, an intruder who is able to piece together enough information may be able to re-identify individuals whose data are contained in one or more of these datasets. To do so, the intruder must possess or be able to secure at least some data on known individuals. Such information is readily available in computerized form in voter registration records, hospital discharge records, commercially marketed databases, and other sources (Rothstein 2010).

Another potential source of information on individuals is the worldwide web. Malin (2005) demonstrates the application of "trail matching" methods to re-identify IP addresses of website visitors. Common patterns in data trails left behind after website visits can be used to discover relationships between them that enable re-identification when some of the locations capture identifying information along with anonymous data. While re-identification in this context

[6] Blocking remains useful as a strategy for reducing the number of false matches. For example, while gender may be recorded incorrectly on occasion, allowing matches that disagree on gender may produce far too many false matches to justify the few additional true matches that it might detect. However, the use of blocking for the sole purpose of reducing computational time is becoming less and less common.

represents a different problem than re-identification of health data released by federal agencies, Malin's results illustrate how re-identification can be accomplished with large amounts of mostly anonymous data when identifiers are attached to some of it.

III. PROTECTING DATA AGAINST DISCLOSURE

There are two general approaches that are used to release microdata in a way that protects the data from disclosure. One is by restricting access to the data, and the other is by restricting the data that are released for public access (National Research Council 2005). The latter approach encompasses a wide range of techniques that include suppressing variables and changing their values. Legal restrictions may be relevant to either approach, although as we note below, federal regulations place much more responsibility upon the data producer than the user.

A common way to view disclosure risk is to express the probability of disclosure as the product of two terms: (1) the probability of a successful re-identification conditional on someone trying to re-identify a record and (2) the probability that someone will try to re-identify a record (Marsh et al. 1991). A data producer can reduce the risk of disclosure by reducing either of these probabilities. For example, charging a high fee for a public use file reduces the probability that a potential intruder will even acquire the file. Sampling reduces the certainty that someone of interest is included on the file, which will also discourage potential intruders. Altering the data values in various ways reduces the likelihood of a re-identification and in so doing may also discourage attempts at re-identification. In addition, altering the data reduces the potential value of the information gained by re-identification, which may further reduce the likelihood that a would-be intruder will attempt a re-identification.

In this chapter we review strategies for restricting access and restricting the data. We also examine the legal environment and discuss approaches to assessing disclosure risk.

A. Restricted Access

There are three basic mechanisms that federal agencies use to provide researchers with restricted access to data that are not released to the public. These include licensing, research data centers (RDCs), and secure remote access. These are discussed in turn below.

1. Licensing

Under licensing arrangements, prospective users request restricted (that is, non-public) data files through a formal application process. To obtain such data, users must demonstrate that the data will be stored and used in a secure environment that meets the issuing agency's standards. This may require an initial agency inspection and a willingness on the part of the user to submit to subsequent inspections. As part of the proposal the user will generally have to explain why the data are needed and how they will be used, and access may be limited to variables and records for which the user can demonstrate a critical need. To receive the data, the user typically has to sign a nondisclosure agreement. If the user's future research is dependent on further use of the agency's data, such an agreement is likely to provide a powerful disincentive to violate its terms. As an example of a well-established licensing program, the Centers for Medicare & Medicaid Services (CMS) provides extensive data from its Medicare and Medicaid programs through its Chronic Conditions Warehouse (Shatto 2014).

2. Research Data Centers

Several federal agencies maintain RDCs, in which approved users can access agency data that are not released to the public. The data never leave the site, and output produced from data held in the RDC cannot be removed without a disclosure review, which can take different forms. For example, RDC staff may be authorized to review output, or the output may have to be screened by an agency disclosure review board. The types of data manipulations allowed to RDC users are limited. Linkages between databases may be prohibited or restricted. Users may not be allowed to attach portable storage devices to the computers or terminals that they use, and even printing of output may not be permitted (one RDC emails output to users after it has been reviewed). Obtaining access to an RDC requires submission of a proposal, and acceptable uses may be restricted to applications that carry potential benefits to the agency. Some agencies impose additional requirements; the Census Bureau, for example, requires that its RDC users undergo a background check and obtain employee-like "Special Sworn Status." The entire approval process may require several months. Despite these restrictions, RDCs are an important venue for access to data that agencies do not release on public use files. Reflecting demand, the Census Bureau has added several new RDCs in the past several years.

3. Secure Remote Access

A number of federal agencies allow users remote access to agency data that are not released on public use files. This can take a number of different forms. For example, the Census Bureau allows users to request tabulations from decennial census files that include more detail than the numerous tabulations that can be obtained from the bureau website (FCSM 2005). The requests are reviewed to ensure that the tabulations do not present a disclosure risk. The National Center for Health Statistics allows approved RDC users to submit programs remotely, although the software that can be used for this purpose is more limited than what is available in the RDC, and certain functions are not accessible. The advantage to the user lies in not having to travel to the RDC, which can be important when the research can be conducted most efficiently with numerous, intermittent program submissions, each requiring extensive review of the results before the next program can be prepared. Some RDCs charge a daily fee for in-person visits, which can make a series of brief visits very costly.[7]

Lane and Schur (2010) discuss the benefits of establishing secure, remote-access entities or "data enclaves" that enable researchers to access confidential data from their desks. Kinney et al. (2009) discuss technical developments with respect to remote, "query-based access," where sophisticated software restricts what the remote user can see or obtain from the data. They view this as an emerging research area.

A major challenge for preserving confidentiality through remote access is preventing users from submitting a sequence of requests that while individually innocuous are able, collectively, to elicit more detailed information from the source data than would be permitted through a single request. For example, Oganian et al. (2009) show that it may be possible to defeat some of the confidentiality protection strategies discussed in the next section through suitable designed

[7] Remote access may not be free, however. The new Virtual Research Data Center at CMS charges $40,000 for an annual "seat," which is defined as an individual user working on one project. The fee includes training, output review, and 500 GB of disk space. Additional users can be added to an existing project at a cost of $15,000 per user.

queries. Because there may be legitimate reasons why a user would submit an extensive sequence of requests, a technical solution would be preferable to simply restricting the number of requests over a period of time and monitoring similar requests from different users.

A variation on this approach that appears to be growing in popularity might be described as a hybrid of restricted access and restricted data. Using methods described in the next section (but generally relying heavily on the synthetic method), an agency creates a public use file with limited or untested analytical value but with the same structure and many of the same variables as the source file. Analysts can use this public file to write sophisticated programs that address all the feature of the data and then submit these programs to be run on the source data. The output will be subject to the same review as other output from remote access, but the process is far more efficient. This approach can also be used to determine whether particular inferences drawn from the public data are valid—that is, supported by the source data.

Illustrating this approach, Borton et al. (2013) used the synthetic method to develop a public use file of Medicare claims data with "pseudo-analytic" utility. The file retained the structure of the original database, but in generating variables by the synthetic method the team deliberately excluded modeling of certain key relationships. Unlike other synthetic files discussed below, this file was designed explicitly as a test file that would not support valid inference. Rather, the file was designed to enable entrepreneurs to develop applications and researchers to become familiar with the source data. Code can be developed and debugged with the synthetic file with the intention that they would later be run on the source data, accessed on a restricted basis.

B. Restricted Data for Public Access

Public use microdata play a critical role in research and policy analysis. Exploratory research and many types of policy analysis do not lend themselves well to the conditions that govern restricted access as described above. The creation of public use data that protect the confidentiality of the subjects begins with de-identification, but depending on the contents of the data and the characteristics of the subjects, it may require the application of a number of additional techniques to reduce the risk of re-identification to a satisfactory level—that is, produce true anonymization.

1. De-identification

HIPAA codified a de-identification process for health records that includes the removal of 18 specific direct and indirect identifiers.8 Following Sweeney's successful re-identification of the Massachusetts governor in a file of hospital discharge data, the protections mandated by HIPAA went well beyond the simpler, informal de-identification practices that were previously common with such data but clearly inadequate. Nevertheless, HIPAA applies to a narrow range of datasets, and even in this context, researchers including Benitez and Malin (2009), discussed in the preceding chapter, have demonstrated the limits of HIPAA de-identification.

[8] The 18 identifiers are listed in Appendix D.

2. The Concept of *k*-Anonymity

Sweeney was able to re-identify a large proportion of the individuals in the Massachusetts hospital discharge data because the combination of ZIP code, sex, and date of birth in many cases pointed to unique individuals in the city of Cambridge voter registration records, which provided their names. In the language of the confidentiality literature, these characteristics defined "population uniques." To protect the individuals in a dataset from re-identification, one must be certain that the characteristics reported on the file do not define unique individuals in a separate, identified database that is accessible to potential intruders.

In theory, the way to achieve this level of protection is to ensure that no combination of characteristics is shared by fewer than some minimum number of persons in the population. This concept is called "*k*-anonymity," where *k* is the chosen minimum number (Sweeney 2002).[9] This is a fundamental concept in protecting public use data from disclosure (Ciriani et al. 2007, El Emam and Dankar 2008). If the data producer has access to a population database containing characteristics that will be reported on the public use file, the application of *k*-anonymity is straightforward and rigorous. Characteristics that in combination define unique individuals can be altered so that, when combined, they point to no fewer than *k* people. Typically, however, the data producer is not able to access population data for this purpose and applies *k*-anonymity to the file that is to be released. This is a conservative approach in that it yields more protection than is necessary to achieve *k*-anonymity at the population level. However, when the files to which it is applied contain a non-trivial proportion of the population, it may not be excessively conservative.

When datasets include only a representative sample of the population, the individuals with unique combinations of characteristics are called "sample uniques." To apply *k*-anonymity in the least conservative way, the data producer would need to know when a sample unique is also a population unique. This is a challenging problem. For example, if a sample is selected with a probability of 1 in 1,000 (which is a relatively high sampling rate for a federal survey), the average respondent represents 1,000 people. If a respondent has a unique combination of the characteristics that would appear in, say, a voter registration database, that combination of characteristics is about as likely to be found in 2,000 people as in just one person in the population. There is a literature on determining or inferring population uniqueness from sample uniqueness (see, for example, Skinner and Elliot 2002), but such formal techniques are not yet widely applied.

3. Statistical Disclosure Limitation

Preventing re-identification is the primary focus of de-identification and the statistical disclosure limitation methods discussed in this section, but a secondary objective (or result, if not necessarily an objective) is to limit what an intruder might learn from an apparent re-identification. Techniques that alter data values or exchange values between respondents contribute to both goals, and when users are informed of these techniques, their application may also discourage potential intruders from attempting to re-identify records.

[9] Sweeney credits Pierangela Samarati with naming *k*-anonymity.

a. Overview of Methods

Statistical disclosure avoidance techniques for microdata have been well developed and widely published in journals, textbooks, and workshop and conference proceedings. The following two sources proide comprehensive accounts of these techniques: (1) Statistical Policy Working Paper 22 (FCSM 2005); and (2) *Handbook on Statistical Disclosure Control* (Hundepool et al. 2010). Techniques to protect microdata for public release include those approaches pertaining to the file in general and those related to variables within the file. For example, Statistical Policy Working Paper 22 identified the following approaches:

1. Include data from only a sample of the population
2. Do not include obvious identifiers
3. Limit geographic detail
4. Limit the number and detailed breakdown of categories within variables on the file
5. Truncate extreme codes for certain variables (top or bottom coding)
6. Recode into intervals or round continuous variables
7. Add or multiply by random numbers (adding noise)
8. Swap or rank swap the values on otherwise similar records (also called switching)
9. Select records at random and blank out selected variables and impute the missing values (also called blank and impute)
10. Aggregate across small groups of respondents and replace each individual's reported value with the average

It is essential for an agency releasing a public use microdata file to remove all obvious identifiers from the individual records. However, some respondents have characteristics or combinations of characteristics that make them stand out from others. As demonstrated in the previous chapter, de-identification alone is not sufficient to eliminate the risk of disclosure.

The methods described next can be used to lower disclosure risk from released microdata. Some of these methods are suitable only for categorical variables, or only for continuous variables, whereas others can be applied to both types of variables.

Nonperturbative methods. These methods do not alter data values; rather, they implement partial suppressions or reductions of detail in the original dataset. These techniques include the following:

- **Sampling:** releasing a subsample of the original microdata
- **Global recoding:** combining several categories to form new, less specific categories
- **Top and bottom coding:** combining values in the upper (or lower) tail of a distribution (a special case of global recoding)
- **Local suppression:** suppressing the values of individual variables for selected records so that no information about these variables is conveyed for these records

Sampling introduces or increases the uncertainty that a particular individual is included in a microdata file, and in doing so it provides a strong disincentive for a would-be intruder to attempt to re-identify records on the file. Sampling can produce a very strong disincentive if the intruder has access to identified records for only a small subset of the population, as there may be no overlap between the two files—that is, no records included in both files. On the other hand, sampling will provide less of a disincentive for attempted re-identification if the intruder has data on the entire population, as the intruder can be nearly certain that every record in the public use microdata file is represented in the population data.

Perturbative methods. With these methods, values in the microdata are distorted, but this is done in such a way that key statistical properties or relationships in the original data are preserved. These techniques include the following:

- **Noise addition:** random noise technique is to add or multiply the original value by random numbers
- **Data swapping:** selecting a sample of records, finding a match in the database on a set of predetermined variables, and swapping all other variables
- **Rank swapping:** unlike regular swapping, in which the match/pair is defined based on exact match, in rank swapping the pair can be defined to be close based on their proximity to each other on a list sorted by the continuous variable; frequently the variable used in the sort is the one that will be swapped
- **Shuffling:** like shuffling a deck of cards, the values of a confidential variable are reordered in a way that preserves the correlation between the confidential variable and a non-confidential variable while also preserving the correlation between the rank order of the confidential variable and that of a non-confidential variable in the original data
- **Rounding:** replace the original values of variables with rounded values
- **Resampling:** for a variable in the original data, a new variable for released data is created in which the values of this new variable are calculated as the average of a set of resampled values from the original variable
- **Blurring:** replacing a reported value (or values) by the aggregate values (for example, the mean) across small sets of respondents for selected variables
- **Microaggregation:** a form of data blurring in which records are grouped based on a proximity measure of all variables of interest, and the same groups of records are used in calculating aggregates for those variables; Domingo-Ferrer and Mateo-Sanz (2002) note that microaggregation provides a way to achieve *k*-anonymity with respect to one or more quantitative attributes
- **Post-randomization method or PRAM** (Gouweleeuw et al. 1997): a probabilistic, perturbative method for a categorical variable; in the masked file, the scores on some categorical variables for certain records in the original file are changed to a different score according to a prescribed probability mechanism, namely a Markov matrix
- **Micro agglomeration, substitution, subsampling and calibration,** or **MASSC** (Singh et al. 2004): this creates sets of identifying variables (called strata) to find records that might

be at risk of disclosure (that is, unique records); calculates a disclosure risk measure for each stratum (unique records are also assigned a disclosure risk associated with that stratum); an overall measure of disclosure risk can be calculated for an individual record and for an entire database by collapsing over the strata

- **Synthetic microdata** (Rubin 1993): Some or all of the variables in the original dataset are replaced with imputed (synthetic) variables developed from models based on the original data; while certain statistics or internal relationships in the original dataset are preserved, the synthetic variables do not represent actual individuals

Even when the microdata have been protected using one or more of these statistical disclosure limitation techniques and are perceived to be safe for release, the risk of re-identification, in all likelihood, is still not zero.[10] The level of risk depends on the amount of information or knowledge available to the intruder and how adept the intruder is at matching this information to the microdata in question.

b. Recent Advances in Protecting Microdata

Much of the recent research on protecting microdata has focused on how the usefulness of the data is affected when methods of statistical disclosure limitation are applied. This topic is addressed in the next chapter. Research on ways to improve the protection afforded to public use microdata has sought ways to enhance existing approaches rather than develop entirely new approaches.

Singh (2009) proposes an enhanced version of MASSC that generalizes the risk measures used in altering the data to encompass cases with "partial risk," defined as having risk scores between 0 and 1. All records with nonzero risk are subject to treatment (that is, alteration of data values), but only a random subset is actually treated. Both disclosure risk and information loss are assessed in developing the final dataset.

Machanavajjhala et al. (2005) show limitations of *k*-anonymity in two situations: (1) one in which the *k* individuals are homogeneous with respect to particular characteristics, resulting in attribute disclosure; and (2) one in which the intruder possesses background knowledge that makes it possible to differentiate between the target individual and the *k*-1 other individuals. To overcome these limitations, the authors propose the concept of *l*-diversity, which requires that the values of sensitive attributes be well-represented in each group. Further work will focus on extending the concept of *l*-diversity to multiple sensitive attributes and to continuous sensitive attributes.

Efforts to improve the quality of synthetic data have received attention as well. Zayatz (2008) notes that this is one of three areas of current research on disclosure avoidance at the Census Bureau (the other two being the use of noise addition for tabular magnitude data and the

[10] Theoretically, a fully synthetic file has no risk of disclosure because none of the records corresponds to an actual person. Concerns about synthetic data focus almost exclusively on their usefulness for analysis, a concept discussed below. Nevertheless, if a synthetic file mimics the original data sufficiently closely, it can still reveal information about the individuals in the original data. In other words, if a synthetic file captures relationships in the original data so well that it is highly useful analytically, it may also carry some disclosure risk.

development of a system for remote microdata analysis). The Census Bureau uses the synthetic method to produce two databases that incorporate data from administrative records and is also applying synthetic methods to produce group quarters microdata from the American Community Survey. One of the challenges in generating synthetic data from real data, Zayatz notes, involves dealing with structurally missing values—for example, children ever born to males. The level of complexity required to incorporate structural zeroes into the modeling exceeded the capacity of the methods used at the time. Instead, values were imputed to the structural zeroes and later removed. Enhancements to the methodology include a multi-level modeling of parent-child relationships that incorporates all of the constraints.

C. The Legal Environment

A 2005 National Academy of Sciences (NAS) panel report on expanding access to research data notes that "at present, the obligation to protect individual respondents falls primarily on those who collect the data, thereby creating a disincentive for providing access to other researchers" (National Research Council 2005). As shown in the companion to this review (see Appendix D), the laws regulating the sharing of federal data provide severe penalties for agency employees in the event that disclosures occur, but these penalties rarely extend to the individuals outside the agency who are actually responsible for the disclosures. The Census Bureau addresses this asymmetry for users of its RDCs. To obtain access to a Census Bureau RDC, prospective users must obtain Special Sworn Status from the Census Bureau, which makes them subject to the same penalties for disclosure as employees. This does not extend to the public use data that the Census Bureau releases, however. The NAS panel notes that a rare exception exists for the National Center for Education Statistics (NCES) in the form of a provision in the Education Sciences Reform Act of 2002 that defines as a felony offense the use of data from the agency to "identify any individual student, teacher, administrator, or other individual" and knowingly discloses this information.[11]

The NAS panel addressed two recommendations to this problem:

> Recommendation 7. All releases of public-use data should include a warning that the data are provided for statistical purposes only and that any attempt to identify an individual respondent is a violation of the ethical understandings under which the data are provided. Users should be required to attest to having read this warning and instructed to include it with any data they redistribute.
>
> Recommendation 8. Access to public-use data should be restricted to those who agree to abide by the confidentiality protections governing such data, and meaningful penalties should be enforced for willful misuse of public-use data.

Achieving these objectives—particularly the second—would require new legislation authorizing agencies to impose penalties (National Research Council 2005).

[11] P.L. 107-279, Education Sciences Reform Act of 2002, Section 183(d)(6).

D. Assessing Disclosure Risk

A critical element in preparing a public use file of microdata is the assessment of disclosure risk, which may involve estimating the probability of re-identification. Often this is an iterative process, in which a preliminary file is tested and if the risk is determined to be too high, additional protective measures are applied. For the MASSC method, described above, risk assessment is incorporated into the disclosure limitation process.

When microdata protection is based on *k*-anonymity, the assessment of disclosure risk involves determining if *k*-anonymity is satisfied. Ideally, this is done with population data, but strategies applicable to sample data exist, as noted above.

When public use files contain numerous variables or include continuous variables, sample uniqueness across the range of variables is almost assured. Under these circumstances a different approach to assessing disclosure risk is required. Commonly, this involves using one or more alternative files with identifiers and attempting to match records on the public use file. NCES is able to exploit this strategy because there are publicly available lists of schools, and these lists include selected characteristics (Federal Committee on Statistical Methodology 2005). The accuracy of unique matches can be measured and, depending on the results, the data producer may decide to exclude high-risk records from the public use file or increase the level of masking on these records to prevent matches. If the accuracy of unique matches is sufficiently low, and there is no indication that correct matches can be differentiated from the vastly greater number of incorrect matches, the data producer may conclude that deleting or further masking the records that were matched correctly is not necessary. However, correct matches from publicly-available data do provide direct evidence of vulnerability.

The most rigorous way to assess disclosure risk is to attempt to identify records in the public use file from the source records in the original or internal file. The rigor in this approach comes from two factors. First, the only data divergence between the public use file and the internal file is that which was created deliberately to reduce disclosure risk. Second, unless the public use file was subsampled from the internal file, the overlap in records between the two files is 100 percent, which is analogous to an intruder knowing with certainty who is included in the public use file. Because these factors can make re-identification rather easy, data producers that use this approach must introduce some limitations on the match attempt to produce a realistic assessment of risk. Typically, this involves first determining what variables from the internal file might be available to a potential intruder, as it is likely that all or nearly all of the records in the public use file could be re-identified if the full set of variables appearing in both files were used in the attempt. Data producers also need to account for the impact of sampling if the internal file is itself a sample of the full population. The reduction in disclosure risk if the public use file is a subsample of the internal file should be reflected in the results of the match attempt.

IV. MAINTAINING THE UTILITY OF DATA

If a dataset that is protected against disclosure can be described as one in which a user cannot determine anything about a given individual from the dataset that could not be determined without the dataset, then by the same principle, preserving utility could be described in the following terms: the data should be no less useful to the users than if statistical disclosure limitation had not been applied. Arguably, no dataset to which any measure of disclosure protection has been applied can meet this standard. Preserving the utility of the data is a prominent topic in the statistical literature but much less so in the public discussion of data security. In the statistical literature, research has focused on measuring the information loss due to the application of the protective measures, but recent research is exploring the problem of minimizing risk and maximizing utility. This is illustrated by Shlomo (2010), who argues for a coordinated analysis of disclosure risk and information loss by data producers in order to maximize the analytic utility of the public use data consistent with providing the desired level of protection.

Purdam and Elliot (2007) classify the impact of statistical disclosure limitation on data utility into two categories: (1) reduction of analytical completeness and (2) loss of analytical validity. The former implies that some analyses cannot be conducted because critical information has been removed from the file. The latter implies that some analyses will yield different conclusions than if they had been conducted on the original data.

Procedures applied to protect the data may seriously reduce their usefulness or, worse, lead to incorrect inferences about a population or subpopulation. A recent example underscores this concern. The Census Bureau makes extensive use of data swapping, in which the values of selected variables are exchanged between respondents. For several years, errors in the program used to swap data in the Current Population Survey Annual Social and Economic Supplement and the American Community Survey produced a significant distortion in the age-sex composition of the elderly population, which was eventually noticed by users (Alexander et al. 2010).

As a general principle, statistics computed from the protected dataset should not differ significantly from the statistics obtained from the original dataset. An approach to measuring information loss is to compare statistics—totals, means, medians, and covariances—between the public use data and the source data. Some of the methods of statistical disclosure limitation discussed in the previous chapter have been shown to protect certain statistics—in particular, totals—or to introduce less distortion into covariances than other methods. Consider, for example, top coding. One can assign top codes in such a way that the original totals are preserved (by assigning the mean of the top coded values as the top code), but this will not extend to variances, which will be reduced.

Shlomo (2010) reviews several approaches to measuring information loss. These include distance metrics, impacts on measures of association, and impacts on regression analyses. Because statistical disclosure limitation introduces error into the data, measures of goodness of fit may capture information loss particularly well. Depending on how disclosure limitation affects the data, the error added may reduce between-group variance and increase within-group variance in regression analysis or ANOVA. Alternatively, it is possible that disclosure limitation

may artificially increase between group variance, creating more association than was present in the original data. Calculating a range of information loss measures will enhance the data producer's understanding of the impact of disclosure limitation on the analytic utility of the data.

Measures of data utility provide a basis for assessing and comparing alternative methods of statistical disclosure limitation as well. Woo et al. (2009) present several global (as opposed to analysis-specific) measures of data utility for masked data. They compare measures based on empirical distribution estimation, cluster analysis, and propensity scores, and they find that the measures based on propensity scores appear to hold the most promise for general use. In their analysis, the measure based on propensity scores exhibits the expected behavior with increasing levels of alteration of the original data, and it differentiates among alternative masking strategies more effectively than the measures based on the other two approaches.

REFERENCES

Alexander, J. Trent, Michael Davern, and Betsey Stevenson. "Inaccurate Age and Sex Data in the Census PUMS Files: Evidence and Implications." *Public Opinion Quarterly*, vol. 74, no. 3, Fall 2010, pp. 551-569.

Barth-Jones, Daniel C. "The 'Re-identification of Governor William Weld's Medical Information: A Critical Re-examination of Health Data Identification Risks and Privacy Protections, Then and Now." Working paper. New York: Columbia University, 2012.

Benitez K. and B. Malin. "Evaluating Deidentification Risks with Respect to the HIPAA Privacy Rule." *Journal of the American Medical Informatics Association,* vol. 17, 2010, pp. 169–177.

Borton, J.M., A. T-C. Yu, A.M. Crego, A.C. Singh, M.E. Davern, and E.Hair. "Data Entrepreneurs' Synthetic PUF: A Working PUF as an Alternative to Traditional Synthetic and Non-synthetic PUFs." Proceedings of the Joint Statistical Meetings. Alexandria: American Statistical Association, 2013.

Cavoukian, Ann, and Daniel Castro. "Big Data and Innovation, Setting the Record Straight: De-identification Does Work." Toronto, Ontario, Canada: Office of the Information and Privacy Commissioner, June 16, 2014.

Ciriani, V., S. De Capitani di Vimercati, S. Foresti, and P. Samarati. "k-Anonymity." In *Secure Data Management in Decentralized Systems*, edited by Ting Yu and Sushil Jajodia. New York: Springer, 2007.

Dalenius, Tore. "Towards a Methodology for Statistical Disclosure Control." *Statistik Tidskrift*, vol. 15, 1977, pp. 429-444.

Domingo-Ferrer, Josep, and Josep Maria Mateo-Sanz. "Practical Data-oriented Microaggregation for Statistical Disclosure Control." *IEEE Transactions on Knowledge and Data Engineering*, vol. 14, no. 1 (2002), pp. 189-201.

Domingo-Ferrer, Josep, and Vicenc Torra. "Disclosure Risk Assessment in Statistical Microdata Protection via Advanced Record Linkage." *Statistics and Computing,* vol. 13, 2003, pp. 343–354.

Duncan, George T., Mark Elliot, and Juan-Jose Salazar-Gonzalez. *Statistical Confidentiality: Principles and Practice.* New York: Springer, 2011.

Duncan, George T., and Diane Lambert. "The Risk of Disclosure for Microdata." *Journal of Business and Economic Statistics*, vol. 7, no. 2, April 1989, pp. 207-217.

Dwork, Cynthia, and Moni Naor. "On the Difficulties of Disclosure Prevention in Statistical Databases or the Case for Differential Privacy." *Journal of Privacy and Confidentiality*, vol. 2, no. 1, 2010, pp. 93-107.

Dwork, Cynthia, and Adam Smith. "Differential Privacy for Statistics: What We Know and What We Want to Learn." *Journal of Privacy and Confidentiality*, vol. 1, no. 2 (2009), pp. 135-154.

El Emam, Khaled, and Fida Kamal Dankar. "Protecting Privacy Using k-Anonymity." *Journal of the American Medical Informatics Association*, vol. 15, no. 5, September/October 2008, pp. 627-637.

El Emam, Khaled, Elizabeth Jonker, Luk Arbuckle, and Bradley Malin. "A Systematic Review of Re-Identification Attacks on Health Data." *PLoS ONE*, vol. 6, no. 12, December 2011, pp. 1-12.

Federal Committee on Statistical Methodology. "Report on Statistical Disclosure Limitation Methodology." Statistical Policy Working Paper 22 (second version). Washington, DC: Office of Information and Regulatory Affairs, Office of Management and Budget, 2005.

Fellegi, Ivan P., and Alan B. Sunter. "A Theory for Record Linkage." *Journal of the American Statistical Association*, vol. 64, 1969, pp. 1183-1210.

Golle, Phillippe. "Revising the Uniqueness of Simple Demographics in the U.S. Population." Palo Alto, CA: Palo Alto Research Center, 2006.

Gostin, L. O. "Health Information Privacy." *Cornell Law Review,* vol. 80, 1995, pp. 451–528.

Gouweleeuw, J.M., P. Kooiman, L.C. Willenborg, and P.P. DeWolf. "Post Randomisation for Statistical Disclosure Control: Theory and Implementation." Research paper no. 9731. Voorburg, Netherlands: Statistics Netherlands, 1997.

Hundepool, A., J. Domingo-Ferrer, L. Franconi, S. Giessing, R. Lenz, J. Naylor, E.S. Nordholt, G. Seri, and P.P. de Wolf. *Handbook on Statistical Disclosure Control*. A Network Excellence in the European Statistical System in the Field of Statistical Disclosure Control (ESSNet SDC). Hoboken, NJ: Wiley, January 2010.

Kinney, Satkartar K., Alan F. Karr, and Joe Fred Gonzalez, Jr. "Data Confidentiality: The Next Five Years Summary and Guide to Papers." *Journal of Privacy and Confidentiality*, vol. 1, no. 2, 2009, pp. 125-134.

Kwok, Peter, and Deborah Lafky. "Harder Than You Think: A Case Study of Re-Identification Risk of HIPAA-Compliant Records." *Proceedings of the Joint Statistical Meetings.* Alexandria, VA: American Statistical Association, 2011.

Lane, Julia and Claudia Schur. "Balancing Access to Health Data and Privacy: A Review of the Issues and Approaches for the Future." Health Services Research, vol. 45, no. 5, Part II, October 2010, pp. 1456-1467.

Machanavajjhala, Ashwin, Johannes Gehrke, Daniel Kifer, and Muthuramakrishnan Venkitasubramaniam. "*l*-Diversity: Privacy Beyond k-Anonymity." Cornell Computer Science Department Technical Report. Ithaca, NY: Cornell University, 2005.

Malin, Bradley. "Betrayed by My Shadow: Learning Data Identify via Trail Matching." *Journal of Privacy Technology*, June 9, 2005.

Marsh, C., C. Skinner, S. Arber, B. Penhale, S. Openshaw, J. Hobcraft, D. Lievesley, and N. Walford. "The Case for Samples of Anonymized Records from the 1991 Census." *Journal of the Royal Statistical Society, Series A*, vol. 154, 1991, pp. 305-340.

Narayanan, A. and V. Shmatikov. "Robust De-anonymization of Large Spare Datasets. *Proceedings of the IEEE Symposium on Security and Privacy*, 2008, pp. 111-125.

National Research Council. *Expanding Access to Research Data: Reconciling Risks and Opportunities.* Panel on Data Access for Research Purposes, Committee on National Statistics, Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press, 2005.

Office of Management and Budget, Executive Office of the President. "Memorandum Number M-13-13: Open Data Policy—Managing Information as an Asset." Washington, DC: OMB, May 9, 2013. Available at http://www.whitehouse.gov/omb/memoranda_default/. Accessed July 15, 2013.

Oganian, A., J. Reiter, and A. Karr. "Verification Servers: Enabling Analysts to Assess the Quality of Inferences from Public Use Data." *Computational Statistics and Data Analysis*, vol. 53, no. 4, 2009, pp. 1475-1482.

Pozen, D.E. "The Mosaic Theory, National Security, and the Freedom of Information Act." *The Yale Law Journal,* December 2005, pp. 628–679.

Prada, S.I., C. Gonzalez, J. Borton, J. Fernandes-Huessy, C. Holden, E. Hair, and T. Mulcahy. "Avoiding Disclosure of Individually Identifiable health Information: A Literature Review." *SAGE Open,* December 2011, pp. 1–16.

Purdam, K. and M.J. Elliot. "A Case Study of the Impact of Statistical Disclosure Control on Data Quality in the Individual UK Samples of Anonymised Records." *Environmental Planning*, Vol. A 39, 2007, pp. 1101-1118.

Purdam, K. and M.J. Elliot. "An Evaluation of the Availability of Public Data Sources Which Could be Used for Identification Purposes—A Europe Wide Perspective." CASC Project. University of Manchester, Manchester, 2002.

Rothstein, M.A. "Is Deidentification Sufficient to Protect Health Privacy in Research?" *The American Journal of Bioethics*, 10:9, 2010, pp. 3-11.

Rubin, Donald B. "Discussion of Statistical Disclosure Limitation." *Journal of Official Statistics*, vol. 9, no. 2, 1993, pp. 461-468.

Shatto, Andy. "CMS Program Data." Presentation to the Council of Professional Associations on Federal Statistics, Washington, DC, June 6, 2014.

Shlomo, Natalie. "Releasing Microdata: Disclosure Risk Estimation, Data Masking, and Assessing Utility." *Journal of Privacy and Confidentiality*, vol. 2, no. 1, 2010, pp. 73-91.

Singh, Avinash C. "Maintaining Analytic Utility while Protecting Confidentiality of Survey and Nonsurvey Data." *Journal of Privacy and Confidentiality*, vol. 1, no. 2, 2009, pp. 155-182.

Singh, A.C., F. Yu, and G.H. Dunteman. "MASSC: A New Data Mask for Limiting Statistical Information Loss and Disclosure." In *Work Session on Statistical Data Confidentiality 2003, Monographs in Official Statistics,* edited by H. Linden, J. Riecan, and L. Belsby, pages 373-394. Luxemburg, Belgium: Eurostat, 2004.

Skinner, C.J. and M.J. Elliot. "A Measure of Disclosure Risk for Microdata." *Journal of the Royal Statistical Society, Series B*, vol. 64, 2002, pp. 855-867.

Sweeney, Latanya. "Weaving Technology and Policy Together to Maintain Confidentiality." *Journal of Law, Medicine & Ethics*, vol. 25 (1997), pp. 98-110.

Sweeney, Latanya. "Uniqueness of Simple Demographics in the U.S. Population." Technical report. Pittsburgh, PA: Carnegie Mellon University, 2000.

Sweeney, Latanya. "k-Anonymity: A Model for Protecting Privacy." *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems*, vol. 10, no. 5, 2002, pp. 557-570.

Sweeney, Latanya, Akua Abu, and Julia Winn. "Identifying Participants in the Personal Genome Project by Name." Manuscript, no date.

Torra, Vicenc. "Towards the Re-identification of Individuals in Data Files with Non-common Variables. In *Proceedings of the Sixth International Conference on Soft Computing*. Iizuka, Japan, 2000.

Torra, Vicenc, John Abowd, and Josep Domingo-Ferrer. "Using Mahalanobis Distance-based Record Linkage for Disclosure Risk Assessment." *Lecture Notes in Computer Science*, vol. 4302, 2006, pp. 233-242.

U.S. Government Accountability Office. "Information Resellers: Consumer Privacy Framework Needs to Reflect Changes in Technology and the Marketplace." Document number GAO-13-663. Washington, DC: GAO, September 2013.

Woo, Mi-Ja, Jerome P. Reiter, Anna Oganian, and Alan F. Karr. "Global Measures of Data Utility for Microdata Masked for Disclosure Limitation." *Journal of Privacy and Confidentiality*, vol. 1, no. 1, 2009, pp. 111-124.

Zayatz, Laura. "New Ways to Provide More and Better Data to the Public While Still Protecting Confidentiality." *Proceedings of the Joint Statistical Meetings, Section on Survey Research Methods*. Alexandria, VA: American Statistical Association, 2008.

www.mathematica-mpr.com

Improving public well-being by conducting high quality, objective research and data collection

PRINCETON, NJ ■ ANN ARBOR, MI ■ CAMBRIDGE, MA ■ CHICAGO, IL ■ OAKLAND, CA ■ WASHINGTON, DC

MATHEMATICA
Policy Research