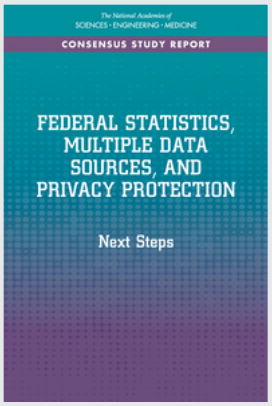


# Federal Statistics, Multiple Data Sources, and Privacy Protection: Next Steps



## DETAILS

194 pages | 6 x 9 | PAPERBACK
ISBN 978-0-309-46537-3 | DOI 10.17226/24893

## CONTRIBUTORS

Robert M. Groves and Brian A. Harris-Kojetin, Editors; Panel on Improving Federal Statistics for Policy and Social Science Research Using Multiple Data Sources and State-of-the-Art Estimation Methods; Committee on National Statistics; Division of Behavioral and Social Sciences and Education; National Academies of Sciences, Engineering, and Medicine

# FEDERAL STATISTICS, MULTIPLE DATA SOURCES, AND PRIVACY PROTECTION

## Next Steps

Panel on Improving Federal Statistics for
Policy and Social Science Research Using
Multiple Data Sources and State-of-the-Art Estimation Methods

Robert M. Groves and Brian A. Harris-Kojetin, *Editors*

Committee on National Statistics

Division of Behavioral and Social Sciences and Education

A Consensus Study Report of

*The National Academies of*
SCIENCES • ENGINEERING • MEDICINE

THE NATIONAL ACADEMIES PRESS
*Washington, DC*
**www.nap.edu**

**THE NATIONAL ACADEMIES PRESS 500 Fifth Street, NW Washington, DC 20001**

This activity was supported by a grant from the Laura and John Arnold Foundation with additional support from the National Academy of Sciences Kellogg Fund. Support for the work of the Committee on National Statistics is provided by a consortium of federal agencies through a grant from the National Science Foundation, a National Agricultural Statistics Service cooperative agreement, and several individual contracts. Any opinions, findings, conclusions, or recommendations expressed in this publication do not necessarily reflect the views of any organization or agency that provided support for the project.

International Standard Book Number-13: 978-0-309-46537-3
International Standard Book Number-10: 0-309-46537-0
Digital Object Identifier: https://doi.org/10.17226/24893

Additional copies of this report are available for sale from the National Academies Press, 500 Fifth Street, NW, Keck 360, Washington, DC 20001; (800) 624-6242 or (202) 334-3313; http://www.nap.edu/.



Printed in the United States of America

Suggested citation: National Academies of Sciences, Engineering, and Medicine. (2017). *Federal Statistics, Multiple Data Sources, and Privacy Protection: Next Steps*. Washington, DC: The National Academies Press. doi: https://doi.org/10.17226/24893.

*The National Academies of*
SCIENCES • ENGINEERING • MEDICINE

The **National Academy of Sciences** was established in 1863 by an Act of Congress, signed by President Lincoln, as a private, nongovernmental institution to advise the nation on issues related to science and technology. Members are elected by their peers for outstanding contributions to research. Dr. Marcia McNutt is president.

The **National Academy of Engineering** was established in 1964 under the charter of the National Academy of Sciences to bring the practices of engineering to advising the nation. Members are elected by their peers for extraordinary contributions to engineering. Dr. C. D. Mote, Jr., is president.

The **National Academy of Medicine** (formerly the Institute of Medicine) was established in 1970 under the charter of the National Academy of Sciences to advise the nation on medical and health issues. Members are elected by their peers for distinguished contributions to medicine and health. Dr. Victor J. Dzau is president.

The three Academies work together as the **National Academies of Sciences, Engineering, and Medicine** to provide independent, objective analysis and advice to the nation and conduct other activities to solve complex problems and inform public policy decisions. The National Academies also encourage education and research, recognize outstanding contributions to knowledge, and increase public understanding in matters of science, engineering, and medicine.

Learn more about the National Academies of Sciences, Engineering, and Medicine at **www.nationalacademies.org**.

*The National Academies of*
SCIENCES • ENGINEERING • MEDICINE

**Consensus Study Reports** published by the National Academies of Sciences, Engineering, and Medicine document the evidence-based consensus on the study's statement of task by an authoring committee of experts. Reports typically include findings, conclusions, and recommendations based on information gathered by the committee and the committee's deliberations. Each report has been subjected to a rigorous and independent peer-review process and it represents the position of the National Academies on the statement of task.

**Proceedings** published by the National Academies of Sciences, Engineering, and Medicine chronicle the presentations and discussions at a workshop, symposium, or other event convened by the National Academies. The statements and opinions contained in proceedings are those of the participants and are not endorsed by other participants, the planning committee, or the National Academies.

For information about other products and activities of the National Academies, please visit www.nationalacademies.org/about/whatwedo.

## PANEL ON IMPROVING FEDERAL STATISTICS FOR POLICY AND SOCIAL SCIENCE RESEARCH USING MULTIPLE DATA SOURCES AND STATE-OF-THE-ART ESTIMATION METHODS

ROBERT M. GROVES (*Chair*), Office of the Provost, Department of Mathematics and Statistics, and Department of Sociology, Georgetown University
MICHAEL E. CHERNEW, Department of Health Care Policy, Harvard Medical School
PIET DAAS, Department of Corporate Services, Information Technology and Methodology, Statistics Netherlands
CYNTHIA DWORK, John A. Paulson School of Engineering and Applied Sciences, and Radcliffe Institute for Advanced Study, Harvard University
OPHIR FRIEDER, Department of Computer Science, Georgetown University
HOSAGRAHAR V. JAGADISH, Computer Science and Engineering, University of Michigan
FRAUKE KREUTER, Joint Program in Survey Methodology, University of Maryland, and Statistics and Methodology, University of Mannheim and Institute for Employment Research
SHARON LOHR, Westat, Rockville, MD
JAMES P. LYNCH, Department of Criminology and Criminal Justice, University of Maryland
COLM O'MUIRCHEARTAIGH, Harris School of Public Policy Studies, University of Chicago
TRIVELLORE RAGHUNATHAN, Institute for Social Research, University of Michigan
ROBERTO RIGOBON, Sloan School of Management, Massachusetts Institute of Technology
MARC ROTENBERG, Electronic Privacy Information Center, Washington, DC

BRIAN HARRIS-KOJETIN, *Study Director*
HERMANN HABERMANN, *Senior Program Officer*
GEORGE SCHOEFFEL, *Research Assistant*
AGNES GASKIN, *Administrative Assistant*

## COMMITTEE ON NATIONAL STATISTICS

# Acknowledgments

This report of the Panel on Improving Federal Statistics for Policy and Social Science Research Using Multiple Data Sources and State-of-the-Art Estimation Methods is the product of contributions from many colleagues, whom we thank for their generous sharing of their time and expertise.

The panel is grateful to the Laura and John Arnold Foundation for funding this study, and to foundation staff Stuart Buck and Meredith McPhail for their help and guidance throughout the study. The panel also is grateful for the supplemental funding provided by the National Academy of Sciences Kellogg Fund.

The panel thanks the many individuals who participated in the panel's workshops and open meetings and shared their research, their challenges, and their creative approaches to using administrative and private-sector data sources. We also thank Steve Eglash (Stanford University) for his work examining issues of data access for private-sector companies.

At the National Academies of Sciences, Engineering, and Medicine, the panel would not have been able to complete its work efficiently without a capable staff. Constance F. Citro, former director of the Committee on National Statistics (CNSTAT), had the vision and perseverance to make this study a reality. The division's Kirsten Sampson-Snyder was extremely helpful in coordinating the review process, and Eugenia Grohman provided meticulous and thorough editing that greatly improved the readability of the report. For CNSTAT, Agnes Gaskin, administrative assistant, provided assistance in managing the logistics of this panel and our meetings. Hermann Habermann, senior program officer, provided valuable feedback

and guidance on drafts of this report. George Schoeffel, research assistant, assisted with every aspect of the study, including creating and managing a database of references, creating figures and tables, researching and drafting items for the report, carefully reviewing drafts, and performing whatever tasks needed to be done for the panel and the report.

This Consensus Study Report was reviewed in draft form by individuals chosen for their diverse perspectives and technical expertise. The purpose of this independent review is to provide candid and critical comments that will assist the National Academies in making each published report as sound as possible and to ensure that it meets the institutional standards for quality, objectivity, evidence, and responsiveness to the study charge. The review comments and draft manuscript remain confidential to protect the integrity of the deliberative process.

We thank the following individuals for their review of this report: Cynthia Z.F. Clark, independent consultant, McLean, VA; Mick P. Couper, Institute for Social Research, University of Michigan; Jeremy Freese, Department of Sociology, Stanford University; Pamela Herd, Robert M. La Follette School of Public Affairs, University of Wisconsin–Madison; Thomas L. Mesenbourg, U.S. Census Bureau (retired); Stephen W. Raudenbush, Department of Sociology, University of Chicago; Jerome P. Reiter, Department of Statistical Science, Duke University; and Larry A. Wasserman, Department of Statistics and Machine Learning Department, Carnegie Mellon University.

Although the reviewers listed above provided many constructive comments and suggestions, they were not asked to endorse the report's conclusions or recommendations, nor did they see the final draft of the report before its release. The review of this report was overseen by Michael Hout, Department of Sociology, New York University, and Alicia L. Carriquiry, Department of Statistics, Iowa State University. They were responsible for making certain that an independent examination of this report was carried out in accordance with the standards of the National Academies and that all review comments were carefully considered. Responsibility for the final content rests entirely with the authoring panel and the National Academies.

Robert M. Groves, *Chair*
Panel on Improving Federal Statistics for
Policy and Social Science Research Using Multiple Data Sources and
State-of-the-Art Estimation Methods
and Brian A. Harris-Kojetin, *Study Director*

# Preface

This is the second Consensus Study Report of the Panel on Improving Federal Statistics for Policy and Social Science Research Using Multiple Data Sources and State-of-the-Art Estimation Methods. Our first report, *Innovations in Federal Statistics: Combining Data Sources While Protecting Privacy*, was released in January 2017. In that report, the panel noted that there has been increasing attention in recent years to using data already collected by government entities for statistical purposes, such as evaluation of government programs. These data include such records as employment and earnings information on state unemployment insurance, income reported on federal tax forms, Social Security earnings and benefits, medical conditions and payments made for services from Medicare and Medicaid records, and food assistance program benefits.

We also noted that after the panel had begun its work, Congress had established an Evidence-Based Policymaking Commission (P.L. 114-140) and charged it with examining arrangements for integrating federal survey and administrative data and making those data available to researchers for program evaluation. The commission issued its final report on September 7, 2017, after the panel had completed its deliberations.

The commission's focus was somewhat different from that of the panel. It addressed using statistical analysis to evaluate government programs and alternative policy options. The panel was more specifically focused on improvement in federal statistics through the use of multiple data sources. However, there was clearly overlap in the two activities.

Since the panel had completed its work when the commission's report was released, we could not consider the similarities and differences between

the commission's recommendations and our own, so we leave that to the readers of the two reports. It is our hope that this report is useful to federal agencies and their stakeholders, as well as to the broader research community. It attempts to identify key challenges to sample surveys, which have long been the mainstay of federal statistics, and offer approaches to using the wealth of administrative and private-sector data that exist and that are being created every day.

Robert M. Groves, *Chair*
Panel on Improving Federal Statistics for
Policy and Social Science Research Using Multiple Data Sources and
State-of-the-Art Estimation Methods

# Contents

SUMMARY 1

1 INTRODUCTION 5
Panel Charge and Foundation, 6
Overview of the Panel's First Report, 10
Overview of This Report, 12
Report Structure, 13

2 STATISTICAL METHODS FOR COMBINING MULTIPLE DATA SOURCES 15
Demands for More Granular Statistics, 16
Statistical Methods for Combining Data, 22
Next Steps for Combining Data Sources, 39

3 IMPLICATIONS OF USING MULTIPLE DATA SOURCES FOR INFORMATION TECHNOLOGY INFRASTRUCTURE AND DATA PROCESSING 45
Issues for Federal Statistical Agency IT Systems, 46
System Architecture, 48
Data Processing Issues, 51
System Migration, 57
Personnel Staffing and Skills, 59

4 LEGAL AND COMPUTER SCIENCE APPROACHES TO PRIVACY 61
Personally Identifiable Information and Privacy Law, 61
Legal View of Privacy in the Context of Statistical Data Analysis, 64
The Scope of PII, 67
Examples Elucidating the PII/Non-PII Issue, 68
Synthesis: A Proposed Liability Rule for PII, 71
Implications for Federal Statistical Agencies, 71

5 PRESERVING PRIVACY USING TECHNOLOGY FROM COMPUTER SCIENCE, STATISTICAL METHODS, AND ADMINISTRATIVE PROCEDURES 79
Two Avenues to a Breach of Privacy, 80
Inference Control Techniques, 91
Implications for Federal Statistical Agencies, 104

6 QUALITY FRAMEWORKS FOR STATISTICS USING MULTIPLE DATA SOURCES 109
A Quality Framework for Survey Research, 109
Broader Frameworks for Assessing Quality, 114
Assessing the Quality of Administrative and Private-Sector Data, 118
The Quality of Alternative Data Sources: Two Illustrations, 127

7 A NEW ENTITY TO PROVIDE VITAL INFORMATION THROUGH ENHANCED FEDERAL STATISTICS 133
Attributes of the New Entity, 135
Implementation, 154

REFERENCES 157

APPENDIXES
A Executive Summary from *Innovations in Federal Statistics: Combining Data Sources While Protecting Privacy* 171
B Biographical Sketches of Panel Members and Staff 175

# Summary

The environment for obtaining information and providing statistical data for policy makers and the public has changed significantly in the past decade, raising questions about the fundamental survey paradigm that underlies federal statistics. New data sources provide opportunities to develop a new paradigm that can improve timeliness, geographic or subpopulation detail, and statistical efficiency. It also has the potential to reduce the costs of producing federal statistics.

The panel's first report (National Academies of Sciences, Engineering, and Medicine, 2017b) described federal statistical agencies' current paradigm, which relies heavily on sample surveys for producing national statistics, and challenges agencies are facing; the legal frameworks and mechanisms for protecting the privacy and confidentiality of statistical data and for providing researchers access to data, and challenges to those frameworks and mechanisms; and statistical agencies' access to alternative sources of data. The panel recommended a new approach for federal statistical programs that would combine diverse data sources from government and private-sector sources and the creation of a new entity that would provide the foundational elements needed for this new approach, including legal authority to access data and protect privacy (see Executive Summary in Appendix A).

This second of the panel's two reports builds on the analysis, conclusions, and recommendations in the first one. In this report we assess alternative approaches for implementing a new approach that would combine diverse data sources from government and private-sector sources, including describing statistical models for combining data from multiple sources;

examining statistical and computer science approaches that foster privacy protections; evaluating frameworks for assessing the quality and utility of alternative data sources; and various models for implementing the recommended new entity.

Together, the two reports offer ideas and recommendations to help federal statistical agencies examine and evaluate data from alternative sources and then combine them as appropriate to provide the country with more timely, actionable, and useful information for policy makers, businesses, and individuals.

**Methods for Combining Data** Statistical methods that are currently available, such as record linkage techniques, dual frame estimation, imputation-based models, and small-area estimation methods can be used to combine data sources and develop statistical estimates for characteristics of interest. The panel recommends that federal statistical agencies redesign current data collection efforts and estimation using multiple data sources, adapt current statistical methods to combine data sources, and develop partnerships with academia and external research organizations to develop the new methods needed for design and analysis using multiple data sources. Federal statistical agencies should also document the processes used to access, combine, and analyze multiple data sources and make that documentation publicly available. Altering current data collection practices for major federal surveys because one is able to combine data from different sources (such as administrative data with survey data) to enhance federal statistics requires substantial research efforts, and such changes should be careful and deliberative.

Adopting and exploiting multiple sources of nonsurvey data for national statistics will require significant changes to the data collection and processing pipelines currently used by federal statistical agencies. Federal statistical agencies will need to create research and production systems capable of using multiple, diverse data sources to create statistics. In doing so, agencies will need to consider the governance, functionality, and flexibility of the system. With the advent of new and different data sources and innovations in statistical products, federal statistical agencies need to provide transparency of their methods in clear communications to the public.

**Privacy** Moving to an environment in which multiple datasets are combined can change the threats to privacy. Federal statistical agencies are subject to a number of privacy and confidentiality laws that apply to their statistical data. New legal and policy issues may arise when linking records from different data sources. Because linked datasets can offer greater privacy threats than single datasets, the panel recommends that federal statistical agencies

develop and implement strategies to safeguard privacy while increasing accessibility to linked datasets for statistical purposes.

It is important to distinguish between two avenues to privacy breach in the context of statistical data analysis: threats to the security of the raw data and threats through the use of statistical findings, aggregations, and conclusions drawn from the confidential data to identify an individual or organization. At this time of transition in the statistical environment, there are weaknesses in the methods for disclosure limitation while the feasibility of implementing new approaches, such as differential privacy, has not been clearly demonstrated. Thus, the panel recommends that statistical agencies engage in collaborative research with academia and industry to develop new techniques to address potential breaches of the confidentiality of their data.

**Data Quality** Survey researchers have developed quality frameworks for classifying and examining different potential sources of error in surveys. However, unlike survey data, nonsurvey data sources are not created with the purpose of creating statistics. Thus, combining data from multiple data sources will also require a new or modified quality framework. Some quality dimensions, such as timeliness and granularity, have often been undervalued as indicators of quality, but they will become increasingly relevant with statistics based on multiple data sources. The panel recommends that federal statistical agencies adopt a framework for statistical information that goes beyond the traditional quality measure of the total survey error. The new framework should include additional dimensions that better capture user needs, such as timeliness, relevance, accuracy, accessibility, coherence, integrity, privacy, transparency, and interpretability. In addition, more attention should be paid to the tradeoffs between different quality aspects of data.

**A New Entity** Although some of the recommendations in this report for improving federal statistics could be carried out by existing agencies or by cooperative agreements among agencies, the panel recommends the creation of a new entity that will provide a secure environment for analysis of data from multiple sources, coordinate acquisition and use of data, and identify and facilitate research on the challenges that are common across statistical agencies. The entity should follow the principles and practices for federal statistical agencies and permit data access only for statistical purposes.

The panel's proposed new entity should assist federal statistical agencies in identifying data sources that can most effectively inform the creation of national statistics, help develop techniques to use those data to compute national statistics while respecting privacy and other protection obligations on the data, and nurture the expertise required for these activities. While

adhering to confidentiality, privacy, and data security requirements, statistical agencies and the new entity should strive to provide both federal and external researchers access to data for exclusively statistical purposes in a timely manner that is not administratively burdensome.

**Staff Development** Current and future staff of the federal statistical agencies will need additional training and skills to combine multiple data sources to enhance federal statistics in several areas. These areas include statistical methods for combining multiple data sources; various aspects of data quality and the appropriate metrics and methods for examining data quality from different sources; and modern computer science technology, including but not limited to distributed computing, database management, cryptography, and privacy-preserving and privacy-enhancing technologies.

# 1

# Introduction

At 8:30 a.m. on the first Friday of every month, the Bureau of Labor Statistics (BLS) announces the employment situation for the United States, which includes the count of new jobs and the unemployment rate. The monthly announcement can result in the movement of more than $150 billion in investments in U.S. stock markets within minutes of release (see, e.g., Saslow, 2012). This economic indicator is one of a variety of indicators produced and released by BLS and other federal statistical agencies on a weekly, monthly, quarterly, or annual basis. These statistics are scrutinized by economists, policy makers, and advocacy groups, and they influence a broad range of decisions by governments, businesses, and individuals.

However, in the not-too-distant future, the release of the employment situation and other economic indicators for the United States may look more like the following: at 8:30 a.m. each business day, a labor market dashboard on the BLS website is updated with information compiled from a multitude of sources that provide various readings on the employment situation in the United States, including the number of jobs, new hires, job openings, layoffs, job leavers, and claims for unemployment insurance, as well as the number of business establishments, new businesses created, and businesses that were dissolved.

In this not-too-distant future, website visitors may see changes since the beginning of the year, beginning of the month, the previous day, or over any time period they select. Rates of unemployment and employment can also be calculated and shown. BLS endeavors to provide timely information as transparently as possible and provides links to the information

on the various sources of data and when they are expected, received, and included in the released statistics as well as their strengths and weaknesses. For example, data from a number of companies and payroll services providers arrive on similar schedules due to monthly, biweekly, or weekly payroll data that is provided on a flow basis to BLS. Similarly, states can transmit updates to their administrative information about their business establishments and unemployment insurance claims on a daily or weekly basis, which are clearly noted. Updates from various Internet job sites are summarized and updated daily.

Although each of these sources provides large amounts of information, each source represents a distinct portion of the universe of the U.S. population. BLS also combines these sources and uses data from ongoing federal surveys to update and calibrate statistical models that provide more timely and geographically detailed information than what is currently available. Technical documentation on these models is readily accessible for users interested in this level of detailed information.

How far away is the above scenario, a 21st-century statistical information infrastructure that can provide near real-time statistics on the U.S. labor market and other aspects of the economy and society? That question underlies the work of this panel.

## PANEL CHARGE AND FOUNDATION

The Committee on National Statistics (CNSTAT), in the Division of Behavioral and Social Sciences and Education (DBASSE) at the National Academies of Sciences, Engineering, and Medicine, received funding from the Laura and John Arnold Foundation to convene an ad hoc panel of experts in social science research, sociology, survey methodology, economics, statistics, privacy, public policy, and computer science. The panel's charge was to consider a possible shift in federal statistical programs, from the current approach of providing users with the output from a single census, survey, or administrative records source, to a new paradigm of combining data sources with state-of-the-art methods. The goal of such a shift would be to give users richer and more reliable datasets that lead to new insights about policy and socioeconomic behavior. The full statement of task for the panel is shown in Box 1-1.

The goal of the panel's first report was to review the changing social and technological environment and its effect on the survey paradigm that underlies many federal statistical programs, as well as the potential of making greater use of other data sources, such as government administrative data and private-sector data for federal statistics. The goal of this report was to examine further what is known and attainable now and what needs further research, resources, and leadership to accomplish. This second

report expands on the issues raised in our first report and discusses what is needed to implement the panel's recommendations.

In its first report, the panel reviewed the importance of federal statistics in providing critical information to the country and serving a key role in the functioning of a democratic government and society. CNSTAT recently updated its *Principles and Practices for a Federal Statistical Agency* (National Academies of Sciences, Engineering, and Medicine, 2017c) with regard to the principles and practices that ensure that these statistics are objective and independent of political influence so that users can trust and rely on them for making decisions. That report offers four principles applicable to our panel's work:

> **Principle 1. Relevance to Policy Issues** A federal statistical agency must be in a position to provide objective, accurate, and timely information that is relevant to issues of public policy.
>
> **Principle 2. Credibility Among Data Users** A federal statistical agency must have credibility with those who use its data and information.
>
> **Principle 3. Trust Among Data Providers** A federal statistical agency must have the trust of those whose information it obtains.
>
> **Principle 4. Independence from Political and Other Undue External Influence** A federal statistical agency must be independent from political and other undue external influence in developing, producing, and disseminating statistics.

These principles are reflected in the operations of national statistical offices around the world (see U.N. General Assembly, 2014) and have similarly been affirmed by the U.S. Office of Management and Budget (OMB; 2014b) for U.S. federal statistical agencies. Because OMB is charged with coordinating the federal statistical system (44 U.S. Code 3504 (e)), the agency's Statistical and Science Policy Branch plays a critical role in communicating these important principles to the Executive Office of the President and supporting statistical agencies within their departments.

To fulfill their missions, federal statistical agencies must uphold and express these principles. *Principles and Practices for a Federal Statistical Agency* (National Academies of Sciences, Engineering, and Medicine, 2017c) also delineates 13 practices that agencies should follow to help achieve and embody these principles (see Box 1-2). Federal statistical agencies are the entities tasked with providing the objective, timely, relevant, accurate information that the country's policy makers, businesses, and individuals need to make decisions and understand the status of the economy and social issues.

**BOX 1-1**
**Statement of Task**

An ad hoc panel of nationally renowned experts in social science research, computing technology, statistical methods, privacy, and use of alternative data sources in the United States and abroad will conduct a study with the goal of fostering a paradigm shift in federal statistical programs. In place of the current paradigm of providing users with the output from a single census, survey, or administrative records source, a new paradigm would use combinations of diverse data sources from government and private-sector sources combined with state-of-the-art methods to give users richer and more reliable statistics leading to new insights about policy and socioeconomic behavior. The motivation for the study stems from the increasing challenges to the current paradigm, such as declining response rates and increasing cost and burden for surveys.

The panel will prepare two reports as part of this study:

**First Report**

The first report will discuss the challenges faced by the federal statistical system; the current paradigm of providing users with the output from a single census, survey, or administrative records source and that paradigm's increasing disadvantages for meeting user needs; and the foundational elements needed for a new paradigm.

More specifically, the first report will discuss

- federal statistical agencies' current paradigm for producing national statistics and challenges to this paradigm;
- federal statistical agencies' legal frameworks and mechanisms for protecting the privacy and confidentiality of their data and challenges to those frameworks and mechanisms;
- federal statistical agencies' legal frameworks and mechanisms for providing access to underlying data to researchers to foster transparency, replicability of statistical series, and for policy and social science research and challenges to those frameworks and mechanisms;
- federal statistical agencies' access to alternative sources of data for federal statistical programs, the organizational structures sustaining access, and the impediments to access;

Although the states and private-sector firms play important roles in working with the statistical agencies, those entities do not have the same mission as that of federal statistical agencies. The panel's outreach and discussions with a variety of private-sector firms revealed that even the large amounts of data firms have that could be useful for federal statistics are unlikely to replace federal statistics; indeed, firms often rely on federal

- the characteristics of a new paradigm for federal statistical programs that would combine diverse data sources from government and private-sector sources with state-of-the-art methods to give users richer and more reliable statistics; and
- the foundational elements needed for a new paradigm.

The first report will contain findings and conclusions from the panel's deliberations and recommendations for steps needed to lay the foundation for a new paradigm.

**Second Report**

The second report will propose approaches for implementing a new paradigm that would combine diverse data sources from government and private-sector sources with state-of-the art methods to give users richer and more reliable statistics.

The second report will

- assess alternative approaches for implementing a new paradigm that would combine diverse data sources from government and private-sector sources;
- evaluate concepts, metrics, and methods for assessing the quality and utility of alternative data sources, analogous to the "total error" framework used for surveys;
- evaluate statistical models for combining data from multiple sources;
- examine metrics and methods for evaluating the quality of combined-information estimates;
- evaluate alternative designs of statistical processes that foster privacy protections, transparency, objectivity, timeliness, replicability, efficiency, and continuity of statistical series; and
- identify priorities for research needed for federal statistical agencies to advance a multiple data-sources paradigm.

The second report will contain findings, conclusions, and recommendations for actions toward implementing a new multiple data-sources paradigm for federal statistics.

statistical information to better use and understand their own data. As we noted in our first report (National Academies of Sciences, Engineering, and Medicine, 2017b, Ch. 3), administrative and private-sector data sources vary in their fitness for use in federal statistics. Administrative data sources are currently being used in the federal statistical system in a variety of ways (see Chapter 2), but private-sector data have been used

**BOX 1-2**
**Practices for a Federal Statistical Agency**

*Principles and Practices for a Federal Statistical Agency: Sixth Edition* (National Academies of Sciences, Engineering, and Medicine, 2017c, p. 3) delineates 13 practices for a federal statistical agency to implement the four principles:

1. A clearly defined and well-accepted mission
2. Necessary authority to protect independence
3. Use of multiple data sources for statistics that meet user needs
4. Openness about sources and limitations of the data provided
5. Wide dissemination of accessible and easy-to-use data
6. Cooperation with data users
7. Respect for the privacy and autonomy of data providers
8. Protection of the confidentiality of data providers' information
9. Commitment to quality and professional standards of practice
10. An active research program
11. Professional advancement of staff
12. A strong internal and external evaluation program
13. Coordination and collaboration with other statistical agencies

in much more limited applications. Private-sector data sources could be part of a multitude of data sources that agencies might use in their modeling. There are issues of obtaining access for these sources, as well as the feasibility of being able to maintain stable access over time. Those issues would need to be addressed to incorporate these kinds of data into federal statistical programs (see National Academies of Sciences, Engineering, and Medicine, 2017b, Chs. 3 and 4).

Reliable, objective statistics for the public good has been inherently a governmental function. However, a more cost-efficient and cost-effective 21st-century information infrastructure can be created for the federal statistical system that would permit greater collaboration among federal agencies, states, and private-sector entities in providing vital information for the common good.

## OVERVIEW OF THE PANEL'S FIRST REPORT

The panel's first report (National Academies of Sciences, Engineering, and Medicine, 2017b) reviewed the current ability of the federal statistical agencies to access and use administrative and other data sources to enhance federal statistics. In our review of the potential of various data sources in our first report, we noted that some administrative and private-sector

sources hold particular promise for enhancing federal statistics and providing vital information to inform policy makers and the public. However, these sources also need a careful examination of their properties before they could be used to produce reliable statistical information. Some data sources, such as those from various Internet sources, require very different processing than the survey data currently collected by federal statistical agencies. We discussed the potential benefits, as well as the risks, of using these data sources in combination with surveys to enhance federal statistics, and we recommended that federal statistical agencies systematically review their programs and these new data sources to assess their use for enhancing federal statistics.

Combining a diversity of sources with different characteristics, strengths, and weaknesses also requires different statistical modeling techniques than producing direct estimates from a single survey or administrative data source. Bringing diverse data sources together and linking them in various ways (at the individual level, at the establishment level, and at various geographic levels) and permitting a variety of useful statistical analyses to be done also requires a focus on privacy and security to ensure that the data are used only for statistical purposes and are protected from disclosure—intentional or unintentional. Thus, we also recommended that agencies examine new approaches from computer science and cryptography to protect the confidentiality of their data and the privacy of those whose information is in their datasets.

Throughout that report, we emphasized the dramatic changes that have taken place in recent years in the amount of government administrative data and private-sector data that are available in electronic form and noted that the current system is not structured to take advantage of this wealth of data. We concluded that the status quo limits the statistical system in providing objective, relevant, timely, and accurate statistics to inform policy makers, businesses, and the public. We provided evidence and examples of the obstacles federal statistical agencies face in obtaining access to federal administrative data. We noted even greater obstacles when data are held outside the federal government by states, local governments, or private entities. We noted how statistical agencies have continued to work creatively under these constraints, but without a standardized process for accessing data, the result is missed opportunities. We further noted that one possible remedy for these difficulties would be the creation of an agency that is directly charged to ensure timely and effective access of program data for statistical purposes. Our analysis and these conclusions led to our overall assessment and recommendation (National Academies of Sciences, Engineering, and Medicine, 2017b, p. 102):

> The panel believes that the nation needs a secure environment where administrative data can be statistically analyzed, evaluated for quality, and linked to surveys, other administrative datasets, and other data sources. Such an environment would need to have the authority to control access for statistical purposes. It would also have to use and continually evaluate and enhance privacy measures. Integration of these efforts into a single entity could achieve many benefits if all statistical agencies could use a secure data-sharing environment. Without a new entity, no scaling of expertise can occur in privacy protection measures, statistical modeling on multiple datasets, and IT [information technology] architectures for data sharing.
>
> **A new entity or an existing entity should be designated to facilitate secure access to data for statistical purposes to enhance the quality of federal statistics.** (Recommendation 6-1, p. 102)

We concluded that such an entity was needed to create a 21st-century statistical information infrastructure given the decentralized nature of the federal statistical system and the difficulties that face statistical agencies in accessing, evaluating, and using a variety of administrative and private-sector data sources for statistical purposes. We did not specify exactly where this entity would be located or precisely how it would operate, but we did describe several prerequisites for the entity to be successful and sustainable, and we noted a variety of issues that would need to be addressed in creating this entity.

We made clear that the recommended new entity would not be intended to serve as a national data center or data warehouse and would not contain massive linked files on individuals or businesses. Indeed, we are confident that new privacy-enhancing developments from computer science could offer greater assurances to the U.S. public about proper protections of the data on them held by agencies. We stressed that any data accessed through the recommended entity could be used only for statistical purposes: the data could not be used by any agency for any administrative, enforcement, or regulatory uses that would affect the rights, privileges, or benefits of any individual, business, or organization. With careful oversight and controls, data would be accessed through the new entity by certified federal statistical agency personnel to create national statistics and by certified researchers conducting approved statistical analyses.

## OVERVIEW OF THIS REPORT

This report builds on the analysis, conclusions, and recommendations in the panel's first report. We describe what is known and attainable now and what needs further research, resources, and leadership to accomplish. Our goal is to help federal statistical agencies examine and evaluate data

from multiple alternative sources and then combine them as appropriate to enhance the timeliness, geographic detail, and scope of federal statistics. Such use of multiple data sources will ultimately benefit the country with more timely, actionable, and useful information for policy makers, businesses, and individuals.

In this report, we describe statistical methods for combining different data sources, privacy-preserving techniques for analysis, and a needed quality framework for different data sources. We also consider legal and computer science views of privacy and implications for statistical agencies, as well as key elements of information technology (IT) infrastructure needed for utilizing multiple data sources. We also further elaborate on various approaches to creating the recommended new entity and the pros and cons of those approaches.

We believe that the recommended new entity in our first report is the foundation for making substantial progress on all these topics and enabling a 21st-century statistical information infrastructure for the country. We consider possible answers to questions raised in the first report regarding how the recommended new entity should be set up and operate, and compare advantages and disadvantages of arrangements for this entity.

Although we have the recommended entity very much in mind throughout this report and believe that it is a much-needed resource for the federal statistical system and the country, it is important to note that many of the conclusions and recommendations in this report are applicable without the recommended entity. Individual federal statistical agencies are already making efforts along all the lines we described in our first report and describe further in this report. This work will progress and needs to progress with or without a new entity. Without a new entity, however, large opportunity costs will be incurred and the benefits from these new data sources will be realized much more incompletely, unevenly across domains, and inefficiently than would be the case with a new entity.

## REPORT STRUCTURE

As detailed in the panel's statement of task (Box 1-1), this second report focuses on implementation of a new approach for producing federal statistics from multiple data sources, including evaluating quality metrics, statistical models for combining data, and methods for preserving privacy. It also provides recommendations for needed research to move forward with a paradigm of using multiple data sources for federal statistics.

As part of its fact-gathering activities, the panel sponsored three pub-

lic workshops[1] and held two open discussions: one with the heads of the principal statistical agencies, and one with statistical agency experts with technical knowledge of IT in the federal statistical system. We also commissioned additional outreach to some private-sector firms to better understand their perceptions and use of federal statistics and their potential interest in providing access to their data for use in federal statistics.

In Chapter 2, we build on the brief overview of statistical methods for combining multiple data sources in the first report and describe issues with linking different data sources, as well as techniques for analyzing combined data sources, and note areas where further research is needed. In Chapter 3, we provide an overview of the issues and requirements related to IT infrastructure that federal statistical agencies will need to consider when implementing a paradigm for using multiple data sources for federal statistics.

In Chapter 4, we bring together the legal and computer science approaches to privacy and confidentiality and discuss the implications for federal statistical agencies for combining multiple data sources. In Chapter 5, we expand on our discussion from the first report for how statistical agencies should deal with the security and privacy issues raised by combining multiple data sources. We suggest techniques and approaches that agencies should consider for their programs and needed research. In Chapter 6, we describe existing quality frameworks and apply these to examples in which statistics could be created by combining data from multiple sources and note areas where further research is needed.

In Chapter 7, we examine in more detail different possible models of the recommended new entity for combining multiple data sources for federal statistics and consider their advantages and disadvantages.

[1]Copies of the workshop presentations are available at http://sites.nationalacademies.org/DBASSE/CNSTAT/DBASSE_170269 [September 2017].

# 2

# Statistical Methods for Combining Multiple Data Sources

In the panel's first report, we described the multiple types of additional data sources—federal and state administrative data, electronic health records, web scrapings, credit card transactions, satellite images, and sensor data, among others—that might be used to improve the level of detail, timeliness, and cost of federal statistics. Federal statistical agencies have long used administrative data to improve the efficiency of the design of probability surveys and to adjust for nonresponse, and we noted how a number of agencies are currently investigating nonsurvey data sources to supplement or replace data from probability surveys. These investigations share common features, in which the information from the different sources needs to be evaluated and combined.

In this chapter, we review statistical methods for combining information, identify research needs, and propose steps that can be taken to facilitate a new paradigm for producing federal statistics. As noted in Chapter 1, that paradigm would shift from sole reliance on probability surveys to a system that relies on probability surveys along with administrative and private-sector data, making use of the strengths of each data source. We begin by describing statistics that are currently produced or might be desired and summarizing some of the features of data sources that might be combined to produce those statistics. We next summarize the statistical methods that have been proposed for combining information, where the choice of method depends on the statistical purpose, the nature of the available data, and privacy and other considerations. When individual records from multiple datasets for each person or entity are available, they can sometimes be linked through statistical models. When aggregate

statistics are available or linkage cannot be done, multiple frame methods or modeling can be used. We also outline research that is needed in the area of statistical methodology and describe a framework for promoting the development of methods for combining data sources. Citro (2014) and Lohr and Raghunathan (2017) provide more detailed discussions of statistical methods and possible research directions.

## DEMANDS FOR MORE GRANULAR STATISTICS

The usefulness of a data source for federal statistics depends on the type of information that is desired, which includes

- Information for the United States as a whole: What is the national unemployment rate? How many people were victimized by violent crime in 2016? How many people have diabetes in the United States, and what are the associated health care costs?
- Information for regions or states: How many children are eligible to receive assistance from the Supplemental Nutrition Assistance Program (SNAP) in Arkansas? What is the forecast yield of the winter wheat crop in Kansas?
- Information for local jurisdictions: What is the violent crime victimization rate in Chicago? What is it in Fresno? What percentage of 4th-grade students in the Chicago Public Schools is at or above the "proficient" level in mathematics? What effect did Hurricane Katrina have on poverty in New Orleans?
- Information for demographic groups or other subpopulations: What is the 2016 unemployment rate among adults without a high school degree? What percentage of people ages 65 and older worked full time in January 2017? What is the job creation rate among businesses that are less than 5 years old? How many business establishments in Hawaii with four or fewer employees closed in 2016? Information may be desired for race or ethnicity groups, men or women, and specific age or education groups. Information may be desired for cross-classifications of demographic and geographic subpopulations.

Large-scale probability sample surveys have long been the foundation for producing many national statistics for the United States. Probability surveys can be designed to measure the specific concepts of interest, but they are expensive, particularly those conducted through face-to-face interviews. As discussed in the panel's first report, both costs and nonresponse rates for probability surveys have increased in recent years.

Policy makers and data users are demanding ever-increasing granular-

ity for statistics, wanting more geographic detail, more frequent releases of statistics, and more information about subpopulations. Some probability surveys have been designed and others modified to allow for the release of more detailed or frequent statistics. Before the American Community Survey (ACS) was launched in 2005, detailed geographic-level information on poverty, disability, employment, family relationships, and other characteristics of the U.S. population was available only from the "long form" of the decennial census at 10-year intervals. The ACS produces direct annual estimates[1] for areas with populations of at least 65,000, and estimates based on the past 5 years of data collection for areas with populations of at least 7,000. These estimates produce 11 billion statistics each year (see Hedrick and Weister, 2016).

Similarly, the Current Population Survey (CPS) publishes monthly estimates of national unemployment and labor participation rates, with separate statistics given for subpopulations that include cross-classifications by race and ethnicity, sex, and age (for an example, see Bureau of Labor Statistics, 2017b). However, labor force estimates for subpopulations of smaller geographic regions—census regions and divisions, states, metropolitan areas, and principal cities—are produced annually by aggregating the monthly surveys (see Bureau of Labor Statistics, 2015).

For the ACS and CPS, as for other probability samples, there is a tradeoff between geographic or subpopulation detail and timeliness.[2] A direct estimate of a subpopulation characteristic from the survey requires a sample size for the subpopulation that is large enough for the statistic to be reliable. In order to do so, data must be accumulated either over time

[1]A direct estimate is one that is produced using the data from the survey. Other data sources may be used to calibrate the weights of the survey for undercoverage and nonresponse, but the data about the characteristic being estimated come from the survey.

[2]Tradeoffs are also made in the surveys' design to enable production of state-level estimates. To produce state-level statistics, the CPS takes a sample of households from every state, with sample sizes ranging from 500 to 4,600 households (Bureau of Labor Statistics, 2012). Some states have a higher share of the sample than their share of the adult population (people ages 16 and older) and this oversampling of smaller states allows the CPS to produce reliable state-level estimates, but it makes the design less efficient for producing national estimates because adults in large states are less likely to be included in the sample than adults in small states.

The design of the National Crime Victimization Survey (NCVS) has also been modified to enable production of selected state-level estimates. The original survey design, tailored to produce national estimates of victimization, gave every household in the United States roughly an equal chance of being selected for the survey, but this design could result in some states having no one in the sample. In response to an increasing need for crime statistics at the state and local level, the Bureau of Justice Statistics redesigned the survey to produce direct estimates of victimization for 22 states (Planty and Langton, 2014; Langton and Fay, 2016; Bureau of Justice Statistics, 2016). This was done by augmenting the sample size in those states as needed to produce 3-year rolling-average estimates of victimization with the desired precision.

(through repeated data collections) or over space (collapsing across different geographic areas).

The designs of the ACS and CPS have been formed or modified so that they can produce direct estimates at more frequent intervals or at finer levels of geography, but the high costs of data collection limit their sample sizes in small geographic areas. By combining these surveys with other data sources, it may be possible to produce reliable estimates for even smaller subpopulations or with greater frequency. In addition, other data sources may measure variables not found in a survey, which may give a richer picture of the relationship between, say, poverty and health outcomes. Administrative and private-sector data sources already exist and the cost to use them for statistical purposes may be lower than the cost to collect additional data from probability surveys.

Nonsurvey data sources can also provide a fresh perspective on the redesign of federal surveys. In some cases, questions can be eliminated from a survey if equivalent and reliable measurements are available from another data source. Nonsurvey data sources can also be used to construct or refine the sampling frame used to draw the samples to improve efficiency and reduce respondent burden. The measurements available in nonsurvey data sources may also be useful in determining the most efficient mode for data collection. Thus, nonsurvey data sources are not just useful for estimation purposes but also for the possible redesign of many federal surveys.

Yet nonsurvey data sources have their own problems (see Chapter 6 for a fuller disussion of quality issues). Administrative data, such as tax records, are collected for a specific purpose that may not match the needs for the statistics being produced. For example, the tax entity represented in the records could be a large business enterprise with multiple locations, which could be different from the statistical entity of interest: a single business establishment location. The administrative data often have limited variables, and these do not necessarily measure the characteristics of interest. Data sources may be missing important parts of the population: for example, electronic medical records may be less likely to contain information about people who do not have health insurance or people who have not recently used any medical services. Data sources such as social media may be vulnerable to external manipulation through "bots" or organized campaigns. The quality of the responses given in data sources may be unknown, and protocols for data collection may change without notice or documentation. Finally, to be useful, an alternative data source must have continued accessibility and availability for federal statistical purposes. Despite these shortcomings, it would be valuable to investigate and implement strategies to combine information from survey and nonsurvey data sources to improve efficiency and meet the ever-growing need for more information.

Ultimately, a framework is needed for combining different data sources that draws on the strengths and counterbalances the weaknesses of each source, resulting in more useful information, or lower costs, than what would be achievable from a single source. For example, Horrigan (2013a, 2013b) describes data sources that the Bureau of Labor Statistics (BLS) uses when producing the Consumer Price Index (CPI) and the Producer Price Index (PPI), which include the following:

- data from the Billion Prices Project (Cavallo and Rigobon, 2016),
- retail scanner data,
- information on used cars from J.D. Power and Associates,
- stock exchange bid and ask prices and trading volume data,
- data on hospitals from the American Hospital Association,
- diagnosis codes from the Agency for Healthcare Research and Quality,
- administrative data on crude petroleum from the Energy Information Administration,
- administrative data on baggage fees from the U.S. Department of Transportation,
- SABRE data on airline pricing, and
- Medicare Part B reimbursement information.

The economic concepts for the CPI and PPI provide the framework for integrating different data sources, and BLS can create more accurate and cost-effective indexes by relying on multiple sources rather than on a single source. In a similar vein, the Medical Expenditure Panel Survey incorporates data from multiple survey and administrative records sources as part of its design (see Box 2-1).

Combining survey data with other data sources, or combining multiple administrative data sources, has many potential advantages over the survey paradigm. A number of recent studies have identified information domains that would benefit from drawing on alternative data sources to provide key statistics beyond what is possible or practical through a federal survey. For example, one study (National Research Council, 2014a) recommended that the National Center for Science and Engineering Statistics (NCSES) engage in a program of research to explore and experiment with a variety of existing alternative datasets quickly and inexpensively to understand aspects of innovation in science and engineering. Similarly, another study that considered measuring social and civic engagement and social cohesion (National Research Council, 2014b) concluded that only a limited number of variables can be included on national surveys and that combining survey data with other sources can provide useful explanatory variables and

**BOX 2-1**
**Use of Multiple Data Sources in the Medical Expenditure Panel Survey**

The Medical Expenditure Panel Survey (MEPS) is sponsored by the Agency for Healthcare Research and Quality (AHRQ). It is designed to give accurate and reliable information about the U.S. population's health care coverage, utilization, expenditures, and access to care. Created in 1996, MEPS is designed by a combination of three different interrelated surveys: the household component (MEPS-HC), the medical provider component (MEPS-MPC), and the insurance component (MEPS-IC). MEPS is cosponsored by the National Center for Health Statistics (NCHS) and uses Westat, Research Triangle Institute (RTI) International, and the U.S. Census Bureau as main data collection organizations. MEPS provides a model for combining data sources, combining information across person, household, and provider level, and using information from parts of one component as a source of information for other components.

The MEPS-HC is designed by selecting a subsample of households from the National Health Interview Survey (NHIS) conducted by NCHS. During the survey, information is collected about health conditions, health status, demographic characteristics, employment, and income, in addition to information regarding health insurance coverage, access to care, and changes and source of payment.

Following the MEPS-HC, the MEPS-MPC is used to collect information from providers for the individuals who provided responses in the MEPS-HC (see Cohen and Cohen, 2013). The information is collected from health care providers, including physicians, hospitals, health agencies, and pharmacies, and includes dates of visits, charges, and medical care services.

These two components are then linked using statistical probabilistic matching procedures. Ideally, these two components should match as they contain information about the same individual, but sometimes there are inconsistencies in the reporting of the same medical events between MEPS-HC and MEPS-MPC. When there is an inconsistency, the MEPS-MPC information is preferred because providers' data are generally considered superior in accuracy to household responses (Cohen et al., 2009). This linked dataset is then used as the primary source of information regarding expenditure estimation.

Finally, the MEPS-IC obtains information from a sample of private- and public-sector employers on the health insurance plans they offer their employees, including health insurance plans offered, premiums, contributions by employer and employees, and employer characteristics. The purpose of this component is to better understand what health insurance is available on both a national and state level; these data are not linked with data from the MEPS-HC.

greater geographic detail needed for research on social capital, social cohesion, and civic engagement.

Citro (2014, p. 152) summarized the advantages of using multiple data sources to produce official statistics, listing eight ways in which administra-

tive data sources could be used to improve the quality of household survey data:

1. Assist in the evaluation of survey data quality by using comparisons with aggregate estimates, appropriately adjusted for differences in population universes and concepts, and by exact matches of survey and administrative records.
2. Provide control totals for adjusting survey weights for coverage errors.
3. Provide supplemental sampling frames for use in a multiple frame design.
4. Provide additional information to append to matched survey records to enhance the relevance and usefulness of the data.
5. Provide covariates for model-based estimates for smaller geographic areas than what the survey can support directly.
6. Improve models for imputations for missing data in survey records.
7. Replace "no" for survey respondents who should have reported an item, replace "yes" for survey respondents who should not have reported an item, and replace reported values for survey respondents who misreport an item.
8. Replace survey questions and use administrative records values directly.

Her arguments can be extended to nonfederal and non-administrative data sources as well. Most household surveys currently use methods 1 and 2, and some surveys use or are exploring methods 3 through 8 to make more efficient use of data from other sources.

**CONCLUSION 2-1 New data sources have emerged during the last few years, providing opportunities to develop a new paradigm for statistical design and analysis systems that can improve timeliness, geographic or subpopulation detail, statistical efficiency, and reduce costs of producing federal statistics.**

**RECOMMENDATION 2-1 Multiple data sources should be used to redesign current data collection efforts and estimation tasks to improve the utility, timeliness, and cost-efficiency of federal statistics.**

In the panel's first report we noted several examples of statistical agencies that are currently making efforts along these lines (see National Academies of Sciences, Engineering, and Medicine, 2017b, Ch. 2), but more research is needed to understand these new approaches and to evaluate specific sources for use in particular applications. We recognize that alter-

ing major federal surveys by combining data sources (such as administrative data and survey data) requires substantial work both in planning and research and in the design phase. Agencies should be careful and deliberative in implementing changes based on this research, to understand the implications of substituting an administrative data source for particular survey data.

In some cases, items currently collected in a survey could be available from an administrative source. The Census Bureau has been exploring the usefulness of tax information from the Internal Revenue Service to replace the income questions in the American Community Survey (see O'Hara, 2016). In other cases, it may be possible to considerably redesign or even discontinue a survey based on the possibilities of obtaining and using administrative data and data from other sources. The National Center for Health Statistics was able to replace the National Nursing Home Survey and the National Home and Hospice Survey beginning in 2012 with administrative data from the Centers for Medicare & Medicaid Services. In yet other situations, it may be possible to combine information from administrative data sources with information from surveys. The remainder of this chapter summarizes statistical methods that can be used for combining data from different sources.

## STATISTICAL METHODS FOR COMBINING DATA

### Record Linkage

Record linkage refers to any method by which records from different data sources that are thought to belong to the same entity are associated, and records that are thought to belong to different entities are distinguished. Linking variables are variables that are used to match or distinguish records from different sources, most commonly: Social Security number (SSN), name, address, date of birth, age, race, sex, family relationships, and use of social services. However, almost any variable that is present on two (or more) data sources can be used as a linkage variable.

Record linkage methods are typically classified as either deterministic or probabilistic, and these methods are described briefly in Box 2-2 and in greater detail in Herzog et al. (2007), Christen (2012), and Harron et al. (2015). In the remainder of this section, we provide examples for which linkage is or could be used and discuss potential problems with using record linkage techniques.

Record linkage can increase the number of variables for records in a survey or administrative data source. For example, the National Center for Health Statistics (NCHS) routinely links the data from the National Health Interview Survey (NHIS) to records from the Social Security Administra-

tion, the Centers for Medicare & Medicaid Services, and the National Death Index (see Box 2-3); this linkage allows researchers to investigate the relationship between health and sociodemographic information reported in the surveys and medical care costs, future use of medical services, mortality, and other variables found in the administrative data sources (National Center for Health Statistics, 2012).

Record linkage can also decrease the time needed to conduct a survey and increase the amount of information obtained for analysis by obtaining information from other sources instead of asking the survey respondent. When faced with a long questionnaire or interview, respondents may stop answering questions before finishing a survey. Cynamon and Blumberg (2016) reported that for every year since 2007, 20 to 30 percent of NHIS interviews have had incomplete data. A shorter survey reduces the burden on the respondents and can also result in fewer uncompleted interviews.

Record linkage can serve to augment the number of records available for study. Ramaprasan (2015) linked records from the tumor registry of Group Health Cooperative, a health insurance company, with records from the Washington State Cancer Registry. The record linkage enabled the researchers to identify and remove duplicated records from the concatenated databases, adding 35,166 new tumor cases from the registry to the Group Health Cooperative database.

Record linkage can validate responses to a survey, fill in values for missing data, or replace survey items. A Housing and Urban Development pilot project (described in the panel's first report, National Academies of Sciences, Engineering, and Medicine, 2017b, Ch. 3) linked American Housing Survey records with tax assessment information. Bucholtz (2015) explored whether tax assessment data could substitute for a respondent's missing data about housing characteristics or replace erroneous information. He also suggested that tax data might be considered for replacing some survey items entirely.

The assessment or improvement of sampling frames is also possible through record linkage. The National Teacher and Principal Survey (NTPS) collects data on teacher and principal preparation, the demographic characteristics of teachers and principals, school characteristics, and other information on elementary and secondary education. The sample of public schools is drawn from the Common Core of Data, which is the U.S. Department of Education's annually updated database on public elementary and secondary schools, and each sampled school is asked to provide a listing of teachers. Brummet et al. (2014) explored using commercial school and teacher lists as an alternative sampling frame for teachers in the NTPS, as these lists could avoid the costs of obtaining the teacher listings from each school. Lists from three vendors were linked to the sampling frame for the Schools and Staffing Survey (the predecessor of the NTPS) to evaluate the

**BOX 2-2**
**Record Linkage Methods**

In deterministic record linkage (DRL), a set of linking variables is specified and records must agree on all of the linking variables in that set to be considered a match. The simplest way to use DRL is to have a single linkage variable, such as a Social Security number. Often, however, a single identifying variable is not available, and multiple variables are used for linkage. For example, two records might be linked if they agree exactly on name, ZIP code, and date of birth; otherwise, they are not linked. Some DRL systems have complex sets of rules specifying that records are linked if, say, they agree on at least four of the six linkage variables. If "nearly exact" matches are allowed on linking variables, rules are needed for specifying how close the variables need to be.

If the linkage variables have no errors or missing values and uniquely identify entities in the population, then DRL works well. But few data sources are without errors, even when there are unique identifiers in the population, and there may be missing values or typographical errors. DRL methods that require exact matches often have missed links.

In practice, every method of linking records is subject to missed or false links. This situation led to probabilistic record linkage (PRL), sometimes called fuzzy matching, in which a quantification is sought for the errors in linking records.

In PRL, an algorithm evaluates the similarity in the linkage variables among records from different sources. Many PRL methods are based on the work of Newcombe et al. (1959) and Fellegi and Sunter (1969). In a simple form of PRL, suppose that there are two data sources, *A* and *B*, and that the linking variables are name, marital status, and date of birth. For each pair of records considered as a potential match, one from source *A* and one from source *B*, the *agreement pattern* is determined for the linking variables. If the records have the same name, marital status, and date of birth, the agreement pattern is (Y, Y, Y); if they have the same name and marital status but different dates of birth, the agreement pattern is (Y, Y, N), and so on. Then two probabilities are calculated: the probability that

lists' coverage of the school and teacher population. Brummet et al. (2014) recommended continuing with the current sampling frame because of its greater coverage but also continuing to investigate the vendor lists for possible future use as sampling frames.

At the same time, record linkage methods come with concerns. Linked records have more information about individuals than the original data sources, which raises privacy concerns. Fellegi (1999, p. 5) noted that record linkage is "intrinsically privacy intrusive, in the sense that information is brought together about a person without his or her knowledge and control." Although records do not need to be physically joined at the same location in order to be linked (see Chapter 3), and encryption can be used

two records would have this agreement pattern if they are a true match, and the probability that two records would have this agreement pattern by chance if they are distinct entities. The ratio (*R)* of these two probabilities is

$$R = \frac{\text{probablity that two records have this agreement pattern if they are a true match}}{\text{probablity that two records have this agreement pattern if they are distinct entities}}.$$

If the pair of records is truly a match and the agreement pattern is (Y, Y, Y), then *R* is expected to be large; that is, the probability is higher that the linking variables agree for two records if they are from the same entity than if they are from two distinct entities. Conversely, if the pair of records are from distinct entities and the agreement pattern is (N, N, N), *R* is expected to be small. PRL uses a decision rule with two cutoff values, $C_U$ and $C_L$, where the pair is deemed to be a match if $R \geq C_U$, the pair is deemed to be from distinct entities if $R \leq C_U$, and further review is needed if *R* is between $C_U$ and $C_L$. The probabilities can be estimated from existing datasets in which the matching status of records is known, or they can be estimated from the data sources of interest as processing is done, with early decisions used to improve the accuracy of matching for later record pairs. Many variations are possible, and the probabilities can depend on the values of the linking variables as well as on the simple agreement/disagreement: it may be desirable to have a higher probability of agreement for nonmatching pairs for a common name, such as Jones, than for a less common name, such as Hoogland.

The probabilities evaluated in the Fellegi-Sunter (1969) method are *not* the probabilities that two records are a true match: they are probabilities that records have a specified agreement pattern if they are a true match (or a true nonmatch). In a Bayesian formulation of the problem (see, e.g., Belin and Rubin, 1995; Tancredi and Liseo, 2011; Steorts et al., 2016), a different probability is calculated: the probability that two records are a true match given that they have a specific agreement pattern. This is a more intuitive formulation of the probability of interest, but calculating this probability from the available data can be challenging.

in the linkage process (see, e.g., Schmidlin et al., 2015), record linkage may represent increased privacy risks to entities in the linked data sources. This issue, and the issue of obtaining consent for record linkage, is discussed further in Chapter 4.

It is often difficult to do record linkage well, particularly when good linkage variables are not available. Linkages can have errors, which can affect conclusions of analyses (see Chapter 6). If records that belong to different entities are mistakenly linked, or if records belonging to the same entity are not linked, then relationships among the variables from different datasets can be distorted.

As the panel described in our first report, the Center for Administra-

**BOX 2-3**
**Linking Records from the National Health Interview Survey: Case Study**

The National Health Interview Survey (NHIS) is the principal source of information on the health of the civilian noninstitutionalized population of the United States. It collects data through in-person interviews from a representative sample of households, adults, and children that covers information on topics, such as health status, medical conditions, health insurance coverage, health care access and utilization, and health behaviors. The 2006–2015 sample design is described in Parsons et al. (2015).

At the end of the interview for the 2010–2013 NHIS, adult respondents were asked a question about the use of their information (Weissman et al., 2016, p. 3):

> To help us link your survey data with vital statistics and health-related records of other government agencies, we would like the last four digits of your Social Security Number. The National Center for Health Statistics uses this information for research purposes only. Providing this information is voluntary. Federal laws authorize us to ask for this information and require us to keep it strictly private. There will be no effect on your benefits if you do not provide this information. What are the last four digits of your Social Security Number?

Respondents eligible for Medicare were also asked for the last four digits of their Medicare Health Insurance Claim number. A survey respondent who did not provide his or her Social Security number (SSN) or Medicare number was then asked if the agency would be allowed to try to link the survey data without the number. A respondent was considered to have consented to the linkage if he or she either provided the SSN or Medicare number or gave permission to link the survey data without it, and 10 to 12 percent of survey respondents refused to allow linkage (Weissman et al., 2016). This result from the 2010–2013 NHIS can be compared with what occurred in the mid-2000s, when respondents were asked to provide all nine digits of their SSN, and approximately 50 percent of respondents did not consent to linkage (Zhang et al., 2016).[a]

The NHIS *Field Representative Manual* provides guidance to interviewers for how to respond to frequently asked questions from survey participants. If a participant asks about why the SSN is needed, the interviewer can respond by outlining some of the uses and benefits of linking records in some detail or providing a less detailed explanation (U.S. Census Bureau, 2017, pp. F-54-F-55):

> NCHS currently links various records from NHIS with death certificate records from the National Death Index (NDI), Medicare enrollment and claims records collected from the Centers for Medicare and Medicaid Services (CMS), and the Old-Age, Survivors, and Disability Insurance (OASDI) and Supplemental Security Income (SSI) benefit records collected from the Social Security Administration (SSA). Files containing the personally identifying information are sent from NHIS to these federal agencies. Personally identifying information used in linkage includes name, date of birth, Social Security Number and/or Medicare number, race, sex, state of birth, and state of residence. If an agency is able to find a survey participant in its own data files, information can be sent back to NHIS and linked with the original survey data. These files contain-

ing detailed health survey data plus information on costs, mortality, or benefits can be used for more complex research, without having to follow up directly with participants.

Alternatively:

We know that this is a long interview and we don't wish to keep you tied up answering more questions. By having your name and Social Security Number or Medicare number, we can combine these health data with other information from Social Security, Medicare and Medicaid, and death records. These records have information about medical conditions and care, and how much they cost. We can join this information to the information that we get during an interview. This allows us to do more complex types of health research without having to come back or ask you more questions. (p. F-56)

The manual says that the interviewer then can give some examples of research that has been done using the linked data: "Predicting the number of disabled persons in the U.S. based on health conditions reported in the NHIS," "Predicting the costs of Medicare based on health conditions reported in the NHIS," or "Studying the health characteristics of people who retire early" (U.S. Census Bureau, 2017, p. F-56).

Golden et al. (2015) described the procedure used to link the 1994–2005 NHIS survey records with administrative records. Because respondents were asked for their SSNs, they were used as the primary variable for matching. When the SSN could not be verified, a probabilistic linkage procedure was used with other information found in both sources.[b]

Because the survey asked for respondents' SSNs, the linkage rates for respondents who consented to linkage were high. However, care must be used when analyzing linked records because people who consent to linkage may differ in some ways that are unknown from those who refuse to consent. Weissman et al. (2016) reported that NHIS respondents with heart disease, stroke, cancer, hypertension, diabetes, chronic obstructive pulmonary disease, or serious psychological distress were significantly more likely to consent to linkage than respondents without those conditions.

A variety of linked data files have been constructed.[c] Many of the linked files are restricted use and may be accessed only at a Research Data Center.[d] Data users are required to abide by the same rules concerning disclosure of confidential information as agency employees. However, to assist researchers in estimating their maximum available sample for analysis, feasibility files containing a limited set of variables are publicly available. The feasibility files contain information about a survey participant's eligibility for linkage and whether a participant was successfully linked to an administrative data source, but do not contain any information about benefits or payments. Public-use linked mortality files containing a limited set of mortality variables for adult survey participants are also available for download from the NCHS Data Linkage website.[e]

Many researchers have used the linked data sources to investigate mortality and health care costs for NHIS respondents. The linked mortality data have been used to investigate the relationship between mortality and strength training, depression, body mass index, smoking, diabetes, alcohol consumption, and height.[f]

Other researchers have used linkages with other datasets. Miller et al. (2016)

*continued*

**BOX 2-3 Continued**

used the linked NHIS/Medicare data to explore differences in health characteristics between people who enrolled in Medicare fee-for-service plans and those who enrolled in Medicare Advantage plans. Gorina et al. (2015) studied hospitalization, readmission, and death rates among Medicare fee-for-service enrollees using linkages among the NHIS, Medicare, and National Death Index files.

Mortality estimates and other research, however, need to account for potential differences in linkage rates: Lariscy (2011) found that the linkage quality for the 1998–2000 linked files was greater for non-Hispanic white adults and adults born in the United States than for Hispanic and foreign-born adults. Failure to account for different linkage error rates might result in too-low estimates of mortality because the matching records in the National Death Index were not found (see Miller et al., 2017).

---

[a]Please note that the methods for obtaining permission to link also changed.

[b]The text describes the linking procedures used with previous NCHS datasets. The linking methodology has been revised, and a publication describing the revised methodology is forthcoming; see https://www.cdc.gov/nchs/data-linkage/medicare-methods.htm [June 2017].

[c]For example, see https://www.cdc.gov/nchs/data/datalinkage/linkagetable.pdf [June 2017].

[d]See https://www.cdc.gov/rdc/ [June 2017].

[e]See https://www.cdc.gov/rdc/data/b4/disclosuremanual.pdf [June 2017].

[f]A list of publications using the linked mortality data is available at https://www.cdc.gov/nchs/data/datalinkage/linked_mortality_files_citation_list_12_2016.pdf [June 2017].

tive Records Research and Applications (CARRA) at the Census Bureau has developed a probabilistic record linkage system in which a protected identification key (PIK) is created for each entity and the PIK is used to link records from different sources behind a secure firewall. Records are matched against a reference file that contains each person's PIK, which is associated with the SSN, name and variants of the name used, date of birth, sex, and current and previous addresses. The linkages provided by CARRA are used in numerous research projects.[3] Jones (2016), for example, used linked data from the CPS and from W-2 records collected by the Internal Revenue Service to study wages of tipped workers in the restaurant industry.

Linkage also allows for the study of entities that are related but not necessarily the same. In a medical study, it may be desired to link electronic medical records of patients with information about their health care providers or with records of other patients of those providers. Baldwin et al. (2015), for example, linked the records of women who had delivered an

[3]See https://census.gov/library/working-papers/series/carra-wp.html [June 2017].

infant to the records of the infant using the surname, address, and dates of birth and delivery for the purpose of evaluating effects of therapeutic interventions during pregnancy.

Hospitals selected to participate in the National Hospital Care Survey are asked to submit electronic health records for all patient discharges and all emergency department and outpatient department visits. NCHS plans to link these records with other data sources, such as the National Death Index and Medicare and Medicaid data, to measure mortality after discharge and other health outcomes (see DeFrances et al., 2012). Such outcomes would be difficult to study without linking records. Levant et al. (2016) illustrated the types of new analyses possible by linking records from a hospital's emergency department to its inpatient treatment records and its outpatient department to show the outcomes of people with traumatic brain injury.

Research conducted for the National Household Food Acquisition and Purchase Survey of the U.S. Department of Agricultire (FoodAPS; see Ver Ploeg et al., 2015)[4] links survey responses from a probability sample of approximately 5,000 households with administrative data on SNAP participation and purchases, as well as information about the food items and prices that are accessible to the surveyed households. The linked information from SNAP is used to determine SNAP eligibility in the 30 days prior to the survey, resolve data discrepancies, and provide information on usage of the electronic benefit transfer card (U.S. Department of Agiculture, 2016).

The U.S. Bureau of Justice Statistics (BJS) is linking records of admissions and releases from state correctional facilities with other administrative record data to better understand why prisoners recidivate. CARRA gives BJS access to numerous data sources that can be used to identify activities and changes in status that can affect both criminal activity and return to prison (Carson, 2015). For example, Social Security data will indicate whether the former inmate has a job, while data from the decennial census or the ACS will indicate whether the former prisoner is married. These data indicate events that can be turning points leading to or away from prison.

All of these examples illustrate the potential benefits of record linkage for more efficient use of information. At the same time, it is not a panacea. Linkage rates vary across studies and for subpopulations within studies. Wagner and Layne (2014) found correct matches for more than 90 percent of the records in the 2010 census and more than 70 percent of the records in two commercial files, but match rates for other sources can be much lower. For example, Bucholtz (2015) found links between American Housing Sur-

[4]Also see http://ers.usda.gov/data-products/foodaps-national-household-food-acquisition-and-purchase-survey/faqs.aspx [June 2017].

vey records and tax assessment information for more than 70 percent of single-family detached homes but for only 13 percent of condominiums in multifamily buildings. Rates of missed links and false links depend in part on the linkage method used, but they depend even more on the quality of the linkage variables. Better statistical methods and algorithms can reduce linkage errors, but their utility is limited if the data sources have little identifying information about the records.

Harron et al. (2014) wrote that linkage errors can lead to biased conclusions, particularly when the linked and unlinked populations differ. Statistical methods have been proposed that account for linkage bias (see, e.g., Lahiri and Larsen, 2005; Hof and Zwinderman, 2012; Judson et al., 2013), but these, like nonresponse adjustments, are not guaranteed to remove the bias in key variables of interest.

### Multiple Frame Methods

Record linkage usually requires that data for individual entities be available from the data sources, along with sufficient identifying information to allow records to be matched. For example, individual property tax records from county assessors are available on the Internet and can be linked with address-based records from survey data. Often, however, even when individual records are available from different data sources, there is not enough identifying information to allow the records to be linked. In other situations, information may be available only at aggregate levels. Although Census Bureau staff and other approved personnel have access to individual data records from decennial censuses and the ACS, the public and agencies without agreements to access the data can see only aggregate statistics that are produced from these surveys.[5] A business collecting data about customers may be willing to distribute summary statistics but not individual records. Thus, statistical methods are needed that can combine aggregate statistics or can combine individual-level information from different sources when records cannot be linked. The multiple frame methods

[5]The Public Use Microdata System makes a sample of individual records from the ACS available to the public; however, all personal information is removed from these records, and other confidentiality protections are used to ensure "that it is impossible to identify individuals who provide any response" (https://www.census.gov/programs-surveys/acs/technical-documentation/pums/confidentiality.html [June 2017]). In addition, the "72-year rule" specifies that the full census records are made available to the public 72 years after the census date (U.S. Census Bureau, 2008).

Selected information from the decennial census is available for census blocks. Statistics from the ACS are available for block groups, which on average contain about 39 census blocks. Other data are available only for census tracts, which generally contain between 1,200 and 8,000 people (U.S. Census Bureau, 2012).

described in this section, as well as some of the statistical modeling techniques described below, can be used to combine statistics from different data sources.

A multiple frame survey draws samples from two or more sampling frames[6] to improve coverage of the population or to decrease costs. In its simplest form, with frames A and B, estimates are calculated for (1) the units in frame A but not in frame B, (2) the units in frame B but not in frame A, and (3) the units in both frames. The units in group 1 could be sampled from frame A or from frame B and thus have a higher chance of being sampled than if they were only in one frame. Lohr (2011) summarized methods that can be used to obtain unbiased estimates from multiple frame surveys, adjusting for the multiple chances of selection. Most of those methods involve reducing the survey weights for observations that are in both frames so that they represent the "overlap" part of the population and are not double-counted in the estimates.

For example, the Behavioral Risk Factor Surveillance System (BRFSS) measures health-related behaviors, health conditions, and use of medical services. It collects more than 400,000 telephone interviews with adults each year, with samples in every state. In the survey's early years, only landlines were called, but pilot studies indicated that, as the number of households with only cell phones increased, limiting the survey to landline households might result in biased estimates of some health characteristics (Hu et al., 2011). In response, in 2011 BRFSS began including cell phone as well as landline data in the public-use datasets. A dual frame design is used, in which one sample is drawn from a sampling frame for landlines and a second sample is drawn independently from a sampling frame for cell phone numbers (Centers for Disease Control and Prevention, 2016). Some households have both a landline and a cell phone, so they could be selected from either or both frames. Adjustments are made to the weights of households with both landline and cell phones so that they represent that part of the population in the combined samples.

Multiple frame surveys are often used in situations in which the frames

[6]A sampling frame is a list of population units from which the sample is drawn or a method for describing the population. The Current Employment Statistics Survey, which provides the establishment survey data in the monthly news releases on the employment situation (see, e.g., Bureau of Labor Statistics, 2017b), surveys about 147,000 businesses and government agencies, representing approximately 634,000 individual worksites. The sample is drawn from a list of Unemployment Insurance accounts (Bureau of Labor Statistics, 2017a, Ch. 2, p. 1). The NCVS samples areas that are formed from individual counties or groups of counties from the list of all U.S. counties and then subsamples households and group quarters within those areas from a sampling frame built from address lists (Bureau of Justice Statistics, 2014, p. 8). In other situations, a sampling frame may be described algorithmically, without assembling a list of the population, such as sampling every 20th visitor to a website.

cannot be consolidated before sampling. The cell and landline frames used for dual frame telephone surveys do not contain enough information to link the records and eliminate duplicates before sampling. Thus, respondents are asked about telephones in their household, and that information is used to determine whether they have a cell phone only (group *a*), a landline phone only (group *b*), or both cell and landline phones (group *ab*). Then, the population total for the characteristic of interest (e.g., the number of smokers in the population) is calculated as the sum of the estimated total number of smokers from groups *a*, *b*, and *ab*. Because group *ab* is sampled from both frames, the total number of smokers from group *ab* may be estimated as (estimated total number of smokers in group *ab* from the cell sample) + $(1 - \lambda)$ (estimated total number of smokers in group *ab* from the landline sample), where $\lambda$ is often chosen to be 0, 1, or 0.5.

The U.S. Department of Agriculture frequently uses multiple frame surveys. The National Agricultural Statistics Service (NASS) maintains a list frame of farm operations, which attempts to list all of the farms in the United States. The list frame is less expensive to sample from and contains most of the large operations, but it is incomplete because farms go in and out of business. To address this situation, NASS surveys often supplement a sample drawn from the list frame with a sample of land segments drawn from an area frame. The area frame for a state contains all of the land in the state and thus is complete, but it is more expensive to sample from (Davies, 2009). Farm operations in the area frame are matched with the list frame, and those found in the list frame are removed before sampling so they have only one chance of being in the sample. The Farm Labor Survey is an example of a NASS survey using this dual frame design.[7]

The 2015 Local Food Marketing Practices Survey was designed to produce statistics on the number of farms that market food directly, for example, through farmers' markets. Two frames were used for the survey. The first frame was the NASS list frame. The second frame, containing potential local food operations, was derived from web-based information and was used to measure coverage of the first frame.[8]

Multiple frame surveys can increase coverage of the population, and they have the potential to reduce costs if one or more of the frames is inexpensive to sample from. In some cases, an incomplete frame may have the information needed so that the entire frame can be used and sampling is not necessary. However, when one or more of the frames is incomplete, it is necessary to determine whether an entity sampled from one frame could also

[7]See https://www.nass.usda.gov/Publications/Methodology_and_Data_Quality/Farm_Labor/05_2017/LABQM_May2017.pdf [September 2017].

[8]See https://www.agcensus.usda.gov/Publications/2012/Online_Resources/Local_Food/quality_measures/2015_LFMPS_Methodology.pdf [September 2017].

have been sampled from the other frames. Although record linkage may be used to determine frame membership, typically there is less privacy intrusion for multiple frame surveys than for record linkage. It is also important to account for potential differences in the data collection procedures among frames—for example, if one sample is conducted in person and another by e-mail—when analyzing the data.

Multiple frame methods have great potential when used with some of the newer data sources. Some current multiple frame surveys rely on an expensive area frame to ensure complete coverage of the population. It may be possible in some cases to obtain better (although perhaps incomplete) coverage with less expense by constructing supplemental frames from alternative sources such as data provided by commercial vendors, web-scraping, or imaging data.

### Imputation-Based Methods

Another way to conceptualize combining different data sources is using a missing data framework and imputing (filling in) the missing data. Different data sources often measure different sets of variables, and linking or adding two or more data sources results in a merged dataset that has missing values. For example, suppose data source *A* has an identification (ID) variable, age, and sex; data source *B* has ID, age, medical expenditures, and smoking status; and data source *C* has ID, sex, and smoking status. Some of the people in source *A* are also represented in source *B*, while source *C* has different people. If sources *A* and *B* are linked by ID and added to the records in source *C*, the resulting merged dataset has "holes," as shown in Table 2-1. In this situation, the problem of combining information can be viewed as a missing data problem, and imputation methods can be used to fill in or impute the missing values in the combined dataset.

**TABLE 2-1** Information from Three Sources, *A*, *B*, and *C*

| Source | ID | Age | Sex | Medical Expenditures | Smoking Status |
|---|---|---|---|---|---|
| Records Linked from *A* and *B* | X | X | X | X | X |
| Records from *A* with No Linked Record from *B* | X | X | X | | |
| Records from *B* with No Linked Record from *A* | X | X | | X | X |
| Records from *C* | X | | X | | X |

Many approaches can be used to impute the missing values, some of which are reviewed by van Buuren (2012) and Kim and Shao (2013). In some approaches, a missing value on an item is replaced by the value from another data record. In other approaches, a multivariate model is used to predict the missing set of values using the information in the observed values. Alternatively, the imputation may proceed variable-by-variable through a sequence of regression models using all the variables other than the variable being imputed as predictors. The variable-by-variable approach simplifies the modeling task to finding a good fitting regression model for every variable to be imputed. Multiple imputations can be used to include the extra variability from the imputation predictions in standard errors for statistics.

Often, one wishes to combine data sources containing different sets of individuals or sources in which individuals cannot be deterministically linked. Statistical matching, also called data fusion, is sometimes recommended for these situations. Suppose that data source *A* contains demographic variables and information on health care expenditures, and data source *B* contains demographic variables and information on exercise habits for a different set of people. Statistical matching methods (Rodgers, 1984; Moriarity and Scheuren, 2001) use the correlations between the demographic variables and health care expenditures from source *A* and the correlations between the demographic variables and exercise habits from source *B* to make inferences about the relationship between exercise habits and health care expenditures. Statistical matching methods typically rely on strong assumptions for these inferences because there are no records that have both variables of exercise habits and health care expenditures. An alternative approach imputes the missing variable to one or both datasets using estimated relationships between the demographic variables and the responses of interest. Fosdick et al. (2016) reviewed recent literature on statistical matching and proposed a new method in which an inexpensive online survey is used to provide additional information relating the variables of interest.

Schenker and Raghunathan (2007) described four examples in which multiple survey data sources were combined to (1) extend and enhance the coverage, (2) handle transitions from one approach of measurement of a variable to another, (3) correct errors in self-reported data, and (4) improve small-area estimation. Most of these examples used multiple imputation or a Bayesian modeling approach. See Lohr and Raghunathan (2017) for several other examples that use both non-Bayesian and Bayesian perspectives.

Implementation of an imputation-based approach for combining data from multiple sources is now feasible because of the availability of several software packages that can create model-based imputations (see van Buuren, 2012). However, using these packages to combine data sources

requires expertise in imputation methods and in evaluating the comparability of the data sources. Agencies that do not already have this expertise on staff may need to develop it.

Given sufficient computational resources, it is conceivable that data from multiple sources could be used to create a large, representative population, perhaps even with a longitudinal component. This could be constructed from various surveys and administrative data sources. Spatial and temporal components could be added by linking satellite imageries, environmental monitors, and weather and climate data. This dynamic linking of multiple surveys and administrative data sources could create spatiotemporal data representing the U.S. population. Given the large number of variables and subjects, there would likely be a good deal of missing data; however, machine learning techniques informed by substantive modeling could be used to predict the missing values and capture the associated uncertainty with the predictions. Thus, the predictions and associated uncertainty may be used to create several copies of the populations to construct inferences for population quantities of interest.

This approach of creating a synthetic or modeled micro dataset from partially observed data has not been tried in the federal statistical system except in two instances, both undertaken to protect confidentiality. Both the Survey of Income and Program Participation (SIPP) and the Longitudinal Business Database (LBD) use modeling to produce synthetic datasets. The population creation described in this section, however, may be a useful strategy for protecting confidentiality when the actual observed data are embedded in modeled data, thus affording protection from disclosure.

Using multiple imputations to combine information from multiple data sources presents challenges, including taking into account the complex design of the survey data sources and incomparability between sources. Lohr and Raghunathan (2017) note a number of potential incomparabilities that may arise when combining multiple survey data sources, including:

- the types of respondents and the source of information: self-reported medical information from respondents to a health care survey may differ from medical records obtained from health care providers;
- mode of interview: in-person versus telephone versus self-administered questionnaire;
- survey context and sponsorship, such as a federal or a private-sector entity;
- differences in survey design and measurement, such as asking about recall of exercise as opposed to having respondents keep a diary or obtaining data from a fitness tracker; and
- different questions, question wordings, or question orderings.

Additional sources of incomparability can arise when combining surveys with nonsurvey data sources, such as administrative records and private-sector data. We review and discuss a number of these issues in Chapter 6.

Another important component of imputation methods is their reliance on the model assumptions about the mechanism producing the missing data and predictive models for the missing values. These model assumptions have to be thoroughly checked (see, e.g., Abayomi et al., 2008; Bondarenko and Raghunathan, 2016), and the sensitivity of the inferences to the underlying assumptions (for both the missing data mechanism and predictive models) needs to be explored (see, e.g., Raghunathan, 2015; Permuutt, 2016; Smuk et al., 2017).

Despite these challenges, there are a number of advantages to using imputation to fill in missing data: imputation can provide a complete dataset without any "holes"; the imputations can take advantage of the relationships among all the variables that are present on the files; it provides a means for inferring beyond the scope of each individual data source; and the modeling framework provides an explicit and transparent means for incorporating differences and incomparabilities among data sources (see Lohr and Raghunathan, 2017). We elaborate on additional modeling techniques in the next section.

## Modeling Techniques

Record linkage and imputation are suitable methods when individual record-level data from multiple sources are available. When data are available only as aggregated statistics at the national, subnational, or subpopulation level, the multiple frame methods described above can be used to combine summary statistics that all measure the same characteristic. In addition, other statistical modeling methods can be used to combine aggregated statistics with each other or with individual record data when the data sources measure different variables.

Small-area estimation methods are examples of statistical models that combine statistics estimated from a probability survey with statistics calculated from administrative data (see National Academies of Sciences, Engineering, and Medicine, 2017b, Box 3-3). In the Small Area Income and Poverty Estimates Program[9] at the U.S. Census Bureau and the Small Area Estimates for Cancer-Related Measures Program at the National Cancer Institute,[10] models are developed that relate direct estimates of the characteristics of interest (poverty rate or cancer rate) to covariates that are available from administrative data. The models are used to predict the

[9]See http://www.census.gov/did/www/saipe/ [June 2017].
[10]See http://www.sae.cancer.gov [June 2017].

poverty or cancer rate in areas where no direct survey estimates are available to improve the precision of estimates in those areas.

For small-area estimates for cancer-related measures, Bayesian hierarchical or multilevel models have been used to model the direct estimates from multiple surveys rather than using a combination of survey- and nonsurvey-based estimates. The models incorporate differing error structures in the estimates (bias and sampling variance) across surveys and also use rich sets of covariates assembled from administrative data. These types of models can also be used to combine survey and nonsurvey estimates.

One example of this method is the small-area estimation of yield or acreage devoted to a particular crop. The estimates from a farm survey, which may be available only for a subset of areas, and the estimates based on area-level satellite imagery, which may be available for all areas, could be combined to improve the accuracy of small-area estimates, especially for the locations that are not sampled in the farm survey (Bellow, 2007; Cruze, 2015; National Academies of Sciences, Engineering, and Medicine, 2017a). The modeling framework provides a means for incorporating differing sources of error structures in the two estimates (in this example, one is subject to mostly sampling and nonresponse errors and the other is subject to mostly measurement errors). There are numerous examples throughout the federal statistical system, as noted above, in such areas as crime and victimization rates, health status, and economic activity: multivariate hierarchical models can be used to combine data from multiple sources to create a systematic program of small-area estimation. Such combining of information can not only benefit estimation, but also provide information useful for redesigning surveys to fully exploit the correlation between various estimates. For example, more survey data could be collected for areas where the measurement error properties of the nonsurvey estimates are high rather than for areas with small measurement errors.

Currently, every federal statistical agency develops its own system for small-area estimation. Even within one agency, small-area estimation may be compartmentalized across divisions. Thus, the current distributed system of developing small-area estimates may not be fully efficient. For example, consider a case in which small-area estimates of smoking status and poverty are needed. To the extent that smoking behavior patterns differ by socioeconomic status, the correlation structure between these two variables can be exploited by jointly modeling the two outcomes using the multivariate Bayesian hierarchical model framework and, hence, deriving the estimates of the prevalence of smoking status and poverty. This modeling technique can be applied, and can be even more useful, when direct estimates for both outcomes are not available in every area. Suppose that for a subsample of areas both outcomes are measured, for some areas only smoking status is available, and for others only poverty is measured. The correlation between

these two outcomes provides information on the missing outcome. Consider another situation in which the precision of the available direct estimates differ by outcomes across areas. Here, too, the correlation between outcomes improves the precision of the model-based estimates. Thus, borrowing strength not just across areas but also across variables may improve the efficiency of small-area estimates, and, hence, a systemic view of the small-area estimation tasks coordinated across federal agencies could leverage aggregated data from multiple sources through joint estimation procedures.

Data from multiple sources may be of a mixed nature, with some having aggregated data and others having individual-level data. Methods for combining such data have been developed using the hierarchical models. For example, Raghunathan et al. (2003) used aggregated data from a large number of small areas or communities and small samples of individual-level data from a few areas to obtain estimates of the parameters in the individual level model (see also, e.g., Haneuse and Wakefield, 2007; Chatterjee et al., 2016). A general hierarchical framework may be used to develop a constrained estimation of individual-level population parameters given the aggregated data from a large number of areas and the individual-level data from a small number of areas.

The methods discussed in this section rely on models to a greater extent than methods currently used for most surveys in the federal statistical system. For most estimates produced from federal probability surveys, it is not necessary to postulate a statistical model relating the quantities being measured, although models are commonly used when adjusting for nonresponse (see Skinner and Wakefield, 2017).[11] However, when combining data from survey and nonsurvey sources, model assumptions may be needed for inference because the nonsurvey data sources lack a probabilistic selection structure for the units in the dataset (Elliott and Valliant, 2017). When statistical models for combining information from survey and nonsurvey data sources are developed, they will need to be empirically tested and substantively justified. These statistical models can then form the basis of constructing estimates and associated measures of uncertainty. In the Bayesian framework, credible intervals from the posterior distribution of the estimand of interest combines information from both survey- and nonsurvey-based estimates. Using statistical models for inference would comport with the practice used in most other areas of statistics.

Modeling plays a central role in developing estimates using the framework described in this section. But what if the model is misspecified? The federal statistical system has traditionally relied on estimates that are based

[11]In practice, models are used in probability sampling inference to adjust for nonresponse and undercoverage, but inference for a survey with a 100 percent response rate could be based solely on the selection probabilities.

on the sampling design rather than specified statistical models to avoid the problem that model-based inferences can be wrong if the model chosen is not appropriate for the data, and the design-based inference approach works well when there are high response rates and low costs. With increasing nonresponse and a need to combine multiple data sources, however, it is necessary to make modeling assumptions. As George Box (1979, p. 2) wrote, "All models are wrong but some are useful." Thus, one may want to change the question, "Is the model reasonable?" and, therefore, useful. The danger lies in using unreasonable models that yield unreasonable estimates. Thus, a transparent description of the underlying assumptions, model checking or diagnostics, and exploration of sensitivity of inferences to the modeling assumptions need to become integral parts of the estimation framework. Such a transparent framework will build trust, open the models and methods for critical review, and minimize the danger of using unreasonable models. The technical documentation, at minimum, needs to include detailed descriptions of the models, the methods used to support the models, and descriptions of the limitations and methods used to explore sensitivity of the derived estimates to the underlying model assumptions. The documentation needs to be accessible at several levels. For the methods described in this report to succeed, staff are needed who are experts in statistical modeling techniques and traditional survey designs.

**CONCLUSION 2-2 Many statistical methods currently available can be adapted for using combined data sources to develop estimates of target population quantities of interest.**

**RECOMMENDATION 2-2 To achieve transparency, federal statistical agencies should document the processes used to collect, combine, and analyze data from multiple sources and make that documentation publicly available.**

## NEXT STEPS FOR COMBINING DATA SOURCES

### Research Needed

This chapter reviews some of the statistical methods that are currently being used or could be adapted to be used to combine information from different data sources to produce official statistics. Many of those methods have been developed to augment data collected from the probability surveys that currently form the backbone of the federal statistical system. Some of the methods—notably, record linkage—can be applied to administrative and commercial data sources as well as to probability surveys. The record

linkage techniques can be applied to join any datasets with common variables that can be used for linkage.

Much of the current federal agency research on using multiple data sources is exploring linking records from different sources. Nearly all of the technical presentations by federal agency personnel at the panel's December 2015 workshop involved data linkage. Much more research is needed on record linkage methods. In particular, more research is needed on estimating the quality of the links and on how to propagate the uncertainty about linkage to analyses of linked datasets.

Most methods assume that some sources of data (or combination of sources) produce approximately unbiased estimates of some characteristics of the population of interest—that is, the expected value of an estimate of a characteristic is approximately equal to the true population value. In theory, when there is no nonresponse or undercoverage, probability samples produce unbiased estimates and, historically, that unbiasedness has been a major reason for their use. But as discussed in the panel's first report (National Academies of Sciences, Engineering, and Medicine, 2017b), decreasing response rates may be threatening that assumption: although survey analysts attempt to adjust for nonresponse through weighting or imputation (usually based on demographic information), there is no guarantee that these methods remove the bias in the key variables from a survey. Administrative or private-sector records can be similarly weighted or imputed; again, it is anticipated that such adjustments will reduce the bias due to records that are not in the data, but it is always possible that the individuals not present in the dataset have different characteristics than the demographically similar individuals who are in the dataset.

Both probability sample survey records and administrative records have large amounts of missing data by design: sample surveys include only those people or entities selected into the sample and who responded, while administrative records include only those in the program (e.g., SNAP recipients). A key area for which more research is needed is on using the information in all of the data sources to identify potential biases and fill in data that are missing from some of the sources. This can be done through record linkage; through multiple frame methods, in which it may be possible to identify the overlapping parts of the population; or through modeling and imputation methods, in which relationships among variables can be used both to study biases in different sources and to fill in missing values. More research is needed on other methods that can deal with missing data from multiple sources. With multiple sources of data come (possibly) multiple estimates of population characteristics. A framework is needed for evaluating the quality of data sources, and this is discussed in Chapter 6.

Even if alternative data sources are used for some purposes, there will likely be continued reliance on probability samples as a primary source of

federal statistics, at least for some indicators. The National Crime Victimization Survey, for example, measures both crimes reported to the police and those not reported to the police. It is difficult to see how the latter could be measured accurately without using a survey, although police agency reports may be helpful in improving estimates of the former. Even if new data sources are integrated into federal statistics, we anticipate that traditional probability surveys will still be needed to cover parts of the population not in the other sources and to provide a check on their quality. The decreasing response rates of surveys continue to be a concern, and ongoing research is needed on ways to promote response and to deal with nonresponse. How does the public view using alternative sources of data for official statistics, and do those views affect willingness to provide data?

Statistical methods in use for surveys typically produce static estimates: for example, the National Crime Victimization Survey produces estimates of victimization rates in each calendar year, and the CPS produces monthly unemployment statistics. Administrative data records and sensor data, however, may be updated much more frequently. Sensor data, in particular, are collected continuously, as are other automatically collected data, such as data collected from smart phones, fitness monitors, and "smart cars." Challenges arise on how to integrate information that is collected in different time frames. In particular, little research has been done to date on statistical methods for combining some of the "big data" sources with administrative records or probability samples.

As we discussed in our first report (National Academies of Sciences, Engineering, and Medicine, 2017b, Ch. 4), many of the private-sector data sources that might be used for the statistics of the future are generated as by-products of electronic activity and are massive. Electronic health records, which may contain information on all utilizations of health care; credit card transactions; traffic sensors; cell phone location records; web-scrapings; smart meters; and other data sources produce exabytes of data every day. Machine learning methods—techniques in which algorithms search for patterns in data—are frequently used with these types of large, organically collected datasets to uncover correlations among variables. New statistical methods are needed for interpreting and merging such data. Machine learning techniques have also been used for record linkage and imputation. More research is needed on using and developing machine learning methods to combine data sources. There are additional research needs on the privacy issues associated with these data sources (Froomkin, 2016), which we cover in Chapter 5.

Many of the statistical models for combining data sources discussed in this chapter start with the structure of the existing data sources and then specify how to combine them. More research is needed on designing new systems of data collection that make use of multiple sources to provide

federal statistics. This would represent a shift from the current framework, where a probability survey is designed to serve as the primary source of information and other sources are used as auxiliary information, to a model in which the "best" data source is used for particular aspects of the data. In some cases, this approach may mean systematically redesigning a data collection so that inexpensive data sources are used for the parts of the population they can capture, and more expensive probability samples are used for the parts of the population that cannot be measured any other way.

Finally, research is needed on the robustness of statistical systems. One advantage of the probability sampling framework is that it is difficult for an external actor to manipulate the system. Participants in the survey are selected randomly, and although people or entities that are sampled may decline to participate, no external actor can decide which units are sampled or which choose to respond. With administrative records, commercial data, or convenience data sources, however, it may be possible for external actors to modify the data; for example, social media data could be flooded with responses if those data were to be used to inform policy. In addition, there is no guarantee that the data sources available today will be available next year or any time in the future. A data vendor may stop collecting data or choose to keep the data private.

Many of the methods described above rely on modeling the relationship among variables in different data sources. If the probability sample is replaced by other sources, safeguards need to be put in place to ensure that the models continue to be valid, as relationships among variables may change.

**CONCLUSION 2-3 Research is needed on designing new systems to collect and process multiple data sources to create and enhance federal statistics.**

**RECOMMENDATION 2-3 Current statistical methods should be adapted to the extent possible and new methods should be developed to harness the statistical information from multiple data sources for analysis.**

### Structure Needed for Implementation

Though statistical methods for record linkage, multiple frame surveys, imputation, machine learning techniques, and hierarchical models for combining data are available, many of them need further research and adaptation. Such research is currently being done at many agencies and by academic researchers. This research can be facilitated by better communica-

tion and, perhaps, coordination of the research projects and the knowledge gained from the projects.

As detailed in Chapter 1, the panel's first report recommended the creation of a new entity with the authority to access multiple data sources for blending. Such an entity could achieve this communication through summarizing data linkage projects and other projects involving the combination of data sources; we discuss this topic in Chapter 7. A publicly accessible website could supply basic information about ongoing research projects through the entity, including their purpose, the datasets being combined, an outline of the statistical methodology, results, and lessons learned.

The statistical methods described in this chapter assume that agencies have access to the data sources needed. As described in the panel's first report (National Academies of Sciences, Engineering, and Medicine, 2017b), obtaining this access is a challenging process. In addition, even with access, the data may not be in a form that is amenable for producing statistical estimates. A partnership among agencies is needed to make such data accessible for combining.

Of the research that is needed for establishing a new paradigm for producing federal statistics, one of the primary areas is systemic redesign of data collection methods that rely on multiple sources to produce federal statistics. Such research will require the resources of multiple agencies, as well as cooperation with academia and businesses. The skills needed for the research include the traditional skills in probability survey design and analysis, but they also include knowledge of record linkage, machine learning, new statistical modeling techniques, and privacy expertise. Training is needed both in the statistical agencies and the broader research community to ensure that research staff have the skills and adaptability needed to advance the field of combining data. In addition, development of algorithms and user-friendly software is needed for implementing some of the methods for combining data. A multidisciplinary approach would be ideal, drawing on and developing expertise in statistics, computer science, economics, engineering, and the fields related to the substance of statistics that are produced.

Federal agencies also need to continue to develop partnerships with research organizations and businesses to develop new data sources and new statistical methods. Two important parts of these partnerships are evaluating the quality of different data sources and the quality of the statistics produced by combining data from different sources.

In Chapter 7, we further discuss how the entity could guide and serve as a resource for research and training. The structure for revolutionizing federal statistics has to be dynamic and innovative, willing to explore new frontiers and modern modeling methods. The federal statistical system needs to be empowered to capitalize on modern computational and statisti-

cal developments and the plethora of emerging data sources and to continually improve the methods for producing statistics.

**CONCLUSION 2-4 Federal statistical agencies are currently combining information from multiple data sources for specific projects. Systematic coordination and dissemination of their results will help advance knowledge and promote the use of appropriate statistical methods.**

**RECOMMENDATION 2-4 Federal statistical agencies should ensure their statistical staff receive training for the new skills needed for combining data from different sources.**

**RECOMMENDATION 2-5 Federal statistical agencies should develop partnerships with academia and external research organizations to develop methods needed for design and analysis using multiple data sources.**

# 3

# Implications of Using Multiple Data Sources for Information Technology Infrastructure and Data Processing

Adopting and exploiting nontraditional sources of data for national statistics will require significant changes to the data collection and processing currently used by many statistical agencies. The design of computer systems to meet the increasing data processing demands is largely well understood, at least in the computing industry, though there remain challenges that continue to be studied in both industry and academia. In addition, there often are situation-specific issues, specific to the multisource system envisaged, that may not be covered by the accepted general solutions.

Many detailed texts describe the relevant design principles and corresponding performance tradeoffs in developing the envisaged computer systems (see, e.g., Kleppmann, 2017; Laudon and Laudon, 2017; Martin, 1981; Özsu and Valduriez, 2011). Here, we provide an overview of the design and implementation issues that can be expected to be encountered when developing information technology (IT) systems architectures for the proposed system. Our intent is not to provide solutions but, rather, to highlight considerations.

We begin with a brief overview of the IT issues that federal statistical agencies face with their current systems. We then review the nature of the architecture that will be needed by statistical agencies and for the panel's recommended new entity. Broadly speaking, there is a choice between a centralized system and a distributed system. Because a centralized structure imposes prohibitive constraints, we focus on a distributed system and discuss various distributed configurations.

The chapter continues with a discussion of data processing issues. That

is, we describe data acquisition, data cleaning and transformation, provenance, and reproducibility. We emphasize the existing and future quality control requirements of the individual statistical agencies and how these will be met.

The chapter concludes with two brief discussions of some considerations in transitioning existing systems toward the future environment and the implications for staffing. We note the implications of supplementing the existing systems not only in terms of architectures, but also in terms of staffing, both retraining and growth. Because a sudden shock would be difficult and is not needed, we discuss the evolution to a supplemented approach rather than a massive one-time change.

## ISSUES FOR FEDERAL STATISTICAL AGENCY IT SYSTEMS

The decentralized nature of the U.S. federal statistical system requires that every statistical agency has its own IT system, both because of tradition and the laws that authorize the agencies. In recent years, with greater efforts toward centralizing IT systems within departments and the passage of the Federal Information Technology Acquisition Reform Act (P.L. 113.291), department chief information officers (CIOs) have a strong role in managing the information systems of all bureaus in their departments. However, the U.S. Office of Management and Budget has also issued guidance that CIOs are to work closely with their statistical agencies to meet statutory obligations to protect the confidentiality of their data and ensure the data are used only for statistical purposes (U.S. Office of Management and Budget, 2015).

This vertical organization and control of IT systems within departments means that individual statistical agencies cannot directly access each other's systems or data. Even statistical agencies in the same department may not be able to access each other's data. For example, the Bureau of Economic Analysis and the Census Bureau are both part of the Department of Commerce and were recently co-located at the Census Bureau's headquarters building. However, they have completely separate IT systems and, given the different statutory protections and authorizations of their datasets, completely separate access. One senior manager described this situation as having a glass wall between the employees of the two agencies but a solid statutory brick wall between their datasets.

There were efforts a few years ago to create a statistical "community of practice" that could serve as platform for statistical agencies to collaborate on common protocols and tools (Bianchi, 2011, p. 5):

> [to] enhance the horizontal, functionally-based integration of IT resources among federal statistical agencies. A statistical enterprise data center would

> be mandated to foster creative approaches for collecting, storing, analyzing, and otherwise processing federal statistical data to meet statistical agencies missions—with significant cost savings—and to more efficiently feed data to Data.gov. The center would house federal and commercial statistical datasets; visualization and dissemination tools; data quality, interoperability and confidentiality tools; and statistical analytical applications and models for cloud-type access by all federal statistical agencies.

However, no specific funding or authorization was ever provided for these kinds of activities.

In most agencies, there are also organizational and programmatic silos for IT systems. It is not uncommon for statistical agencies to have separate systems for collecting and processing data for each of their survey programs, or a system may be shared by only a couple of related programs. However, there have been recent changes. The National Agricultural Statistics Service recently consolidated its own highly decentralized survey processing architecture from 46 field offices into a centralized system (see Nealon and Gleaton, 2013). The Census Bureau has embarked on a new census enterprise data collection and processing system in conjunction with the reengineering of the 2020 census: the goal is to attempt to reduce the more than 100 systems it operates for data collection and processing to a single unified approach.[1] The Census Bureau is similarly seeking to streamline 30 different applications used to disseminate information into a unified approach by creating a new Center for Enterprise Dissemination Services and Consumer Innovation in the Census Bureau that will centrally disseminate information to application program interfaces as well as interactive web tools, data visualizations, and mapping tools.

The number and diversity of IT systems within and across federal statistical agencies will pose challenges as agencies move from being focused on processing a single survey or administrative data source to integrating and using data from multiple different sources. National statistical offices in other countries have been facing similar issues. Struijs et al. (2013) note that most statistics are produced on separate production lines, each with its own methodology and IT systems. Even countries with centralized statistical offices have been working recently to integrate their systems into an overall enterprise architecture because of the higher costs of developing and maintaining separate systems and the difficulties in combining and reusing data across systems when needed (see Borowik et al., 2012; Struijs et al., 2013).

As part of these modernization efforts, there have been international collaborations to develop a common metadata framework and a generic

[1] See https://www.census.gov/library/video/cedcap_cedsci.html?ncid=edlinkushpmg00000313 [July 2017].

statistical business process model to describe the common processes that all organizations use in producing statistics (see Vale, 2009). A generic statistical information model has also been created to provide internationally agreed-on definitions for information objects (e.g., data, metadata, rules, parameters) that flow through the various processes in the production of statistical information. These frameworks facilitate communication across statistical offices and can help harmonize architectures and the sharing of statistical software across organizations, nationally and internationally (see Eltinge et al., 2013). The U.N. Economic Commission for Europe (2015) has also created the Common Statistical Production Architecture initiative to provide reference architecture for the statistical industry to support the facilitation, sharing, and reuse of statistical services across and within statistical organizations. Adopting this common reference architecture would make it easier for organizations to standardize and combine components of statistical production, thus enabling sharing components across agencies or even countries.

## SYSTEM ARCHITECTURE

Traditionally, a computer system is designed as a single system, controlled by its owner and designed according to the owner's choice. Databases may be stored and managed on this computer and access provided to others as needed. This access could be restricted to local access, that is, only to others who can physically visit a "safe room," but more commonly access is provided across a network. Such access can be provided to selected authenticated parties or to the public.

When data from multiple data sources are aggregated into a database, one popular paradigm in the computing industry follows the centralized model described above: the traditional "data warehouse." This was the expected structure of the National Data Center proposed many years ago (Kraus, 2013). In contrast, the committee believes that it is possible to obtain many of the benefits of a national data center without privacy risks incurred by storing so much data in one place. Specifically, we would like to provide access to aggregated statistical data obtained through fusion of multiple sources, but not direct access to individual-level data (identifiable information about a person, household, or business), and to do so with careful attention to privacy (see Chapter 5 for a more detailed discussion). In this section, we look at architectural alternatives.

As the number of users, the sizes of databases, and the processing performed on the collected data are scaled up, it may no longer be feasible for a single centralized system to manage the load. In this case, a set of systems can be used in parallel to perform this task. It would still be a single cen-

tralized system in terms of the system architecture, even if implemented as a room full of machines.

Instead of being placed in a single machine room or data center, the set of machines could be distributed across multiple locations. Such distribution may be desirable for a variety of reasons, including proximity to users, resilience to a disastrous event, and availability and cost of space for a machine room. With a distributed physical structure, it would be necessary to decide whether to reflect that distribution in the data placement design. That is, the data could be partitioned across sites, with each site handling some of the data, or the data could be replicated so that multiple complete copies of the data are created, one at each site. A mix of the two approaches is also possible, with some popular data replicated at each site, but most data are held only at one site. The choice depends on various factors, including the desired performance requirements or objectives. Another decision to make with a distributed system is whether to expose the distribution at the logical level: should users know where the servers are located? Do they need to know? Often, but not always, the answer to these questions is "no."

Traditionally, businesses and other private-sector enterprises have developed their own data centers to meet their storage and processing needs. As in so many aspects of business, it sometimes makes sense to outsource this responsibility to a service provider that has particular expertise in this task. For data processing in particular, this outsourcing has been made particularly easy through the "cloud." The basic idea is that the data centers are owned and managed by service providers that often share the same data center facilities across multiple enterprises. Enterprises rent needed capacity and services from the service provider. Arrangements vary greatly, from fixed capacity to variable on-demand plans, with various quality of service guarantees. The service contracted for could also range from the bare bones, compute and storage, up through data management and web hosting, to sophisticated software capabilities.

Outsourcing responsibility for some tasks is not the same thing as transferring ownership. Even when an enterprise obtains services from a third-party service provider, it is still the owner of the data and responsible for all aspects of the data, including database design and data quality. The owner could continue to specify every aspect of the system. However, a looser federated design is possible, particularly when there are multiple sources of data: for example, instead of integrating all the data into a single warehouse, datasets could remain under the control of the providers of the original data or of intermediaries. A thin layer of software could provide users with the illusion of a centralized system while actually providing requested data from multiple places as needed (see Contreras and Reichman, 2015). In the case of derived data products, such as national statistics,

the same principle would apply to the derived data creation process. For a user, one could combine data from multiple owners, at multiple locations, and use these combined data to generate national statistics.

A key difference between a federated and a distributed architecture is in the logical control and ownership: in a federated system, each member of the federation designs and owns its own data; in a distributed architecture, there is a single owner in charge. This difference has an impact on data integration; if one party controls the design and structure of multiple databases, it can also determine the required protocols to combine databases. However, if databases are independently designed and structured, required translations are usually possible only through negotiated specifications and interfaces. Such negotiations are often cumbersome and can cause systems to become brittle and unresponsive to changes that may be necessary or desirable. However, incompletely-agreed-to standards and interfaces can become a barrier to data integration and can result in errors. Since in practice there are many scenarios in which one may encounter data from a system that does not adhere to desired or standard structures, there is a need to develop abilities to perform ad hoc integration.

As we described in Chapter 2, protocols for linking individual, household, establishment, and enterprise data records have been well developed by federal statistical agencies. Such handling of multiple datasets would be a key feature of the panel's recommended new entity. If statistical systems use ad hoc integration technologies, care must be taken to validate results and manage any errors. If statistical systems use engineered integration technologies, they must develop processes for incrementally adapting integration rules as data sources evolve. The questions that need to be addressed include not only developing new mappings for the modified source data structure, but also how a statistical system will even know that the source has updated its data representation. Will the source system reliably convey information regarding updates to the statistical system? Will the statistical system perform some checks on the data supplied by the source system to validate assumptions regarding structure, representation, value encoding, and other characteristics before ingesting the data? The personnel responsible for performing such validation will need to have skills both in statistics and in computer science.

The panel noted in its first report that data breaches and identity theft pose risks to the public and that a continuing challenge for federal statistical agencies is to produce data products that safeguard privacy. Even if strong access and data release practices are designed to satisfy privacy requirements, it is difficult to *guarantee* against a data breach. We discuss security issues and protecting privacy in more depth in Chapter 5, but we note here in the context of systems design that privacy loss from a data breach can

be greatly ameliorated by distributing the data among different places and making it difficult for an attacker to access those locations.

**CONCLUSION 3-1 Moving to a paradigm of using multiple data sources requires a new and different information technology architecture than a paradigm based on a single data source. Federal statistical agencies will need to create research and production systems capable of using multiple, diverse data sources to create statistics.**

**CONCLUSION 3-2 A range of possible computing environments could enable use of multiple data sources for statistics. Federal statistical agencies will need to consider the governance, functionality, and flexibility of the system, as well as the implications for protecting privacy and addressing data providers' concerns regarding privacy.**

## DATA PROCESSING ISSUES

Moving to a paradigm of integrating multiple data sources for federal statistics will necessitate a greater focus on data curation by federal statistical agencies, which requires the "processes and activities needed for principled and *controlled data* creation, maintenance, and *management*, together with the capacity to add value to data" (Miller, 2014, p. 4). As noted in the panel's first report, agencies are used to using administrative data in a variety of ways to enhance their design, collection, and analysis of survey data, as well as to produce some statistics directly (see National Academies of Sciences, Engineering, and Medicine, 2017b). However, agencies generally directly collect much of the data they use to produce statistics, are used to having a good deal of control over design and collection of those data, and know what happens at each stage of collection, editing, imputation, and analysis. In the system we envisage, agencies will not have control and may have limited knowledge or documentation of all of these processes for some of the data they acquire. It will be essential that federal statistical agencies, including the panel's recommended new entity, carefully document all of the operations that they perform on datasets they acquire or access for federal statistics (see section, "Provenance").

### Data Acquisition

There are two main paradigms for software to obtain data from the source: "push" and "pull." In a push paradigm, the data source pushes data to the statistical agency or other entity. This push could be periodic, say, once a month; it could be in response to an event occurring, such as the accumulation of 100 updates; or it could be on any other basis chosen by

the data source, such as whenever the data source has spare computational and network bandwidth resources.

In a pull paradigm, the statistical agency or other entity pulls the data from the source when it needs the data. This pull could be based on the issuance of a service order; the pull could be periodic, just as in the case of a push. The pull can be whenever the agency or entity needs the data to perform some computation. In practice, the data source may not give free access to its data to the requesting (consumer) agency or entity. So a pull is typically implemented as a request from the consumer to the source, which the source then responds to. The key point is who controls the timing. However, a data pull can also be implemented without requiring the explicit cooperation of the data source: for example, the consumer system could scrape data from a website put up by the source.

Although data can be pulled without the explicit cooperation of the generating source, several issues regarding such a pull need to be addressed. First, there may be legal restrictions on the frequency or volume of access. Second, there may be no guarantee of continued access in the future. In addition, the guarantees regarding the quality of the data are unknown. Data quality is always of concern, and obtaining the data without coordination with the data generator further exacerbates the issue since no explicit contractual guarantees are provided by the data generator to the data collector.

Another design parameter to consider is how each data transfer (whether push or pull) relates to what has been previously transferred. A transfer could be a complete refresh, meant to overwrite the previous data; it could be an addition, comprising only new data from the current period; or it could be a change log, comprising not only new data but also other changes (such as updates and deletes). In many uses of multiple data sources for the federal statistical system, the panel assumes that updates of data sources will be to update statistics. When data sources send updates, statistical agencies will need to keep track of the multiple versions of data. It is possible that various statistical products will have been computed using different versions of data. Reconciling these statistical results will require keeping track of specific versions used (see sections, "Provenance" and "Reproducibility").

### Data Cleaning and Transformation

Generally, data that are obtained for statistical analyses for the first time will require statistical attention prior to analysis. Such attention may be required for many reasons, including: there could be recording errors in the original source; there could be mistakes in understanding or interpreting metadata; there could be errors in data linkage (see Chapters 2 and 6); and

there could be missing data in fields needed for the computations. Regardless of the cause, such errors can be propagated and result in bad statistics if they are not corrected. As such, data cleaning is a critical function, which has to be performed when new data are received and possibly again after processing stages.

Data cleaning techniques range from actual removal of erroneously detected data to replacement of or additions to data items through extrapolation, harmonization, or approximation. Rules might be imposed on the data—such as data domain ranges; averaging or mode selection; and comparing and augmenting through external sources.

Data might be enhanced or completed. One such example is the completing of addresses by adding four-number ZIP code suffixes to the originally provided five-digit codes or simply adding a missing city or state to the ZIP code provided. External information might be used to obtain the appropriate code suffixes.

The data might be harmonized. For example, state names might all be converted to the two-letter state name coding. Another example is the conversion of U.S. phone numbers, that is, stripping any additional characters other than the 10-digit numbers. In a quite different realm, for medical data, missing body temperatures might be assumed as "normal" if only fever ratings are recorded. Regardless of the cleaning techniques used, caution is needed to ensure that the cleaning process itself does not introduce error or bias.

Federal statistical agencies are well acquainted with data cleaning and transformation in the context of survey data. A major difference between what they have been doing and what would be required in the envisaged new system is that they currently often build in data cleaning checks at the acquisition stage so that they can collect more accurate information from the household or business respondent directly. For example, there may be specified ranges built into an Internet questionnaire or the interviewer's computer-assisted personal interviewing (CAPI) instrument that do not permit respondents to enter values that are out of the range. Consistency checks are often also built in to make sure that data are consistent and to catch potential errors. However, what is done in the survey context often cannot be done in the same way with data acquired from other sources, which results in more work to clean the data and a shift in costs from data collection to preprocessing and preparing the data.

In the panel's first report, we noted that there are a wide variety of structured, semi-structured, and unstructured data available that could have the potential to enhance federal statistics. Much of the data from these sources will not be available in the desired form and structure; it will need to be transformed. Usually, such a transformation is straightforward to perform, though it may not always be easy to specify the transformation

correctly. Furthermore, it may not be possible to perform all required data cleaning at the time of acquisition. For example, if a selected data source has some critical missing values in some records, it may not be feasible to insist that these be completed, as could be done in a CAPI survey. Instead, these missing values may need to be imputed, and the most efficient way to perform such imputation is with additional context obtained through record linkage.

Current surveys often collect detailed descriptions of jobs and industries, which coders then review and classify into the North American Industry Classification System or standard occupation codes. Federal statistical agencies have been developing and using sophisticated tools to streamline these kinds of coding tasks and will need to develop and apply similar tools with new data sources.

### Provenance

There is a well-developed notion of metadata associated with surveys, including capturing and recording paradata—auxiliary information obtained during data collection that provides data about the data collection process itself. Similarly, there is also a well-developed notion of reproducibility in software, by recording the specific version of a program run and all parameters used. (Recording of the statistical methods used, such as imputation of missing values, removal of outliers, and the like constitutes a subset of the issue of software reproducibility, discussed below). In the computer science field, these notions are referred to under the concept of "provenance."

Provenance, a term most often associated with a work of art, refers to its origin and provides confidence that it is not a fake. For data, provenance serves the same purpose. For data obtained from nontraditional sources, such as repurposed administrative data or private-sector data, it will be critically important to carefully specify what the equivalent would be to survey metadata and paradata. Since data are being repurposed, the meanings of particular values are likely to be subtly different. Having a precise understanding of the provenance of the data will be critical for correct interpretation, but it will be difficult due to varying metadata recording standards across data sources. Similarly, population coverage and sampling bias are also of concern for repurposed data. Therefore, understanding and documenting what is known about these domains will be a necessary step to ensuring correctness. Finally, repurposed data will often require considerable manipulation. Therefore, recording the editing and cleaning processes applied to the data, as well as the statistical transformations and the software run, will also be critically important.

There are multiple types of metadata, and all of these have to be

recorded at the source. However, mere recording is not in itself enough: metadata will also be needed as data are transformed and new data products are derived, so that the dependencies associated with any data product of interest can be fully understood, including, in particular, the final reported statistics. Data provenance methods, which allow researchers to track their data through all transformations, analyses, and interpretations, have been developed for this purpose.

There are many different ways in which provenance can be recorded. Perhaps the most important distinction is between set-level (or process) provenance and item-level (or database) provenance. The former is captured automatically by workflow systems: for a given dataset or a statistic, it provides information on the sequence of operations for its creation. However, if a particular item in a newly developed dataset surprises a user, knowing the creation process for the dataset is not necessarily helpful. In such cases, the alternative is to record provenance for individual data items, recording how each one was derived. Such fine-grain recording of provenance can involve significant time and, consequently, costs, and ways to do this efficiently is an active area of research.

For national statistics, the panel assumes that users, for the most part, care only about the aggregate averages so one may think that process-level provenance will suffice. However, specific values are often manipulated individually, and a detailed record is required of the individual manipulations. For instance, outliers may be eliminated and erroneous entries may be manually corrected. It is not enough, for example, to record a manual review and error correction step at the level of a dataset: information needs to be recorded on the individual manual corrections applied.

The panel recognizes that federal statistical agencies currently have thorough documentation and good metadata, often including paradata, for their surveys. For many ongoing major surveys, there is a wealth of information and research that has been accumulated over many years. However, statistical agencies typically do not have or retain some of the fine-grained detailed records discussed above as part of data provenance. For example, analysts who review and edit business survey data may not retain a record of every action they took in cleaning the data. Edits and imputations performed on survey datasets may not be fully documented in a user-friendly manner except at a very high level (e.g., if hot deck imputation is used). Often, only a small number of people are familiar with the code that performs these operations, which also may not be clearly documented. Thus, instituting the more comprehensive and detailed documentation of these processes and activities that will be required for new data sources will be new to agencies and may be seen as burdensome, but we believe it is worthwhile.

The complete provenance associated with any dataset can be over-

whelming. Even if the provenance contains all the requisite information, actual utilization is not easy. This "fitness-for-use" issue is much more complex when using multiple data sources, as multiple data sources often provide more information than traditional single data sources. Frequently, the user may not be interested in the provenance of the entire dataset, but only in a particular value. Much of the provenance recorded may not be relevant to that particular value, so a much less complete provenance may suffice for the user's needs. However, it may not be sufficient simply to identify the data sources from which the particular result was derived; one may wish to identify the specific source values that contributed to it. The concept of fine-grain provenance will be valuable in such cases.

Often, when a user seeks provenance for a dataset or item, the user has a particular purpose in mind: for example, to answer a specific question. Rather than providing such a user with the provenance for the dataset, should the user be provided with only the provenance components that are relevant to the question? This question and similar ones are currently topics of an on-going stream of research, with some good ideas being investigated.

### Reproducibility

In theory, a complete provenance record should permit the entire data production and computational process to be reproduced. In practice, a few additional considerations arise. One is dependence on secondary inputs. For example, suppose that one aggregates input from a credit card processor to determine spending by category. This process requires that every merchant be mapped to a category. If the category of a merchant changes, then the computation loses reproducibility unless the entire mapping table is also recorded. A second consideration is dependence on a software version. Small, supposedly innocuous changes in software can cause different results to be produced if rounding is done differently.

Software does not exist in a vacuum; it relies on an operating environment composed of both hardware and software. If the operating environment changes—for example, if the hardware platform is changed—results might change. The precision of a computation can change if the operating environment changes even though the application software system remains unaltered.

There are examples from the decennial census of how there have been changes over time in edit rules and coding. At one time, "tailor" was an occupation with male connotations and "seamstress" an occupation with female connotations. Therefore, a census form response identifying a woman as a tailor quietly had her occupation changed to seamstress at that time (Conk, 1980). In the 2000 census, same-sex couples who identified as

married were recorded as unmarried as there were no states at that time that legally recognized same-sex marriages.

In short, issues for reproducibility are mostly well understood, but perfect reproducibility can be very difficult to implement in practice. Departures from attaining full reproducibility may become necessary, sometimes for reasons of cost, but they should be undertaken with care as reproducibility is of growing importance throughout scientific work (National Academies of Sciences, Engineering, and Medicine, 2016b). Reproducibility will be key to helping researchers both in and outside the federal statistical system understand what was done with different data sources and the quality implications for the resulting statistics (see Chapter 6). Allowing internal and qualified external researchers to access raw granular data and to examine the provenance and perform appropriate analyses will permit useful evaluations of what was done and sharing of good practices across agencies. The panel recognizes that to the extent that reproducibility requires recording detailed, record-level information, there are implications for privacy (see Chapters 4 and 5).

Being able to explain exactly what process was used is also important to maintaining the credibility and trust of an agency with the users of its data. This consideration increases in importance as more complex and more computational processes are used to generate statistical products.

**CONCLUSION 3-3 Creating statistics using multiple data sources often requires complex methodology to generate even relatively simple statistics. With the advent of new and different sources and innovations in statistical products, federal statistical agencies need to figure out ways to provide transparency of their methods and to clearly communicate these methods to users.**

## SYSTEM MIGRATION

As we noted above, computing statistics from diverse data sources will require a system architecture that differs substantially from what many statistical agencies have today. This requirement raises the question of the migration path for data both within an agency and to the panel's recommended new entity. We note that this migration occurs not just for the computing systems, but also for the business processes used.

In general, a gradual migration introduces less risk. However, in many instances an agency may not have the luxury of being able to migrate gradually. For example, if credit card transaction data are to be used to compute some statistics of economic activity, there may be no reliable way to take a portion of the reported transactions and use traditional data collection techniques for the rest.

**BOX 3-1**
**Pilot Study in Migration: Palantir Technologies**

In 2016, Palantir Technologies conducted an 8-month pilot study to report on the "health" of the U.S. economy. The findings of this study indicated that by combining credit card transaction reports with public measures, accurate and timely economic insights were obtained.

The pilot study integrated data from the Bureau of Labor Statistics, the Census Bureau, the Energy Information Agency, and the Bureau of Economic Analysis (BEA), along with credit card transaction information from the database of First Data (one of the largest payment processing companies in the United States). The data integration and analysis engine enabled BEA and Census Bureau analysts to incorporate various other datasets into the system to support independent investigations at multiple levels of granularity, with varying aggregation duration lengths.

The underlying motivation for the pilot was the creation of a system to carefully track the indicators of retail trade as measured by the Monthly Retail Trade Surveys. While traditionally collecting and processing the monthly retail trade indicators takes weeks, the pilot system took hours, with nearly a perfect (0.96) correlation between the results from the surveys and the credit card transaction data.

Traditional survey-based data collection is not only difficult and suffers from ever-declining response rates, but processing at a fine granular level is prohibitively expensive. In contrast, such processing at a geographically refined sector and duration-specified intervals using the pilot solution resulted in a nearly 0.9 correlation of findings as compared to the findings obtained using traditional sources without increasing cost.

Architecturally, Palantir Technologies' platform consists of a *data hub* and a set of *report generators*. The data hub logically is a set of storage platforms and corresponding analytical models that organize and manipulate the data on the storage platforms. Report generators of various types process the data according to their respective business rules. Key to the acceptance of the effort is Palantir Technologies' data integration approach, which retains complete data provenance throughout the entire data creation and manipulation cycle.

Palantir Technologies supported a variety of core capabilities, including security and auditing mechanisms, collaboration environments, and report generation tools. The open architecture enabled third parties to develop and integrate specific applications. Like all successful analytical infrastructures, Palantir Technologies touts and demonstrates that its engine supports a flexible data management configuration, reliable and easily parsed auditing tools, a highly secure environment, and a wide variation of collaboration options.

To soften the possible impact of and concerns about an abrupt migration, it would be advisable to initially run traditional approaches simultaneously with the new approaches, which is currently common practice in the statistical agencies. Comparing the findings of the two approaches

would verify the correctness of the new approach and instill confidence in it. It may also be possible to create "sandbox" spaces where new computational streams can be experimented with and tested before being moved into production. In addition, agencies can use "rollback" mechanisms, in which some "old" processing modes can be used if difficulties are found in the new mode.

When transitions are made, the changes need to be carefully logged. Thus, any errors or undesired changes can be detected, and users will at least have the beginning of an explanation in many situations. Such a log may even permit a rollback of changes that have been applied.

## PERSONNEL STAFFING AND SKILLS

The use of new technologies and new methodologies will require staff at federal statistical agencies and the panel's recommended new entity to have appropriate new skills. This is true even if the bulk of the computational work is outsourced to a private contractor. For the existing statistical agencies, we believe this need can be met with a judicious use of training programs and a shift in the skill profile for new staff over time. The agencies have many staff with strong technical skills and experience that is relevant for dealing with data from any source. We believe that with additional training, many staff will be able to adopt the new paradigm, provided that key technical steps are outsourced.

The key requirement for moving to the new paradigm is a smooth transition. Agency history and domain knowledge need to be preserved. Hence, it is important not only that current staff do not experience hardships due to a migration to a new approach, but that they are also incentivized to be involved with the new approach so that they can capture the needed domain knowledge and improve the chances of meaningful new data being obtained and properly utilized. See Box 3-1 for a brief description of a pilot study that illustrates an exemplary migration.

**RECOMMENDATION 3-1 Because technology changes continuously and understanding those changes is critical for the statistical agencies' products, federal statistical agencies should ensure that their information technology staff receive continuous training to keep pace with these changes. Training programs should be set up to meet the current and expected future training needs for technology, and recruitment plans should account for future technology demands.**

# 4

# Legal and Computer Science Approaches to Privacy

In the panel's first report, we briefly reviewed some of the major privacy issues related to combining multiple data sources, including the relevant laws and approaches that statistical agencies have taken to fulfill those laws and protect their data while providing access to researchers. In this chapter we expand on that discussion, focusing on legal issues that arise in collecting, acquiring, and combining multiple data sources. We discuss privacy from a legal perspective and a computer science perspective, attempting to reconcile these different views. We also discuss how moving to a world of combined multiple data sources changes threats to privacy and introduces new threats. We then address the implications for federal statistical agencies, including the additional privacy and confidentiality laws that apply to statistical data, as well as the legal and policy issues that arise with linking records from different data sources. We continue discussion of privacy issues in the next chapter, expanding on the discussion in our first report on how federal statistical agencies can use security measures, computer science technologies, statistical methods, and administrative procedures to protect data and permit access for statistical purposes.

## PERSONALLY IDENTIFIABLE INFORMATION AND PRIVACY LAW

Almost all federal household and economic surveys assure respondents that the information they provide will be protected and will not be used to harm them. For example, respondents to the American Community

Survey are told: "We never reveal your identity to anybody else. Ever."[1] The website explains that "[a]ll Census Bureau employees take an oath of nondisclosure and are sworn for life to protect all information that could identify individuals." Information that could identify individuals is referred to as "personally identifiable information" (PII), and each federal agency has regulations and procedures for protecting PII.

Combining data sources has the potential to reveal more information about individuals in the data sources. For example, if records from a survey of college graduates were linked with university records and information about the subsequent work history of the respondents, the original, limited information from the survey now is joined with more detailed information. It is possible that a public-use dataset published from the linked dataset might have enough information to allow for the identification of individual respondents even if all information such as names, addresses, dates of birth, and names of universities and employers had been deleted from the records. The additional information available through the linkage also makes it possible to publish statistical summaries, such as cross-tabulations on more variables, and, in some instances, those tables taken together might be used to identify individuals in the survey even though the individual tables contain only statistical information on groups of records.

The phrase "personally identifiable information" is central to the development of modern privacy law. The phrase appears frequently in federal law, judicial opinions, and legal scholarship. It is also an imprecise term that has led to confusion, particularly between legal scholars and technology experts. When legal scholars use the phrase, they anticipate that a determination will be made, within the context of laws and legal institutions, as to what constitutes PII. As with many terms in law, the phrase takes on meaning in the context of specific use: data that may be considered PII in some circumstance may not be considered PII in other circumstances. Computer scientists, in contrast, have a different view of the phrase. They would argue that all information could be viewed as PII, and, as a consequence, threats to individual privacy are not adequately addressed by the PII/non-PII dichotomy. Moreover, computer scientists would argue that in a networked world, the protection of privacy will require mathematically rigorous notions that can be translated with algorithms into numerical outcomes.[2]

The core differences, and the source of much confusion, may be understood as the difference between the central place of PII in modern privacy law and the ability of modern computer science to breach individual pri-

[1]See https://www.census.gov/programs-surveys/acs/about/is-my-privacy-protected.html [August 2017].

[2]See the section "Examples Elucidating the PII/Non-PII Issue" for a more detailed discussion of both views.

vacy, that is, to turn non-PII into PII in situations in which it is not obvious. There is no simple way to resolve these two views of PII: one relies on legal constructs, the other on scientific specifications.

However, there may be a way to integrate the insights of both disciplines to inform current understanding of PII. In law, the concept of PII carries with it legal rights and responsibilities, often described as fair information practices. The aim is to ensure the protection of PII, which is a legal obligation typically assigned to the entity in possession of the PII. There is a good reason to assign this responsibility to the data holder and not the data subject: the entity in possession of the personal data is in a better position to reduce the risks that might result from adverse use or a security breach. As explained in the famed 1973 report that led to establishment of the Privacy Act (Turn and Ware, 1976, p. 1):

> Privacy is an issue that concerns the computer community in connection with maintaining personal information on individual citizens in computerized record-keeping systems. It deals with the rights of the individual regarding the collection of information in a record-keeping system about his personal activities, and the processing, dissemination, storage, and use of this information in making determinations about him.

The corollary is that such obligations do not apply if the dataset does not contain PII. In recent years, computer scientists have helped make clear that what may not appear to be PII is in fact PII when new techniques or additional data are considered.[3]

The current situation has led some experts to suggest that PII is no longer a workable category because PII and non-PII are no longer readily distinguished. But if the legal purpose of PII—to assign rights and responsibilities in the collection and use of data—is combined with the scientific ability to reveal the existence of individual privacy compromise when it is not obvious, then the better solution is to recognize that the legal definition of PII should include both data that are obviously PII and "latently PII," that is, data that can be transformed into PII or, more broadly, that enable individual privacy compromise.

In essence, the panel believes that the legal PII category remains relevant and that the insight of scientists should inform how the law understands the term. One obvious implication is that the concept of PII becomes more important in a world of simultaneous use of multiple data sources.

[3]A simple example of this is provided by a Social Security number. By itself, an SSN may appear not to be PII because the actual identity of the person associated with the SSN is not clear. However, if there is a lookup table that matches SSNs to individuals, the problem becomes trivial. The law understood this problem from the outset and always treated SSNs as PII (see U.S. Department of Health, Education, and Welfare, 1973; also see more detailed discussion in Chapter 5).

## LEGAL VIEW OF PRIVACY IN THE CONTEXT OF STATISTICAL DATA ANALYSIS

Our starting point for a discussion of the legal context of statistical data analysis in the federal government begins with the language of the federal Privacy Act (see Box 4-1). That law sets out a wide range of responsibilities for federal agencies that collect, use, and disclose personal information, namely, PII. That information can include everything from employment records for agency personnel to license applications for pilots to the investigative records of law enforcement agencies. When a record is contained in a system of records, many legal obligations are created, including obligations to ensure the accuracy and integrity of the record, to ensure its security, and to make it available to those individuals to whom it pertains. However, an important exception for records maintained by federal agencies is made for "statistical records." It is these records that are the focus of our discussion.

The Privacy Act describes a statistical record as "a record in a system of records maintained for statistical research or reporting purposes only and not used in whole or in part in making any determination about an identifiable individual" (5 U.S.C. 552a(a)(6)) with a few exceptions. Other sections of the Privacy Act limit matching of datasets except those that "produce aggregate statistical data without any personal identifiers" (5 U.S.C. 552a(a)(8)(B)(i)) or "performed to support any research or statistical project, the specific data of which may not be used to make decisions concerning the rights, benefits, or privileges of specific individuals" (5 USC 552a(a)(8)(B)(iii)). Another provision of the law limits disclosure of personal records maintained by federal agencies except "to a recipient who has provided the agency with advance adequate written assurance that the record will be used solely as a statistical research or reporting record, and the record is to be transferred in a form that is not individually identifiable" (5 USC 552a(b)(5)). Records are also excluded from certain obligations, including privacy and accuracy if they are "required by statute to be maintained and used solely as statistical records" (5 USC 552a(k)(4)).

It is noteworthy that these legal constructs are quite distinct from those common in statistics. In statistics, a statistical record is an aggregate of individual records: the notion of a single statistical record conflicts with the fact that statistics are based on aggregates of multiple records. Statisticians more commonly refer to "statistical uses" of record systems, implying that the statistics are summaries of attributes of many records in a record system.

Under federal privacy law, statistical data may be widely gathered, exchanged, and disseminated with the understanding that the federal agency does not have the ability to make determinations about "identifi-

**BOX 4-1**
**Findings and Purposes of the Privacy Act**

**Findings**

(a) The Congress finds that—

(1) the privacy of an individual is directly affected by the collection, maintenance, use, and dissemination of personal information by Federal agencies;

(2) the increasing use of computers and sophisticated information technology, while essential to the efficient operations of the Government, has greatly magnified the harm to individual privacy that can occur from any collection, maintenance, use, or dissemination of personal information;

(3) the opportunities for an individual to secure employment, insurance, and credit, and his right to due process, and other legal protections are endangered by the misuse of certain information systems;

(4) the right to privacy is a personal and fundamental right protected by the Constitution of the United States; and

(5) in order to protect the privacy of individuals identified in information systems maintained by Federal agencies, it is necessary and proper for the Congress to regulate the collection, maintenance, use, and dissemination of information by such agencies.

**Purposes**

(b) The purpose of this Act [enacting this section and provisions set out as notes under this section] is to provide certain safeguards for an individual against an invasion of personal privacy by requiring Federal agencies, except as otherwise provided by law, to—

(1) permit an individual to determine what records pertaining to him are collected, maintained, used, or disseminated by such agencies;

(2) permit an individual to prevent records pertaining to him obtained by such agencies for a particular purpose from being used or made available for another purpose without his consent;

(3) permit an individual to gain access to information pertaining to him in Federal agency records, to have a copy made of all or any portion thereof, and to correct or amend such records;

(4) collect, maintain, use, or disseminate any record of identifiable personal information in a manner that assures that such action is for a necessary and lawful purpose, that the information is current and accurate for its intended use, and that adequate safeguards are provided to prevent misuse of such information;

(5) permit exemptions from the requirements with respect to records provided in this Act; and

(6) be subject to civil suit for any damages which occur as a result of willful or intentional action which violates any individual's rights under this Act.

SOURCE: The Privacy Act of 1974, 5 U.S.C. § 552a; available: https://www.justice.gov/opcl/privacy-act-1974 [September 2017].

able individuals," to gather "personal identifiers," and that the records are not "individually identifiable."

Critical to understanding the significance of the term "statistical data" in the context of federal agency systems is the recognition that many of the responsibilities assigned to federal agencies for the collection and use of personal data are relaxed for the category of statistical data. To better understand the current situation, we turn to a bit of history.

Prior to the enactment of the Privacy Act, there was a lengthy review of federal record-keeping systems that resulted in a major report. The U.S. Department of Health, Education, and Welfare (HEW) closely examined the issue of statistical data, and many of the insights in that report are reflected in the law that followed.

The report noted that, with few exceptions (U.S. Department of Health, Education, and Welfare, 1973):

> [T]here is little to prevent anyone with enough time, money, or perseverance from gaining access to a wealth of information about identifiable participants in surveys or experiments. This should not be the case . . . (p. 93)
>
> Social scientists and others whose research involves human subjects are vocal about the importance of being able to assure individuals that information they provided for statistical reporting and research will be held in strictest confidence and used only in ways that will not result in harms to them *as individuals* [emphasis in original]. (p. 93)

At the same time, the report noted the value of statistical data:

> The obverse of the problem of data confidentiality is the need to make basic data more accessible for reuse or reanalysis by all qualified persons or institutions. Personal data systems for statistical reporting and research are largely in the hands of institutions that wield considerable power in our society. Hence, it is essential that data which help organizations to influence social policy and behavior be readily available for independent analysis. (p. 94)

The report even anticipated some of the current issues:

> In principle, there should be no conflict between informing the public about how the government conducts its business and protecting the individual data subject from harm. If data cannot be made available for reuse or reanalysis without disclosing the identity of data subjects, special precautions may have to be taken before making basic data accessible to qualified persons outside the collecting organization, but such precautions should be taken. (p. 95)

Overall, the HEW report describes many proposed safeguards for statistical record systems that were eventually adopted in the federal law.

## THE SCOPE OF PII

As we explain below, the concept of PII varies among different laws that relate to privacy, which has added to further skepticism among computer scientists about the use of the term. For example, it is not at all obvious why PII should create a different boundary condition for medical records than it does for video viewing records. However, many definitions of PII include both what is obviously PII and what could, through additional steps, be PII. The recently adopted General Data Protection Regulation of the European Union, which will likely be enormously influential in the years ahead, defines "personal data" as:[4]

> . . . any information relating to an identified or identifiable natural person ('data subject'); an identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person.

This definition contains the key phrase "identified or identifiable," which conveys the view that actual identification may not be immediately apparent. This interpretation is strengthened by the subsequent modifier that a person may be identified "directly or indirectly." Many privacy laws adopt this view of PII: that is, information that is both personally identifiable and information that could be personally identifiable.

As noted above, computer scientists look at privacy and statistical data through a different lens, questioning if there even exists a meaningful boundary dividing PII from non-PII. Given the tools of cryptography, they ask if such a distinction exists given the wealth of possible side ("auxiliary") information that is available, such as other accessible datasets, last year's statistics, newspapers, blog posts, and tracking information.[5]

Privacy laws and their implementing regulations include numerous examples of specific rules for identifying concrete characteristics as PII, as well as more open-ended decision tools (standards) for categorizing data as PII or non-PII. Thus, while PII is a core concept in modern privacy law, there is wide variation in the definition of PII (and similar terminology, such as "individually identifiable information" and "personal information") and how it is interpreted in practice, even among federal agencies. For example, the Family Educational Rights Privacy Act and the Health Insurance Portability and Accountability Act provide different approaches to protecting data and enabling statistical use by external researchers. A

[4]See https://www.privacy-regulation.eu/en/4.htm [August 2017].

[5]Statisticians would also be concerned about available auxiliary information and would apply statistical disclosure limitation methods.

clean mathematical separation between PII and non-PII could pave the way for unfettered access to the teachings of non-PII data on an Internet scale.

Thus, to computer scientists, the relevant questions about the appropriate meaning of PII are, "What is the law trying to promote? What is it trying to proscribe?" Equipped with answers to these questions, not only can one evaluate a proposed definition and treatment of PII and non-PII in light of these goals, but one can also begin to address these goals directly, bypassing the definitional question.

A focus on distinguishing between inferences about individuals and inferences about groups provides essential tools for reasoning about privacy when the data are not collections of records, each belonging to a single individual. For example, the inference approach permits reasoning about the privacy implications of sets of statistics computed from individual records, even if these records have subsequently been destroyed. The "post-PII" question of what can be inferred from the statistics is a persistent privacy concern even in this extreme setting. We cannot overemphasize this point: the mere fact that there is no record to "re-identify" or to associate with a unique individual (because the records have been destroyed) does *not* mean there is no residual risk of disclosure specific to an individual through inference from the statistics that are available (together with auxiliary information). Thus, for such a collection of statistics, PII sounds like a misnomer.

At this point in the argument, it is useful to note that "inference" in this context is itself distinct from the word as used in much of statistics and therefore common to the federal statistical system. Much of the descriptive power of sample surveys is based on the use of probability samples of large populations. The use of probability sampling provides mathematical bases for the relationship between a statistical aggregate in a sample and the corresponding aggregate for the whole population. In this sense, sample-based statistics have known inferential characteristics in relationship to the full population from which the sample was drawn. In contrast, "inference" with regard to the possibility of identifying a specific individual in a once-existing record system concerns the probability of inferring something about an individual based on information extractable from the record system. In this case, an "inference" is from a set of information to the identity of an individual record.

## EXAMPLES ELUCIDATING THE PII/NON-PII ISSUE

In this section we explore the PII/non-PII issue through the lens of inference. Our goal will be to distinguish the case in which it is possible to infer a sensitive attribute about a single individual from the case in which it is only possible to infer that attribute about members of a group as a whole. One can call this the difference between an individual privacy breach and

a group privacy loss. Generally, statistical analysis accepts group privacy loss. For example, one may learn that people with a specific gene have an increased risk of developing a particular illness, which is a fact about people in general. In one sense, "group privacy loss" may accurately be viewed as "scientific discovery." Suppose the data from the study about this increased risk contain the medical history of an individual, Alice, who has been diagnosed with the illness. If the study enables one to infer that Alice has been diagnosed with this illness, then it is an individual privacy breach. It is not a fact about the population as a whole; an individual diagnosis does not logically follow from increased risk (or even illness). Anything that can be learned about Alice as a result of her participation in the study that could not have been learned had she not participated in the study is an individual privacy breach. In contrast, however, if Bob—who may or may not have been in the study—publishes his genetic data, and the study allows one to infer that Bob is at increased risk of the illness, it would not be considered as an individual privacy breach for Bob.[6]

With the distinction between group privacy loss and individual privacy breach in mind, it is useful to consider such subjects as water salinity data, ice shelf measurements, and location of the jet stream, which do not appear to be about people at all or have any implications for individual privacy. In the legal view, these data are not PII. However, consider air quality statistics that summarize levels of a pollutant produced only by automobiles in a small town. Since this is a direct measurement of something produced by human-driven vehicles, it is clearly "about" people's driving patterns. Indeed, given enough information about the driving of all the inhabitants but one, and given the measured pollutant level, it is possible to learn how much the "final" inhabitant drove. There are ways of getting at this personal information; for example, by comparing measurements during days in which she is ill (and off the road) to measurements during days in which she is healthy. Although this leads, in theory, toward classifying the pollutant level as PII (especially in small towns), it is logical to adopt a "watch and wait" approach for this scenario because a number of factors suggest that breaching individual privacy may be difficult. Two such factors are "noise" in the measurements caused by atmospheric conditions such as wind and precipitation, and, the possible difficulty of obtaining repeated measurements of the type (currently) known to be useful, when combined, for an individual privacy breach.

Taking this example one step further, consider a dataset obtained by linking the pollutant measurements to health records. Assume further that

[6]This example also highlights the danger of categorizing information: when Bob chose to publish his genetic data, he might not have anticipated the (future) scientific discovery of his increased risk of disease; had he done so, he might not have published the information.

the data exist for a large city, rather than a small town. A public health goal might be to learn about correlations between pollutant levels and the incidence of chronic bronchitis. Such correlations are aggregate statistics about the population and would not constitute PII. However, if one could learn that an individual in the dataset experiences chronic bronchitis, it would be a personal privacy breach unless the same conclusion could be drawn if the individual was not in the dataset. Thus, the individual health records should be viewed as PII, while the link between pollution level and chronic bronchitis is a statistical fact about the population. Moreover, in this example, the records could be queried in a way designed to breach individual privacy, something that appears harder to do in the example of the atmospheric measurements in a small town.

We close with a compelling example of the subtlety of the individual privacy breach determination: allele frequency statistics in genome wide association studies:[7]

> A genome-wide association study is an approach that involves rapidly scanning markers across the complete sets of DNA, or genomes, of many people to find genetic variations associated with a particular disease. Once new genetic associations are identified, researchers can use the information to develop better strategies to detect, treat and prevent the disease. Such studies are particularly useful in finding genetic variations that contribute to common, complex diseases, such as asthma, cancer, diabetes, heart disease and mental illnesses.

The large number of measurements in a genome-wide association study present a privacy problem that has only relatively recently come to be understood (Homer et al., 2008; Dwork et al., 2015): If one has *only the statistics* for the case group (people diagnosed with the illness) and control group (healthy individuals), together with the DNA of an individual, it is possible to determine if this individual is a member of the case group. Since the members of the case group have been diagnosed with the illness, this determination is an individual privacy breach. This situation is different from learning that someone's DNA suggests an increased risk of the disease, which is what can be inferred for someone not in the study. In our approach to the PII/non-PII problem, learning the markers associated with an illness is a scientific fact about the population: learning that an individual has been diagnosed with an illness is an individual privacy breach.

[7]See https://www.genome.gov/20019523/ [March 2017].

## SYNTHESIS: A PROPOSED LIABILITY RULE FOR PII

The formulation set out in this chapter respects the perspectives of both law and computer science. The panel's aim is to uphold the approach set out in the original Privacy Act of 1974 for the treatment of statistical data but to recognize that, over time, new techniques have emerged that have changed non-PII statistical data as defined in the original law to PII because of the availability of auxiliary information. We are not the only ones to attempt to bridge these views. Nissim et al. (2016, p. 33) wrote:

> Legal and computer science concepts of privacy are evolving side by side, and it is becoming increasingly important to understand how they can work together. The field of computer science can benefit from an understanding of legal thinking on privacy, and the legal field can similarly be influenced by computer science thinking. The influence of one discipline on another can be very valuable in the future.

One way to resolve these two perspectives is simply to acknowledge that data that at one time may have been viewed as statistical data (i.e., non-PII data in the legal sense) are no longer statistical data. The practical consequence that follows from this acknowledgment is that the data must be protected in line with the higher standards typically associated with PII. And the acknowledgment has a further import: it establishes that the privacy status of data is dynamic over time, that datasets that are not individually identifiable today may in the future become individually identifiable.

There are at least two policy consequences of this acknowledgment. First, a determination that a dataset is statistical data should likely now include a date of certification that establishes when the data were deemed to be non-PII. Since it is a federal agency that is responsible for the management of the data systems, the federal agency should likely be responsible for this certification. These determinations should be periodically reviewed as more auxiliary data and new techniques are developed.

Second, if one wishes to establish that a dataset will remain statistical data as long as one can foresee, it should be provably so. This characterization can be seen in such data as water salinity on the Chesapeake: as the data never contain PII, there is no auxiliary information or technique that would make it PII at any time.

## IMPLICATIONS FOR FEDERAL STATISTICAL AGENCIES

To this point in the chapter, we have contrasted legal and computer science definitions of PII and emphasized that the common legal interpretation of the PII status of data is not a simple, invariant function. Rather, the PII status of a record is a dynamic feature, not a static feature, of a record.

Procedures to protect the privacy of individuals are therefore an ongoing responsibility of the holder of a record system, which, for federal statistics, will involve a federal agency. Thus, this section considers the implications for federal agencies of the dynamic features of record.

This chapter represents a significant part of the panel's assessment because when multiple record systems are combined, new issues of privacy protection may arise. Consider, for example, the case of a public-use file that is released by a statistical agency (agency A) for statistical analysis after careful attempts to anonymize the data. Now consider another record system with personal identifiers that has been kept totally confidential and in a program agency (agency B), never subjected to statistical analysis, and not released to the public. Since no information was ever disseminated from agency B, the work by agency A to protect the identity of the public-use file was not informed by the information held by agency B. If, however, the agency B dataset were combined with the dataset from agency A that generated the public-use file and statistical analyses disseminated on the combined set, the probability of re-identification of an individual in agency A's dataset might be altered.

In short, moving into a world in which multiple datasets are combined can change threats to privacy. In the next sections of this chapter, we examine other privacy and confidentiality laws that apply to statistical data, as well as the legal and policy issues that arise with linking records from different data sources.

**RECOMMENDATION 4-1 Because linked datasets offer greater privacy threats than single datasets, federal statistical agencies should develop and implement strategies to safeguard privacy while increasing accessibility to linked datasets for statistical purposes.**

### Other Laws Protecting Statistical Information

Administrative records systems on individuals are likely covered by the Privacy Act and subject to the permitted statistical uses described above. When statistical agencies use these records as a sampling frame for their surveys and append the survey data to that frame, the entire dataset is covered by the Privacy Act. In addition to the Privacy Act, federal statistical agencies are subject to protecting the confidentiality of identifiable information that they collect or acquire by the Confidential Information Protection and Statistical Efficiency Act of 2002 (CIPSEA) and their own organic statutes. The confidentiality of Census Bureau data is governed by Title 13, Section 9, of the U.S. Code, which specifies that:

> (a) Neither the Secretary, nor any other officer or employee of the Department of Commerce or bureau or agency thereof, or local government census liaison, may . . .
> (1) use the information furnished under the provisions of this title for any purpose other than the statistical purposes for which it is supplied; or
> (2) make any publication whereby the data furnished by any particular establishment or individual under this title 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.
> No department, bureau, agency, officer, or employee of the Government, except the Secretary in carrying out the purposes of this title, shall require, for any reason, copies of census reports which have been retained by any such establishment or individual. Copies of census reports which have been so retained shall be immune from legal process, and shall not, without the consent of the individual or establishment concerned, be admitted as evidence or used for any purpose in any action, suit, or other judicial or administrative proceeding.

Similarly, CIPSEA Subtitle A, Section 512, requires that data be used only for statistical purposes and not be disclosed in identifiable form without consent:

> (a) Use of Statistical Data or Information.—Data or information acquired by an agency under a pledge of confidentiality and for exclusively statistical purposes shall be used by officers, employees, or agents of the agency exclusively for statistical purposes.
> (b) Disclosure of Statistical Data or Information.—Data or information acquired by an agency under a pledge of confidentiality for exclusively statistical purposes shall not be disclosed by an agency in identifiable form, for any use other than an exclusively statistical purpose, except with the informed consent of the respondent.

In addition to protecting PII, statistical agencies must protect identifiable information from businesses, schools, and health care providers, and many other organizations from which they collect or acquire data. Although the Privacy Act generally does not apply to these respondents, CIPSEA and the agency's organic statutes do apply and impose strict requirements on agencies to ensure that they do not disclose identifiable information (e.g., see U.S. Office of Management and Budget, 2007). In addition, other laws, such as the Trade Secrets Act or exemptions in the Freedom of Information Act, may protect some of the information that statistical agencies collect from these organizations.

## Record Linkage

As we discuss in Chapter 2, combining some data sources will likely involve using record linkage techniques to match records from two different data sources on the same individuals or entities. These linkages could involve linking survey responses with administrative records, linking two or more administrative records sources, or linking private-sector information, such as credit reports or credit card transactions, to survey or administrative data.

Ivan Fellegi traced the field of record linkage methods to the 1960s, with three simultaneous developments: accumulation of large data files about businesses and individuals, new computing capabilities that enabled processing those files, and increased demand for more detailed information. These developments simultaneously resulted in increased demands for privacy safeguards. He wrote about the development of linkage policy at Statistics Canada (Fellegi, 1999, p. 12):

> As a society we did not want comprehensive population registers, largely because we did not want a large scale and routine merging of information contained in government files. But we did not want to rule out *some* merging for *some* well justified purposes. So, as a matter of conscious public policy, we made linkage very difficult. However, we allowed the development of record linkage methodology for use in exceptional circumstances. The applications were indeed important, often requiring a high level of accuracy, so we refined the methodology, and also made it vastly more efficient.

Statistics Canada (2017, p. 2) recently updated its directive on microdata linkage, which acknowledges the "inherent privacy-invasive nature of the activity" of record linkage and expects that (1) the linked data results in information for the public good, (2) confidentiality will be maintained and the information will be used only for statistical purposes, and (3) the linkages offer demonstrable cost or respondent burden savings over other alternatives, or are "the only feasible option to meet the project objectives." The directive includes omnibus authority for linkages for specific purposes, which can include linking surveys or administrative data.

In the United States, linking records has historically raised privacy concerns. The Computer Matching and Privacy Protection Act was enacted in 1988 as an amendment to the Privacy Act to create procedures to prevent any use of computer matching that could end program benefits without notifying individuals of the matching program or illegitimate uses of computer matching. Computer matching refers to the comparison of information that often includes PII data between two or more systems, which can be used between multiple agencies to ensure that federal ben-

efits are distributed properly. For example, records from the Temporary Assistance for Needy Families program can be matched with information from the National Directory of New Hires to see if program participants have acquired a job and therefore are no longer in need of program benefits. Computer matching on average has saved New York an estimate of $62 million annually (U.S. Government Accountability Office, 2014). But computer matching has also raised significant concerns about the protection of individuals. New York City recently sought to delete data about New Yorkers that could be used to prosecute immigration cases.[8]

The requirements of the Computer Matching and Privacy Protection Act do not apply to all federal agencies; exemptions include statistical or research purposes, law enforcement investigation, and some tax-specific matching, so this does not directly affect statistical agencies. Statistical and research exemptions are provided in 5 U.S. Code § 552a section 8, Records maintained on individuals:

> (B) but does not include—
> (i) matches performed to produce aggregate statistical data without any personal identifiers;
> (ii) matches performed to support any research or statistical project, the specific data of which may not be used to make decisions concerning the rights, benefits, or privileges of specific individuals.

### Consent for Record Linkage

In reviewing the privacy risks inherent in record linkage and discussing obtaining consent from respondents to link records, the U.S. Government Accounting Office (2001, p. 57) noted:

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

However, the report also stated that consent to linkage may not be necessary (p. 58):

> If certain safeguards are in place, such as review by a group with the interests of the data subjects in mind or use of appropriate confidentiality and security protections.

The original Fair Information Practices described above do not explicitly refer to consent for record linkage. A subsequent version of Fair Infor-

[8]National Public Radio: "City Officials Go to Court to Protect New Yorkers with Municipal IDs." Available: http://www.npr.org/2016/12/20/506285207/city-officials-go-to-court-to-protect-new-yorkers-with-municipal-ids [September 2017].

mation Practices, set out by OECD in 1981, contemplates consent in two specific instances: (1) to allow the collection of personal information and (2) to use personal data for purposes other than those originally stated.[9] Recently, the U.S. Office of Management and Budget updated its Circular A-130 (U.S. Office of Management and Budget, 2016), "Managing Information as a Strategic Resource," which provides policies and requirements for federal agencies to follow for the management of federal information. This circular included the following "Fair Information Practices Principle":[10]

> Individual Participation. Agencies should involve the individual in the process of using PII and, to the extent practicable, seek individual consent for the creation, collection, use, processing, storage, maintenance, dissemination, or disclosure of PII. Agencies should also establish procedures to receive and address individuals' privacy-related complaints and inquiries.

This requirement seems to reflect the notion of individual control noted by the U.S. Government Accounting Office (2001).[11]

In the United States, there is no uniform policy that guides consent requirements for linking records. Some statistical agencies are in departments that are signatories to the Common Rule for the protection of human subjects (45 CFR Part 46), while others are not. Under the Common Rule, organizations must have an Institutional Review Board (IRB) determine whether the risks to human subjects have been minimized and informed consent has been obtained. However, even federal statistical agencies subject to the Common Rule may receive a waiver from the IRB or may not be required to go through an IRB or obtain consent from respondents for linking data because of the strong confidentiality protections they have for data they collect or acquire.

Currently, there are differences in policies and practices across statistical agencies regarding consent. For some surveys, including the National Health Interview Survey sponsored by the National Center for Health Statistics, interviewers ask respondents for explicit consent for record linkage. In contrast, the Survey of Income and Program Participation, sponsored by the Census Bureau, sends survey respondents an advance letter that states

[9] *OECD Guidelines on the Protection of Privacy and Transborder Flows of Personal Data*. Available: http://www.oecd.org/sti/ieconomy/oecdguidelinesontheprotectionofprivacyandtransborderflowsofpersonaldata.htm [September 2017]. This elaborates the discussion in the panel's first report.

[10] Available: https://www.whitehouse.gov/omb/circulars_a130_a130trans4 [September 2017].

[11] We note, however, that the individual participation principles in the *OECD Privacy Guidelines* (see fn. 9, above) do not address the issue of consent, but focus instead on the right of the individual to obtain information about the personal data that are held by others and to seek correction or deletion if requested.

the Census Bureau will obtain administrative records from other agencies and asks respondents to "opt out" if they don't want this to occur:[12]

> To be efficient, the Census Bureau attempts to obtain information you may have given to other agencies if you have participated in other government programs. We do so because it helps to ensure your data are complete, and it reduces the number of questions you are asked on this survey. The same confidentiality laws that protect your survey answers also protect any additional information we collect (Title 13, U.S.C., Section 9). If you wish to request that your information not be combined with information from other agencies, we ask that you notify the field representative at the time of the interview.

The Census Bureau has a provision in the Privacy Act that permits it to receive identifiable information from other agencies (5 USC Section 552a(b)(4)) for use in censuses, surveys, or other activities under Title 13 and does not need specific consent from the respondents to those censuses, surveys, or other activities.

Whether a statistical agency requires consent for linkage can have implications for the quality of the resulting linked data. Surveys that require consent to link records have reported declines in the percentage of respondents giving that consent, paralleling the decrease in response rates for surveys (Fulton, 2012; Kreuter et al., 2016). As with survey response rates, declines in consent rates do not necessarily introduce bias, but the potential for bias increases with declining consent rates. Bias in the linked data arises if there are differences between people who consent and those who do not consent.

Because of the many new data sources available, including those in the private sector and from the Internet, further questions have been raised about the feasibility of asking for consent or what informed consent means when all the uses of these data cannot be known (see e.g., Barocas and Nissenbaum, 2014). There have been international efforts to produce a set of ethical principles and discuss governance, legal frameworks, and other key issues such as privacy, consent, and data sharing for research using these new forms of data (OECD, 2016). This work goes beyond the scope of this panel, but we believe that federal statistical agencies should be aware of and follow these efforts and adapt from them any policies and best practices that they believe would be appropriate.

[12]See https://www.census.gov/programs-surveys/sipp/information/sipp-faqs.html [September 2017].

### Public Attitudes About Record Linkage

Public concerns about record linkage activities will need to be carefully monitored and addressed by federal statistical agencies as they move forward with combining multiple data sources. *The New York Times*[13] recently noted the public outcry in reaction to a decree from the French government to merge information from passports and identity cards into one large database containing photographs, names, addresses, marital status, weight, and fingerprints. This database was to be used for identity verification, not statistical purposes, but there may not be a clear delineation between the two purposes in public perception. If publicity about record linkage leads to greater public mistrust, that mistrust can carry over to other aspects of the federal statistical system.

[13]See http://www.nytimes.com/2016/11/11/opinion/the-risks-of-frances-big-new-database.html?_r=0 [August 2017].

# 5

# Preserving Privacy Using Technology from Computer Science, Statistical Methods, and Administrative Procedures

In this chapter we examine how statistical agencies implement the legal requirements described in Chapter 4 to protect the privacy of the data they collect or acquire. We also build upon the discussion from our first report regarding restricting data, restricting access, and using computer science and cryptography to protect the privacy of statistical data in the context of using multiple data sources (National Academies of Sciences, Engineering, and Medicine, 2017b, Ch. 5). We begin by drawing a distinction between two avenues of privacy breach in the context of statistical data analysis: (1) threats to the security of the raw data, and (2) the use of statistical findings, aggregations, and conclusions drawn from the confidential data to learn about an individual or organization. As in the previous chapter, we focus only on individual privacy breaches to the data providers: that is, breaches that could not have occurred if the individual's or the organization's data had not been in the confidential dataset. Inferences based on facts about the population as a whole are not viewed as individual privacy breaches.

For each of the two avenues, we describe specific vulnerabilities and some of the technologies that address or mitigate these risks. We note some of the costs of privacy protection and discuss approaches for statistical agencies to begin to evaluate new technologies and methods for protecting privacy. It is important to note that the technologies and methods have different purposes in protecting data. In addition, some techniques have been used for a longer time in statistical agencies than others so their uses are at least partly understood; others are not in widespread use.

## TWO AVENUES TO A BREACH OF PRIVACY

There are two distinct avenues to an individual privacy breach: security threats and inferential disclosure. Security threats comprise most usual concerns and occur in many ways. Most simply, there may be a data breach (e.g., someone breaks in or eavesdrops) or data loss (e.g., a laptop containing data is left on a train). Other, more deliberate, ways can involve data integrity problems (e.g., an outsider or an insider has tampered with the data, deleted it, or added spurious records) or bribery (e.g., an insider with authorized access sells data access or sells a copy of the data). There are also a variety of side-channel attacks, such as determining which data are accessed, and potentially the values of these data, based on physical measurements—acoustical properties, power consumption—of the computing device and the time used to carry out certain steps of the computation. All of these are threats to the raw data, the individual records.

The second avenue to privacy breach is more subtle. As noted in Chapter 4, this avenue involves being able to learn something about an individual in a dataset. This kind of breach is often referred to as *inferential disclosure* in the statistics literature, but this term is sometimes used in a very different sense. To avoid confusion, we use the term *individual privacy breach*. The concept is easily explained through an example. Suppose statistical analysis of a medical dataset shows that smoking causes cancer. The scientific conclusion of the study allows one to infer that any smoker is at increased risk of cancer. This is true whether or not the data of a particular smoker are in the medical dataset. So far, this is not an individual privacy breach—learning that smoking causes cancer was the point of the study. An individual privacy breach occurs when one can learn something about a participant in the study that cannot be learned about someone not in the study.

For example, if the holders of the medical data were to release poorly "anonymized" individual medical records of the study participants and if it was possible to determine from the records, say, information about the date in which someone entered the study, it might be possible to learn specific details about the person's medical history by tracing individual records back to the corresponding participants or through more mathematically subtle means. Surprisingly, individual privacy breaches can occur even when only simple statistics about the study participants are made public. Attacks on individual privacy can occur by combining the legitimate and correct outputs of a statistical analysis with auxiliary information from other sources—such as other studies (including replications), information people publish on their websites, previous measurements for the same area or population, product recommendations, blogs, social networking sites, and information one has about colleagues and peers—to learn facts about individual data providers that could not have been deduced had the

individual not contributed data. In other words, *by definition*, only data providers are at risk of inferential threats of privacy breaches. It is important to note that these individual privacy threats persist even if all security threats are eliminated.

## Security Threats

### Securing Data

Many security threats can be addressed by maintaining data in a secure form—partitioned or encrypted—which includes protection against data loss. Data loss can occur for many reasons. One is due to physical loss of the storage medium because of fire or flooding. To mitigate this danger, all data of value need to be backed up or replicated. Another reason for data loss is due to tampering, such as an unauthorized update being applied. Replication may be helpful in this case because an intruder is forced to tamper with multiple copies to install the update. However, computer security methods are much more likely to prevent tampering and intrusions in the first place.

Security threats are also addressed by access protocols and technologies, such as multifactor authentication, to prevent unauthorized access to data. As described in Chapter 4, datasets may be protected by different laws, and access needs to be provided to only those people authorized to use the data, who are typically sworn under penalty of law not to disclose the information or use it for anything other than a statistical purpose. Administrative procedures, such as access lists, are used to ensure that only authorized users are permitted to log into particular sites and access particular data repositories. That is, specific users or Internet protocol addresses are safelisted (if allowed access) or blacklisted (if forbidden). Alternatively, rather than authorizing users, roles can be authorized, that is, access is controlled through a role-based access-control paradigm. In this approach, users having a given role are provided with access authorized for that particular role. However, it might be necessary to add a role-based context-constrained access control to guarantee that external constraints are additionally imposed on the data access.[1] That is, while a role-based access-control paradigm provides capabilities based on roles, a role-based context-constrained access control accounts for additional context constraints, potentially externally extracted, imposed on the given role. For example, a public employee (a role) may be able to access a range of sensitive government information about individuals to perform her job, but she

[1]This constraint is sometimes referred to as a context primitive.

would be prohibited from accessing such information on family members (contextual constraint).

Unfortunately, data breaches still may occur, and so efforts have to be made to minimize their effects. Often, these breaches could be due to guessing or stealing a password from an authorized user rather than through a successful attack on a system's defenses. As such, systems need to be designed to minimize the harm caused by any breach. Best practices in this regard include storing the data in encrypted form, so that an additional decryption step has to be performed by anyone who manages to get unauthorized access to the encrypted data. One such practice is storing the data in distinct partitions so that any one breach exposes only one partition rather than all the data. Another practice is breaking data into "shares" and storing the shares separately, so that reconstructing even a single data element requires obtaining most or even all of the shares.

Ideally, data are partitioned so that what is in any one partition is not particularly sensitive even if a combination of partitions contains very sensitive data. A simple example is to have an identification (ID) to name and address mapping in one partition and an ID to income mapping in a second partition. Only by putting the two partitions together could an intruder learn the mapping between names and incomes.

In the case of statistics generated from multiple sources, it is possible that joined data across sources are far more sensitive than data in any one source. But the joined data may only be an ephemeral intermediate result, which is aggregated down to statistical products that can be made public. In such situations, it may be desirable to avoid ever storing the full granular joined data. Technology for secure multiparty computing could, in some situations, even permit a statistical agency to compute the desired aggregate result without ever actually learning all the detailed data in each of the data sources.

### Securing Computation

As we describe in Chapter 3, statistical analyses using multiple data sources can be accomplished in a wide variety of computing architectures. All (and, if possible, only) the data required for a given analysis or task could be copied to a single location ("centralized") and retained only while needed for the statistical computations. This is an example of the general concept of principle of data minimization: security risks are mitigated both by constraining the scope of the collected information and the duration for which it is held. Indefinitely housing multiple data sources together exacerbates the risks of disclosure—there is more time for an adversary to find a successful attack—and it also presents a richer, and therefore more tempting, target.

At the opposite end of the spectrum, computation on multiple data sources can be decentralized: the data never leave their individual original locations, and the multiple locations engage in a cryptographic protocol to cooperatively compute the outcome of the statistical computations.[2] The protocol ensures that no party to the analysis (i.e., none of the individual sources) learns about the data held by the other parties, other than what can be inferred from the output and the party's own input. This kind of secrecy guarantee is easily illustrated with a simple example: suppose Alice and Bob wish to determine who read more books in the previous week, but they don't wish to reveal how many books they read. Suppose further that both will engage in a cryptographic protocol for achieving this comparison without cheating—they are honest but curious (Goldreich, 2004). Say that Alice read seven books, and Bob read five. The output of the protocol simply says "Alice," since Alice read more books than Bob. From this output, Bob can infer a little extra: he knows that Alice read at least six books, but he cannot determine whether Alice read seven books (reality) or 17 books. Bob has learned no more than what can be determined from the output of the protocol ("Alice") and his own input (five). This example also shows that what Bob can infer about Alice from the output of the protocol depends also on Bob's own inputs.

Cryptographic protocols of this type are instances of secure multiparty computation, also known as secure function evaluation. To study or implement secure multiparty computation, one needs to consider three factors:

1. What types of bad behavior does the protocol need to protect against? Are the parties honest but curious? Might the parties deviate from the protocol, either deliberately or if the host machine is invaded by a virus?
2. What fraction of the parties would need to collude to circumvent the protections of the protocol?
3. How many (total) parties are participating in the protocol?

The past decade has seen tremendous progress in secure multiparty computation for the two-party, honest but curious (also known as semi-honest) case (see, e.g., the optimizations noted in Gueron et al., 2015). Secure versions of computations whose nonprivate versions require upwards of a trillion Boolean gates (the simplest computational element in a digital computer) are now routine. In the future, it may become practical for a physical

[2]A cryptographic protocol permits two or more parties to cooperatively make decisions based on, or learn specifics about, the combination of their individual information, without explicitly revealing their information to one another.

data warehouse to be replaced by a virtual data warehouse through secure multiparty computation and other advanced cryptographic technology.

Whether in a centralized setting or at each individual data source, outsider threats should be reduced by encrypting data both at rest and in transit. Other cryptographic techniques are also available for ensuring that stored data have not been tampered with (memory checking).

In the future, it is expected that advanced cryptographic techniques, for which proofs of concept already exist, will permit computation on encrypted data without the need to decrypt (Gentry, 2009; also see Canetti et al., 2017). Such technology can mitigate many of the concerns with centralization without the complexity of secure multiparty computation. In a related vein, proofs of concept also exist for encrypting data so as to permit only the results of certain prespecified computations to be decrypted without a special key (Boneh et al., 2010). Although they are not yet mature, these emerging technologies would be powerful enablers of securely combining multiple data sources in both centralized and decentralized settings. Using current statistical techniques, analysts often examine data, for example, to identify outliers and data errors. If and when these processes can be described in algorithmic terms, they too could be carried out using advanced cryptographic techniques. Other analytical goals, such as assessing whether or not there are sufficient numbers of cases of interest to support conclusions, can already be achieved using current cryptographic techniques.

Modern cryptography is founded on the principle that certain kinds of computational problems are too complex to be carried out in practice. For example, encryption systems require fast algorithms for encrypting data and fast algorithms for decrypting encrypted messages with the help of the secret decryption key, but the systems also require that there is no practical algorithm that would allow someone intercepting the encrypted message to decrypt it *without* the secret key. In other words, decrypting (without the secret decryption key) is computationally hard. The famous RSA cryptosystem,[3] for example, rests on the assumption that factoring a number, $n = pq$, where $p$ and $q$ are large primes, is computationally hard.

In cryptography, algorithms (e.g., for encryption) are often modeled as "black boxes" whose internal states and inputs—such as secret keys and (intentionally) random bits—are not visible to an adversary. Algorithms can then be proven to be secure under various assumptions about the hardness of certain computations, such as factoring. This is not the end of the story, however. The algorithms are often run in poorly secured settings, and "side-channel attacks" can exploit the physical properties of the

[3]A system in which the encryption key is public and different from the decryption key, which is kept private. It is named after the three authors who first described it.

devices on which the cryptographic algorithms are running. The physical implementation might enable "leakage" of the internal state of the algorithm, which is defined as observations and measurements of the internal characteristics of a cryptographic algorithm, including random choices made by the algorithm and secret keys, such as a decryption key for an encryption algorithm. For example, the RSA encryption algorithm requires a series of operations that depend on the individual bits in the secret key, but the operation associated with "0" is faster than the operation associated with "1." There are also differences in the levels of power consumed by the two operations. Attacks exploiting these differences can and have broken systems with a mathematical security proof without violating any of the underlying mathematical principles; for timing attacks, see Kocher (1996); for differential power analysis, see Kocher et al. (1998); for acoustic cryptanalysis, see Genkin et al. (2013). However, there is leakage-resilient cryptography that defends against such attacks; see Akavia et al. (2009) and Goldwasser and Rothblum (2010).

Other security threats arise if algorithms are run on untrusted servers or in the cloud. For example, one may have a large amount of sensitive data, such as genomic data, stored on the untrusted server. Even if the data are encrypted, the server can observe data access patterns, potentially allowing the server to infer which medical test is being applied to the genomic data. Or, consider a program that tests a given location in the DNA and branches according to whether the value is a "C" or a "T." That is, the program examines the data to see if it is a "C" or a "T" and, according to the output, proceeds with one of two possible computations. Suppose further that these two computations access data stored in different parts of the computer memory. Then, an untrusted server may be able to infer from the program's data access pattern which of the two possible computations was carried out, allowing inference of the correct protein, "C" or "T." These kinds of threats are addressed in the oblivious random access machine literature (Goldreich et al., 1987; Wang et al., 2015).

Even if a server is trusted, for example, in certain cloud computing environments, information can pass from one application to another running on the same server in what is known as a cross-VM (virtual machine) attack. A VM operates as a complete computer system, providing the complete functionality of a computer. Virtualization permits unrelated programs to be run simultaneously on a single computer, allowing multiple computing environments to behave as if each is run in isolation. Services, such as Microsoft's Azure, Google's Compute Engine, and Amazon's EC2, allow users to instantiate VMs on demand. As a result, VMs instantiated by distinct cloud customers share a physical infrastructure. In a cross-VM attack, a malicious program on one VM can learn information about co-resident instances. For example, (time-shared) caches allow an attacker to measure

when other instances are experiencing computational load. Leaking such information is less innocuous than it appears, permitting "co-residence detection (agnostic to network configuration), surreptitious detection of the rate of web traffic a co-resident site receives, and even timing keystrokes by an honest user (via SSH) of a co-resident instance" (Ristenpart et al., 2009, p. 9).

Risks from "inside jobs" are also problematic, although these too can be mitigated, for example, with untamperable and auditable logs and possibly with cryptographic copyright protection methods to discourage selling the data or access to the data.

Of the two classes of risks—security and individual privacy—the former is better understood. Altman and colleagues (2015, p. 28) note that the formal guidance agencies are given for analyzing and mitigating related information security risks is "voluminous, proscriptive, specific, actionable, frequently updated, and integrative into systems of legal audit and certification." In contrast, the guidance for identifying and mitigating individual privacy risks is "general, abstract, infrequently updated, and self-directed," which can often lead to "inconsistent identification of privacy risks and ineffective application of privacy safeguards" (p. 28).

Modern cryptography has studied aspects of security problems in detail, starting in the late 1970s (see Diffie and Hellman, 1976; Rivest et al., 1978) with a rigorous theory beginning in the early 1980s (Goldwasser and Micali, 1982; Goldwasser et al., 1988). The science of inferential disclosure risk dates back to the 1970s (see Fellegi, 1972), but a rigorous theory of privacy loss began only in the mid-2000s (Dwork, 2006; Dwork et al., 2006). The widespread availability of auxiliary information in the Internet age, together with the availability of vast computational power even in a single laptop, completely transformed the landscape of security and privacy, giving substance to what had previously been considered merely theoretical threats.

In summary, if one pictures a system for statistical analysis of confidential data as having the data and all computing devices operating on these data in a vault, with only the results of the analyses emanating from the vault, then security threats are the moral equivalent to breaking into the vault. These threats range from the physical security of the data—from fire, floods, theft of storage devices—through more esoteric information channels, such as analysis of very fine-grained time and power analysis consumption profiles of computations and data accesses. Security threats can often be addressed by appropriate use of back-ups and cryptographic techniques.

## Inferential Disclosure: Threats to Individual Privacy

Attacks on individual privacy use legitimately obtained statistical findings, aggregations, and conclusions drawn from confidential data to obtain information about an individual or organization. As in Chapter 4, we focus here only on individual privacy breaches to the data providers—that is, breaches that could not have occurred had the individual's or organization's data not been in the confidential dataset. Inferences based on facts about the population as a whole are not viewed as individual privacy breaches.

Different adversarial goals may require different resources, so to fully specify an attack, one also has to specify the resources to which the adversary has access. Examples of resources include computational capabilities and additional, or auxiliary, information that cannot be obtained by interacting with the dataset. Examples of useful auxiliary information might be personal details about an individual known, for example, to a sibling or coworker, such as the approximate dates on which one has watched a few movies on Netflix (Narayanan and Shmatikov, 2008) and blog entries describing a favorite recent purchase (Calandrino et al., 2011).

### Re-identification and De-anonymization

The technical literature and popular press frequently refer to re-identifying data. Such references refer to an approach to privacy protection in which individual data records, containing explicit identifying information, have presumably been de-identified or anonymized, and re-identification refers to reversing this step, tracing an individual record back to its human source. Thus, in a re-identification attack, the goal of the adversary is to trace an individual record to its source. While re-identification may seem difficult for those who do not have access to the underlying data, that is not the case: anyone looking at supposedly de-identified data who is also in possession of auxiliary information about a member of the dataset may well be in a position to re-identify. One example is linking public records, such as voter registration rolls, with de-identified data (Sweeney, 1997; Narayanan and Shmatikov, 2008). Indeed, the richer the dataset, the greater the set of possibilities for useful auxiliary information, and a host of results suggest that de-identified data are either not de-identified or no longer can serve as data. In the words of the President's Council of Advisors on Science and Technology (2014, p. 38):

> Anonymization of a data record might seem easy to implement. Unfortunately, it is increasingly easy to defeat anonymization by the very techniques that are being developed for many legitimate applications of big data.

We now turn to privacy attacks that can be launched against collections of statistics. We formalize the computational model for interacting with a dataset. In this model, a data analyst can be viewed as the adversary, for several reasons. First, this approach fosters the democratization of statistical research, in theory permitting anyone—a journalist, a concerned citizen, a lawyer in a class action lawsuit—to analyze data. Thus, the computational model does not need to understand the motives of those accessing the data and how their different goals might interact. Second, it recognizes that even the best intentioned data analysts may fail to understand the privacy risks inherent in their own statistical analyses. Third, the approach captures the fact that the choice of measurements selected that is based on exploratory data can result in a privacy breach.[4] For example, suppose the dataset contains detailed demographic information about individuals and their charitable contributions. Suppose further that a member of a specific demographic group is an extremely generous outlier. The analyst, intrigued by the data of the outlier, may choose to examine the correlation between membership in this specific demographic group and charitable donations. Thus, the choice of measurement (correlation between group membership and charitable donation) depends on the data of a specific individual in the dataset, which has the potential to result in an individual privacy breach when combined with other sources of information. Finally, data analysis results in observable actions, such as the publication of statistics and technical papers describing the findings. Taken together, statistics and findings obtained by multiple perfectly trustworthy data analysts can, in combination, compromise privacy.

The problem, then, is how to permit analysts to carry out their work while at the same time understanding that they may be (intentionally or otherwise) a privacy adversary. One solution is for raw data to be hidden from the data analyst, whose access is restricted to repeatedly posing a "query" and receiving a possibly noisy response. Here the word "query" means a function, an algorithm, or a statistical estimation that takes data as input and produces a numerical or categorical output. The noise comes from randomness introduced by the query-answering mechanism, which is introduced in order to protect privacy. Sometimes, the data themselves are altered prior to analysis. This is a generalization of the well-known randomized response technique developed in the 1960s to facilitate surveying of embarrassing or even illegal behavior (Warner, 1965; Erlingsson et al., 2014; Apple, 2016). In this approach, the query-answering mechanism receives data for which privacy has already been "rolled in" through the

[4]Exploratory data analysis, in which the questions asked depend on the data themselves, has other problems, such as overfitting a model to the dataset, meaning that the model does not generalize to the population from which the data were drawn.

randomization process. Often, the data remain unaffected prior to the analysis: that is, the query-answering mechanism has access to the raw, unadulterated data. The queries may be specified ahead of time, for example, when a government agency decides on a set of tables to release and the statistical estimates themselves have added noise. Alternatively, given a specified set of quantities of the dataset to be revealed—such as means and variances—while preserving privacy, synthetic data may be generated whose approximate means and variances (and other prespecified quantities) closely match those of the original dataset. That is, a synthetic dataset can be constructed from existing records on the basis of statistical models that induce noise in statistics relative to those from the original data.

We next turn to queries of the form, "What fraction of the records in the dataset satisfy property *P*?" These are called fractional counting queries or statistical queries. For example, in a database containing information about people's height, weight, and age, a statistical query might ask, "What fraction of the members of the dataset are over 6 feet tall?" The mechanism might compute the true answer to the query and produce as output the sum of the true answer and some random noise. Thus, the mechanisms need only to provide approximate answers to these queries. This approximation is a source of "inaccuracy." Of course, there are many sources of inaccuracy in statistics based on sample surveys, such as sampling error. Probability sampling techniques provide the ability to measure the nature and extent of uncertainty in inferences to the full population. So too with the privacy-enhancing statistical techniques described here. When the inaccuracies are systematic and well behaved, they can be addressed with statistical techniques. This is an important point: many privacy-enhancing techniques introduce errors to protect individual privacy. When this is done using ad hoc and secret techniques, the data analyst cannot compensate and cannot, for example, form correct confidence intervals. In contrast, when the randomization is done using publicly known techniques, a statistician can understand the extent of uncertainties in estimates from the data.

Deliberately introducing inaccuracies is necessary because failure to do so while providing estimates from statistical analyses results in vulnerability to several kinds of attacks on individual privacy. We next describe two of these: reconstruction and tracing attacks.

### Reconstruction Attacks

Suppose each data record contains a good deal of nonprivate identifying information and a secret binary digit (bit), one per individual, for example, indicating whether or not the individual has one of the genes associated with Alzheimer's disease. The goal in a reconstruction attack is to determine the secret bits for nearly everyone in the dataset. Reconstruc-

tion attacks succeed, for example, if an adversary can reconstruct the secret bits for all but 1 percent of the records in the dataset.

Reconstruction attacks can be launched against a subset of the rows of a dataset, for example, on the members of an extended family (e.g., parents, children, grandparents, cousins) by proper formulation of the query, such as: "Among the rows in the dataset corresponding to members of the extended family F, what fraction correspond to people over 6 feet tall?"

There is now conclusive research showing that any mechanism providing overly accurate answers to too many counting queries (like the computation of related percentages) succumbs to a reconstruction attack. This collection of results is known informally as the fundamental law of information recovery (or reconstruction). In fact, there is a single attack strategy that succeeds against all such overly accurate answering of too many queries. Here, "too many" is quite small: only $n$ queries are required, where $n$ is the size of the set under attack (which can be as large as the whole dataset or as small as the extended family F). "Overly accurate" means having error on the order of the square root of $N$, the size of the set under attack.[5] In fact, when $n$ is very small, it is possible to launch a simple attack requiring $2^n$ queries; in this case, the attack works even if the noise is on the order of $n$ itself.[6] For example, 11 billion estimates are produced from the American Community Survey; it is worth considering the possibility that these estimates can be used to carry out a reconstruction attack on some portion of the respondents.

### Tracing Attacks

Reconstruction represents success on the part of the adversary, or, conversely, failure of a privacy mechanism. Tracing—that is, determining whether or not a specific individual is a member of a given dataset—is a much more modest adversarial goal: there are settings in which tracing attacks are possible, but reconstruction attacks are provably impossible. Tracing entered the public consciousness when a group of genomics researchers showed how to use allele frequency statistics in a genome-wide association study, together with the DNA of a target individual and allele frequency statistics for the general population (or a control group), to determine the target's presence or absence in the study (Homer et al., 2008). In response, the National Institutes of Health and the Wellcome Trust changed the access policy to statistics of this type in the studies they

[5]One can think of this as errors on the scale of the sampling error in an opinion poll. An attack requires polling large and overlapping subgroups of the dataset and combining the results in clever fashion.

[6]For a survey of reconstruction results, see Dwork et al. (2017).

fund (see Dwork et al., 2017, for a survey of tracing results; also see further discussion below).

## INFERENCE CONTROL TECHNIQUES

Previous National Research Council studies (2005, 2007) have examined techniques and approaches agencies have used for protecting the confidentiality of data while providing access to researchers, and these are briefly described in the panel's first report (National Academies of Sciences, Engineering, and Medicine, 2017b). In this section, we elaborate on that discussion with additional perspectives and frameworks for protecting data as a foundation for our discussion of implications and next steps in the final section of this chapter. Specifically, we review the major approaches that statistical agencies have used, including applying statistical disclosure limitation methods and perturbing the data and creating synthetic datasets for analysis, as well as noting the weaknesses of these approaches. We also review different approaches that limit access to data depending on its sensitivity and that require researchers to go to data enclaves in order to analyze the data. We conclude the section with a discussion and review of differential privacy and its potential to meet the needs of statistical agencies.

Statistical agencies and their contractors currently use a variety of statistical disclosure control techniques to attempt to protect the confidentiality of respondents' information (for a summary, see Federal Committee on Statistical Methodology, 2006). Many of them are combinations of techniques. Typical methods used in statistical disclosure control (see Hunderpool et al., 2012; Karr and Reiter, 2014; Skinner, 2009) include

- only releasing information at high levels of aggregation;
- top- or bottom-coding data, by reporting an income of say, $150,000 for all respondents with incomes higher than that figure;
- swapping data, in which variables from pairs of different records are interchanged;
- adding noise to individuals' responses or to computed statistics; and
- creating synthetic data, in which models fit to the real data are used to generate artificial data that are then released.

Many statistical agencies or their contractors also have disclosure review boards, which review data products—including tables, reports, and microdata—prior to release to ensure that the releases do not reveal any information about respondents. The Census Bureau provides instructions to

program managers and a checklist to complete and submit to its disclosure review board.[7]

### De-identification Through Perturbing the Original Data

"De-identification" typically refers to a situation in which there is a collection of records, each essentially belonging to or being about a specific individual. In this technique, certain data fields in an individual record are eliminated, coarsened, or otherwise modified, leaving the remaining fields untouched (or perhaps more lightly modified). De-identification does not prevent linkage attacks, in which a second database is used as auxiliary information to compromise privacy in the first database. This type of attack is at the heart of the vast literature on hiding small cell counts in tabular data (cell suppression). Naïve de-identification resulted in perhaps the world's most famous linkage attack, in which the medical records of the governor of Massachusetts were re-identified by linking voter registration records to anonymized Massachusetts Group Insurance Commission medical encounter data, which retained the date of birth, gender, and ZIP code of the patient (Sweeney, 1997). Sweeney proposed an antidote, known as *k*-anonymity. In *k*-anonymity, a syntactic condition requires that every "quasi-identifier" (essentially, a combination of nonsensitive attributes) must appear at least *k* times in the published database if it occurs at all. This can be achieved by coarsening attribute categories, for example, replacing 5-digit ZIP codes with their 3-digit prefixes. In addition to potentially being hard to compute (in the sense discussed above), the choice of category coarsenings may reveal information about the database. That is, had the data been different, the coarsenings chosen would have been different, thus the choice itself may constitute an individual privacy breach. Perhaps more concerning is the fact that if a given individual is known to be in the dataset, then *k*-anonymity permits an attacker to narrow down to a set of the size of no more than *k* the set of records possibly corresponding to the individual. If, for example, all *k* of these records share a diagnosis, then the individual necessarily received this diagnosis. Difficulties of this sort led to a series of works and additional syntactic constraints on anonymized datasets (Machanavajjhala et al. [2006] on *l*-diversity; Xiao and Tao [2007] on *m*-invariance; Li et al. [2007] on *t*-closeness).

In general, syntactic conditions may fail to capture the semantics of privacy: that is, suppressing and aggregating various data fields is not formally tied with a mathematical definition of privacy. The methodological problem with syntactic considerations is understanding when these syntactic condi-

[7]See https://www.census.gov/srd/sdc/wendy.drb.faq.pdf and https://www.census.gov/srd/sdc/drbchecklist51313.docx [August 2017].

tions protect against individual privacy breaches. It is also important to note that the damage done to the data in the anonymization process can completely destroy utility (Brickell and Shmatikov, 2008).

### Synthetic Data

Using synthetic data is an approach to privacy protection in which a model is first specified and estimated from the data and then fictitious samples are generated according to the model (Reiter and Raghunathan, 2007; Rubin, 1993).[8] For example, if one assumes the data are drawn from a normal distribution, one learns from the data a mean $\mu$ and a variance $\sigma^2$ and then generates random samples drawn from the normal distribution $N(\mu,\sigma^2)$. These random samples are assembled into a dataset that is publicly released. The Census Bureau has released synthetic datasets for the Survey of Income and Program Participation linked to Social Security income records (Benedetto et al., 2013) and the Longitudinal Business Database (Kinney et al., 2011).

Although at first glance this approach appears to protect privacy—after all, only random noise is released—the parameters $\mu$ and $\sigma^2$ themselves leak information about the original dataset. Synthetic data do not protect privacy merely by virtue of being synthetic; the process by which they are generated must also be protective.[9] Another limitation is that synthetic data can tell no more about a dataset than can be encapsulated by the model. The model provides assurance that a class of models, which are distinct from those used in generating the synthetic datasets, can be estimated with expected statistical error properties. Furthermore, in order to obtain estimates of the added uncertainty of statistical values arising from the synthetic generation of data, multiple synthetic datasets are often generated, complicating the statistical estimation and, potentially, increasing risk of privacy loss.

### Weaknesses of Statistical Disclosure Limitation Methods

Although the methods discussed may protect the data against some attacks, they cannot be guaranteed to protect the privacy of respondents, and they all negatively affect the utility of the data for researchers (see, e.g., Reiter, 2012). As argued above, aggregated tables can be assembled

[8]Synthetic data is a general term. The model-based, multiple imputation approach of Rubin (1993) and others is one way to create synthetic data; other approaches can be found in the differential privacy literature.

[9]See, for example, https://obssr.od.nih.gov/synthetic-data-protecting-data-privacy-in-an-era-of-big-data/ [August 2017].

to provide information about ever-smaller segments of the population. As noted above, 11 billion estimates are published every year from the American Community Survey, with the data sliced and diced on multiple dimensions for many different geographic areas (see Hedrick and Weister, 2016). An adversary can assemble different statistics to learn about individuals in the dataset. Data swapping can distort multivariate relationships in the data and result in unusual patterns, for example, if the income of a fast food worker is swapped with that of a neurosurgeon (although in practice restrictions would be placed on preserving certain univariate and multivariate relationships while swapping). Synthetic data convey only the information in the model used to generate them, and that information can be disclosive. In addition, synthetic data cannot be used to discover any information not in the model, and it might therefore be questioned why one would not just publish the model.

### Approaches for Tiered Access

In addition to the methods described above, the other major approach statistical agencies use to protect the privacy of data is to restrict access to and use of the data with legally binding contracts and administrative procedures. Given the different levels of legal protection for different data (see Chapter 4), there have also been efforts to tailor the protections for a particular dataset to the sensitivity of the information and the potential harms that could result from disclosure. Sweeney and her colleagues (2015) have "tagged" data with different grades of security measures needed to access the data (see Box 5-1). By using a tiered access system, data are able to be matched with appropriate privacy control techniques. By using tiered access systems, organizations are able to grant different levels of access based upon the merit of individuals using the data for research.

Altman and colleagues (2015) elaborate on several privacy controls, including procedural, technical, educational, economic, and legal means, and they note how these can be applied throughout the information life cycle of collection/acceptance, transformation, retention, access and release, and post-access (see Box 5-2 and Table 5-1). Their approach to tiered access is to evaluate identifiability against potential harms from uncontrolled use of the data (see Figure 5-1). In the lowest risk category, data that are not harmful or could result in minor harm would be available for public use. The next tier, either data that contain identifiers but are not harmful or data that are de-identified that could result in significant harm, would be de-identified and require additional steps, such as consent and terms of service. The final two tiers cover data that could result in major harms and would work in similar ways, involving formal application and data use

**BOX 5-1**
**Datatagging: Example**

The datatag system (Sweeney et al., 2015) was created to maximize data sharing while minimizing risk of harm from sensitive information. Through the notion of datatags, data are allowed to be accessed and handled with different security and credential measures. The system contains six different grades of security described by numbers or colors, from one (blue) to six (crimson), where one contains the least sensitive information.

- A blue datatag (one) requires no access credentials to access. The data typically do not contain any personal information. Examples include non-personal research data that has already been published and public-use files.
- A green datatag (two) is for controlled public data. These data can be accessed by the public with minimal agreement, but access often requires verification, similar to providing a valid email address.
- A yellow datatag (three) is the first level for which a data use agreement is required. This agreement often is in click-through form, and a password and approval are required in order to access data. This is also the first stage at which data are encrypted when transmitted.
- An orange datatag (four) is the first level for which a written signed agreement is required for data access. At this stage, the data are also encrypted in storage.
- A red datatag (five) usually requires two-factor authentication for access, such as a phone number and email address.
- A crimson datatag (six) is similar to a red datatag, except data are multi-encrypted in storage.

There are multiple methods to determine at what level data should be tagged. One possibility includes having an interview with a panel, such as an Institutional Review Board, which includes privacy experts. It is also possible to have automated data tagging by computer programs. It is recommended to have some sort of human-assisted approaches in combination with computer techniques, however, to tag data. Sweeney and colleagues (2015) draw a comparison with TurboTax, in which a user goes through a series of "yes" or "no" questions for most determinations while experts are available to address questions.

**BOX 5-2**
**Privacy Control Techniques: Example**

In trying to create a new privacy framework for government data, Altman and colleagues (2015) delineated five different privacy controls that are available to policy makers: procedural, technical, educational, economic, and legal. These privacy controls can also be used in conjunction with five stages of data use: collection and acceptance, transformation, retention, access and release, and post-access (see Table 5-1 for more information).

1. Procedural controls involve internal procedures, such as implementation notices, creating inventories, or vetting internal or external access to an organization's database. Across data life cycles, examples of procedural means include data minimization and access controls.
2. Technical controls are defined as statistical methods, computational methods, and human factors analysis. Examples include synthetic data, encryption, and differential privacy.
3. Educational controls include providing information to data subjects, collectors, controllers, and recipients about privacy practices and risks. Techniques that can be used for education include notices and transparency.
4. Economic controls include any intervention intended to change the economic incentives of stakeholders, such as fees to access data, provision of insurance, and property rights agreements.
5. Legal controls ensure data are properly protected through legal rights among stakeholders. Legal means can include data rights, terms of service, and civil or criminal punishment.

agreements. The most sensitive data would be held in enclaves and require strict data logs, as major harms could result if the person was identified.

### Data Enclaves

Federal statistical agencies provide access to some microdata files through the Federal Statistical Research Data Centers (FSRDCs).[10] In these controlled environments, researchers can conduct statistical analysis on specific datasets for their approved projects. Prior to being permitted to use an FSRDC, researchers are required to go through screening and training on preserving the confidentiality of the data and are sworn to protect the data, with criminal penalties for disclosure. Researchers seeking to link external data to data in the FSRDC must provide the data to the Census Bureau,

[10]See https://www.census.gov/fsrdc [August 2017].

which does the linkage. No data are brought into or out of the FSRDC itself because the hardware used is a thin client, which is built for remote access to a server, with a virtual private network. Researchers can use a variety of statistical software, but procedures that would permit viewing individual records are disabled. No results are permitted to be taken outside of an FSRDC until they have gone through disclosure review. Similar procedures are used at other data enclaves (see Box 5-3). There have been no known security breaches of these secure enclaves; however, the limited access can make it difficult to replicate the research from their use (Altman et al., 2015).

### Differential Privacy

Differential privacy is a definition of privacy tailored to the statistical analysis of large datasets. Differential privacy precisely captures the difference between learning about an individual in the dataset and learning about a population (as represented by a dataset) as a whole. Differentially private algorithms ensure that the only possible harms are group harms: the outcome of any analysis is almost equally likely independent of whether any individual joins, or refrains from joining, the dataset. If the same output occurs whether or not a particular person is in the dataset, then that person cannot suffer any privacy loss.

This seems paradoxical. If the behavior of the algorithm is the same even when one changes the data on which it is operating, how can one understand the results? The answer to the seeming paradox is through the careful introduction of randomness into the computation. For example, if 52 percent of the population supports a candidate and one randomly samples a large number of people in the population, one can expect that the fraction of the sample supporting the candidate will be close to 52 percent. The sampling process introduces randomness in the answer, but the answer is still very likely to be roughly the same, as long as the privacy loss parameter is relatively small.

Differential privacy is a property of an algorithm, not an output of the algorithm. Differentially private algorithms are randomized, meaning that they intentionally introduce randomness into the computation and therefore the statistical estimates produced. Randomness is used in essentially all cryptographic algorithms: for example, randomness is used in choosing secret keys for digitally signing documents; the fact that they are generated randomly for each signer is what makes them secret, preventing forgeries.

In looking at differential privacy, one also needs the concept of adjacent datasets: $x$ and $y$ are adjacent datasets if one is contained in the other and the larger one contains exactly one additional record. For example, the records in the dataset might be individual hospitalization histories: the

**TABLE 5-1** Privacy Controls over Data Life Cycle

| Life Cycle Stage | Procedural | Economic | Educational | Legal | Technical |
|---|---|---|---|---|---|
| Collection/ Acceptance | Collection limitation; data minimization; data protection officer; Institutional Review Boards; notice and consent procedures; purpose specification; privacy impact assessments | Collection fees; markets for personal data; property right assignment | Consent education; transparency; notice; nutrition labels; public education; privacy icons | Data minimization; notice and consent; purpose specification | Computable policy |
| Transformation | Process for correction | | Metadata; transparency | Right to correct or amend; safe harbor de-identification standards | Aggregate statistics; computable policy; contingency tables; data visualizations; differentially private data summaries; redaction; statistical disclosure limitation techniques; synthetic data |
| Retention | Audits; controlled backups; purpose specification; security assessments; tethering | | Data asset registries; notice; transparency | Breach reporting requirements; data retention and destruction requirements; integrity and accuracy requirements | Computable policy; encryption; key management (and secret sharing); federated databases; personal data stories |

| | | | | | |
|---|---|---|---|---|---|
| Access/Release | Access controls; consent; expert panels; individual privacy settings; presumption of openness vs. privacy; purpose specification; registration; restrictions on use by data controller; risk assessment | Access use fees (for data controller or subjects); property rights assignment | Data asset registries; notice; transparency | Integrity and accuracy requirements; data use agreements (contract with data recipient, terms of service) | Authentication; computable policy; differential privacy; encryption (including functional and homomorphic); interactive query systems; secure multiparty computation |
| Post-Access (audit, review) | Audit procedures; ethical codes; tethering | Fines | Privacy dashboard; transparency | Civil and criminal penalties; data use agreements (terms of service); private right of action | Computable policy; immutable audit logs; personal data stores |

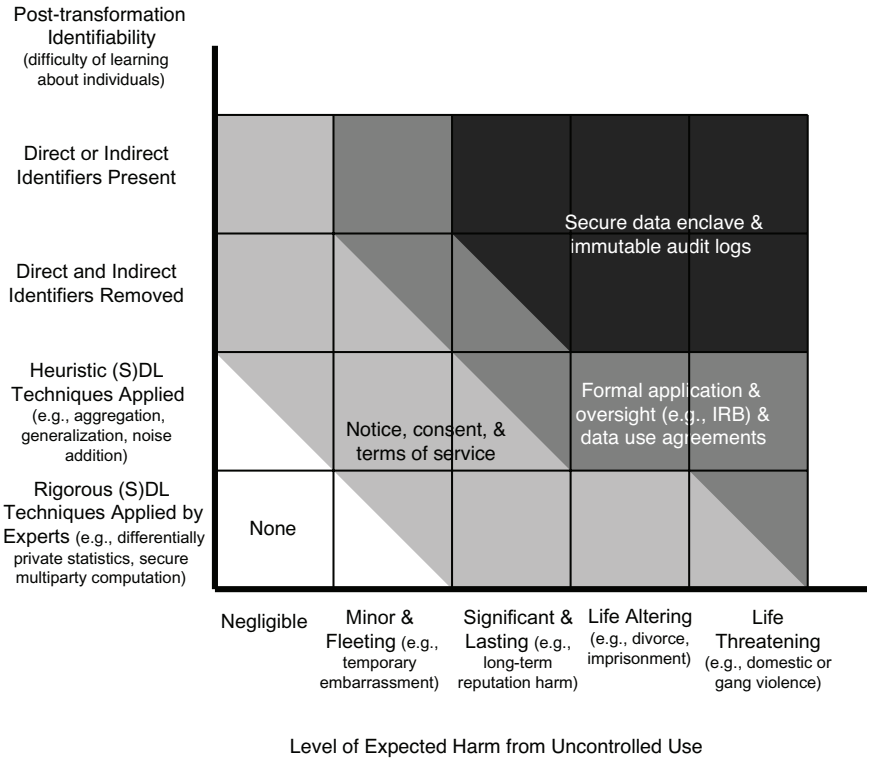


**FIGURE 5-1** Privacy controls needed given the identifiability of the data and the expected harm from uncontrolled use.
SOURCE: Adapted from Altman et al. (2015, p. 2046).

larger dataset will contain the hospitalization history of exactly one additional person. Differentially private algorithms ensure that the probability of any event is essentially equally likely when the algorithm is running on adjacent databases. "Essentially equally likely" is measured by a privacy loss parameter, usually denoted as epsilon, $\varepsilon$. Smaller values of $\varepsilon$ give better privacy protection than larger values.

Think of $\varepsilon$ as a small number, say, 0.01. Let $U$ denote the universe of theoretically possible data records. Formally, letting $M(x)$ denote the result when algorithm $M$ is run on database $x$, one can define differential privacy as follows (Dwork et al., 2006): a randomized algorithm $M{:}U^* \rightarrow \text{Range}(M)$ is $\varepsilon$-differentially private if for all adjacent datasets $x,y$ and all events $S \subseteq \text{Range}(M)$, we have $\Pr[M(x) \in S] \leq e^{\varepsilon} \Pr[M(y) \in S]$. When $\varepsilon$ is small, $e^{\varepsilon} \approx 1+ \varepsilon$. In this case, $\Pr[M(x) \in S] \leq (1 + \varepsilon) \Pr[M(y) \in S]$, meaning that the

**BOX 5-3**
**The Five Safes Framework: Example**

The five safes framework was developed by the data service in the United Kingdom and is used in many research labs, including the country's Office for National Statistics and the Administrative Data Research Network (Administrative Data Research Network, 2017). The purposes of the five safes are to have protections at different levels to minimize the risk of sensitive and confidential data being used incorrectly. The five safes are safe projects, safe people, safe settings, safe data, and safe outputs.

**Safe projects:** Researchers submit proposals that must be approved in order to begin research. A proposal must meet a public good criterion and must hold a certain amount of value in order to receive acceptance.

**Safe people:** Once a project is approved, researchers must explain why they are suitable for data access, must be affiliated with a respected academic institution, and must sign a user agreement for data use. They must complete a mandatory training course about the legal and ethical use of confidential and sensitive data, statistical disclosure control, and using the secure labs.

**Safe settings:** After training, researchers can access data only through safe settings, such as remote access through Citrix secure remote access technology, technology that is currently used by both military and banking sectors. Ethical hackers are employed in order to test the security of the servers, in order to ensure that unauthorized access to data is as limited as possible.

**Safe data:** Safe data attempts to limit disclosure risk of the data as much as possible. Due to all previous precautions, data at this stage are the most vulnerable for re-identification. Therefore, no data are able to be taken out of the safe settings.

**Safe outputs:** After using data, researchers cannot take any results out of the safe settings without statistical disclosure techniques being applied to the output to reduce the risk of re-identification.

All of the five safes aim to maximize researcher access to important data while preventing certain forms of individual privacy breaches or threats.

two probabilities are nearly equal. The definition says that the probability of $M$ observing any output is nearly equal, independent of whether the database is $x$ or $y$.

For example, suppose $x$ is a database of medical records and $y$ is the database consisting of $x$ and the data of one additional person, "George." Assume that $M$ is an algorithm for computing correlations between the pollutant in a small town and chronic bronchitis. Now, George is in $y$ but not in $x$. Thus, the output $M(x)$ is not "about" George since his data were not inputs to $M$. Note, however, that since $M$ might be about people in general, it can have implications for George. For example, it could be used to inform

a doctor of basic scientific information that will be useful in diagnosing George's respiratory difficulties. Differential privacy ensures that when the algorithm $M$ is run on database $y$ (which does contain George's medical record), essentially nothing more is learned about George than was learned when $M$ is run on database $x$ (which does not contain George).

As a further illustration, let $M$ be a 0.1-differentially private mechanism (i.e., an $\varepsilon$-differentially private mechanism where $\varepsilon$ has value 0.1) reporting the number of people in the database suffering from chronic bronchitis. Suppose we tell a gambler all of database $x$ as well as the entirety of George's medical record, and suppose further that George suffers from chronic bronchitis. This means that the true answer to the question, "How many people in the database suffer from chronic bronchitis?" is larger—by exactly 1—when the database is $y$. We then flip an unbiased coin and if it comes up heads, we run $M$ on input $x$, while if it comes up tails, we run $M$ on input $y$. Suppose $M$ outputs the number 501. If we do not give the outcome to the gambler, he has a probability of exactly 0.5 of guessing which input—$x$ or $y$—we used. If we tell him the output his probability of correctly guessing which input we used increases from 0.5 to at most 0.525. By hiding the data of each individual in this way, differentially private algorithms ensure that anything that is learned is about people in general and not about individuals in the dataset. Differentially private reporting of allele protein frequencies will prevent even someone who has obtained a sample of George's DNA from determining whether or not George is in the case group.

### Weaknesses of Differential Privacy

Differential privacy is a mathematically rigorous concept, so algorithms can be designed and proven to provide it. The approach automatically provides protection against linkage attacks, as well as present and future auxiliary information. It also provides a formal measure of privacy loss and a calculus for computing how privacy losses compound over multiple data analyses. In addition, it yields a collection of simple privacy-preserving building blocks that can be combined to yield differentially private versions of sophisticated analytical techniques from statistics and machine learning.

Nonetheless, differential privacy is not a panacea in many ways. First, neither differential privacy nor any other technique can circumvent the fundamental law of information recovery: overly accurate estimates of too many statistics can completely destroy privacy. Thus, differential privacy is not an immediate solution to the problem uncovered in the genome-wide association study discussed above; it must be combined with sophisticated techniques for false discovery rate control so as to limit the number of (noisy) statistics actually revealed. Again, there are limits for *any*

technique that provides any reasonable notion of accuracy (Dinur and Nissim, 2003; Dwork et al., 2007, 2015; Muthukrishnan and Nikolov, 2012; Kasiviswanathan et al., 2013). Differential privacy often achieves the best possible bounds, and it does provide provable guarantees. Unlike other techniques, differential privacy provides the tools needed to measure and bound the cumulative risk as the data are used and reused, permitting the informed enforcement of a privacy "budget." In other words, all systems are eventually in trouble: differential privacy allows one to monitor the trouble and decide when to stop.

Conceptually, there is no difference between multiple releases of synthetic datasets and interactive query systems (sometimes the query is simply "give me a synthetic dataset capturing the following attributes"). Nonetheless, the engineering and social challenges should not be underestimated. To adhere to a privacy budget, it will be necessary to optimize the choice of queries to be handled and to determine how to use the privacy budget as effectively as possible while ensuring fairness of access to many users of the data, not all of whom are known in advance. For example, one needs to protect against attacks intended to deplete the privacy budget just to shut down access to an agency's data. One also needs to remove side-channel attacks and other threats. All of these challenges exist for the other approaches to protecting privacy; however, there are no tools to measure them.

Second, in traditional data analysis an analyst looks at the dataset. Differential privacy limits interaction with the dataset, not allowing the analyst to see the raw data (although she may have auxiliary information about the dataset, as in our example above about George and the sets $x$ and $y$). Moreover, data analysts are not trained to operate in this scenario. They do not know about cumulative privacy loss, nor are they experienced in considering a computation to minimize privacy loss while maximizing utility.

Differential privacy also hides outliers. Indeed, it is often the outliers who most need protection. Note that many kinds of analyses make use of outliers (e.g., income distributional statistics).

Differential privacy may require larger sample sizes in order to learn things about the full population that, without privacy concerns, would require fewer sample cases.

Differential privacy intentionally introduces statistical noise. This is essential, but using currently known techniques the noise may be too much to permit useful learning unless the dataset is very large. In some kinds of analyses, differential privacy introduces exactly the minimum amount of noise needed to prevent reconstruction and tracing attacks; this means no technique exists for adding less noise in these cases. However, not all sets of analyses are so well understood. In these cases, the situation may be better than assumed: not only is it possible that better differentially private algo-

rithms exist than those currently known, but there might even be a different and meaningful definition of privacy that can achieve better accuracy than anything known. It is also possible that the situation is worse than assumed: it is possible that any algorithm that adds less noise than the existing ones could be shown to be vulnerable to some new method of statistical attack.

Differential privacy algorithms come equipped with a privacy parameter ($\varepsilon$). The meaning of this parameter is not well understood. The choice of the parameter is currently not assigned to any actor in the federal statistical system.

Differential privacy is a young field, and good algorithms—with high accuracy and low privacy loss—for many key analytical tasks are still under development. The federal statistical system produces statistics that are simple means and totals, such as sums of weighted products of variables, but also correlations, regression model coefficients, and cumulative statistics on multivariate distributions. Much remains to be developed to apply differential privacy techniques to some of these statistics.

Differential privacy does not distinguish between the types of information protected: hair color implicitly receives the same treatment as sexual identity. The privacy guarantee is framed in terms of a comparison between the behavior of the algorithm when one person (with any color hair and any sexual preference) is in the dataset compared with the behavior when this same person is absent. This may seem silly: Why go to such efforts to protect hair color? But it frees the algorithm designer from the need to know whether or not specific attributes are considered sensitive, either at all or by a specific user. For example, it protects the information about all genetic markers, even those not currently known to be correlated with, say, Alzheimer's disease. If, as has happened in the past, certain genotypes are later discovered to have such a correlation, differential privacy will have protected against that future information.

## IMPLICATIONS FOR FEDERAL STATISTICAL AGENCIES

As noted in Chapter 4, federal statistical agencies are required to follow prescriptive guidance from the U.S. Office of Management and Budget and the National Institute of Standards and Technology related to the Federal Information Security Management Act and other requirements for maintaining the security of their information systems. In contrast, agencies draw on general best practices and their staff's professional judgment in protecting their statistical data products from inferential disclosure individual priavacy breaches. Statistical agencies currently use a number of statistical disclosure methods to protect the confidentiality of their data; however, these methods are increasingly susceptible to privacy breaches given the proliferation of external data sources and the availability of high-powered computing that could enable inferences about people or entities in a dataset,

re-identification of specific people or entities, and even reconstruction of the original data. Thus, in our first report we drew the following conclusions:

> **A continuing challenge for federal statistical agencies is to produce data products that safeguard privacy. This challenge is increased by the use of multiple data sources.** (National Academies of Sciences, Engineering, and Medicine, 2017b, Conclusion 5-5, p. 90)

> **As federal statistical agencies move forward with linking multiple datasets, they must simultaneously address quantifying and controlling the risk of privacy loss.** (National Academies of Sciences, Engineering, and Medicine, 2017b, Conclusion 5-6, p. 95)

As noted above and in our first report, differential privacy provides a framework and a measure of privacy loss and tools for tracking this loss over analyses of the data, creating a "privacy loss budget." Federal statistical agencies typically seek to maximize the statistical use of their data within the constraints of ensuring the statistical products do not reveal anything about individual respondents. Agencies have only recently begun to examine the implications of a privacy loss budget for their own production of statistical products and secondary analyses by external users (see, e.g., Abowd et al., 2017). Use of these methods will require new algorithms to be developed, tested, and reviewed not only by the scientific community, but the many stakeholders for federal statistics. Therefore, the panel also made the following recommendations in its first report:

> **Statistical agencies should engage in collaborative research with academia and industry to continuously develop new techniques to address potential breaches of the confidentiality of their data.** (National Academies of Sciences, Engineering, and Medicine, 2017b, Recommendation 5-1, p. 96)

> **Federal statistical agencies should adopt modern database, cryptography, privacy-preserving, and privacy-enhancing technologies.** (National Academies of Sciences, Engineering, and Medicine, 2017b, Recommendation 5-2, p. 96)

However, the panel recognizes that the current era is one of transition: there is awareness of weaknesses of current statistical disclosure limitation methods, but the feasibility for federal statistical agencies of implementing new technologies, such as differential privacy, has not been clearly demonstrated. So far, the only application of differential privacy in federal

statistical products is the Census Bureau's OnTheMap application,[11] which implements a variant of differential privacy allowing users to see where people work and where workers live (see Machanavajjhala et al., 2008). More research and development will be needed to show how this can be done and the costs and benefits of implementation of these methods. It also will require statistical agencies to hire and train staff to use these technologies.

**RECOMMENDATION 5-1 Federal statistical agencies should ensure their technical staff receive appropriate training in modern computer science technology including but not limited to database, cryptography, privacy-preserving, and privacy-enhancing technologies.**

Overall, much work, interaction, and collaboration will be needed across the various disciplines and stakeholders as agencies seek to move forward to provide stronger privacy protection for the data they either collect from respondents or acquire access to from other administrative and private-sector sources for statistical purposes. It will be critical for there to be robust discussions of the implications of this approach for all stakeholders and these discussions will need to be informed by concrete examples to help everyone understand how use of these technologies will affect them. Pilot studies or test cases will be valuable in identifying the variety of issues that affect agencies and the users of their data, including effects on timeliness of production, the scope of statistical products produced, the utility of the resulting estimates, and the usability of microdata by external researchers.

Comparisons will need to be made using agencies' current procedures with state-of-the-art differentially private algorithms and various levels of epsilon from a variety of federal statistical datasets to evaluate the effects on accuracy of results and utility of the resulting data. Some efforts along these lines have been conducted (see, e.g., McClure and Reiter, 2012), and additional pilot studies using different federal survey datasets would be beneficial and help enhance understanding by the communities involved. More specifically, the panel thinks there is potential value in pilot studies that look retrospectively, as well as prospectively, at the current uses of agency survey datasets and how a privacy budget would have been used over time, across users, and for various activities. For example, agencies could look back at uses over the past 5 years using internal information, as well as literature searches, to assess the volume and types of analyses that have been conducted and how the use of a privacy budget could have affected internal and external analyses of the data or, conversely, the effects on privacy loss given the uses.

The panel also thinks that much could be learned by generating syn-

[11] See https://onthemap.ces.census.gov/ [August 2017].

thetic data in a differentially private manner and encouraging researchers and students to conduct their analyses on this dataset and use verification servers (see Karr and Reiter, 2014) to assess the fidelity of the results with the original dataset. Note, however, that even synthetic data are subject to the fundamental law of information recovery: if a synthetic dataset permits overly accurate estimates of too many statistics, then privacy could be destroyed. For this reason, query-response systems may deliver more accuracy for queries actually of interest than what can be achieved using synthetic datasets, at the same level (epsilon) of privacy loss.

The panel suggests the use of an epsilon registry to encourage those who work with or profit from the use of personal information to take a greater interest in the algorithms and share their experience in setting this parameter so that this work moves beyond the community of privacy researchers (Dwork and Mulligan, 2014). A typical element in the registry might describe a use of differential privacy, including the value of $\varepsilon$ used (this could be done in several ways, such as per calculation, together with the number of analyses run, or a "burn rate" of privacy loss per day or week); a discussion of the factors leading to this choice; the granularity at which differential privacy is applied (per data attribute or per individual record, which typically contains many attributes); whether or not data are "retired" when a privacy loss limit is reached; and the variant of differential privacy used.

Ultimately, the adoption of new technologies will be driven either by necessity or the broad embrace of the communities of researchers, data users, and privacy advocates. The panel hopes that agencies will engage in collaborative research programs with the academic and user communities to promote greater use and understanding of new and emerging privacy-protection methods. We also hope that workshops and other public discussions will be held on the results of the research and the policy issues associated with setting and allocating privacy budgets. These issues go beyond the purview of a single statistical agency and could have dramatic effects on the uses, users, and providers of statistical data. The implications of these issues need to be broadly discussed and a transparent and participatory process outlined for moving forward.

# 6

# Quality Frameworks for Statistics Using Multiple Data Sources

A recurrent theme in the panel's first report (National Academies of Sciences, Engineering, and Medicine, 2017b), as well as the previous chapters of this report, is that the quality of administrative and private-sector data sources needs careful examination before being used for federal statistics. The theme and the caution have been driven by the relatively recent novelty of the simultaneous use of multiple data sources and the fact that some potential new sources of data present new issues of data quality.

We begin this chapter with a discussion of quality frameworks for survey data and then briefly review additional quality features and some extensions of these frameworks for administrative and private-sector data sources. We then consider how different data sources have been or could be combined to produce federal statistics, with examples to illustrate some of the quality issues in using new data sources.

## A QUALITY FRAMEWORK FOR SURVEY RESEARCH

The science of survey research has its origins in the provision of quantitative descriptive data for the use of the state beginning in the 17th century. In general, the approach was to enumerate the whole population (censuses) to provide a description of the population to rulers and administrators. Over time, the increasing demands for information by states for both planning and administrative functions placed greater strains on the capacity of national statistical offices to provide the data, particularly in a timely manner.

The International Statistical Institute, founded in the second half of the 19th century, brought together the chief statistical officers of developed countries, including some leading academic scholars. At its convention in 1895, the chief statistician of Norway, Anders Kiaer, proposed a radical innovation—to collect information on only a subset of the population, a representative sample. Over a period of 30 years, statisticians refined the ideas he presented and developed the form of the sample survey that remained the foundation of the collection of policy-related statistical data throughout the 20th century.

Beginning in the 1930s, statisticians worked on identifying sources of error in survey estimates, and when possible, measuring their effects. All these components are combined in the total survey error framework shown in Figure 6-1. The basic inferential structure of the sample survey involves two separate processes. The first is a measurement inference, in which the questions answered or items sought from a sample unit are viewed as proxies for the true underlying phenomenon of interest. For example, every month interviewers asked a sample person: "what were you doing

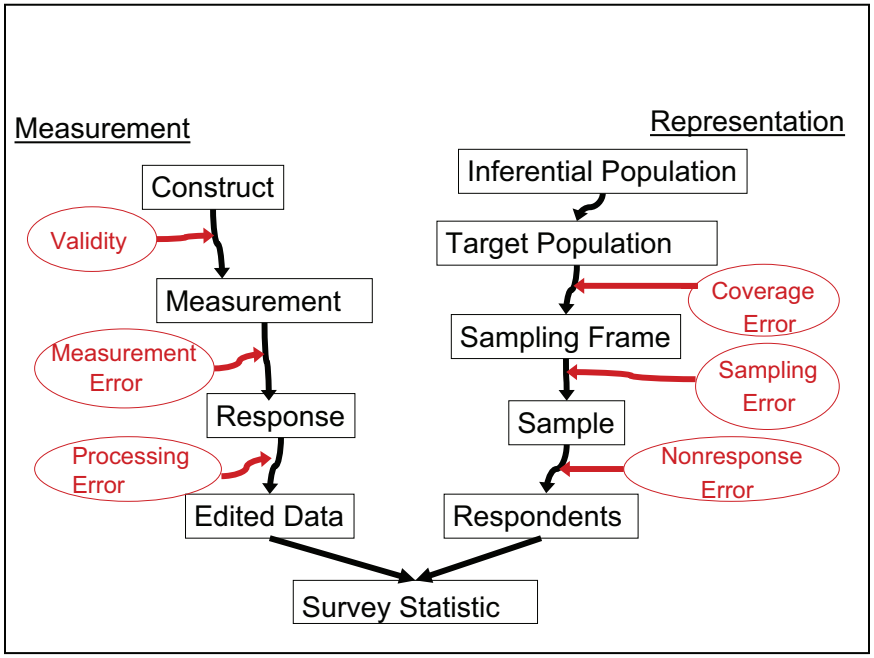


**FIGURE 6-1** Total survey error framework.
SOURCE: Adapted from Groves et al. (2009, p. 48).

last week (the week of the 12th)?" as a key item for the statistic that is labeled as the monthly employment rate. There is a difference between the measurement of behavior for the week of the 12th and the concept of an unemployment rate for the whole month. The inference being made is complicated and threatened in months in which labor strikes occur in the middle of the month as the status of that week is unlike that of other weeks in some other way.

Another example, and one that provides a clearer notion of measurement inference, can be seen when the underlying concept is clearly unobservable without a self-report. Various attributes of human knowledge fit this case well. For example, the Trends in International Mathematics and Science Study (TIMSS) assesses mathematics knowledge among students in grades 4 and 8. Knowledge of mathematics is often measured through problems that can be solved only by those people with the requisite knowledge. There are many types of mathematics problems and thousands of instances of every type of mathematical knowledge. In some cases, there are an infinite number of ways to measure a type of mathematics knowledge. Each problem, therefore, might be viewed as one sample from that infinite population. And since each problem exhibits its own variability over conceptually repeated administrations, measuring mathematics knowledge with multiple problems increases the stability of an estimate. In this case, an inference is drawn from a question or set of questions to the underlying unobservable mathematics knowledge.

The second inferential step concerns the measurement of a subset of units of the target population. In this case, inference based on probability sampling is the foundation of government statistical agencies throughout the world. The inference, however, needs careful descriptions of the target population and the frame population: in an ideal situation, the frame has a one-to-one correspondence to the full population of interest, the target population. Government statistical agencies strive to create and maintain such universal frames for households and businesses. Some countries attempt to construct frames of people through population registers.

For each error source shown in Figure 6-1, one can further distinguish two kinds of errors: (1) biases, which represent systematic errors that are integral to the process and would therefore not be ameliorated without a change in the process; and (2) variances, which represent instability in the estimates and depend on the number of units of a particular kind included in the data collection. Variance can be reduced by increasing the number of units of that kind.

Federal statistical agencies are very sensitive to the potential errors in their surveys. Their primary focus has been on the fact that data are collected on only a sample of the population, as it is this feature that distinguishes surveys from a complete population census. Many of the most

widely used sampling and estimation procedures for surveys were developed at the U.S. Census Bureau in the 1930s and 1940s. Other statistical agencies have also been leaders in the provision of detailed information on sampling errors (measures of precision for survey estimates), and most federal surveys produce detailed tables of standard errors for their estimates. Other aspects of data quality have been given less emphasis, though agencies have conducted important investigations of some nonsampling errors in major surveys.

There has also been some work to apply the same ideas of statistical inference to some other forms of measurement error (response error in surveys) particularly, models of interviewer effects (interviewer variance), using the concepts of simple and correlated response variance (by analogy with simple and correlated sampling variance), and looking at instability in responses (reliability). These models were refined by U.S. and other statistical agencies in the 1960s, 1970s, and 1980s. A major conceptual push toward the examination of measurement error occurred as a result of work on cognitive aspects of survey methods, which examined the psychological mechanisms that influence respondents' understanding of questions and retrieval of information (see National Research Council, 1984). The major statistical agencies set up cognitive laboratories to incorporate these principles into the design of questions used in their surveys (see Tanur, 1999).

Concern about nonresponse has also been a feature of quality control in statistical agencies, and the emphasis has most often been on maximizing the response rate. The standards and guidelines for statistical surveys issued by the U.S. Office of Management and Budget (2006) stipulated standards for the collection of statistical information that included a high threshold for response rates (80 percent) and a detailed protocol for evaluating errors if the response rate is below this threshold. However, there is generally no reporting mechanism that quantifies the effect of nonresponse in a way that corresponds to the routine publication of standard errors. This reflects a general emphasis on variances (stability) rather than biases, as well as the ability to calculate standard errors directly from the sample, while determining nonresponse bias requires information external to the survey.

Some federal statistical agencies have created "quality profiles" for some major surveys to bring together what is known about the various sources of error in a survey. This was first done for the Current Population Survey in 1978 (Federal Committee on Statistical Methodology, 1978) and was also done for other surveys (see, e.g., National Center for Education Statistics, 1994; U.S. Census Bureau, 1996, 1998, 2014b). The Census Bureau currently includes this kind of information in design and methodology reports for the Current Population Survey and American Community Survey (see U.S. Census Bureau, 2006, 2014a).

**CONCLUSION 6-1 Survey researchers and federal statistical agencies have developed useful frameworks for classifying and examining different potential sources of error in surveys, and the agencies have developed careful protocols for understanding and reporting potential errors in their survey data.**

Effectively using frameworks for the different sources of error in surveys requires agencies to provide information or metrics for reporting on survey quality. A subcommittee of the Federal Committee on Statistical Methodology (2001) reviewed the different kinds of reports produced by statistical agencies and the amount of information provided about the different sources of error in each and further provided guidance to agencies on measuring and reporting sources of error in surveys. It recommended the minimum amount of information that should be provided, which depended on the length and detail of the report.

Measures of sampling error (estimates of the variability in the estimates due to sampling) are the most commonly reported metric of quality across all agency reports (see Federal Committee on Statistical Methodology, 2001). These measures include standard errors or coefficients of variation, and they are often the only quantitative indicator of quality reported by the agency. Other error sources are noted in more general narrative form, such as statements that surveys are also subject to nonsampling errors. Since standard errors are typically the only quantitative metric available at the estimate level, it is easy for users to conclude that this measure conveys the overall quality of the estimate.

The second principal metric of quality often used in official statistics is the response rate, which is also a very imperfect metric of quality. Low response rates do not, by themselves, indicate poor-quality statistics. Instead, lower rates indicate a higher risk of nonresponse error. Similarly, high response rates provide important protection against the potential for nonresponse bias. However, this is an area in which the direct collection of ancillary data (paradata) could facilitate a more comprehensive assessment of the danger of the vulnerability of survey results to bias arising from nonresponse. Furthermore, access to other data sources (administrative or private sector) could also provide validating (or invalidating) evidence. And in some cases paradata could provide a basis for imputing data when a survey failed to obtain information directly. In this instance, the incorporation of additional data would supplement or validate, rather than replace, the survey data.

**CONCLUSION 6-2 Commonly used existing metrics for reporting survey quality may fall short in providing sufficient information for evaluating survey quality.**

## BROADER FRAMEWORKS FOR ASSESSING QUALITY

There is considerable inertia in any long-established system. Consequently, the evaluation of the quality of any statistic tends to be seen through the lens of the current dimensions of focus and concern. In the new environment in which emerging data sources with quite different provenance and characteristics provide alternative ways of measuring the underlying phenomena, it may be necessary to rethink the weight that is placed on traditional measures of quality relative to other quality measures.

### BOX 6-1
### Eurostat Quality Assurance Framework

The *Eurostat Quality Assurance Framework* (European Statistical System Committee, 2013) describes activities, methods, and tools that can provide guidance to national statistical offices to fulfill the principles in the *European Statistics Code of Practice* (European Commission, 2011) for the principles on commitment to quality and statistical processes. In the framework, there are three main categories under which the data principles fall: institutional environment, statistical processes, and statistical output. Institutional environment is mainly composed of a principle of commitment to quality, including a quality commitment statement for agencies, a structure for managing quality, quality guidelines, infrastructure for documentation, and training courses. Statistical processes involve sound methodology, appropriate statistical procedures, burden, and cost.

Statistical output comprises five principles—(1) relevance, (2) accuracy and reliability, (3) timeliness and punctuality, (4) coherence and comparability, and (5) accessibility and clarity—that are further described.

1. **Relevance.** In order to have relevance, statistics must meet the needs of users. In order to ensure that user needs are met, it is important to monitor the relevance of existing statistics while considering emerging needs and priorities, ensure that priorities are reflected in the work program and being met, and monitor user satisfaction on a regular basis.

2. **Accuracy and Reliability.** Statistics should accurately and reliably portray reality. The combination of source data, intermediary data, and statistical outputs should be regularly assessed. Additionally, sampling and nonsampling errors need to be measured and documented and revisions regularly analyzed in order to improve statistical processes.

There are broader quality frameworks that emphasize the granularity of the data and the estimates in ways that were previously not possible. An influential quality framework has been developed by Eurostat, the Statistical Office of the European Union (European Statistical System Committee, 2013) (see Box 6-1). This framework has five major output quality components:

1. Relevance
2. Accuracy and Reliability
3. Timeliness and Punctuality
4. Accessibility and Clarity
5. Coherence and Comparability

3. **Timeliness and Punctuality.** Statistics should be released in a timely and punctual manner. Eurostat recommends that agencies ensure that they follow proper release dates and times and make sure to publish the time that outputs will be published. In creating the times for public release, it is also recommended that agencies consider user needs. If there are any divergences, those should also be published. Finally, any preliminary estimates created to help users should be published only after they are determined to hold information useful to users.

4. **Accessibility and Clarity.** Statistics should be presented in a clear and understandable form, released in a suitable and convenient manner, and available and accessible on an impartial basis with supporting metadata and guidance. It is recommended that metadata are preserved and properly archived. It is also recommended that metadata are standardized according to systems, dissemination services use proper communication and current technology, custom-designed analysis is provided when feasible, and public microdata files are available to researchers for specific purposes following protocols. Additionally, the public should be informed as to the current methodologies for statistical processes.

5. **Coherence and Comparability.** Statistics should be consistent internally and over time and comparable among regions and countries. It should be possible to combine and make joint use of related data from different data sources. In order to ensure coherence and comparability, it is recommended that standards are followed with respect to scope, definitions, units, and classifications. Statistics over time should also be as comparable as possible. Agencies should do the best job possible to ensure that statistics from different sources are compared and reconciled.

It is worth noting that all the technical aspects of the total survey error model discussed above are encompassed by just one of the above five output quality components—accuracy and reliability. Of course, many of the others are critical in determining what information is collected and disseminated, in particular, relevance. The criterion of coherence also makes important demands on the process by requiring that the estimates are consistent internally and comparable across subsets of a population. Maintaining relevance and coherence may be a particular challenge in moving from one system or configuration to another. This challenge is exacerbated in the case of surveys by the length of time it takes to develop and test survey components, leading to a lack of nimbleness in response to changes in circumstances and policy requirements.

Two aspects of data quality in particular warrant emphasis here: timeliness and spatial and subgroup granularity.

### Timeliness

Though timeliness is described in the European Statistical System Committee's quality framework in terms of the timing and punctuality of reports, it is important to recognize that existing systems have tailored their reporting mechanisms and their reporting requirements to the practical constraints and limitations in place at the time the systems were established. For example, the unemployment rate, calculated from the Current Population Survey (CPS), is published monthly based on information collected about a single week in the preceding month. When the system was established in the 1940s, this schedule (and basis) was at the outer limits of feasibility and practicality for a national survey. Changes in the data environment have made the range of possibilities much broader, and, therefore, the assessment of quality should now incorporate this new context.

For policy purposes, it is particularly important that information be available to decision makers in time to incorporate it into the decision-making process. Data collected through surveys tend to have a minimum interval between the beginning of data collection and the production of the estimate. Even with the CPS, there is a lag of about 3 weeks between the reference period for the survey responses and the production of the estimate. With some private-sector data that is captured electronically, the time lapse between the event being measured and the availability of the data is, in principle, negligible (though in practice this may not be the case).

In this context, one needs to think about the value of this aspect of statistical estimates and the particular use to be made of the estimate. With prices (see the discussion in Chapter 4 of the panel's previous report), the current (old-fashioned) process of visiting stores to collect price information has the virtue of providing almost comprehensive coverage (through

probability sampling) of the population of stores, but it suffers from a considerable time lag in reporting. Data from the Internet, in contrast, can be harvested in a timely manner but may miss differentially key sectors of the population of stores. Researchers and analysts need to be able to evaluate the relative importance of timeliness and coverage for each particular purpose of a statistic. If the purpose is to provide early warning of price changes, one might argue that the Internet data, though deficient in coverage terms, would function better than the more comprehensively representative survey data. If the purpose is to provide an estimate that gives an unbiased measure of change in prices for the population as a whole, the argument would swing in favor of the probability sample of stores. Combining the sources offers the potential to improve the timeliness of estimates from probability samples and reduce the bias present in Internet data.

### Spatial and Subgroup Granularity

A second important aspect of quality is the degree to which estimates can be obtained for small subdivisions of the population, such as spatial subdivisions or different socioeconomic status categories. In census and survey statistical terminology, these estimates are usually referred to as small-area statistics. The practical limitation on the size of the sample that can realistically be afforded for traditional surveys places a severe limit on the ability to provide reliable small-area statistics. With private-sector data, such refinement of the estimates may be accomplished at low marginal cost once the system processes have been put in place. The value to policy makers at a national level of having information at this more detailed level can be considerable and could compensate for some loss of quality on the dimensions of accuracy and reliability.

All of these considerations point to the importance of recognizing the need to evaluate quality in a more broad-based way. By combining data sources (as described in Chapter 2 and in the examples in this chapter, below), hybrid estimates can be produced that come close to possessing the positive quality aspects of the components used to construct them.

**CONCLUSION 6-3 Timeliness and other dimensions of granularity have often been undervalued as indicators of quality; they are increasingly more relevant with statistics based on multiple data sources.**

**RECOMMENDATION 6-1 Federal statistical agencies should adopt a broader framework for statistical information than total survey error to include additional dimensions that better capture user needs, such as timeliness, relevance, accuracy, accessibility, coherence, integrity, privacy, transparency, and interpretability.**

## ASSESSING THE QUALITY OF ADMINISTRATIVE AND PRIVATE-SECTOR DATA

### Administrative Data

Administrative and private-sector data have their own challenges and errors. These errors arise for multiple reasons, such as mistakes in understanding or interpreting metadata, errors in entity linkage, and incomplete or missing information. Unlike handling data from traditional surveys, these errors usually need to be dealt with after the data have been obtained and been through cleaning and processing (see Chapter 3). That is, the errors can usually not be avoided during data gathering or generation because the processes that generate the data have their own purposes: that is, unlike surveys, the production of the statistic is secondary to another objective. In contrast, survey designers spend a great deal of time and effort developing and pretesting survey instruments to ensure they are obtaining the information they want from respondents and minimizing measurement errors. Electronic survey instruments often include consistency checks and acceptable ranges of responses to further ensure that potential problems with data entry or responses are resolved at the point of collection.

### Data Linkage and Integration

The use of administrative and private-sector data not only shifts the focus of reducing errors into the postdata gathering stage, it also adds a new error source that is not usually encountered in surveys: linkage errors. In many instances, it is necessary to match data related to the same real-world entities, even if they are identified in different ways. As we note in Chapter 2, many linkage variables have variants: the same person might be listed in different data sources as "Susan Johnson Wright," "Suzy Johnson," "Sue Wright," or "S.J. Wright." Failure to recognize these as belonging to the same person will result in records being declared distinct when they ought to be linked—a *missed link*. Conversely, there may be multiple Susan Wrights in the population, and two records may match exactly on the identifying variables yet represent two different people—a *false link*. Missed links and false links can distort relationships among variables in the data sources or result in inaccurate measures of the population size when linkage is used to augment the number of records available for study.

In the case of a potential false link, one would need to look at other evidence, such as an address, to determine whether it is the same person. However, this is complicated, because people do move, so one would have to distinguish between the same person who has changed addresses and a different person who just happens to have the same name. And in the case

of matching individuals to themselves, there is the complication of the difficulty to match an individual after a name change, say, due to marriage or simply an interest in changing one's name. Additional difficulties of linking individuals could also include joint bank accounts, accounts that have only one spouse's name, and parents or others paying for a service for a child or other relative.[1] In short, good entity matching is very hard. However, there is much active research work in this area and a significant body of technology that can be exploited for this work.

Most data of interest to statistical agencies is likely to have a temporal component. Each fact will represent either an instant in time or a period of time. It is likely that the time frames for data reporting will not perfectly line up when integrating data sources. In such circumstances, techniques are needed to proceed with integration and, if at all possible, without introducing too much error. For instance, where reported aggregates change smoothly over time, some type of interpolation may be appropriate. For example, where certain secondary information, such as North American Industry Classification System codes or dictionary terms, is largely static, one may reasonably assume it has the same value at the time of interest as at another time.

Similar issues also arise for geographical alignment: if some data are reported by ZIP code and other data by county, one can integrate them only if one can transform one of the two datasets into a report by the other. Such transformation is complicated by the fact that ZIP code to county is a many-to-many relationship: many ZIP codes can include more than one county and vice versa. Techniques for small-area estimation (see Chapter 2) may be helpful in achieving such temporal and spatial alignment.

**CONCLUSION 6-4 Quality frameworks for multiple data sources need to include well-developed treatments of data linkage errors and their potential effects on the resulting statistics.**

### Concepts of Interest and Other Quality Features

The kinds of statistical models discussed in Chapter 2 when using multiple data sources underscore the need for more attention to quality features that have not received as much attention in traditional statistics (see also Groves and Schoeffel, in press). We note here several issues with regard to administrative data that will need attention.

One core difference with respect to quality for administrative and private-sector data is tied to the concept of interest. Since data are being

[1]Additional concerns arise if consent is required in order to link different datasets (see Chapter 4).

repurposed, the meanings of particular values are likely to be subtly different. A precise understanding of those differences is critical for correct interpretation and difficult due to varying metadata recording standards across data sources. Similarly, population coverage and sampling bias are also of particular concern when repurposing data. Understanding exactly what has been done is a necessary step to ensuring correctness. Finally, repurposing data will often require considerable manipulation. Therefore, recording the editing and cleaning processes applied to the data, the statistical transformations used, and the version of software run become particularly important (see Chapter 3).

It is likely that the population covered by one dataset will not match the population covered by another. Administrative records exist for program participants but not for nonparticipants, and some people may not be in any record system. Some administrative records contain people outside of the population of inference of a survey dataset to which records are to be linked (e.g., voter records may contain dead people formerly eligible to vote). There may be duplicate records in some sources, and it is possible that records from multiple datasets simply cannot be linked. Moreover, it will not be possible to separate linkage errors with coverage errors.

The level of measurement in the multiple data sources may vary. One dataset may be measured on the person level, another on the household level, another on the consumer unit level, another on the address level, and another on the tax unit level, and other data, such as credit card purchases, may be at a transaction level. Errors in statistics can arise due to mismatches when combining data from such different datasets.

The temporal extent of the data may vary. Some administrative data systems are updated on a relatively haphazard schedule, depending on interaction with the client. The result is that time can be variable over records in the same administrative dataset. The value reflects the "latest" version, sometimes with no metadata on the date the value was entered. The data may also contain information on an individual only for the time during which the individual was in a program. As individuals leave the program, they will not have new data recorded about them, but they may also not be removed from the system. Combining records to produce statistics that are designed to describe a population at a given time point is thus problematic.

The underlying measurement construct may differ across datasets. In the total survey error framework, this occurrence is sometimes labeled as an issue of validity or a gap between concept and measurement. For example, in one dataset the value of an economic transaction may capture only the goods or service provided, but in another it may also involve cash given to the customer by the provider but charged to the account of the customer.

The nature of missing data in records may vary across surveys and other sources. In survey data, questions may be skipped for several rea-

sons: because a respondent chooses to terminate participation before the question is posed; because of a refusal to answer the question because the respondent does not want to answer it; or because the respondent does not know the answer.

In addition, in administrative record systems, some data fields may be empty because of their lack of importance to the program: for example, a service clerk did not need the data to execute the task at hand. A field may also be empty because new processes fail to capture the datum. And less important information may not be captured as accurately or as carefully as information that is critical to the immediate administrative task.

As with surveys, some recorded items may be inaccurate because they are misunderstood by the clerk entering the information or by the respondent providing the information. To compensate for missing data, survey-based imputation schemes use patterns of correlations of known attributes to estimate values missing in the records. Those kinds of techniques may need some reassessment for administrative data. For example, unstructured text data appear in many medical record systems. Does the absence of the mention of some attribute mean that the attribute is nonexistent for the patient or that the health provider failed to record the attribute?

### Frameworks for Assessing Quality of Administrative Data

In the panel's first report (National Academies of Sciences, Engineering, and Medicine, 2017b) we concluded that "administrative records have demonstrated potential to enhance the quality, scope and cost efficiency of statistical products" (p. 35), and that "not enough is yet known about the fitness for use of administrative data in federal statistics" (p. 48). Therefore, we made the following recommendation:

> **Federal statistical agencies should systematically review their statistical portfolios and evaluate the potential benefits and risks of using administrative data. To this end, federal statistical agencies should create collaborative research programs to address the many challenges in using administrative data for federal statistics.** (National Academies of Sciences, Engineering, and Medicine, 2017b, Recommendation 3-1, p. 48).

As part of their review of administrative data, agencies need to assess the quality of the data and whether the data are fit for use for their intended statistical purposes (see, e.g., Iwig et al., 2013). Brackstone (1987, p. 32) notes that the quality of administrative records for statistical purposes depends on at least three factors:

(i) the definitions used in the administrative system;

(ii) the intended coverage of the administrative system;
(iii) the quality with which data are reported and processed in the administrative system.

In the United States, Iwig and his colleagues (2013) created a tool for assessing the quality of administrative data for statistical uses to help statistical agencies systematically obtain the relevant information they need at each stage of the data-sharing process (see Box 6-2). Although the tool is based on quality frameworks from a number of national statistical offices, it is intended to help guide the conversation between a statistical agency and a program agency, rather than provide a comprehensive framework itself.

**BOX 6-2**
**The Data Quality Assessment Tool for Administrative Data**

The Data Quality Assessment Tool for Administrative Data, commonly referred to as the Tool, was developed by the Federal Committee on Statistical Methodology's Data Quality Working Group (Iwig et al., 2013). The Tool is designed both to help users better understand data's attributes so that data can be used more appropriately in new applications and to promote better data quality over time for administrative and statistical purposes. Although the Tool does not result in a single overall numerical measure, it does provide users with a set of questions that help them consider key data quality attributes.

The Tool consists of six dimensions: relevance, accessibility, coherence, interpretability, accuracy, and institutional environment. Although other frameworks also contain other dimensions, such as timeliness and comparability, the Tool covers these topics under the six dimensions: timeliness is covered under relevance, and comparability is covered under coherence.

The Tool provides 43 questions organized by three different phases: initial (discovery) phase, initial acquisition phase, and repeated acquisition.

The initial phase contains 12 questions to determine the feasibility and desirability of the data in order to help develop a memorandum of understanding. This phase covers the dimensions of relevance, accessibility, and interpretability.

The initial acquisition phase begins after a memorandum of understanding has been approved. This phase contains 29 new questions that deal with such issues as recording methods, known sources of error, and missing values. This phase covers all six dimensions.

The repeated acquisition phase covers repeated installments of data. This phase has 11 questions: 2 are new, and 9 are repeated questions from the two previous phases. It includes the dimensions of interpretability, coherence, accuracy, and institutional environment; the questions of institutional environment are new.

Administrative records are used by many national statistical offices for producing statistics; however, there have not been statistical theories developed for assessing the uncertainty from administrative data as there have been for surveys (Holt, 2007). There have been efforts to examine the processes that generate administrative data and population registers (see Wallgren and Wallgren, 2007), as well as examinations of measurement errors in both administrative records and survey reports (e.g., see Oberski et al., 2017; Abowd and Stinson, 2013; Groen, 2012). Recently, Zhang (2012) has extended the Groves et al. (2009) model of survey error sources (shown in Figure 6-1) to create a two-phase (primary and secondary) life cycle of integrated statistical microdata. In this model, the first phase covers the data from each individual source, while the second phase concerns the integration of data from different sources, which typically involves some transformation of the original data (see Zhang [2012] for a complete discussion of the model).

## Private-Sector Data

### Different Types of Data

In our first report (National Academies of Sciences, Engineering, and Medicine, 2017b, p. 57) we noted the enormous amounts of private-sector data that are being generated constantly from a wide variety of sources. We distinguished between structured data, semi-structured data, and unstructured data:

- Structured data, such as mobile phone location sensors, or commercial transactions, are highly organized and can easily be placed in a database or spreadsheet, though they may still require substantial scrubbing and transformation for modeling and analysis.
- Semi-structured data, such as text messages or emails, have structure but also permit flexibility so that they cannot be placed in a relational database or spreadsheet; the scrubbing and transformation for modeling and analysis is usually more difficult than for structured data.
- Unstructured data, such as in images and videos, do not have any structure so that the information of value must first be extracted and then placed in a structured form for further processing and analysis.

The challenges with the quality of these different data sources led to a conclusion and recommendation:

> **The data from private-sector sources vary in their fitness for use in national statistics. Systematic research is necessary to evaluate the quality, stability, and reliability of data from each of these alternative sources currently held by private entities for their intended use.** (National Academies of Sciences, Engineering, and Medicine, 2017b, Conclusion 4-2, p. 70)

> **The Federal Interagency Council on Statistical Policy should urge the study of private-sector data and evaluate both their potential to enhance the quality of statistical products and the risks of their use. Federal statistical agencies should provide annual public reports of these activities.** (National Academies of Sciences, Engineering, and Medicine, 2017b, Recommendation 4-2, p. 70)

The panel thinks it likely that the data to be combined with traditional survey and census data will increasingly come from unstructured text, such as web-scraped data. Converting such data to a more structured form presents a range of challenges. Although errors arise in coding open-ended responses to surveys, the issues with unstructured text compound those errors given the ambiguity of the context and the development and use of coding algorithms, which can result in a special type of processing error. The ambiguity of words—does "lost work" refer to a job termination or a hard disk crash?—will be a constant challenge. The possibility of coding words on multiple dimensions (e.g., meaning and effect) arises. From a quality or error standpoint, a coding error might create a new source of bias if one coding algorithm is used, or a new source of variance in statistics if multiple coding algorithms are used. Training these algorithms using human-coded data essentially builds in any biases that were present in the original human coding. The emergence of computational linguistics since the early days of survey coding may offer help in this aspect of big data quality.

Some of the data used in federal statistics could be combinations of data arising from sensors. For example, traffic sensor data are useful in traffic volume statistics. From time to time, sensors fail, creating missing data for a period of time until the sensor is repaired or replaced. If the probability that the sensor fails is related to volume of traffic or traffic conditions (e.g., when heavy rain or snow occurs), then the existence of missing data can be correlated with the very statistic of interest, creating what survey researchers would label a type of item nonresponse error.

The panel's first report also discussed the possibility that social media

data might be combined in certain circumstances with survey or census data. Social media data have an error source uncommon in surveys: the possibility that data are generated by actors outside the population of inference. For example, software bots are known to create Twitter posts. The software bots might have handles and profiles that appear to be people, with revealed geographic positions, under the Twitter protocols. However, the data generated by the software bots are not a person- or business-measurement unit eligible for a survey or census measurement. Although this might be viewed as a type of coverage error in traditional total survey error terms, it has such a distinct source that it needs its own attention.

### Risks of Private-Sector Data

Although there is considerable potential in the enormous volume of private-sector data, these data carry considerable risks. First, as they are primarily administrative or transactional data collected for purposes related to the transactions they cover, they tend to be less stable in definition and form than federal survey data that are collected specifically for statistical purposes. Consequently, statistical agencies need to be cautious in relying on these data as a primary (or, especially, sole) source of information; private-sector data are vulnerable to being changed or discontinued without notice. Second, statistical agencies do not have control over the creation and curation of the data; there is the possibility of deliberate manipulation or "front-running" by private-sector data companies providing data to government agencies: that is, if the data supplied by a private-sector entity constituted a sufficiently influential portion of a statistic so that the entity itself could predict the agency's results, the entity could profit by selling this information to others or by acting on it directly. Third, the data themselves could be subject to manipulation for financial gain.

### Quality Frameworks

There have been some recent efforts to extend the total survey error framework to include big data (see Biemer, 2016; Japec et al., 2015; U.N. Economic and Social Council, 2014; U.N. Economic Commission for Europe [UNECE], 2014). The UNECE model is a multidimensional hierarchical framework that describes quality at the input, throughput and output phases of the business process (see Box 6-3).

Biemer (2016) and Japec et al. (2015) similarly distinguish between the phases of generation of the data, the extraction/transformation/load of the data, and the analysis of the data. In this scheme, data generation errors are analogous to survey data collection errors and may result in erroneous, incomplete, or missing data and metadata. Extraction/transformation/

**BOX 6-3**
**UNECE Big Data Quality Framework**

The U.N. Economic Commission for Europe (UNECE, 2014) has developed a complex quality framework for big data. UNECE recommends three general principles when evaluating data: fitness for use, generic and flexible, and the tradeoff between effort and gain. Fitness for use covers whether the data are appropriate for their intended use. Generic and flexible covers whether the quality framework is adaptable to the specific data sources used to create statistical products. The effort-gain tradeoff covers how to ensure that the benefits gained from using a new data source outweigh the time and effort required to use the data.

The quality framework itself focuses on three main phases, three hyperdimensions, and three principles. The phases are the input process that acquires data and begins pre-analysis of the data, the throughput process that begins transforming and analyzing the data, and the output process that reports the quality of statistical outputs.

Each phase has three hyperdimensions: source, consisting of characteristics of the data obtained and the governance under which the data are administered; metadata, consisting of information about the data itself, including the contents of a dataset, the processes that apply to it, and the information available in the dataset; and the data themselves, that is, their quality.

load errors are similar to survey processing errors and include errors in specification, linking, coding, editing, and integration. Analysis errors are analogous to modeling and estimation errors in surveys and also include errors in adjustments and weighting, which may reflect errors in the original data or in filtering and sampling the data. Both models attempt to provide an overall conceptual view of errors in a broad array of data sources and a language to describe these errors; however, neither of the models has yet been applied to a variety of sources by researchers or analysts.

Hsieh and Murphy (2017) have taken a more specific approach and described error sources for a specific social media data source, creating a "total Twitter error." The authors posit that major classes of errors occur during the data extraction and the analysis process.

**CONCLUSION 6-5 New data sources require expanding and further development of existing quality frameworks to include new components and to emphasize different aspects of quality.**

**CONCLUSION 6-6 All data sources have strengths and weaknesses, and they need to be carefully examined to assess their usefulness for a given statistical purpose.**

**RECOMMENDATION 6-2 Federal statistical agencies should outline and evaluate the strengths and weaknesses of alternative data sources on the basis of a comprehensive quality framework and, if possible, quantify the quality attributes and make them transparent to users. Agencies should focus more attention on the tradeoffs between different quality aspects, such as trading precision for timeliness and granularity, rather than focusing primarily on accuracy.**

As we note in previous chapters, expanding the use of data sources for federal statistics also requires expanding the skills of the statistical agency staff to address the new issues that arise when using these new data sources. Researchers and analysts who are trained in survey methodology have a solid foundation for conceptualizing and measuring different error sources, but expertise and training is also needed in computer science for processing, cleaning, and linking datasets and the errors that can arise in these operations. In addition, specific expertise is needed in the data generating mechanisms and current uses of different administrative and private-data sources that an agency is considering for use for statistical purposes.

**RECOMMENDATION 6-3 Federal statistical agencies should ensure their statistical and methodological staff receive appropriate training in various aspects of quality and the appropriate metrics and methods for examining the quality of data from different sources.**

## THE QUALITY OF ALTERNATIVE DATA SOURCES: TWO ILLUSTRATIONS

In this section we provide two examples of the current approach for generating federal statistics and of existing different data sources that could be used. The two examples are measuring crime and measuring inflation. We discuss how some key quality characteristics of these sources might interact and permit enhanced federal statistics if these sources were combined. Our goal here is illustrative rather than prescriptive, and the responsible federal statistical agencies would need to conduct a much more in-depth review of the alternative sources and the methods for combining them than we can do here.

### Measuring Crime

#### Current Approach

The National Crime Victimization Survey (NCVS) of the Bureau of Justice Statistics (BJS) is the nation's primary source of information on

criminal victimization.[2] Since 1973, data have been obtained annually from a nationally representative sample of about 90,000 households, comprising nearly 160,000 people, on the frequency, characteristics, and consequences of criminal victimization in the United States. The NCVS screens respondents to identify people who have been victims of nonfatal personal crimes (rape or sexual assault, robbery, aggravated and simple assault, and personal larceny) and property crimes (burglary, motor vehicle theft, and other theft), and it includes crimes whether or not they have been reported to the police. For each crime incident, the victim provides information about the offender, characteristics of the crime, whether it was reported to the police, and the consequences of the crime, including the victim's experiences with the criminal justice, victim services, and health care systems. In addition, there is a wealth of additional information collected—including a victim's demographic characteristics, educational attainment, labor force participation, household composition, and housing structure—that is related to risk of victimization (see Langton et al., 2017; National Academies of Sciences, Engineering, and Medicine, 2016a).

Despite the appearance of being comprehensive, the NCVS has its quality challenges:

- Not everyone can be reached for an interview, and if people who spend large amounts of time outside their homes are more likely to be at risk for victimization, crimes can be missed.
- Of the people who are reached, not all are willing to participate in the survey, introducing the possibility for bias if those who choose not to participate differ from those who do.
- Not everyone is willing to report crimes that happen to them.
- Some respondents may misunderstand the questions asked.
- Careful testing of measurement instruments and long redesign processes lead to lags in capturing emerging crimes, such as identity theft.
- Data collection takes considerable time, so that annual estimates are only available months after the end of the year in which the crimes occurred.

### Alternative Approaches

Alternative data sources on crime include the Uniform Crime Reports (UCR), which is based on police administrative records. The UCR Summary System covers jurisdiction-level counts of seven index crimes that are collated and published by the FBI. The scope of this collection has remained

[2]See https://www.bjs.gov/index.cfm?ty=dcdetail&iid=245#Collection_period [August 2017].

essentially constant since 1929. Several features make the UCR an attractive alternative. The UCR is intended to be a census of about 18,000 police departments covering the whole country; however, there are lots of missing data. It provides consistent classification over time, with only modest periodic changes: arson was added in 1978, and hate crimes in the 1980s. The UCR also can provide monthly data, allowing for finer granularity in time, as well as granularity in geography because of the jurisdiction-level counts of the major crime categories.

However, the UCR also has its own errors and limitations (Biderman and Lynch, 1991; Lynch and Addington, 2007):

- It is restricted to crimes reported to the police, and the NCVS has found that more than 50 percent of crimes reported by victims are not known to the police.
- The data are reported at the jurisdiction level and cannot be disaggregated. That is, one cannot obtain counts of offenses or rates of crimes for a subjurisdictional area or any subpopulation in the jurisdiction.
- There is virtually no information provided on the characteristics of victims or offenders or the circumstances of the incident.
- Police jurisdictions map poorly onto census classifications of places that serve as the denominator of crime rates.
- There is little enforcement of standards of reporting. Although the FBI provides training and audits to check for appropriate counting and classification of offense, there is no way to ensure that all eligible crimes are included. (The FBI collates the reported data but does not provide a warrant of quality.) Occasional audits are performed on a small number of agencies annually, but they focus on counting rules and proper classification and not on the completeness with which eligible incidents are reported. Police agencies may also misreport or misclassify crimes: for example, the *Los Angeles Times* found that nearly 1,200 violent crimes had been misclassified by the Los Angeles Police Department as minor offenses, resulting in a lower reported violent crime rate.[3]
- The system is voluntary at the national level. There are mandatory reporting laws in states that require that local police report to the states, but there are no requirements to report to the UCR. There are also few instances in which a state mandatory reporting law has been invoked for a jurisdiction's failure to report. In addition, unit missing data is assessed on the basis of jurisdictions that have

[3]See http://www.latimes.com/local/la-me-crimestats-lapd-20140810-story.html [August 2017].

previously reported to the UCR, so units that have never participated are not counted as missing.

In addition to the data available in the Summary System, the UCR also includes the National Incident Based Reporting System (NIBRS), which contains data on crime incidents rather than jurisdiction-level counts. Currently, however, this data system is available only in a relatively small proportion of police agencies, and it substantially underrepresents the larger jurisdictions that contain most of the crime.

This would appear to be a situation in which combining data from the different sources (see Chapter 2) could greatly enhance the value of the estimates. BJS's small-area estimation program (see National Academies of Sciences, Engineering, and Medicine, 2017b, Ch. 2) illustrates how the UCR Summary System and the NCVS could be used jointly to increase understanding of crime in states and other subnational areas (see also Li et al., 2017). As NIBRS, through the National Crime Statistics Exchange (NCS-X),[4] is implemented more broadly across the nation, the blending of NCVS and NCS-X data could be done at the incident level, which would substantially increase the ability to leverage the two data sources. The fact that both NCS-X and NCVS are sample based may pose problems for such estimates, but the emphasis on large cities in both data collections may afford sufficient overlap for useful estimation.

## Measuring Inflation

### Current Approach

The Consumer Price Index (CPI), produced by the Bureau of Labor Statistics (BLS), provides monthly data on changes in the prices paid by urban consumers for a representative basket of goods and services. The CPI has three major uses. First, the CPI is an economic indicator and the most widely used measure of inflation. Second, the CPI is a deflator of other economic series: the CPI and its components are used to adjust other economic series for price inflation and to translate them into inflation-adjusted dollars. Third, the CPI is used to adjust wages: more than 2 million workers are covered by collective bargaining agreements that tie wages to inflation.[5]

BLS produces thousands of component indexes, by areas of the country,

[4]The NCS-X program is designed to generate nationally representative incident-based data on crimes reported to law enforcement agencies. It comprises a sample of 400 law enforcement agencies to supplement the existing NIBRS data by providing their incident data to their state or the federal NIBRS program. See https://www.bjs.gov/content/ncsx.cfm [September 2017].

[5]See https://www.bls.gov/cpi/cpifaq.htm#Question_6 [August 2017].

by major groups of consumer expenditures, and for special categories, such as services. In each case the index is based on two components—the "representative basket of goods" and the price changes in that basket of goods. The market basket of goods is based on the Consumer Expenditure Survey (CE), which is a household survey conducted each year in a probability sample of households selected from all urban areas in the United States. In one part of the survey, 7,000 households are interviewed each quarter on their spending habits; an additional 7,000 families complete diaries for a 2-week period on everything they bought during those 2 weeks. The CE produces a picture of the expenditure patterns of households and selected subgroups, in terms of the quantities of different products and services they bought during the period. These are combined across the households in the sample to provide a picture of the total expenditure patterns of the U.S. household population and for subsets of that population. The weight for an item is derived from reported expenditures on that item divided by the total of all expenditures. The most recent CPI market basket is based on data from CE for 2013 and 2014.

Separate surveys and visits to retail stores are conducted by BLS to obtain prices for goods and services used to calculate the CPI, which are collected in 87 urban areas throughout the country, covering about 89 percent of the total U.S. population and about 23,000 retail and service establishments. Data on rents are collected from about 50,000 landlords or tenants.

These massive data collection efforts are facing challenges. As we described in our first report, response rates for the Consumer Expenditure interview and diary surveys have declined (National Academies of Sciences, Engineering, and Medicine, 2017b, Ch. 2). Another major source of error in the CE is the quality of the measurement of expenditures, given that respondents are exposed to very lengthy questionnaires with many recall tasks. Facing many questions in a row about their expenditures and details about each expenditure item, respondents are likely to underreport and so shorten the interview (or skip recording items in the diary) in order to reduce the reporting burden (see National Research Council, 2013a). And, as for the NCVS, some people are not willing to participate in the survey, reducing the precision of the resulting estimates and possibly introducing bias (National Research Council, 2013b).

### Alternative Approaches

The explosion in the availability of online consumer data suggests that this is an area in which there is enormous potential to use alternative data to supplement or replace the current measurement of inflation. One example of such an effort is the Billion Prices Project (BPP): it was initi-

ated in 2007 to provide an alternative inflation index for Argentina, given widespread distrust of the official level reported by the national statistical office, and has since expanded to cover almost 80 countries (Cavallo and Rigobon, 2016). The objective of BPP was to substitute the collection of prices using web-scraping instead of visiting retail stores in person to collect prices. Although the data are dispersed across hundreds of websites and thousands of web pages, advances in automated scraping software now allow design and implementation of large-scale data collections on the web.

The main advantage of web-scraping is that sampling and nonresponse errors are minimized—and in some circumstances they might go to zero.[6] Furthermore, detailed information can be collected for each good, and new and disappearing products can be quickly detected and accounted for. Online data collection is cheap, fast, and accurate, making it an ideal complement to traditional methods of collecting prices, particularly in categories of goods that are well represented online, such as food, personal care, electronics, clothing, air travel, hotels, and transportation. In addition to being well-represented online, these sectors reflect prices that are close to the prices from all transactions (i.e., offline transactions are the same prices), making the data of high quality. In some sectors the data quality is less desirable. Gasoline is an example: gasoline is not bought online, and the prices are collected by third parties and then shown online. The quality of this procedure depends dramatically on how the data are collected and curated before they are shown. Because information about prices on the web may be very good in some sectors but seriously deficient in others, there is a continuing challenge to develop a described, researched error structure for a hybrid approach, which would combine information from different sources to calculate an index of inflation.

To create an inflation index, consumption quantities are needed in addition to prices. Information on quantities is essentially nonexistent online (at least not accessible through web-scraping). For the BPP, the consumption quantities (i.e., the data to construct the weights) come from a combination of the CE and inferences on how the web pages are organized. Therefore, the construction of online-based inflation indexes uses a hybrid approach of survey and alternative data sources, both the prices and quantities.

[6]Nonresponse errors would be zero if all selected websites were able to be scraped. Sampling errors would be zero if the relevant universe was completely covered.

# 7

# A New Entity to Provide Vital Information Through Enhanced Federal Statistics

In the panel's first report we summarized our finding regarding the need for a new entity to meet the nation's need for statistics (National Academies of Sciences, Engineering, and Medicine, 2017b, p. 102):

> The panel believes that the nation needs a secure environment where administrative data can be statistically analyzed, evaluated for quality, and linked to surveys, other administrative datasets, and other data sources. Such an environment would need to have the authority to control access for statistical purposes. It would also have to use and continually evaluate and enhance privacy measures. Integration of these efforts into a single entity could achieve many benefits if all statistical agencies could use a secure data-sharing environment. Without a new entity, no scaling of expertise can occur in privacy protection measures, statistical modeling on multiple datasets, and IT [information technology] architectures for data sharing.

On the basis of that finding, we made the following recommendation:

> **A new entity or an existing entity should be designated to facilitate secure access to data for statistical purposes to enhance the quality of federal statistics.** (National Academies of Sciences, Engineering, and Medicine, 2017b, Recommendation 6-1, p. 102)

Although some of the recommendations in this report for improving federal statistics could be carried out by individual agencies, or by cooperative agreements among agencies, the panel believes that the best way forward is to create a new entity that will provide a secure environment

for analysis of data from multiple sources, coordinate acquisition and use of data, and identify and facilitate research on the challenges that are common across agencies.

In this chapter, we elaborate the potential different ways this entity might operate and the pros and cons of those approaches. There are many questions that need to be addressed in the creation of this new entity, and many are outside the scope of the panel. However, in some areas we do believe there is a clear approach to follow, so we offer recommendations.

There are many stakeholders, including the federal statistical agencies, data providers, and data users who are vital to the success of an endeavor like the one the panel is recommending. In addition, the Commission on Evidence-Based Policymaking is currently studying ways to make survey and administrative data accessible for program evaluation purposes.[1] We view our efforts in this domain as complementary to and informative to those of the commission. Our goal is for this report to help initiate a more detailed discussion among the stakeholders to identify the best path forward for the federal statistical system to provide the objective and reliable information that the country needs to inform decisions by policy makers, businesses, and individuals.

First and foremost, the panel intends this new entity to be a reinforcement and enhancement of ongoing and increasing efforts of the federal statistical agencies. The mission of the new entity would be to assist federal statistical agencies to reduce the costs and increase the value of national statistics by integrating data from multiple data sources. The entity would be a service provider to federal statistical agencies, providing increased access to data from surveys; federal, state, and local administrative data; and private-sector data. The panel believes that the recommended entity would need the same legal protections and secure environment as a federal statistical agency. Furthermore, any data accessed through the entity would be used only for statistical purposes: specifically, data accessible through the entity would not be used by any agency for any administrative, enforcement, or regulatory purpose that would affect the rights, privileges, or benefits of any individual, business, or organization.

Given current technological capabilities and concerns about privacy, the entity would likely store minimal data itself; rather, it would use secure software technology to seamlessly access and link data from other owners without burdening users. The panel does not envision this new entity as a new data warehouse or national data center, in part because the privacy loss from a data breach can be ameliorated by not collecting and storing all the data in one place or by carefully partitioning and encrypting the data (see Chapters 3 and 4). Administrative procedures would reinforce the

[1]See https://www.cep.gov [August 2017].

privacy-preserving analyses and strictly statistical uses that are permitted on the data. Staff of the entity would have the necessary legal authority to have access to key data sources for the entity's statistical agency clients. The staff would have the technical expertise to clean, curate, and link data for privacy-preserving analyses. The entity would also provide technical assistance to federal statistical and program agencies, as well as state and local program agencies and external researchers. Finally, the entity would be constantly evaluating new information security practices in an ever-changing world to ensure that the information technology used to link and analyze data is among the strongest and safest methods currently available.

The next section details the panel's conception of the recommended new entity.

## ATTRIBUTES OF THE NEW ENTITY

In our first report, we noted that this new entity could be successful and sustainable for sharing data only if it met the following prerequisites (National Academies of Sciences, Engineering, and Medicine, 2017b, p. 105):

1. It has to have legal authority to access data that can be useful for statistical purposes. The legal authority needs to span cabinet-level departments and independent agencies.
2. It has to have strong authority to protect the privacy of data that are accessed and prevent misuse. At minimum, that authority needs to be commensurate with existing laws (CIPSEA [the Confidential Information Protection and Statistical Efficiency Act of 2002], the Privacy Act), but it may also require new legislation.
3. It has to have authority to permit appropriate uses for the extraction of statistical information from the multiple datasets relevant to program evaluation and the monitoring of policy-relevant social and economic phenomena. The authority needs to delimit what uses are forbidden as well as what uses are encouraged.
4. It needs to be staffed with personnel whose skills fit the needs of the recommended entity, including advanced IT architectures, data transmission, record linkage, statistical computing, cryptography, data curation, cybersecurity, and privacy regulations.

In our first report, we identified the key questions that would need to be addressed in creating an entity that would respond to the challenges that statistical agencies have had in accessing, evaluating, and using administrative and private-sector data sources for federal statistics. In this section, we discuss the following attributes of the recommended entity: organizational

location, the environment for data access, the functions of the entity, access for external researchers, transparency, privacy protections, governance of the entity, and financing. The requisite skills of the staff are discussed throughout the section.

### Organizational Location

One of the most fundamental questions to be addressed is the organizational location of the recommended entity. One option is for it to be a part of the federal government. In this case, would it be an existing federal statistical agency or existing unit within a statistical agency, a new unit within an existing statistical agency, or a new free-standing statistical agency? Another option is for it to exist outside the federal government, as a new Federally Funded Research and Development Center (FFRDC) or as a new private-government-academic institution with shared governance. Whether the entity is part of the federal government or not carries a host of legal implications for accessing and providing access to data covered by federal privacy and statistical confidentiality laws. There are also implications for how the entity works with existing federal statistical agencies, funding, and staffing.

#### Option: A Federal Statistical Agency

***Legal Authorities and Protections*** All federal statistical agencies are covered by CIPSEA. Many of these agencies' authorizing statutes also provide confidentiality protections and restrictions on using information they acquire for exclusively statistical purposes. This common legal framework and culture of statistical uses and data confidentiality supports the option designating an existing statistical agency or unit as the recommended new entity or creating a new unit or statistical agency that would be covered by this framework.

The entity needs to be collaborative with federal statistical agencies while providing a platform for data sharing and enhancement of statistical programs, as well as for facilitating much needed collaborative research with administrative and other new sources of data. Because federal statistical agencies currently collaborate with each other on specific surveys, and more generally through the Interagency Council on Statistical Policy, the Federal Committee on Statistical Methodology, and other interagency working groups, the entity would be an integral part of the statistical system. The panel believes that the entity has to be within the federal statistical system. The panel believes that if the recommended entity is created as a federal agency, it should also be a federal statistical agency or unit. The

panel does not believe a federal agency outside of the federal statistical system could adequately fulfill the mission of the entity.

To fulfill the goals of the recommended entity, existing administrative and legal barriers limiting access to useful federal administrative data would need to be altered to permit the entity to access those data for statistical purposes. These barriers would need to be addressed regardless of the entity's location—in an existing federal statistical agency or unit, as a new unit in an existing statistical agency, or as a new statistical agency.

If the new entity is created as a new free-standing federal statistical agency, its authorizing statute would need to provide a legal framework that would give it the authorities listed above to access the necessary data, protect the privacy and confidentiality of the data, and ensure that the data are used only for statistical purposes. Creating the entity as a new federal statistical agency covered under CIPSEA would cover the protection of data, but new legislation giving the entity authority to acquire data would also be needed.

***Advantages and Disadvantages*** In considering whether an existing statistical agency or unit should be designated as the recommended entity, it is important to keep in mind the large variation in the size and capabilities of the statistical agencies in the decentralized federal statistical system. Table 7-1 shows the fiscal 2016 budgets and the number of staff for the 13 principal statistical agencies (U.S. Office of Management and Budget, 2017). Many of the statistical agencies have very small budgets and few staff, and many rely on the Census Bureau or private-sector contractors for data collection and other statistical activities to support their mission. Therefore, if an existing statistical agency or unit is designated as the entity, one of the larger statistical agencies would be better able to realistically meet the needs of all of the other statistical agencies.

Two possible candidates are the Census Bureau and the bureau's Center for Administrative Records Research and Applications (CARRA). The Census Bureau has invested substantial resources for and has amassed considerable technological infrastructure and technical staff for linking and processing survey and administrative data in CARRA. Because the Census Bureau is the largest federal statistical agency and currently collects survey data for many of the other statistical agencies, it has a large staff with extensive expertise and could be a natural home for accessing other data sources as well. The Census Bureau also created a network of 24 research data centers around the country, now known as Federal Statistical Research Data Centers (FSRDCs), which include more datasets and active participation by other statistical agencies so that they can also provide access to their data for external researchers. This established infrastructure would be a valuable foundation for the new entity.

**TABLE 7-1** Fiscal 2016 Budgets and Staffing for the 13 Principal Statistical Agencies

| Agency | Budget (in millions of $) | Staffing Levels[a] |
|---|---|---|
| Bureau of Economic Analysis | 105.1 | 517 |
| Bureau of Justice Statistics | 50.2 | 55 |
| Bureau of Labor Statistics | 609.0 | 2,569 |
| Bureau of Transportation Statistics | 26.0 | 84 |
| Census Bureau[b] | 1,368.4 | 13,625 |
| Economic Research Service | 85.4 | 365 |
| Energy Information Administration | 122.0 | 347 |
| National Agricultural Statistics Service | 168.4 | 1,118 |
| National Center for Education Statistics | 332.6 | 120 |
| National Center for Health Statistics | 160.4 | 554 |
| National Center for Science and Engineering Statistics | 58.2 | 52 |
| Office of Research, Evaluation, and Statistics | 26.1 | 68 |
| Statistics of Income, Internal Revenue Service | 36.9 | 122 |

[a]Staffing is full-time equivalents.
[b]Includes funds for the decennial census.
SOURCE: Data from U.S. Office of Management and Budget (2017).

However, there could be challenges in designating any existing agency or unit as the recommended new entity. It could be a challenge for an existing statistical agency (such as the Census Bureau) or a unit within an agency (such as CARRA) to serve all other agencies fairly: the entity may be inclined to be more responsive to the needs of its own statistical programs. Also, an existing agency has statutory and budget ties to its parent department, which may make it difficult to serve other statistical agencies equally and fairly, to understand their data needs, and to expend efforts and resources to acquire access to datasets of particular use to other agencies.

Creating a new federal statistical agency as the entity could level the playing field and address the uncertainties of designating an existing agency or unit as the entity; however, it would introduce many other challenges and potential areas of concern. Creating a new agency would require considerable time, resources, and skilled personnel to set up and operate, as well as new appropriations, at least initially. It would also need to establish relationships with existing federal statistical agencies, as well as federal program agencies that have administrative data useful for federal statistics. It would need to create an organizational culture of service to other agencies. As noted above, its authorizing legislation would need greater authority and

rights than any statistical agency currently has to acquire administrative data from federal program agencies.

***Governance*** Wherever the recommended entity is located within the federal statistical system, there will need to be a structure for governance of its activities that ensures service to all federal statistical agencies to maximize the benefits the entity can provide across the decentralized statistical system (see further discussion below). Achieving meaningful shared governance of the entity could be difficult to accomplish given the nine different cabinet departments that house statistical agencies or units. If a free-standing new federal statistical agency is created, careful planning would be needed for its governance structure and for appropriate authority for the head of the entity as part of the legislation authorizing the new agency.

### Option: A Federally Funded Research and Development Center

Although the reasoning above notes many advantages of locating the new entity as a part of the federal government and in an existing statistical agency, concerns have been raised in recent years that there are a variety of cultural and institutional barriers to innovation in the nation's statistical agencies (see National Research Council, 2011). Federal statistical agencies focus most of their attention and resources on producing reliable statistics and meeting demanding schedules for data collection, processing, and release. Research and development of new methods, data sources, and statistical techniques, which is needed to initiate new processes and new products, often has schedules that can be difficult to integrate with a production culture (see, e.g., Dillman, 1996). Furthermore, relevant research is currently scattered across the decentralized system, without a central focused agenda (see National Research Council, 2011), and research capacity within agencies is also spread very thin outside of the larger agencies (see National Research Council, 2014b).

The federal government's ability to attract and retain people with the needed skills is also a factor to be considered. Indeed, we noted in previous chapters key areas in which more skilled staff or additional training for existing staff will be needed to undertake more analyses with multiple datasets. A recent study (U.S. Government Accountability Office, 2017) found mission critical skills gaps at the Census Bureau, putting the 2020 census on the high-risk list. In addition, the federal government overall is facing the potential loss of highly skilled staff, with 30.8 percent of the workforce eligible for retirement by 2019, and the percentage of those potential retirees at several of the major statistical agencies is even higher (U.S. Government Accountability Office, 2015). Furthermore, attracting statisticians, data scientists, and IT specialists with the needed skills will be difficult given

the high demand for these professions in academia and the private sector[2] and the fixed nature of the federal pay scale, which is often not competitive with market rates for these occupations. For example, the latest salary survey conducted by the American Statistical Association showed lower salaries across all percentiles of income (from $5,000 to $123,000) for federal government statisticians than for industry statisticians (Hall and George, 2016). Another challenge in recruiting highly qualified staff is the requirement that federal employees be U.S. citizens, which is problematic because the majority of new Ph.Ds in statistics in the United States are not U.S. citizens (National Center for Science and Engineering Statistics, 2015).

These factors led Habermann (2010) to propose creating an FFRDC for innovation for the federal statistical system. FFRDCs, which include facilities such as the Jet Propulsion Laboratory (sponsored by the National Aeronautics and Space Administration) and the Los Alamos National Laboratory (sponsored by the Department of Energy),[3] are hybrid organizations designed to meet federal needs through private organizations (Kosar, 2011). They are more flexible than federal agencies and are not restricted by civil service rules and wages. Kosar (2011) notes that a great strength of FFRDCs is their ability to assemble teams of technical experts on a project basis. Habermann notes that an FFRDC would also promote stronger ties between the federal statistical agencies and the academic community, which would help bring the problems of the federal statistical agencies to the attention of academic researchers and provide a pipeline for students to learn more about federal statistics and the statistical system.

A number of issues would need to be addressed if the panel's recommended new entity is an FFRDC. The legal framework for the acquisition, protection, and use of data only for statistical purposes is a fundamental requirement for the entity, and it is unclear whether an FFRDC could operate like a statistical agency and have the authority to acquire and protect data and permit only statistical uses of the information. For the entity to be successful, it would need to have even broader authority to acquire data than that of any statistical agency. It is also unclear which agency or department would sponsor and fund the recommended entity as an FFRDC and how other statistical agencies would work with, participate in the governance of, and benefit from it.

[2]See https://www.bls.gov/ooh/math/statisticians.htm [September 2017], https://www.bls.gov/ooh/computer-and-information-technology/computer-systems-analysts.htm [September 2017], and https://www.bls.gov/ooh/management/computer-and-information-systems-managers.htm [September 2017].

[3]For a complete list, see https://www.nsf.gov/statistics/ffrdclist/ [August 2017].

### Option: A University-Based Public-Private Research Center

Another potential model for the recommended new entity could be outside of the government in a public-private research center managed by a university. There have been other research and data enclaves established at universities for the purpose of creating a platform for providing greater access to administrative and private-sector data sources for research to benefit the public good, and some have relationships with federal statistical agencies: for an example, see Box 7-1. The Institute for Research on Innovation and Science at the University of Michigan has an agreement with the Census Bureau, which permits linking of university administrative data with the demographic and business data from the Census Bureau.[4] The linked data can then be accessed and analyzed through FSRDCs.

A public-private research center managed by a university would have many of the advantages of the FFRDCs in terms of attracting highly skilled personnel outside of the constraints of the federal civil service regulations, and also offer a pipeline for attracting students to work for the entity or federal statistical agencies. A public-private entity could more easily be sponsored and supported by a number of agencies or departments rather than a single one, as is typical for FFRDCs.

However, a number of issues would need to be addressed to determine whether this approach would be able to fulfill the requirements for the entity. The legal framework for the acquisition, protection, and use of data for statistical purposes only is a fundamental requirement for the entity, and it is unclear whether a university-based public-private research center could operate like a statistical agency and have the authority to acquire and protect data and permit only statistical uses of the information. For the entity to be successful, it would need to have even broader authority to acquire data than any statistical agency currently has. Given the large variation in the size and resources of the different statistical agencies and units, there could be concerns that smaller agencies would not be able to participate or benefit as much from this approach as larger agencies.

### Conclusion

Each of the three location choices for the recommended new entity has advantages, and they will all need to address potential challenges. The tradeoffs will need to be carefully considered to best meet the needs of the stakeholders while fulfilling the primary mission of the new entity. Wherever the entity is located, the mission of the entity should remain focused on using any data for statistical purposes only.

[4]See http://iris.isr.umich.edu/ [August 2017].

**BOX 7-1**
**Example of a University-Based Public-Private Research Center: New York University Administrative Data Research Facility**

In 2017, the Census Bureau and the Laura and John Arnold Foundation funded the development of a pilot administrative data research facility that allows local, state, and federal employees (and selected researchers) access to data from different agencies that otherwise would not meet. Managed by New York University, the Administrative Data Research Facility (NYU/ADRF) exemplifies how a computing platform might meet the needs of statistical agencies.

The NYU/ADRF provides a secure platform to host confidential microdata. The platform makes use of Amazon's cloud-based utility web services and has completed a Federal Risk and Authorization Management Program (FedRAMP) authorization, which is a U.S. government-wide program that delivers a standard approach to the security assessment, authorization, and continuous monitoring for cloud products and services.

The NYU/ADRF provides a set of data analysis components that can be combined in different ways, within a secure NYU/ADRF boundary, to meet a wide range of analytical needs. In addition, the NYU/ADRF has developed a library of reusable programs that implement algorithms and strategies that can be used for de-identified comparison across datasets.

The NYU/ADRF uses a data model that pulls largely from metadata standards, such as Dublin Core, the Data Catalog Vocabulary, and the Project Open Data Metadata Schema. This approach to describing datasets facilitates their discoverability across domains and research uses. The data stewardship module provides dataset management to NYU/ADRF. The module controls who has access to which data and what content is related to that data to answer questions that data stewards typically ask, such as: "Which projects use my data?" "How are my data being used?" "Which by-products were generated by whom?"

The pilot ADRF contains data from the Illinois Department of Corrections, the Illinois Department of Employment Services, the Illinois Department of Human Services, and the New York City Center for Innovation through Data Intelligence, as well as Title 13 protected data, such as microdata from the American Community Survey and the Longitudinal Employer Household Dynamics program. About 100 employees from 40 different local, state, and federal agencies have learned to use the environment and analyzed the data jointly to document the postrelease employment of formerly incarcerated individuals, examine the postrelease returns to value-added education and training while incarcerated, predict recidivism, and investigate the value of place-based interventions.

**RECOMMENDATION 7-1 The recommended new entity for meeting the statistical needs of the nation should follow the principles and practices for federal statistical agencies and permit information accessed through it to be used for statistical purposes only.**

## Functions

### Approaches

One could create an entity that is simply an environment in which access is provided to data for statistical purposes. In this case, the entity would be staffed by data curators and experts in data merging, matching, and dataset construction, as well as experts in IT, cryptography, cybersecurity, and privacy regulations. However, the data analysis would be conducted by others who have been authorized to access data through the entity. This approach would focus the functions of the new entity on the minimum necessary to provide access to datasets and provide the services that users would need. Most of these would have minimal overlap with statistical agencies. In this approach, users would need relevant statistical and subject matter expertise to get useful results.

Alternatively, the entity could be a "full-service" research institute, and its staff would include not only those noted above, but also statisticians, economists, and other substantive experts who can analyze the data accessed through the entity and provide technical assistance to users. This approach would expand the functionality of the entity and provide more support to outside users. Potentially, the entity could also build staff capacity to produce statistical products or to provide products to external clients on a reimbursable basis. This functionality could supplement existing statistical agency capacity and capabilities, though it could also result in the entity providing services that were formerly contracted to the private sector.

There are many possibilities between these two ends of a continuum of functionality for the new entity. The primary advantage of creating an entity that provides only the minimal services necessary to provide an environment for accessing data is the limited scope and therefore simplicity of the mission. Such a scope would retain the expertise and independence of the federal statistical agencies while ensuring the entity operates as a service provider. This would serve to keep the entity tightly focused on specific issues related to data linkage, security, privacy, and access that apply to the entity itself. It would help address issues of access and would operate effectively when federal statistical agencies have expertise in their subject matter areas and are best equipped to examine and determine the quality of different data sources for their domains.

In contrast, a major advantage of a full-service entity is that it could provide more support and services to a number of statistical agencies (as well as outside researchers, as discussed below) in a variety of beneficial ways. For example, a researcher working for the new entity could develop considerable expertise with different administrative datasets and be able to conduct the analysis and provide the results a statistical agency needs,

rather than the agency having to invest the time it would take to train its staff to use the new data source. Some statistical agencies could contract with the entity to combine multiple existing data sources as they currently do for survey data collection and estimation. In this way, an agency could operate similarly as it currently does, but potentially be more effective and efficient with the resources and work of the new entity. A full-service entity would also have the potential to create more dynamic partnerships and collaborations both with academics and with researchers in the agencies and improve communication and application of research findings to a broad array of statistical programs.

A potential drawback of the full-service approach is that it could lead to the entity growing considerably and expanding beyond a service provider, taking on the role of a federal statistical program or agency itself. Another potential drawback is that the entity's work and research could be more attractive and interesting to various experts than the work and research at the statistical agency and so it would draw people from the agencies, especially if the entity can offer higher salaries than federal agencies, as would be the case for an FFRDC or university-based research center.

The panel believes that the optimal mix of services provided by the recommended new entity can best be determined by the federal statistical agencies and stakeholders, with attention to how their needs can best be met. However, we do recommend that the entity act as a service provider to federal statistical agencies.

**RECOMMENDATION 7-2 The recommended new entity should assist federal statistical agencies in identifying data sources that can most effectively inform the creation of national statistics, help develop techniques to use data from these sources to compute national statistics while respecting privacy and other protection obligations on the data, and nurture the expertise required to perform these functions.**

As described in more detail below, we also recommend a phased implementation plan for the new entity that would permit regular and recurring review of what functions the new entity can best perform for federal statistical agencies to tailor the scope over time to maximize advantages and minimize disadvantages.

### Technological Environment for Data Access

One key aspect of the IT environment needed for combining multiple data sources is whether there would be a single centralized system in one location storing data from multiple sources, a distributed system using machines across many locations, or a federated system in which data are

in multiple locations under the control of the original data sources' owners or intermediaries (see Chapter 3). To users, these systems may appear to be the same, with software in the distributed and federated environments performing the required pulls of data from multiple places as needed. In all these cases, data may be combined from multiple owners and from multiple locations to generate national statistics. The key difference between a federated and a distributed architecture is in the logical control and ownership: in a federated system, each member of the federation designs and owns its own data; in a distributed architecture, there is a single owner in charge.

The panel does not envision this new entity as a new data warehouse or national data center, as the privacy loss from a data breach can be ameliorated by not collecting and storing all the data in one place or by carefully partitioning and encrypting the data (see Chapters 3 and 5). For these reasons, we expect that a distributed or federated architecture for the proposed integration of multiple data sources would be a better approach than a centralized approach and would still address the issues of access for administrative data by federal statistical agencies.

From an engineering standpoint, there are multiple ways that one could design and build a new entity from an IT perspective, and these will change as technology changes. Ultimately, the IT infrastructure needs to be driven by the functions that the entity is intended to perform.

From a practical perspective, a completely federated model may not be practical for some data providers because their systems cannot easily be queried directly to obtain the necessary data, or the owner would prefer to provide an extract to the new entity on some periodic basis rather than permitting remote access. Because some data sources will be linked together, there may also be a need for a secure work space (whether physical or virtual) that contains the linked datasets rather than recreating the linkage every time an analysis is needed. Since some data sources change constantly, researchers may need to create a static extract or linked file that can be used for analysis.

Similar issues arise for reproducibility of research studies, which requires that the code and data be preserved, as well as with studies that involve longitudinal analyses that require the preservation of historical editions of a dataset. These issues need to be anticipated and appropriately handled by the entity.

The system design has implications not only for where the data are stored, but also for storing metadata. We anticipate that the data owners would most often be the best source for maintaining the most current and complete metadata. However, some datasets may not be well documented, and personnel working at or through the new entity may enhance the metadata, which would need to be stored by the new entity for potential future uses. As noted above, additional datasets may be created within the

new entity that need to have adequate provenance, including documenting linkage methods, data cleaning, and editing, which will also need to be retained by the new entity.

IT solutions currently exist to create an effective entity for combining multiple data sources from different owners for statistical purposes, and we repeat here our conclusion from Chapter 3:

**CONCLUSION 3-2 A range of possible computing environments could enable use of multiple data sources for statistics. Federal statistical agencies will need to consider the governance, functionality, and flexibility of a system, as well as the implications for protecting privacy and addressing data providers' concerns regarding privacy.**

### Access by Outside Researchers

As detailed in in the panel's first report (National Academies of Sciences, Engineering, and Medicine, 2017b), the broad use of federal statistical data through applied social science research and policy analysis has greatly benefited U.S. society. Broad access and statistical uses of data by external researchers is important because multiple investigations are often needed to evaluate the status of the economy or society. Having multiple teams using different strategies and challenging each other's findings is crucial to the scientific process. Moreover, investigators often generate important questions that otherwise would not have arisen. Therefore, the creation or designation of a new entity raises questions not only about how federal statistical agency staff would be able to access data through this new entity, but also whether and how external researchers would be able to similarly obtain access to data for statistical purposes.

There are currently a variety of approaches for external researchers to access and analyze data from a statistical agency: collaborating with agency staff who themselves conduct the analyses, becoming a research affiliate of the agency subject to legal restrictions of all employees, and applying for data access outside the agency on a project-by-project basis. Some agencies, such as Statistics of Income of the Internal Revenue Service, have active programs pairing agency staff with outside researchers for statistical studies using their data. Other agencies provide nonfederal researchers access as an affiliate through a fellowship program, such as the fellows programs managed by the National Science Foundation and the American Statistical Association for some federal statistical agencies, including the Census Bureau and the Bureau of Labor Statistics.[5] A number of agencies provide

[5]See https://www.amstat.org/ASA/Your-Career/ASA-Fellowships-and-Grants.aspx?hkey=7fa08de3-ecd5-4697-bd1c-b66470d21c9f [September 2017].

access for statistical purposes to a limited number of external researchers who apply on a project-by-project basis using a variety of arrangements.

Some agencies have provided online analysis systems for accessing their data, while others have used licensing arrangements that permit researchers to have access to the data at their own institutions, with legally binding agreements that describe the necessary security plans, inspections, and training required. FSRDCs and nongovernmental data enclaves are other options that provide either secure facilities or secure technological approaches to accessing microdata for approved statistical purposes.

Although there are many options, as described above and in the panel's first report, some researchers have noted that it can take a very long time to get approval for research projects and access to the data they need (Card et al., 2010). Although statistical agencies have strict protocols to ensure the confidentiality and appropriate use of their data, administrative applications and review processes can take considerably longer than necessary to meet those requirements. Such delays reduce the utility and use of federal statistical datasets for valuable research purposes.

As described above, the recommended new entity is intended to act as a service provider for the federal statistical agencies. Its services could include some related to handling requests for access by outside researchers, including review of research proposals, training in confidentiality requirements, and other procedures currently handled individually by the agencies whose data a researcher is seeking to access. Some or all of these tasks could be delegated to the new entity to implement on behalf of the agencies, reducing the burden on the statistical agencies and potentially imposing less burden on external researchers.

As stated above, the primary goal of the new entity should be to provide access to data held by federal statistical agencies. A number of issues will have to be addressed to permit appropriately designated staff from different agencies to access and analyze survey and administrative data from other agencies with the appropriate controls, oversight, privacy protections, and governance. There are currently variations in how these are implemented by different statistical agencies: the agencies and the U.S. Office of Management and Budget will need to consider either a common approach that would meet the needs of all the agencies or different tiers of requirements tailored to the restrictions tied to particular datasets (see Chapter 5). Once these procedures have been adequately worked out, consideration needs to be given to appropriately adapting them for external researchers.

**RECOMMENDATION 7-3 Statistical agencies and the recommended new entity should strive to provide federal agency researchers and external researchers access to data for exclusively statistical purposes, in a timely manner, in a way that is not administratively burdensome**

**and with strict adherence to confidentiality, privacy, and data security requirements.**

## Privacy

Access to data by federal statistical agency personnel and external researchers is predicated on the ability to adequately protect the privacy of the data and ensure that it is used for statistical purposes only. The panel's first report made clear that privacy protections must be at the forefront of the design and administration of the recommended new entity, using technological, statistical, and administrative approaches to secure data, along with up-to-date privacy-preserving and privacy-enhancing techniques. In Chapter 4 of this report, we reviewed the legal and computer science views of privacy and noted the implications for federal statistical agencies. Throughout our discussion of the entity, we have noted the fundamental importance of the legal framework protecting data for federal statistics and the restrictions on using these data for statistical purposes only, and we repeat the recommendation from Chapter 4:

**Recommendation 4-1 Because linked datasets offer greater privacy threats than single datasets, federal statistical agencies should develop and implement strategies to safeguard privacy while increasing accessibility to linked datasets for statistical purposes.**

We further elaborated in our first report how federal statistical agencies and the recommended entity will need to address both security threats and inference threats resulting from the use of multiple data sources. We noted that all federal agency IT systems are required to meet standards of the Federal Information Security Management Act of 2002,[6] but the panel's recommendations and suggestions likely exceed the current requirements in some areas. Federal statistical agencies also use a variety of inference control techniques. However, we noted in our first report that the techniques currently used do not provide a sufficient framework for addressing cumulative privacy loss or for using a privacy loss budget (see also Abowd, 2016; Abowd and Schmutte, 2017), nor can they circumvent the fundamental law of information reconstruction (National Academies of Sciences, Engineering, and Medicine, 2017b, Ch. 5). In addition, as we note in Chapter 5 of this report, staff with skills in cryptography and computer science will be needed to research and use new privacy-preserving and privacy-enhancing

[6]See https://www.gpo.gov/fdsys/pkg/STATUTE-116/pdf/STATUTE-116-Pg2899.pdf [August 2017].

techniques for survey and linked datasets, and we repeat the recommendation from that chapter:

> **RECOMMENDATION 5-1 Federal statistical agencies should ensure their technical staff receive appropriate training in modern computer science technology including but not limited to database, cryptography, privacy-preserving, and privacy-enhancing technologies.**

The recommended new entity could serve as a valuable center for coordinating research across the federal statistical system and the academic community on the application and evaluation of privacy-preserving and privacy-enhancing techniques for federal statistics. The entity would need to hire and continually train staff in state-of-the-art privacy protections. The environment of the entity and the data accessible through it should provide rich opportunities for exploring these issues, as well as providing opportunities to leverage expertise for the benefit of the entire federal statistical system.

### Transparency

For the recommended new entity to be sustainable, it will be critical that it acknowledges people's right to know how their data are being used and that the concerns of the public and data providers guide its practices. As we noted in our first report, transparency and continuously improving privacy protections will need to be the hallmark of the entity as threats to privacy and confidentiality can be expected to continuously evolve. Transparency will be fundamental to building the trust of those who provide data to the entity and those whose data may be accessed through the entity.

The new entity will need to carefully consider how best to communicate to the public useful information about its activities, the way data are accessed, the uses permitted of the data, and the privacy and data security protocols that the entity employs. The Administrative Data Research Network in the United Kingdom has made strides in this area that are worth consideration (see Box 7-2).

Federal agencies are currently required by the Privacy Act to publicly issue a notice for every system of records that they hold containing information covered by the act, describing the contents and permitted uses of that information. Agencies have also recently been required by the U.S. Office of Management and Budget (2015) to produce an inventory of their datasets; one use of these inventories is to review existing administrative data holdings for potential statistical uses (U.S. Office of Management and Budget, 2014a).

Whether or not the new entity is located in a federal agency, is an

**BOX 7-2**
**The Administrative Data Research Network in the United Kingdom**

The Administrative Data Research Network (ADRN) resulted from research commissioned by the Economic and Social Research Council and the Office for National Statistics to explore public understanding and views of administrative data, linking, and social research (Cameron et al., 2014). That research resulted in policy and communications recommendations that included explaining why the work is necessary, focusing on the societal benefits of social research, showing how projects have led to policy change or service improvement, and including public representation in the decision-making process. There is a listing on the ADRN website of all of the datasets that can be accessed through the network with a heading that states:

> **We do not hold any datasets.** A **personal service** is provided to all approved researchers, where the ADRN negotiates with government departments to make access to specific datasets available on a case-by-case basis.[a]

There are also FAQs that cover the network, the data, the researchers, linking data, the process, and privacy and security tailored to a general audience.[b] There is an approvals board that is composed of data holders, senior academics, an expert in data privacy and confidentiality, and lay members. This board reviews all research proposals planning to use the ADRN as a resource. There are also lay members on the overall governing board for the ADRN.

---

[a]See https://adrn.ac.uk/get-data/catalogue/ [August 2017].
[b]See https://adrn.ac.uk/public-engagement/pe-across-the-network/faq/ [September 2017].

FFRDC, or is a university-based public-private partnership, it is critical that the new entity strive for transparency in all of its activities. It will also need to give careful consideration to the best way to communicate with various audiences, including both its processes and the results of statistical programs and research projects that are conducted through the new entity.

**RECOMMENDATION 7-4 The recommended new entity should endeavor to maximize the transparency of its statistical activities by posting a summary of the data sources accessed through the entity on a public website. The summary should include the purpose and public benefit of the study, the data sources used, a brief description of the methodology, and links to resulting statistical products.**

As we discuss in Chapter 3, it will be important to provide provenance

for reproducing statistics and maintaining trust in federal statistics. In addition to providing external transparency, the new entity can also serve as a valuable scientific function, promoting the replication and reproducibility of statistics produced and the research conducted through the entity, as well as facilitating the creation of new statistics and research by maintaining metadata, code, and appropriate documentation for other users. This information would be retained within the entity and only accessible to those authorized to access the specific data sources given the potential risks to privacy or confidentiality that might be ascertained from this detailed documentation.

**RECOMMENDATION 7-5 The recommended new entity should strive to facilitate replicability of the linkage, processing, and analyses conducted through the entity by compiling and storing metadata and documentation for authorized data users.**

### Financing

Finding an appropriate ongoing funding source will be key to the sustainability of the recommended new entity. The main source of funding is also clearly linked to the organizational location of the entity. If the new entity is located in the federal government, then it would presumably either receive direct appropriations from Congress or receive funding outside of the congressional appropriations process, such as through the Federal Reserve Board of Governors. An FFRDC would need to be sponsored by an existing federal agency and would therefore require funding from that agency's budget.

This primary source of funding could be supplemented by additional reimbursable agreements with federal statistical or program agencies if the entity is legally permitted to enter into these arrangements. Whether the entity could charge fees of outside users (and retain those fees for its own use rather than turn them over to the U.S. Treasury) would also depend on the legal authority for the entity.

If the new entity is a public-private partnership based at a university, it could be funded by a federal statistical agency (or a consortium of agencies), or it might receive some funding from other federal agencies supporting scientific research, such as the National Science Foundation and the National Institutes of Health. Similar to the arrangements noted above, supplemental funding could be obtained through reimbursable agreements with federal statistical or program agencies and charging fees of outside users.

In the panel's first report, we noted that the new entity would not take over federal statistical agency programs or authorities nor draw heavily on

the current limited federal statistical system resources. We expect the new entity to result in more cost-effective statistical programs and eventual cost efficiencies, but up-front investments will be needed to build the infrastructure for the new entity; establish agreements for accessing useful datasets; conduct research on the quality of the data sources for specific statistical purposes; develop statistical methods for linking, combining, and analyzing data from multiple sources; and develop techniques for preserving and enhancing privacy and confidentiality while permitting statistical uses. It will be essential to keep a longer term perspective in mind when considering the entity's financing: initial investments will pay dividends in better and more useful federal statistics and information for the country, as well as more cost-effective programs in the future.

The panel recognizes that seeking additional funding for a new entity in the current fiscal environment will not be easy. As we describe below, we propose a phased implementation of the new entity so that it can demonstrate its value and utility to a wide range of stakeholders and build support for additional funding for continuing and then expanding. The private sector is well aware of the value of data (Manyika et al., 2011), and some state and local governments have provided clear examples of the growing value of the ability to analyze and integrate large volumes of data (Fantuzzo and Culhane, 2015). Applying this same value proposition to federal statistics would allow better information not only for policy makers, but also for businesses, researchers, and the public and would further encourage and bolster state and local government efforts.

### Governance

In some sense, the governance of the entity will be driven by the location of the organization and the authorizing legislation. If the recommended entity is created and established as a federal agency or unit, one would expect it to be run by a director, who reports to the umbrella agency or department and is also accountable to Congress. One would also expect an FFRDC or a university-based public-private partnership to be led by a director accountable to the funding agency or university or a board of directors.

However, given the mission and nature of the recommended new entity, consideration should be given to additional structures and mechanisms for governance of the entity. Because its role as a service provider to federal statistical agencies is a fundamental rationale for existing, it is essential that the federal statistical agencies have a strong role in governing its activities. As we describe above, there are a range of functions and activities that the entity might conceivably adopt, and these may evolve by expanding or contracting over time depending on the needs of the federal statistical agencies.

Given the decentralized nature of the federal statistical system, the structure needs to ensure that input is obtained from all of the statistical agencies and that the entity fairly addresses their needs.

The recommended new entity will also serve and have responsibilities to data providers and data users. Although strong authority is needed for the new entity to be able to obtain data from different programs, it does not imply that strong partnerships are not needed with the program agencies. The entity needs to be not only a strong steward of any data that are accessed through the entity, but also ensure that its staff, federal researchers, and external researchers working with the program agencies' data provide useful feedback about the properties of the data and have an ongoing dialogue with program agencies about improvements.

It will also be important for internal and external researchers to be able to examine the aspects of the operation of agency programs and potential effects of the program through the data and linking to other data sources to help improve the effectiveness and efficiency of the program, as well as to provide policy makers with valuable information to inform decisions. Indeed, integrated data systems created by some cities and states are directed primarily at improving the efficiency and effectiveness of services being provided. To realize these benefits, it is vital that research access be provided in a reasonable manner while meeting all necessary requirements (as we recommend above), and researchers should have a voice in governance to ensure the entity is fulfilling its obligations in this regard.

**RECOMMENDATION 7-6 The director of the recommended new entity should report to a board of directors that includes representatives of the federal statistical agencies, experts on privacy, holders of data used in the entity, and users of statistical data.**

As we stress throughout this report, privacy is fundamental to the operation and sustainability of the recommended entity. Because of the diverse perspectives on privacy that need to be considered, the entity and the federal statistical system could benefit from regular discussions and advice in this domain. Furthermore, because data linkage may raise concerns from the public about privacy, efforts should be made to illustrate the benefits of the analyses of linked data. The recommended new entity should engage in ongoing dialogue with people and groups whose data are being analyzed and strive to develop case studies for which data linkages can improve data subjects' lives or the economy.

**RECOMMENDATION 7-7 The recommended new entity should have an advisory committee on privacy to inform and advise the federal statistical system on policies and current best practices. The advisory committee should include privacy advocates, data users, and members**

**of the public whose data may be accessed, as well as experts from statistics, computer science, and the legal profession.**

Finally, because the entity will serve federal statistical agencies in providing information for the public good and uphold principles and practices for a federal statistical agency (see Recommendation 7-1, above), it should have strong authority to ensure the integrity of its statistical operations.

**RECOMMENDATION 7-8 The legal foundation of the recommended new entity should foster independence from political and other undue external influence in providing access to data, linking and analyzing data, and producing and disseminating statistical information.**

## IMPLEMENTATION

As we note in Chapter 1, there may be concerns about creating a new entity that would provide greater access to data at a time of heightened privacy concerns over data breaches and potential misuse of data. Therefore, the data accessed through the entity will need to evolve over time with careful oversight and demonstrated results. A strategic plan will be needed to describe milestones for expanding the data sources accessible through the entity. This plan will need to be carefully structured in phases, detailing outcomes for each phase and decision point. The first phase might cover 5 years, at which time it would be useful to have a comprehensive review. Further expansion of the entity's access and capabilities will then be predicated on successful stewardship during the first phase and demonstrated benefits for federal statistics. In this way, stakeholders can ensure that the entity is serving its intended purposes and that any concerns are being adequately addressed.

The first phase needs, at a minimum, to include a broader statistical use of data collected and acquired by one federal statistical agency by other statistical agencies than is currently done. Access to specific datasets will need to be controlled on a project-approved basis, but the uses do not necessarily need to be limited to a single project or a single statistical agency. This access would include survey data collected by the statistical agency as well as federal administrative data acquired by the agency. Currently, a federal statistical agency may only be able to access administrative data from a state or another federal agency for one statistical program in its portfolio even though other programs could benefit: such expanded access would further improve the cost efficiency of the agency and the utility of its statistical products. And other statistical agencies, with the same legal protections and requirements for safeguarding the privacy and confidentiality of the data and similarly secured computing environments, also cannot

currently access those data for their statistical programs. As we noted in our first report, the country can no longer afford these costly restrictions, and we noted that legal or administrative changes may be needed to change this situation (National Academies of Sciences, Engineering, and Medicine, 2017b, Ch. 6).

The first phase also needs to include expanded access to federal administrative and operational data that could be useful for federal statistics. For example, the Census Bureau has arrangements with a number of federal program agencies to obtain access to their data for statistical purposes. Other statistical agency programs would benefit from the same secure access to these same sources for statistical purposes. These arrangements would also need to include administrative data required for the administration of federal programs that is collected and owned by the states, such as the Supplemental Nutrition Assistance Program. In addition, it would also be valuable for data from other federal programs that produce administrative records that could provide useful statistical information for the country to be accessible through the new entity.

As we described in our first report, states and local governments also have other administrative data that have the potential to be used to provide valuable statistics for the country, and federal statistical agencies have made important steps in using some of these sources that should continue. However, these data might best be considered for the new entity in the second phase, after an evaluation can be made of the uses of federal administrative data for federal statistics. We expect that expanded data sharing with states will take more planning and strategic efforts than required for federal data sharing, including identifying appropriate incentives for states and local governments to provide access to their administrative data. There are a variety of arrangements that currently provide mutual benefits to the states and federal statistical agencies—such as the Longitudinal Employer Household Data system (see description in National Academies of Sciences, Engineering, and Medicine, 2017b, Ch. 3)—and we assume these arrangements will continue. However, given the potential greater complexity and additional concerns that might accompany including these efforts in the new entity, the panel believes they would be more appropriate for the second phase of implementation.

Similarly, while the panel anticipates that some private-sector data will ultimately be part of the portfolio of the new entity (and statistical agencies are currently exploring some sources), we believe these data could be included as part of the new entity in a later phase. There are a wide variety of types of data available from private-sector sources, and these will further need to be prioritized in terms of their likely utility for federal statistics that would be most beneficial for the country. Some sources, such as scanner data and some credit card transactions, are currently being used by some

statistical agencies, and we assume this work will continue. More broadly, private firms cannot provide the objective national statistics currently produced by the federal statistical agencies, although we think that some private-sector data sources could contribute to enhancing the timeliness and geographic detail of some federal statistics. However, a good deal of research and development will be needed to evaluate and use these sources for federal statistics. As noted in our first report, there are also other fundamental issues with private-sector data sources that will need to be addressed, as well as with the public-private partnerships or other arrangements that agencies enter into with private firms to access their data (see National Academies of Sciences, Engineering, and Medicine, 2017b, Ch. 4).

As we noted in our first report, the creation of a new entity will not by itself solve the many challenges facing the federal statistical system. As detailed above, the authority and mission of the recommended new entity will need to be clearly delineated. How this entity is created and how it functions will determine its ability to be an effective resource of and for the federal statistical system.

We describe above the advantages and disadvantages of determining the location, functions, and other attributes of the recommended new entity, and there are many ways forward that would benefit federal statistics and the country. All possibilities have strengths and weaknesses and what might be optimal depends on the weight given to different factors. We can envision viable entities being created by giving greater independence and authority to CARRA or a statistical agency or by creating a new entity at a university through a public-private partnership or a new FFRDC. Each arrangement poses some slightly different challenges and requirements that will need to be addressed. What is most important is that the key stakeholders embrace a viable approach and work together to create it and make it successful. We believe the broad federal statistical system welcomes the opportunities to innovate and is eager to work with the broad community of stakeholders to address the challenges ahead.

# References

Abayomi, K., Gelman, A., and Levy, M. (2008). Diagnostics for multivariate imputations. *Applied Statistics*, *57*(3), 273–291.

Abowd, J.M. (2016). *Why Statistical Agencies Need to Take Privacy-Loss Budgets Seriously, and What It Means When They Do*. Available: http://digitalcommons.ilr.cornell.edu/ldi/32 [May 2017].

Abowd, J.M., and Schmutte, I. (2017). *Revisiting the Economics of Privacy: Population Statistics and Confidentiality Protection as Public Goods*. Available: https://www2.census.gov/ces/wp/2017/CES-WP-17-37.pdf [August 2017].

Abowd, J.M., and Stinson, M.H. (2013). Estimating measurement error in annual job earnings: A comparison of survey and administrative data. *The Review of Economics and Statistics, 95*(5), 1451–1467. Available: http://www.mitpressjournals.org/doi/pdf/10.1162/REST_a_00352 [July 2017].

Abowd, J.M., Alvisi, L., Dwork, C., Kannan, S., Machanavajjhala, A., and Reiter, J. (2017). *Privacy-Preserving Data Analysis for the Federal Statistical Agencies*. Available: https://arxiv.org/ftp/arxiv/papers/1701/1701.00752.pdf [July 2017].

Administrative Data Research Network. (2017). *Security Standards*. Available: https://adrn.ac.uk/provide-data/standards-governance/security-standards/ [August 2017].

Akavia, A., Goldwasser, S., and Vaikuntanathan, V. (2009). *Simultaneous Hardcore Bits and Cryptography against Memory Attacks*. Available: https://people.csail.mit.edu/vinodv/AGV-proc.pdf [May 2017].

Altman, M., Wood, A., O'Brien, D.R., Vadhan, S., and Gasser, U. (2015). Toward a modern approach to privacy-aware government data releases. *Berkeley Technology Law Journal, 30*(3), 1967–2072.

Apple, Inc. (2016). *Apple Worldwide Developer's Conference (WWDC16), Keynote Address by Craig Federighi*. Available: https://www.apple.com/apple-events/june-2016 [June 2017].

Baldwin, E., Johnson, K., Berthoud, H., and Dublin, S. (2015). Linking mothers and infants within electronic health records: A comparison of deterministic and probabilistic algorithms. *Pharmacoepidemiology and Drug Safety, 24*(1), 45–51.

Barocas, S., and Nissenbaum, H. (2014). Big data's end run around anonymity and consent. In J. Lane, V. Stodden, S. Bender, and H. Nissenbaum (Eds.), *Privacy, Big Data, and the Public Good: Frameworks for Engagement* (pp. 44–75). New York: Cambridge University Press.

Belin, T.R., and Rubin, D.B. (1995). A method for calibrating false-match rates in record linkage. *Journal of the American Statistical Association, 90*(430), 694–707.

Bellow, M. (2007). Improved County-Level Estimation of Crop Yield Using Model-Based Methodology with a Spatial Component. Available: https://ww2.amstat.org/meetings/ices/2007/proceedings/ICES2007-000110.PDF [September 2017].

Benedetto, G., Stinson, M.H., and Abowd, J.M. (2013). *The Creation and Use of the SIPP Synthetic Beta*. Available: https://www.census.gov/content/dam/Census/programs-surveys/sipp/methodology/SSBdescribe_nontechnical.pdf [July 2017].

Bianchi, R.F. (2011). *The Challenges of Consolidating Federal Government Information Systems and the Genesis of Scope*. Available: https://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.50/2011/wp.12.e.pdf [June 2017].

Biderman, A.D., and Lynch, J.P. (1991). *Understanding Crime Incidence Statistics: Why the UCR Diverges from the NCS*. New York: Springer-Verlag.

Biemer, P. (2016). Errors and inference. In I. Foster, R. Ghani, R.S. Jarmin, F. Kreuter, and J. Lane (Eds.), *Big Data and Social Science: A Practical Guide to Methods and Tools* (pp. 265–298). Boca Raton, FL: CRC Press.

Bondarenko, I., and Raghunathan, T. (2016). Graphical and numerical diagnostic tools to assess suitability of multiple imputations and imputation models. *Statistics in Medicine*, *35*(17), 3007–3020.

Boneh, D., Sahai, A., and Waters, B. (2010). *Functional Encryption: Definitions and Challenges*. Available: https://eprint.iacr.org/2010/543.pdf [April 2017].

Borowik, J., Henden, M., and Fraser, B. (2012). *Riding the Big Data Wave to Streamline Acquiring and Managing Data*. Available: https://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.50/2012/15_Australia.pdf [August 2017].

Box, G.E.P. (1979). *Robustness in the Strategy of Scientific Model Building*. Available: http://oai.dtic.mil/oai/oai?verb=getRecord&metadataPrefix=html&identifier=ADA070213 [July 2017].

Brackstone, G. (1987). *Statistical Issues of Administrative Data: Issues and Challenges*. Presented at Statistical Uses of Administrative Data—An International Symposium, Statistics Canada, November 23–25.

Brickell, J., and Shmatikov, V. (2008). *The Cost of Privacy: Destruction of Data-Mining Utility in Anonymized Data Publishing*. Available: https://www.cs.cornell.edu/~shmat/shmat_kdd08.pdf [May 2017].

Brummet, Q., Masterton, M., and Smith, D. (2014). *Evaluation of Commercial School and Teacher Lists to Enhance Survey Frames*. CARRA Working Paper 2014-07. Washington, DC: U.S. Census Bureau. Available: https://www.census.gov/content/dam/Census/library/working-papers/2014/adrm/carra-wp-2014-07.pdf [August 2017].

Bucholtz, S. (2015). *Matching the American Housing Survey to Tax Assessment Data: Some Preliminary Results*. Available: http://sites.nationalacademies.org/cs/groups/dbassesite/documents/webpage/dbasse_171490.pdf [February 2017].

Bureau of Justice Statistics. (2014). *National Crime Victimization Survey: Technical Documentation*. NCJ-247252. Available: https://www.bjs.gov/content/pub/pdf/ncvstd13.pdf [May 2017].

Bureau of Justice Statistics. (2016). *NCVS Design: Subnational*. Available: www.bjs.gov/index.cfm?ty=tp&tid=911 [May 2017].

Bureau of Labor Statistics. (2012). *Geographic Profile of Employment and Unemployment*. Available: www.bjs.gov/index.cfm?ty=tp&tid=911 [May 2017].

Bureau of Labor Statistics. (2015). *Geographic Profile of Employment and Unemployment, 2014*. Available: www.bls.gov/opub/gp/laugp.htm [May 2017].

Bureau of Labor Statistics. (2017a). *Handbook of Methods*. Available: https://www.bls.gov/opub/hom [May 2017].

Bureau of Labor Statistics. (2017b). *The Employment Situation—December 2016*. News Release. Available: https://www.bls.gov/news.release/archives/empsit_01062017.pdf [May 2017].

Calandrino, J.A., Kilzer, A., Narayanan, A., Felten, E., and Shmatikov, V. (2011). "You might also like:" Privacy risks of collaborative filtering. In *Proceedings of the 2011 IEEE Symposium on Security and Privacy* (pp. 231–246). Washington, DC: IEEE Computer Society. doi:10.1109/SP.2011.40.

Cameron, D., Pope, S., and Clemence, M. (2014). *Dialogue on Data*. Available: http://www.esrc.ac.uk/files/public-engagement/public-dialogues/dialogue-on-data-exploring-the-public-s-views-on-using-linked-administrative-data-for-research-purposes [May 2017].

Canetti, R., Raghuraman, S., Richelson, S., and Vaikuntanathan, V. (2017). Chosen-ciphertext secure fully homomorphic encryption. In *Public-Key Cryptography—PKC 2017* (pp. 213–240). Berlin/Heidelberg, Germany: Springer.

Card, D., Chetty, R., Feldstein, M., and Saez, E. (2010). *Expanding Access to Administrative Data for Research in the United States*. Arlington, VA: National Science Foundation. Available: https://www.nsf.gov/sbe/sbe_2020/submission_detail.cfm?upld_id=112 [December 2017].

Carson, E.A. (2015). *Linking Administrative BJS Data: Better Understanding of Prisoners' Personal Histories by Linking the National Corrections Reporting Program (NCRP) and CARRA Data*. Available: https://s3.amazonaws.com/sitesusa/wp-content/uploads/sites/242/2016/03/A1_Carson_2015FCSM.pdf [August 2017].

Cavallo, A., and Rigobon, R. (2016). *The Billion Prices Project: Using Online Prices for Measurement and Research*. NBER Working Paper Series, Working Paper 22111. Cambridge, MA: National Bureau of Economic Research. Available: http://www.nber.org/papers/w22111 [November 2016].

Centers for Disease Control and Prevention. (2016). *Behavioral Risk Factor Surveillance System. Weighting BRFSS Data: BRFSS 2015*. Available: https://www.cdc.gov/brfss/annual_data/2015/pdf/weighting_the-data_webpage_content.pdf [February 2017].

Chatterjee, N., Chen, Y.H., Maas, P., and Carroll, R.J. (2016). Constrained maximum likelihood estimation for model calibration using summary-level information from external big data sources. *Journal of the American Statistical Association, 111*(513), 107–117.

Christen, P. (2012). *Data Matching: Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection*. New York: Springer.

Citro, C. (2014). From multiple modes for surveys to multiple data sources for estimates. *Survey Methodology, 40*(2), 137–161.

Cohen, J.W., Cohen, S.B., and Banthin, J.S. (2009). The Medical Expenditure Panel Survey: A national information resource to support healthcare cost research and inform policy and practice. *Medical Care, 47*(7), S44–S50.

Cohen, S.B., and Cohen, J.W. (2013). The capacity of the Medical Expenditure Panel Survey to inform the Affordable Care Act. *Inquiry, 50*(2), 124–134.

Conk, M.A. (1980). *The United States Census and Labor Force Change: A History of Occupation Statistics, 1870–1940*. Ann Arbor: University of Michigan Research Press.

Contreras, J.L., and Reichman, J.H. (2015). Sharing by design: Data and decentralized commons. *Science, 350*(6266), 1312–1314.

Cruze, N.B. (2015). Integrating survey data with auxiliary sources of information to estimate crop yields. In *Proceedings of the Survey Research Methods Section* (pp. 565–578). Washington, DC: American Statistical Association.

Cynamon, M., and Blumberg, S. (2016). *National Health Interview Survey Questionnaire Redesign*. Available: http://www.copafs.org/UserFiles/file/NHISRedesignforCOPAFSrev.pdf [February 2017].

Davies, C. (2009). *Area Frame Design for Agricultural Surveys*. Available: https://www.nass.usda.gov/Publications/Methodology_and_Data_Quality/Advanced_Topics/AREA%20FRAME%20DESIGN.pdf [February 2017].

DeFrances, C., Anand, K., Williams, S., and Woodwell, D. (2012). *Designing the National Hospital Care Survey*. Available: http://www.amstat.org/meetings/ices/2012/papers/302080.pdf [April 2017].

Diffie, W., and Hellman, M.E. (1976). New directions in cryptography. *IEEE Transactions on Information Theory, 22*(6), 644–654.

Dillman, D. (1996). Why innovation is difficult in government surveys. *Journal of Official Statistics, 12*(2), 113–124.

Dinur, I., and Nissim, K. (2003). *Revealing Information while Preserving Privacy*. Available: http://www.cse.psu.edu/~ads22/privacy598/papers/dn03.pdf [November 2016].

Dwork, C. (2006). Differential privacy. In *Proceedings of the 33rd International Conference on Automata, Languages and Programming, Part II (ICALP)* (pp. 1–12). Berlin/Heidelberg, Germany: Springer-Verlag. doi:10.1007/11787006_1. Available: http://research.microsoft.com/pubs/64346/dwork.pdf [November 2016].

Dwork, C., and Mulligan, D. (2014). *Differential Privacy in Practice: Expose Your Epsilons!* Paper presented at Privacy Law Scholars Conference (PLSC 2014). Cambridge, MA: Harvard University.

Dwork, C., McSherry, F., Nissim, K., and Smith, A. (2006). Calibrating noise to sensitivity in private data analysis. In *TCC'06 Proceedings of the Third Conference on Theory of Cryptography* (pp. 265–284). Berlin/Heidelberg, Germany: Springer-Verlag. doi:10.1007/11681878_14.

Dwork, C., McSherry, F., and Talwar, K. (2007). The price of privacy and the limits of LP decoding. In *Proceedings of the 39th ACM Symposium on Theory of Computing* (pp. 85–94). New York: ACM. Available: http://dl.acm.org/citation.cfm?id=1250804 [November 2016].

Dwork, C., Smith, A., Steinke, T., Ullman, J., and Vadhan, S. (2015). Robust traceability from trace amounts. In *IEEE 56th Annual Symposium Foundations of Computer Science (FOCS)* (pp. 650–669). Berkeley, CA: IEEE. Available: http://privacytools.seas.harvard.edu/files/privacytools/files/robust.pdf?m=1445278897 [August 2017]. doi:10.1109/FOCS.2015.46.

Dwork, C., Smith, A., Steinke, T., and Ullman, J. (2017). Exposed! A survey of attacks on private data. *Annual Review of Statistics and Its Application, 4*(1), 61–84.

Elliott, M.R., and Valliant, R. (2017). Inference for nonprobability samples. *Statistical Science*, *32*(2), 249–264.

Eltinge, J.L., Biemer, P.P., and Holmberg, A. (2013). A potential framework for integration of architecture and methodology to improve statistical production systems. *Journal of Official Statistics, 29*(1), 125–145.

Erlingsson, Ú., Pihur, V., and Korolova, A. (2014). RAPPOR: Randomized Aggregatable Privacy-Preserving Ordinal Response. In *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security* (pp. 1054–1067). New York: Association for Computing Machinery. doi:10.1145/2660267.2660348.

European Commission. (2011). *European Statistics Code of Practice*. Brussels, Belgium: European Statistical System Committee. Available: http://ec.europa.eu/eurostat/documents/3859598/5921861/KS-32-11-955-EN.PDF/5fa1ebc6-90bb-43fa-888f-dde032471e15 [August 2017].

European Statistical System Committee. (2013). *Quality Assurance Framework*. Available: http://ec.europa.eu/eurostat/documents/64157/4392716/ESS-QAF-V1-2final.pdf/bbf5970c-1adf-46c8-afc3-58ce177a0646 [June 2017].

Fantuzzo, J., and Culhane, D.P. (2015). *Actionable Intelligence: Using Integrated Data Systems to Achieve a More Effective, Efficient, and Ethical Government*. New York: Palgrave Macmillan.

Federal Committee on Statistical Methodology. (1978). *An Error Profile: Employment as Measured by the Current Population Survey*. Statistical Policy Working Paper 3. Available: https://s3.amazonaws.com/sitesusa/wp-content/uploads/sites/242/2014/04/spwp3.pdf [July 2017].

Federal Committee on Statistical Methodology. (2001). *Measuring and Reporting Sources of Error in Surveys*. Statistical Policy Working Paper 31. Available: https://s3.amazonaws.com/sitesusa/wp-content/uploads/sites/242/2014/04/spwp31.pdf [July 2017].

Federal Committee on Statistical Methodology. (2006). *Report on Statistical Disclosure Limitations Methodology*. Statistical Policy Working Paper 22. Available: https://fcsm.sites.usa.gov/files/2014/04/spwp22.pdf [November 2016].

Fellegi, I.P. (1972). On the question of statistical confidentiality. *Journal of the American Statistical Association*, *67*(337), 7–18.

Fellegi, I.P. (1999). Keynote address: Record linkage and public policy—a dynamic evolution. In *Record Linkage Techniques—1997: Proceedings of an International Workshop and Exposition* (Chapter 1). Washington, DC: The National Academies Press. Available: https://www.nap.edu/read/6491/chapter/4 [February 2017].

Fellegi, I.P., and Sunter, A.B. (1969). A theory for record linkage. *Journal of the American Statistical Association, 64*(328), 1183–1210.

Fosdick, B.K., DeYoreo, M., and Reiter, J.P. (2016). Categorical data fusion using auxiliary information. *Annals of Applied Statistics*, *10*(4), 1907–1929.

Froomkin, A.M. (2016). *Privacy Issues with Sensor Data Collection*. Available: http://sites.nationalacademies.org/cs/groups/dbassesite/documents/webpage/dbasse_173124.pdf [February 2017].

Fulton, J.A. (2012). *Respondent Consent to Use Administrative Data*. Available: http://drum.lib.umd.edu/handle/1903/13601 [February 2017].

Genkin, D., Shamir, A., and Tromer, E. (2014). *RSA Key Extraction via Low-Bandwidth Acoustic Cryptanalysis*. Available: http://www.cs.tau.ac.il/~tromer/papers/acoustic-20131218.pdf [May 2017].

Gentry, C. (2009). *A Fully Homomorphic Encryption Scheme*. Available: https://crypto.stanford.edu/craig/craig-thesis.pdf [November 2016].

Golden, C., Driscoll, A.K., Simon, A.E., Judson, D.H., Miller, E.A., and Parker., J.D. (2015). Linkage of NCHS population health surveys to administrative records from Social Security Administration and Centers for Medicare & Medicaid Services. *Vital and Health Statistics, 1*(58), 1–53.

Goldreich, O. (2004). *Foundations of Cryptography* (Vol. 2). Cambridge, UK: Cambridge University Press.

Goldreich, O., Micali, S., and Wigderson, A. (1987). How to play any mental game. In *STOC '87 Proceedings of the Nineteenth Annual ACM Symposium on Theory of Computing* (pp. 218–229). New York: Association for Computing Machinery. doi: 10.1145/28395.28420.

Goldwasser, S., and Micali, S. (1982). Probabilistic encryption & how to play mental poker keeping secret all partial information. In *STOC '82 Proceedings of the Fourteenth Annual ACM Symposium on Theory of Computing* (pp. 365–377). New York: Association for Computing Machinery. doi:10.1145/800070.802212.

Goldwasser, S., and Rothblum, G. (2010). Securing computation against continuous leakage. In T. Rabin (Ed.), *Advances in Cryptology—CRYPTO 2010* (pp. 59–79). Berlin/Heidelberg, Germany: Springer. doi:10.1007/978-3-642-14623-7_4.

Goldwasser, S., Micali, S., and Rivest, R.L. (1988). A digital signature scheme secure against adaptive chosen-message attacks. *SIAM Journal on Computing, 17*(2), 281–308.

Gorina, Y., Pratt, L.A., Kramarow, E.A., and Elgaddal, N. (2015). *Hospitalization, Readmission, and Death Experience of Noninstitutionalized Medicare Fee-for-Service Beneficiaries Aged 65 and Over*. Available: https://permanent.access.gpo.gov/gpo62706/nhsr084.pdf [February 2017].

Groen, J. (2012). Sources of error in survey and administrative data: The importance of reporting procedures. *Journal of Official Statistics, 28*(2), 173–198.

Groves, R.M., and Lyberg, L. (2010). Total survey error: Past, present, and future. *Public Opinion Quarterly, 74*(5), 849–879.

Groves, R.M., and Schoeffel, G. (in press). Use of administrative records in evidence-based policymaking. *The ANNALS of the American Academy of Political and Social Science*.

Groves, R.M., Fowler, F., Couper, M., Singer, E., and Tourangeau, R. (2009). *Survey Methodology* (2nd ed.). New York: Wiley.

Gueron, S., Lindell, Y., Nof, A., and Pinkas, B. (2015). *Fast Garbling of Circuits under Standard Assumptions*. Available: https://eprint.iacr.org/2015/751.pdf [May 2017].

Habermann, H. (2010). Future of innovation in the Federal Statistical System. *The ANNALS of the American Academy of Political and Social Science, 631*(1), 194–203.

Hall, P.H., and George, V. (2016). *2015 Salary Survey of Business, Industry, and Government Statisticians*. Available: http://www.amstat.org/asa/files/pdfs/YCR-SPAIGsalarysurvey15.pdf [May 2017].

Haneuse, S., and Wakefield, J. (2007). Geographic-based ecological correlation studies using supplemental case-control data. *Statistics in Medicine, 27*(6), 864–887.

Harron, K., Wade, A., Gilbert, R., Muller-Pebody, B., and Goldstein, H. (2014). Evaluating bias due to data linkage error in electronic healthcare records. *BMC Medical Research Methodology, 14*(1), 36. doi:10.1186/1471-2288-14-36.

Harron, K., Goldstein, H., and Dibben, C. (2015). *Methodological Developments in Data Linkage*. Hoboken, NJ: John Wiley & Sons.

Hedrick, S., and Weister, T. (2016). *2015 American Community Survey (ACS) 1-Year Estimates*. Available: https://www.census.gov/content/dam/Census/newsroom/press-kits/2016/20160908_acs_1year_webinar.pdf [May 2017].

Herzog, T.N., Scheuren, F.J., and Winkler, W.E. (2007). *Data Quality and Record Linkage Techniques*. New York: Springer Science+Business Media.

Hof, M.H.P., and Zwinderman, A.H. (2012). Methods for analyzing data from probabilistic linkage strategies based on partially identifying variables. *Statistics in Medicine, 31*(30), 4231–4242.

Holt, T. (2007). The official statistics Olympic challenge: Wider, deeper, quicker, better, cheaper (with discussions). *The American Statistician, 61*(1), 1–8. doi:10.1198/000313007X168173.

Homer, N., Szelinger, S., Redman, M., Duggan, D., Tembe, W., Muehling, J., Pearson, J., Stephan, D., Nelson, S., and Craig, D. (2008). Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays. *PLOS Genetics*. Available: http://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1000167 [November 2016].

Horrigan, M. (2013a). *Big Data: A Perspective from the BLS*. Available: http://magazine.amstat.org/blog/2013/01/01/sci-policy-jan2013 [July 2017].

Horrigan, M. (2013b). *Big Data and Official Statistics: International Year of Statistics November 13, 2013*. Washington, DC: Bureau of Labor Statistics. Available: https://www.bls.gov/osmr/symp2013_horrigan.pdf [July 2017].

Hsieh, Y.P., and Murphy, J. (2017). Total Twitter error. In P.P. Biemer, E. de Leeuw, S. Eckman, B. Edwards, F. Kreuter, L.E. Lyberg, N.C. Tucker, and B.T. West (Eds.), *Total Survey Error in Practice* (pp. 23–46). Hoboken, NJ: Wiley.

Hu, S.S., Balluz, L., Battaglia, M.P., and Frankel, M.R. (2011). Improving public health surveillance using a dual-frame survey of landline and cell phone numbers. *American Journal of Epidemiology, 173*(6), 703–711.

Hundepool, A., Domingo-Ferrer, J., Franconi, L., Giessing, S., Nordholt, E.S., Spicer, K., and de Wolf, P.P. (2012). *Statistical Disclosure Control*. Chichester, UK: John Wiley & Sons.

Iwig, W., Berning, M., Marck, P., and Prell, M. (2013). *Data Quality Assessment Tool for Administrative Data*. Available: http://www.bls.gov/osmr/datatool.pdf [August 2017].

Japec, L., Kreuter, F., Berg, M., Biemer, P., Decker, P., Lampe, C., Lane, J., O'Neil, C., and Usher, A. (2015). *AAPOR Task Force Report on Big Data*. Oakpark Terrace, IL: American Association for Public Opinion Research. Available: http://www.aapor.org/Education-Resources/Reports/Big-Data.aspx [August 2017].

Jones, M.R. (2016). *Measuring the Effects of the Tipped Minimum Wage Using W-2 Data*. CARRA Working Paper 2016-03. Washington, DC: U.S. Census Bureau. Available: https://www.census.gov/content/dam/Census/library/working-papers/2016/adrm/carra-wp-2016-03.pdf [February 2017].

Judson, D.H., Parker, J.D., and Larsen, M.D. (2013). *Adjusting Sample Weights for Linkage-Eligibility Using SUDAAN*. Hyattsville, MD: National Center for Health Statistics. Available: http://www.cdc.gov/nchs/data/datalinkage/adjusting_sample_weights_for_linkage_eligibility_using_sudaan.pdf [July 2017].

Karr, A.F., and Reiter, J. (2014). Using statistics to protect privacy. In J. Lane, V. Stodden, S. Bender, and H. Nissenbaum (Eds.), *Privacy, Big Data, and the Public Good: Frameworks for Engagement* (pp. 276–295). New York: Cambridge University Press.

Kasiviswanathan, S.P., Rudelson, M., and Smith, A. (2013). *The Power of Linear Reconstruction Attacks*. Proceedings of the 45th Annual ACM Symposium on the Theory of Computing, Palo Alto, CA, June 2–4. Available: https://arxiv.org/pdf/1210.2381v1.pdf [November 2016].

Kim, J.K., and Shao, J. (2013). *Statistical Methods for Handling Incomplete Data*. Boca Raton, FL: CRC Press.

Kinney, S.K., Reiter, J.P., Reznek, A.P., Miranda, J., Jarmin, R.S., and Abowd, J.M. (2011). *Towards Unrestricted Public Use Business Microdata: The Synthetic Longitudinal Business Database*. Available: https://www.census.gov/ces/pdf/CES-WP-11-04.pdf [August 2017].

Kleppmann, M. (2017). *Designing Data-Intensive Applications: The Big Ideas Behind Reliable, Scalable, and Maintainable Systems*. Sebastopol, CA: O'Reilly Media.

Kocher, P.C. (1996). *Timing Attacks on Implementations of Diffie-Hellman, RSA, DSS, and Other Systems*. Available: http://www.cse.msstate.edu/~ramkumar/TimingAttacks.pdf [May 2017].

Kocher, P., Jaffe, J., and Jun, B. (1998). *Differential Power Analysis*. Available: http://www.cse.msstate.edu/~ramkumar/DPA.pdf [May 2017].

Kosar, K.R. (2011). *The Quasi Government: Hybrid Organizations with Both Government and Private Sector Legal Characteristics*. Available: https://fas.org/sgp/crs/misc/RL30533.pdf [July 2017].

Kraus, R. (2013). Statistical déjà vu: The National Data Center proposal of 1965 and its descendants. *Journal of Privacy and Confidentiality, 5*(1), 1–37.

Kreuter, F., Sakshaug, J.W., and Tourangeau, R. (2016). The framing of the record linkage consent question. *International Journal of Public Opinion Research, 28*(1), 142–152.

Lahiri, P., and Larsen, M.D. (2005). Regression analysis with linked data. *Journal of the American Statistical Association, 100*(469), 222–230.

Langton, L., and Fay, R.E. (2016). *Developmental Efforts of Subnational Crime Rates Based on the National Crime Victimization Survey*. Available: http://www.jrsa.org/webinars/presentations/ncvs-subnational-est.pdf [February 2017].

Langton, L., Planty, M., and Lynch, J.P. (2017). The second major redesign of the National Crime Victimization Survey. *Criminology & Public Policy, 16*(4), 1049–1074.

Lariscy, J.T. (2011). Differential record linkage by Hispanic ethnicity and age in linked mortality studies: Implications for the epidemiologic paradox. *Journal of Aging and Health, 23*(8), 1263–1284.

Laudon, K.C., and Laudon, J.P. (2017). *Management Information Systems: Managing the Digital Firm* (15th ed.). London, UK: Pearson.

Levant, S., Chari, K., and DeFrances, C. (2016). *National Hospital Care Survey Demonstration Projects: Traumatic Brain Injury.* National Health Statistics Reports, No. 97. Available: http://www.cdc.gov/nchs/data/nhsr/nhsr097.pdf [August 2016].

Li, J., Diallo, M.A., and Fay, R.E. (2017). *Rethinking the NCVS: Small Area Estimation Approaches to Estimating Crime*. Available: https://www.researchgate.net/publication/266463975_Rethinking_the_NCVS_Small_Area_Estimation_Approaches_to_Estimating_Crime [August 2017].

Li, N., Li, T., and Venkatasubramanian, S. (2007). T-closeness: Privacy beyond k-anonymity and *l*-diversity. In *2007 IEEE 23rd International Conference on Data Engineering* (pp. 106–115). doi:10.1109/ICDE.2007.367856. Available: http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=4221659 [April 2017].

Lohr, S. (2011). Alternative survey sample designs: Sampling with multiple overlapping frames. *Survey Methodology, 37*(2), 197–213.

Lohr, S., and Raghunathan, T. (2017). Combining survey data with other sources. *Statistical Science*, *32*(2), 293–312.

Lynch, J.P., and Addington, L.A. (2007). *Understanding Crime Statistics: Revisiting the Divergence of the UCR and NCVS.* New York: Cambridge University Press.

Machanavajjhala, A., Gehrke, J., and Kifer, D. (2006). *Diversity: Privacy Beyond k-Anonymity*. Available: http://www.cs.uml.edu/~ge/pdf/papers_685-2-2/ldiversity-icde06.pdf [April 2017].

Machanavajjhala, A., Kifer, D., Abowd, J., Gehrke, J., and Vilhuber, L. (2008). *Privacy: Theory Meets Practice on the Map*. Available: http://www.cse.psu.edu/~duk17/papers/PrivacyOnTheMap.pdf [July 2008].

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., and Byers, A.H. (2011). *Big Data: The Next Frontier for Innovation, Competition, and Productivity.* McKinsey & Company, Digital McKinsey, May. Available: http://www.mckinsey.com/business-functions/digital-mckinsey/our-insights/big-data-the-next-frontier-for-innovation [December 2017].

Martin, J. (1981). *Design and Strategy for Distributed Data Processing*. Englewood Cliffs, NJ: Prentice-Hall.

McClure, D., and Reiter, J.P. (2012). Differential privacy and statistical disclosure risk measures: An investigation with binary synthetic data. *Transactions on Data Privacy, 5*(3), 535–552.

Miller, E.A., Decker, S.L., and Parker, J.D. (2016). Characteristics of Medicare advantage and fee-for-service beneficiaries upon enrollment in Medicare at age 65. *The Journal of Ambulatory Care Management, 39*(3), 231–241.

Miller, E.A., McCarty, F.A., and Parker, J.D. (2017). Racial and ethnic differences in a linkage with the National Death Index. *Ethnicity & Disease*, 27(2), 77–84.

Miller, R. (2014). *Big Data Curation*. Available: http://comad.in/comad2014/Proceedings/Keynote2.pdf [June 2017].

Moriarity, C., and Scheuren, F. (2001). Statistical matching: A paradigm for assessing the uncertainty in the procedure. *Journal of Official Statistics*, *17*(3), 407–422.

Muthukrishnan, S., and Nikolov, A. (2012). Optimal private halfspace counting via discrepancy. In *Proceedings of the Forty-Fourth Annual ACM Symposium on Theory of Computing* (pp. 1285–1292). New York: ACM. Available: http://dl.acm.org/citation.cfm?id=2214090&dl=ACM&coll=DL&CFID=698830182&CFTOKEN=71109411 [November 2016].

Narayanan, A., and Shmatikov, V. (2008). Robust de-anonymization of large sparse datasets (how to break anonymity of the Netflix prize dataset). In *Proceedings of the 2008 IEEE Symposium on Security and Privacy* (pp. 111–125). Washington, DC: IEEE Computer Society. doi:10.1109/SP.2008.33. Available: https://www.cs.cornell.edu/~shmat/shmat_oak08netflix.pdf [November 2016].

National Academies of Sciences, Engineering, and Medicine. (2016a). *Modernizing Crime Statistics: Report 1: Defining and Classifying Crime*. Committee on National Statistics. J. Lauritsen and D. Cork (Eds.). Panel on Modernizing the Nation's Crime Statistics, Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press.

National Academies of Sciences, Engineering, and Medicine. (2016b). *Statistical Challenges in Assessing and Fostering the Reproducibility of Scientific Results: Summary of a Workshop*. M. Schwalbe (Rapporteur). Committee on Applied and Theoretical Statistics. Division on Engineering and Physical Sciences. Washington, DC: The National Academies Press.

National Academies of Sciences, Engineering, and Medicine. (2017a). *Improving Crop Estimates by Integrating Multiple Data Sources*. Panel on Methods for Integrating Multiple Data Sources to Improve Crop Estimates, M.E. Bock and N. Kirkendall (Eds.). Committee on National Statistics, Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press.

National Academies of Sciences, Engineering, and Medicine. (2017b). *Innovations in Federal Statistics: Combining Data Sources While Protecting Privacy*. Panel on Using Multiple Data Sources and State-of-the-Art Estimation Methods in Federal Statistics: Frameworks, Methods, and Assessment, R.M. Groves and B.A. Harris-Kojetin (Eds.). Committee on National Statistics, Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press.

National Academies of Sciences, Engineering, and Medicine. (2017c). *Principles and Practices for a Federal Statistical Agency: Sixth Edition*. Committee on National Statistics. C. Citro (Ed.). Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press.

National Center for Education Statistics. (1994). *Quality Profile for SASS: Aspects of the Quality of Data in the Schools and Staffing Surveys (SASS)*. Available: https://nces.ed.gov/pubs94/94340.pdf [August 2017].

National Center for Health Statistics. (2012). *Linkages between Survey Data from the National Center for Health Statistics and Medicare Program Data from the Centers for Medicare & Medicaid Services*. Available: http://www.cdc.gov/nchs/data/datalinkage/cms_medicare_methods_report_final.pdf [February 2017].

National Center for Science and Engineering Statistics. (2015). *Doctorate Recipients from U.S. Universities: 2013*. Available: https://www.nsf.gov/statistics/sed/2013/digest/nsf15304a.pdf [May 2017].

National Research Council. (1984). *Cognitive Aspects of Survey Methodology: Building a Bridge between Disciplines*. Committee on National Statistics. T.B. Jabine, M.L. Straf, J.M. Tanur, and R. Tourangeau (Eds.). Division of Behavioral and Social Sciences and Education. Washington, DC: National Academy Press. doi:10.17226/930.

National Research Council. (2005). *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. doi:10.17226/11434.

National Research Council. (2007). *Understanding Business Dynamics: An Integrated Data System for America's Future*. Panel on Measuring Business Formation, Dynamics, and Performance, J. Haltiwanger, L.M. Lynch, and C. Mackie (Eds.). Committee on National Statistics, Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press. doi:10.17226/11844.

National Research Council. (2011). *Facilitating Innovation in the Federal Statistical System: Summary of a Workshop*. H. Habermann (rapporteur). Committee on National Statistics, Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press. doi:10.17226/13168.

National Research Council. (2013a). *Measuring What We Spend: Toward a New Consumer Expenditure Survey*. Panel on Redesigning the BLS Consumer Expenditure Surveys, D.A. Dillman and C.C. House (Eds.). Committee on National Statistics, Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press. doi:10.17226/13520.

National Research Council. (2013b). *Nonresponse in Social Science Surveys: A Research Agenda*. Panel on a Research Agenda for the Future of Social Science Data Collection, R. Tourangeau and T.J. Plewes (Eds.). Committee on National Statistics, Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press.

National Research Council. (2014a). *Capturing Change in Science, Technology, and Innovation: Improving Indicators to Inform Policy*. Panel on Developing Science, Technology, and Innovation Indicators for the Future, R.E. Litan, A.W. Wyckoff, and K.H. Fealing (Eds.). Committee on National Statistics, Division of Behavioral and Social Science and Education. Board on Science, Technology, and Economic Policy, Division of Policy and Global Affairs. Washington, DC: The National Academies Press. doi:10.17226/18606.

National Research Council. (2014b). *Civic Engagement and Social Cohesion: Measuring Dimensions of Social Capital to Inform Policy*. Panel on Measuring Social and Civic Engagement and Social Cohesion in Surveys, K. Prewitt, C.D. Mackie, and H. Habermann (Eds.). Committee on National Statistics, Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press. doi:10.17226/18831.

Nealon, J., and Gleaton, E. (2013). Consolidation and standardization of survey operations at a decentralized federal statistical agency. *Journal of Official Statistics*, *29*(1), *5*–28. doi:10.2478/jos-2013-0002. Available: https://www.degruyter.com/downloadpdf/j/jos.2013.29.issue-1/jos-2013-0002/jos-2013-0002.pdf [June 2017].

Newcombe, H.B., Kennedy, J.M., Axford, S.J., and James, A.P. (1959). Automatic linkage of vital records. *Science, 130*(3381), 954–959.

Nissim, K., Bembenek, A., Wood, A., Bun, M., Gaboardi, M., Gasser, U., O'Brien, D.R., Steinke, T., and Vadhan, S. (2016). *Bridging the Gap between Computer Science and Legal Approaches to Privacy*. Available: https://privacytools.seas.harvard.edu/files/privacytools/files/bridging_cs_law_privacy.pdf [July 2017].

Oberski, D.L., Kirchner, A., Eckman, S., and Kreuter, F. (2017). Evaluating the quality of survey and administrative data with generalized multitrait-multimethod models. *Journal of the American Statistical Association*. doi:10.1080/01621459.2017.1302338.

OECD. (2016). *Research Ethics and New Forms of Data for Social and Economic Research*. OECD Science, Technology and Industry Policy Papers, No. 34. Paris: OECD. doi:10.1787/5jln7vnpxs32-en. Available: http://www.oecd-ilibrary.org/docserver/download/5jln7vnpxs32-en.pdf?expires=1503241383&id=id&accname=guest&checksum=9C34738238F81374246539588B536F12 [August 2017].

O'Hara, A. (2016). *Preliminary Research for Replacing or Supplementing the Income Question on the American Community Survey with Administrative Records*. Available: https://www.census.gov/content/dam/Census/library/working-papers/2016/acs/2016_Ohara_01.pdf [August 2017].

Özsu, M.T., and Valduriez, P. (2011). *Principles of Distributed Database Systems* (3rd Edition). New York: Springer Science+Business Media, LLC. doi:10.1007/978-1-4419-8834-8.

Parsons, V.L., Moriarty, C., Jonas, K., Moore, T.F., Davis, K.E., and Tompkins, L. (2015). *Design and Estimation for the National Health Interview Survey, 2006–2015*. Hyattsville, MD: U.S. Department of Health and Human Services. Available: https://www.cdc.gov/nchs/data/series/sr_02/sr02_165.pdf [May 2017].

Permutt, T. (2016). Sensitivity analysis for missing data in regulatory submissions. *Statistics in Medicine*, *35*(17), 2876–2879. doi:10.1002/sim.6753.

Planty, M., and Langton, L. (2014). National Crime Victimization Survey Subnational Program—An update. *JRSA Forum, 32*, 1–3.

President's Council of Advisors on Science and Technology. (2014). *Big Data and Privacy: A Technological Perspective*. Available: https://bigdatawg.nist.gov/pdf/pcast_big_data_and_privacy_-_may_2014.pdf [August 2017].

Raghunathan, T.E. (2015). *Missing Data Analysis in Practice*. Boca Raton, FL: CRC Press.

Raghunathan, T.E., Diehr, P.K., and Cheadle, A.D. (2003). Combining aggregate and individual level data to estimate an individual level correlation coefficient. *Journal of Educational and Behavioral Statistics*, *28*(1), 1–19.

Ramaprasan, A. (2015). Record linkage with Washington State Cancer Registry. *Journal of Patient-Centered Research and Reviews*, *2*(2), 117–118. doi:10.17294/2330-0698.1144.

Reiter, J.P. (2012). Statistical approaches to protecting confidentiality for microdata and their effects on the quality of statistical inferences. *Public Opinion Quarterly, 76*(1), 163–181.

Reiter, J.P., and Raghunathan, T.E. (2007). *The Multiple Adaptations of Multiple Imputation*. Available: http://www2.stat.duke.edu/~jerry/Papers/jasa07.pdf [June 2017].

Ristenpart, T., Tromer, E., Shacham, H., and Savage, S. (2009). *Hey, You, Get Off of My Cloud: Exploring Information Leakage in Third-Party Compute Clouds*. Available: https://cseweb.ucsd.edu/~hovav/dist/cloudsec.pdf [April 2017].

Rivest, R.L., Shamir, A., and Adleman, L. (1978). *A Method for Obtaining Digital Signatures and Public-Key Cryptosystems*. Available: https://web.williams.edu/Mathematics/lg5/302/RSA.pdf [August 2017].

Rodgers, W.L. (1984). An evaluation of statistical matching. *Journal of Business & Economic Statistics*, *2*(1), 91–102.

Rubin, D.B. (1993). Discussion: Statistical disclosure limitation. *Journal of Official Statistics, 9*(2), 462–468.

Saslow, E. (2012). "Jobs Day." Monthly release of employment data: An economic, political obsession. *The Washington Post*, March 9. Available: https://www.washingtonpost.com/national/jobs-day-an-economic-and-political-obsession/2012/03/09/gIQADZPW1R_story.html [November 2016].

Schenker, N., and Raghunathan, T.E. (2007). Combining information from multiple surveys to enhance estimation of measures of health. *Statistics in Medicine, 26*(8), 1802–1811.

Schmidlin, K., Clough-Gorr, K.M., and Spoerri, A. (2015). Privacy Preserving Probabilistic Record Linkage (P3RL): A novel method for linking existing health-related data and maintaining participant confidentiality. *BMC Medical Research Methodology, 15*(1), 46.

Skinner, C.J. (2009). Statistical disclosure control for survey data. In D. Pfeffermann and C. Rao (Eds.), *Handbook of Statistics—Sample Surveys: Design, Methods and Applications* (vol. 29, pp. 381–396). Oxford, UK: North-Holland.

Skinner, C.J., and Wakefield, J. (2017). Introduction to the design and analysis of complex survey data. *Statistical Science*, *32*(2), 165–175.

Smuk, M., Carpenter, J.R., and Morris, T.P. (2017). What impact do assumptions about missing data have on conclusions? A practical sensitivity analysis for a cancer survival registry. *BMC Medical Research Methodology*, *17*(21), 1–7.

Statistics Canada. (2016). *Approved Record Linkages*. Available: http://www.statcan.gc.ca/eng/record/summ [May 2017].

Statistics Canada. (2017). *Directive on Microdata Linkage*. Available: http://www.statcan.gc.ca/eng/record/policy4-1 [August 2017].

Steorts, R.C., Hall, R., and Fienberg, S.E. (2016). A Bayesian approach to graphical record linkage and deduplication. *Journal of the American Statistical Association, 111*(516), 1660–1672.

Struijs, P., Camstra, A., Renssen, R., and Braaksma, B. (2013). Redesign of statistics production within an architectural framework: The Dutch experience. *Journal of Official Statistics, 29*(1), 49–71. doi:10.2478/jos-2013-0004. Available: https://www.degruyter.com/downloadpdf/j/jos.2013.29.issue-1/jos-2013-0004/jos-2013-0004.pdf [August 2017].

Sweeney, L. (1997). Weaving technology and policy together to maintain confidentiality. *The Journal of Law, Medicine & Ethics, 25*(2–3), 98–110.

Sweeney, L., Crosas, M., and Bar-Sinai, M. (2015). *Sharing Sensitive Data with Confidence: The Datatags System*. Available: https://techscience.org/a/2015101601 [June 2017].

Tancredi, A., and Liseo, B. (2011). A hierarchical Bayesian approach to record linkage and population size problems. *The Annals of Applied Statistics, 5*(2B), 1553–1585.

Tanur, J.M. (1999). Looking backwards and forwards at the CASM movement. In M.G. Sirken, D.J. Herrmann, S. Schechter, N. Schwarz, J.M. Tanur, and R. Tourangeau (Eds.), *Cognition and Survey Research* (pp. 13–19). New York: Wiley.

Turn, R., and Ware, W.H. (1976). *Privacy and Security Issues in Information Systems*. Available: https://www.rand.org/content/dam/rand/pubs/papers/2008/P5684.pdf [November 2016].

U.N. Economic and Social Council. (2014). *Report of the Global Group on Big Data for Official Statistics*. Available: https://unstats.un.org/unsd/statcom/47th-session/documents/2016-6-Big-data-for-official-statistics-E.pdf [August 2017].

U.N. Economic Commission for Europe. (2014). *A Suggested Framework for the Quality of Big Data*. Available: https://statswiki.unece.org/download/attachments/108102944/Big%20Data%20Quality%20Framework%20-%20final-%20Jan08-2015.pdf?version=1&modificationDate=1420725063663&api=v2 [August 2017].

U.N. Economic Commission for Europe. (2015). *Common Statistical Production Architecture*. Available: https://statswiki.unece.org/pages/viewpage.action?pageId=120128614 [September 2017].

U.N. General Assembly. (2014). *Fundamental Principles of Official Statistics*. Available: https://unstats.un.org/unsd/dnss/gp/FP-New-E.pdf [June 2017].

U.S. Census Bureau. (1996). *American Housing Survey: A Quality Profile*. Available: https://www.census.gov/content/dam/Census/programs-surveys/ahs/publications/h12195-1.pdf [August 2017].

U.S. Census Bureau. (1998). *SIPP Quality Profile*. SIPP Working Paper 230. Available: https://www.census.gov/sipp/workpapr/wp230.pdf [August 2017].

U.S. Census Bureau. (2006). *Current Population Survey Design and Methodology*. Technical Paper 66. Available: http://www.census.gov/prod/2006pubs/tp-66.pdf [May 2017].

U.S. Census Bureau. (2008). *Factfinder for the Nation: Availability of Census Records About Individuals*. Available: www.census.gov/history/pdf/cff2.pdf [May 2017].

U.S. Census Bureau. (2012). *2010 Census Summary File 1: Technical Documentation*. Available: www.census.gov/prod/cen2010/doc/sf1.pdf [May 2017].

U.S. Census Bureau. (2014a). *American Community Survey Design and Methodology*. Available: https://www.census.gov/programs-surveys/acs/methodology/design-and-methodology.html [August 2017].

U.S. Census Bureau. (2014b). *U.S. Merchandise Trade Statistics: A Quality Profile*. Available: https://www.census.gov/foreign-trade/aip/quality_profile10032014.pdf [August 2017].

U.S. Census Bureau. (2017). *National Health Interview Survey CAPI Manual for Field Representatives*. Available: ftp://ftp.cdc.gov/pub/Health_Statistics/NCHS/Survey_Questionnaires/NHIS/2017/frmanual.pdf [August 2017].

U.S. Department of Agriculture. (2016). *National Household Food Acquisition and Purchase Survey (FoodAPS): User's Guide to Design, Data Collection, and Overview of Datasets*. Available: https://www.ers.usda.gov/media/8591/userguidefoodaps.pdf [August 2017].

U.S. Department of Health, Education, and Welfare. (1973). *Records, Computers, and the Rights of Citizens*. Report of the Secretary's Advisory Committee on Automated Personal Data Systems. Available: https://www.justice.gov/opcl/docs/rec-com-rights.pdf [August 2017].

U.S. Government Accountability Office. (2015). *Human Capital: Update on Strategic Management Challenges for the 21st Century*. GAO-15-619T. Available: http://www.gao.gov/assets/680/670294.pdf [August 2017].

U.S. Government Accountability Office. (2017). *High-Risk Series: Progress on Many High-Risk Areas, While Substantial Efforts Needed on Others*. GAO-17-317. Available: http://www.gao.gov/assets/690/682765.pdf [May 2017].

U.S. Government Accounting Office. (2001). *Record Linkage and Privacy: Issues in Creating New Federal Research and Statistical Information*. GAO-01-126SP. Available: http://www.gao.gov/assets/210/201699.pdf [February 2017].

U.S. Government Accounting Office. (2014). *Computer Matching Act: OMB and Selected Agencies Need to Ensure Consistent Implementation*. GAO-14-44. Available: https://www.gao.gov/assets/670/660140.pdf [August 2017].

U.S. Office of Management and Budget. (2006). *Standards and Guidelines for Statistical Surveys*. Available: https://obamawhitehouse.archives.gov/sites/default/files/omb/inforeg/statpolicy/standards_stat_surveys.pdf.

U.S. Office of Management and Budget. (2007). *Implementation Guidance for Title V of the E-Government Act, Confidential Information Protection and Statistical Efficiency Act of 2002 (CIPSEA)*. Available: https://obamawhitehouse.archives.gov/sites/default/files/omb/assets/omb/fedreg/2007/061507_cipsea_guidance.pdf [November 2016].

U.S. Office of Management and Budget. (2014a). *Guidance for Providing and Using Administrative Data for Statistical Purposes*. Memorandum M-14-06. Available: https://obamawhitehouse.archives.gov/sites/default/files/omb/memoranda/2014/m-14-06.pdf [August 2017].

U.S. Office of Management and Budget. (2014b). *Statistical Policy Directive No. 1: Fundamental Responsibilities of Federal Statistical Agencies and Recognized Statistical Units*. Available: https://www.gpo.gov/fdsys/pkg/FR-2014-12-02/pdf/2014-28326.pdf [August 2017].

U.S. Office of Management and Budget. (2015). *Management and Oversight of Federal Information Technology*. Memorandum M-15-14. Available: https://obamawhitehouse.archives.gov/sites/default/files/omb/memoranda/2015/m-15-14.pdf [August 2017].

U.S. Office of Management and Budget. (2016). *Managing Information as a Strategic Resource*. Circular No. A-130. Available: https://obamawhitehouse.archives.gov/sites/default/files/omb/assets/OMB/circulars/a130/a130revised.pdf [August 2017].

U.S. Office of Management and Budget. (2017). *Statistical Programs of the United States Government: Fiscal Year 2017*. Available: https://obamawhitehouse.archives.gov/sites/default/files/omb/assets/information_and_regulatory_affairs/statistical-programs-2017.pdf [August 2017].

Vale, S. (2009). *Generic Statistical Business Process Model*. Available: www1.unece.org/stat/platform/download/attachments/8683538/GSBPM+Final.pdf [July 2017].

van Buuren, S. (2012). *Flexible Imputation of Missing Data*. Boca Raton, FL: CRC Press.

Ver Ploeg, M., Mancino, L., Todd, J.E., Clay, D.M., and Scharadin, B. (2015). *Where Do Americans Usually Shop for Food and How Do They Travel to Get There? Initial Findings from the National Household Food Acquisition and Purchase Survey*. Economic Information Bulletin Vol. 138. Washington, DC: U.S. Department of Agriculture.

Wagner, D., and Layne, M. (2014). *The Person Identification Validation System (PVS): Applying the Center for Administrative Records Research and Applications' (CARRA) Record Linkage Software*. CARRA Working Paper 2014-01. Washington, DC: U.S. Census Bureau. Available: https://www.census.gov/content/dam/Census/library/working-papers/2014/adrm/carra-wp-2014-01.pdf [July 2016].

Wallgren, A., and Wallgren, B. (2007). *Register-Based Statistics: Administrative Data for Statistical Purposes*. Hoboken, NJ: Wiley.

Wang, X., Chan, T.-H., and Shi, E. (2015). Circuit ORAM: On tightness of the Goldreich-Ostrovsky lower bound. In *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security* (pp. 850–861). New York: Association for Computing Machinery.

Warner, S.L. (1965). Randomized response: A survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association, 60*(309), 63–69.

Weissman, J., Parker, J.D., Miller, D.M., Miller, E.A., and Gindi, R.M. (2016). The relationship between linkage refusal and selected health conditions of survey respondents. *Survey Practice*, *9*(5). Available: http://www.surveypractice.org/index.php/SurveyPractice/article/view/342 [February 2017].

Xiao, X., and Tao, Y. (2007). M-invariance: Towards privacy preserving re-publication of dynamic datasets. In *Proceedings of the 2007 ACM SIGMOD International Conference on Management of Data* (pp. 689–700). New York: Association for Computing Machinery. doi:10.1145/1247480.1247556.

Zhang, G., Parker, J.D., and Schenker, N. (2016). Multiple imputation for missingness due to nonlinkage and program characteristics: A case study of the National Health Interview Survey linked to Medicare claims. *Journal of Survey Statistics and Methodology, 4*(3), 319–338.

Zhang, L.-C. (2012). Topics of statistical theory for register-based statistics and data integration. *Statistica Neerlandica, 66*(1), 41–63.

# Appendix A

# Executive Summary from *Innovations in Federal Statistics: Combining Data Sources While Protecting Privacy*

Federal government statistics provide critical information to the country and serve a key role in a democracy. For decades, sample surveys with instruments carefully designed for particular data needs have been one of the primary methods for collecting data for federal statistics. However, the costs of conducting such surveys have been increasing while response rates have been declining, and many surveys are not able to fulfill growing demands for more timely information and for more detailed information at state and local levels.

The Panel on Improving Federal Statistics for Policy and Social Science Research Using Multiple Data Sources and State-of-the-Art Estimation Methods was charged to conduct a study to foster a paradigm shift in federal statistical programs that would use combinations of diverse data sources from government and private-sector sources in place of a single census, survey, or administrative records source. This first report discusses the challenges faced by the federal statistical system and the foundational elements needed for a new paradigm.

In addition to surveys, some federal statistics are derived from government administrative records, that is, data collected by government entities for program administration, regulatory, or law enforcement purposes. Because these administrative records exist, there is interest in using them much more—both alone and in combination with surveys—to try to enhance the quality, scope, and cost-efficiency of statistical products and to reduce response burden on the public.

Not enough is known about the quality of these new sources of data, and considerable work is required to assess their usefulness for producing

statistics. Some may be useful as is; others may require scrubbing or statistical transformation. Furthermore, for statistical purposes, it may be necessary to combine or blend multiple data sources, which is more complex than working with a single dataset. However, there are statistical methods and models for combining information from multiple data sources.

Some administrative records held by federal agencies are prohibited from being shared among agencies. And for some records held by states and localities, there is no mandate and limited incentive to share them with federal statistical agencies.

**CONCLUSION 3-4** Legal and administrative barriers limit the statistical use of administrative datasets by federal statistical agencies.

**CONCLUSION 3-5** State and local governments may respond to incentives from the federal government to provide access to their administrative data by federal statistical agencies for statistical purposes.

**RECOMMENDATION 3-1** Federal statistical agencies should systematically review their statistical portfolios and evaluate the potential benefits and risks of using administrative data. To this end, federal statistical agencies should create collaborative research programs to address the many challenges in using administrative data for federal statistics.

Large amounts of private-sector data—such as credit card transactions, scanner data, cell phone data, and Internet searches—are generated for commercial use. These sources hold the potential to improve the timeliness and level of detail of national statistics. These data are extremely diverse, and there are many issues of access, quality, and usability that would have to be addressed to consider them for federal statistical use.

**RECOMMENDATION 4-1** Federal statistical agencies should systematically review their statistical portfolios and evaluate the potential benefits of using private-sector data sources.

**RECOMMENDATION 4-2** The Federal Interagency Council on Statistical Policy should urge the study of private-sector data and evaluate both their potential to enhance the quality of statistical products and the risks of their use. Federal statistical agencies should provide annual public reports of these activities.

Any consideration of expanding the use of data must have privacy as a core value. Federal privacy laws have established clear limitations on the

collection and use of personally identifiable information, and statistical agencies have a strong tradition of data confidentiality and stewardship. Nonetheless, data breaches pose real risks to the public. As federal statistical agencies seek to combine multiple datasets, they need to simultaneously address how to control risks from privacy breaches. Privacy-enhancing techniques and privacy-preserving statistical data analysis can be valuable in these efforts and enable the use of private-sector and other alternative data sources for federal statistics.

**RECOMMENDATION 5-1** Statistical agencies should engage in collaborative research with academia and industry to continuously develop new techniques to address potential breaches of the confidentiality of their data.

**RECOMMENDATION 5-2** Federal statistical agencies should adopt modern database, cryptography, privacy-preserving, and privacy-enhancing technologies.

In the decentralized U.S. statistical system, there are 13 agencies whose mission is primarily the creation and dissemination of statistics and more than 100 agencies that engage in statistical activities. However, there is currently no agency directly charged with facilitating access to and the use of multiple data sources for the benefit of the entire statistical system. There is a need for stronger coordination and collaboration to enable access to and evaluation of administrative and private-sector data sources for federal statistics.

**RECOMMENDATION 6-1** A new entity or an existing entity should be designated to facilitate secure access to data for statistical purposes to enhance the quality of federal statistics.

Privacy protections would have to be fundamental to the mission of this entity.

**CONCLUSION 6-1** For the proposed new entity to be sustainable, the data for which it has responsibility would need to have legal protections for confidentiality and be protected, using the strongest privacy protocols offered to personally identifiable information while permitting statistical use.

**RECOMMENDATION 6-2** The proposed new entity should maximize the utility of the data for which it is responsible while protecting pri-

vacy by using modern database, cryptography, privacy-preserving, and privacy-enhancing technologies.

There are many questions about how the entity would function and who would be able to access data for statistical purposes. The panel's second report will examine organizational models for a new entity, quality frameworks for multiple data sources, statistical techniques for combining data from multiple sources, privacy-enhancing and privacy-preserving techniques, as well as the information technology implications for implementing a new paradigm that would combine diverse data sources.

# Appendix B

# Biographical Sketches of Panel Members and Staff

**Robert M. Groves** (*Chair)* is the provost, Gerard Campbell professor in the Department of Mathematics and Statistics, and a professor in the Department of Sociology, all at Georgetown University. His research focuses on the effects of the mode of data collection on responses in sample surveys, the social and political influences on survey participation, the use of adaptive research designs to improve the cost and error properties of statistics, and how public concerns about privacy affect attitudes toward statistical agencies. Previously, he served as director of the U.S. Census Bureau, director of the University of Michigan Survey Research Center, and research professor at the Joint Program in Survey Methodology at the University of Maryland. He is an elected member of the National Academy of Sciences, the National Academy of Medicine, the American Academy of Arts and Sciences, and the International Statistical Institute and an elected fellow of the American Statistical Association. His 1989 book, *Survey Errors and Survey Costs*, was named one of the 50 most influential books in survey research by the American Association of Public Opinion Research. He has a bachelor's degree from Dartmouth College, master's degrees in statistics and sociology from the University of Michigan, and a doctorate in sociology from the University of Michigan.

**Michael E. Chernew** is a professor of health care policy in the Department of Health Care Policy at Harvard Medical School. He is also a research associate of the National Bureau of Economic Research. His research examines areas related to controlling health care spending growth while maintaining or improving the quality of care, including consumer incentives to

align patient cost sharing with clinical value. Related research examines the effects of changes in Medicare Advantage payment rates, as well as the causes and consequences of rising health care spending and geographic variation in spending, spending growth, and quality. He is a member of the Medicare Payment Advisory Commission, an independent agency that advises Congress. He is a recipient of the John D. Thompson Prize for Young Investigators given by the Association of University Programs in Public Health and of the Alice S. Hersh Young Investigator Award from the Association of Health Services Research. He has a B.A. from the University of Pennsylvania and a Ph.D. in economics from Stanford University.

**Piet Daas** is a senior methodologist in the Department of Corporate Services, Information Technology, and Methodology and a data scientist in the Center for Big Data Statistics of Statistics Netherlands. His work focuses on the use of secondary (nonsurvey) data for official statistical purposes, which began with the use of administrative data, and more recently has focused on studies in which Internet and other big data sources are used for official statistics. At Statistics Netherlands, he is a member of the big data core team, which oversees all big data activities of production, information technology, research, management, and training. He teaches the big data component of the European Master of Official Statistics track at the University of Utrecht, is involved in the big data courses of the European Statistical Training Programme, and is a member of the team organizing DataCamps ('hackatons') at the University of Twente. He is active in various European, U.N., and U.N. Economic Commission for Europe big data initiatives. He has an M.S. and a Ph.D. in the natural sciences with honors from the University of Nijmegen in The Netherlands.

**Cynthia Dwork**, on leave from Microsoft Research, is the Gordon McKay professor of computer science at the John A. Paulson School of Engineering and Applied Sciences and a Radcliffe alumnae professor at the Radcliffe Institute for Advanced Study, both at Harvard University. Her work focuses on placing privacy-preserving data analysis on a mathematically rigorous foundation: a cornerstone of this work is differential privacy, a strong privacy guarantee frequently permitting highly accurate data analysis. She also does work in cryptography and distributed computing, including work on the first public-key cryptosystem for which breaking a random instance is as hard as solving the hardest instance of the underlying mathematical problem on combating e-mail spam by requiring a proof of computational effort (the technology that underlie hashcash and bitcoin). She is a recipient of the PET Award for Outstanding Research in Privacy Enhancing Technologies given by Microsoft and of the Edsger W. Dijkstra Prize, awarded jointly by the ACM Symposium on Principles of Distributed Computing of

the Association for Computing Machinery and the European Association for Theoretical Computer Science Symposium on Distributed Computing. She is a member of the National Academy of Sciences and the National Academy of Engineering and a fellow of the American Academy of Arts and Sciences. She has a B.S.E. from Princeton University and a Ph.D. in computer science from Cornell University.

**Ophir Frieder** is the Robert L. McDevitt, K.S.G., K.C.H.S. and Catherine H. McDevitt, L.C.H.S., chair in computer science and information processing at Georgetown University. He is also professor of biostatistics, bioinformatics, and biomathematics in the Georgetown University Medical Center and the chief scientific officer for UMBRA Health Corporation. He previously served as chair of the Department of Computer Science at Georgetown University. His research interests focus on scalable information retrieval systems spanning search and retrieval and communications issues in multiple domains, systems that are deployed worldwide in commercial and governmental production environments. He is a fellow of the American Association for the Advancement of Science, the Association for Computing Machinery, the Institute of Electrical and Electronics Engineers, and the National Academy of Inventors.

**Brian Harris-Kojetin** (*Study Director*) is the director of the Committee on National Statistics (CNSTAT) and served as the study director for this project. Previously, he served as deputy director of CNSTAT. Prior to that, he worked at the U.S. Office of Management and Budget (OMB), where he served as senior statistician in the Statistical and Science Policy Office. He chaired the Federal Committee on Statistical Methodology and was the lead at OMB on issues related to standards for statistical surveys, survey nonresponse, measurement of race and ethnicity, and confidentiality of statistical data. He also previously was senior project leader of research standards and practices at the Arbitron Company and a research psychologist in the Office of Survey Methods Research in the Bureau of Labor Statistics. He is a fellow of the American Statistical Association. He has a B.A. in psychology and religious studies from the University of Denver and a Ph.D. in social psychology from the University of Minnesota.

**H.V. Jagadish** is the Bernard A. Galler collegiate professor of electrical engineering and computer science and distinguished scientist at the Institute for Data Science at the University of Michigan in Ann Arbor. Previously, he was head of the Database Research Department at AT&T Labs in Florham Park, New Jersey. He works widely in information management and holds numerous patents in the field. He is a fellow of the Association for Computing Machinery (ACM), serves on the board of the Computing

Research Association, and was a trustee of the VLDB [very large database] Foundation. He is a recipient of the SIGMOD Contributions Award from the ACM and of the David E. Liddle Research Excellence Award from the University of Michigan. He has a B.Tech. from the Indian Institute of Technology, Delhi, and an M.S. and a Ph.D. from Stanford University, all in electrical engineering.

**Frauke Kreuter** is a professor in the Joint Program in Survey Methodology at the University of Maryland and professor of statistics and methodology at the University of Mannheim, Germany. She is also affiliated with the Maryland Population Research Center, the Institute for Social Research in Michigan, and the German Institute for Employment Research. Previously, she held positions at the Institute for Statistics at the Ludwig-Maximilians University in Munich, Germany, and in the Department of Statistics at the University of California, Los Angeles. Her research focuses on nonresponse errors, paradata and responsive designs, record linkage, and, recently, issues of linkage consent and generalizability for nonprobability samples. She is an elected fellow of the American Statistical Association, and she is a recipient of the Gertrude Cox Award from the Washington Statistical Society. She serves on the advisory boards of Statistics Canada, Statistics Sweden, the U.S Bureau of Labor Statistics, the U.S. Energy Information Association, and the ad hoc committee for the Germany 2021 census. She has a B.A. and an M.A. in sociology and a Ph.D, all from the University of Konstanz in Germany.

**Sharon Lohr** is a vice president and senior statistician at Westat in Rockville, Maryland. Previously, she was dean's distinguished professor of statistics at Arizona State University. Her research has focused on survey sampling, hierarchical models, small-area estimation, missing data, and design of experiments. She is a fellow of the American Statistical Association, and an elected member of the International Statistical Institute. She was the inaugural recipient of the Washington Statistical Society's Gertrude M. Cox Statistics Award for contributions to the practice of statistics and a recipient of the society's Morris Hansen Lecture Award. She was recently selected to present the Deming Lecture at the Joint Statistical Meetings. She has a Ph.D. in statistics from the University of Wisconsin–Madison.

**James P. Lynch** is a professor and chair of the Department of Criminology and Criminal Justice at the University of Maryland. Previously, he served as the director of the Bureau of Justice Statistics at the U.S. Department of Justice and was a distinguished professor in the Department of Criminal Justice at John Jay College of the City University of New York. He also previously was a professor and chair of the Department of Justice, Law and Society at

American University. His research focuses on victim surveys, victimization risk, the role of coercion in social control, and crime statistics. He has been vice president of the American Society of Criminology and served on the Committee on Law and Justice Statistics of the American Statistical Association. He has a B.A. from Wesleyan University and an M.A. and a Ph.D. in sociology from the University of Chicago.

**Colm A. O'Muircheartaigh** is a professor and former dean of the Harris School of Public Policy Studies and a senior fellow at NORC, both at the University of Chicago. Previously, he was the first director of the Methodology Institute and a faculty member of the Department of Statistics at the London School of Economics and Political Science. The primary focus of his work is on the design of complex surveys across a wide range of populations and topics and on fundamental issues of data quality, including the impact of errors in responses to survey questions, cognitive aspects of question wording, and latent variable models for nonresponse. He is a fellow of the Royal Statistical Society and of the American Statistical Association and an elected member of the International Statistical Institute. He has served as a consultant to a wide range of public and commercial organizations around the world, including the OECD and the United Nations. He received his undergraduate education at University College Dublin and his graduate education at the London School of Economics.

**Trivellore Raghunathan** is the director of the Survey Research Center and a research professor at the Institute for Social Research at the University of Michigan, where he is also a professor of biostatistics and an associate director of the Center for Research on Ethnicity, Culture and Health in the School of Public Health. He is also a research professor in the Joint Program in Survey Methodology at the University of Maryland. Previously, he was on the faculty in the Department of Biostatistics at the University of Washington. His research interests are in the analysis of incomplete data, multiple imputation, Bayesian methods, design and analysis of sample surveys, combining information from multiple data sources, small-area estimation, confidentiality and disclosure limitation, longitudinal data analysis, and statistical methods for epidemiology. He has developed SAS-based software for imputing the missing values for a complex dataset. He has a Ph.D. in statistics from Harvard University.

**Roberto Rigobon** is the Society of Sloan Fellows professor of management and professor of applied economics at the Sloan School of Management at the Massachusetts Institute of Technology (MIT). He is also a visiting professor at the Instituto de Estudios Superiores de Administración (Institute of Advanced Studies in Administration) in Venezuela and a research associate

of the National Bureau of Economic Research. His research has addressed the causes of balance-of-payments crises, financial crises, and the propagation of them across countries. He is currently studying the properties of international pricing practices and how to produce alternative measures of inflation. He is one of the two founding members of the Billion Prices Project, as well as a cofounder of PriceStats. He is a member of the Census Bureau's Scientific Advisory Committee and the current president of the Latin American and Caribbean Economic Association. He has a B.S. in electrical engineering from Universidad Simon Bolivar (Venezuela), an M.B.A. from Instituto de Estudios Superiores de Administración (Venezuela), and a Ph.D. in economics from MIT.

**Marc Rotenberg** is a president of the Electronic Privacy Information Center (EPIC) in Washington, D.C., and teaches information privacy law and open government at Georgetown University Law Center. He has testified before Congress on more than 60 occasions and authored more than 50 amicus briefs on emerging privacy and civil liberties issues. He has served on several national and international advisory panels, including the expert panels on Cryptography Policy and Computer Security for the OECD and the Legal Experts on Cyberspace Law for UNESCO. He is founding board member and former chair of the Public Interest Registry, which manages the .ORG domain. He is a fellow of the American Bar Foundation, a member of the Council on Foreign Relations, and the recipient of several awards, including the World Technology Award in Law from the World Technology Network. He has an A.B. from Harvard College, a J.D. from Stanford Law School, and an LL.M. in international and comparative law from Georgetown University.

## COMMITTEE ON NATIONAL STATISTICS

The Committee on National Statistics was established in 1972 at the National Academies of Sciences, Engineering, and Medicine to improve the statistical methods and information on which public policy decisions are based. The committee carries out studies, workshops, and other activities to foster better measures and fuller understanding of the economy, the environment, public health, crime, education, immigration, poverty, welfare, and other public policy issues. It also evaluates ongoing statistical programs and tracks the statistical policy and coordinating activities of the federal government, serving a unique role at the intersection of statistics and public policy. The committee's work is supported by a consortium of federal agencies through a National Science Foundation grant, a National Agricultural Statistics Service cooperative agreement, and several individual contracts.