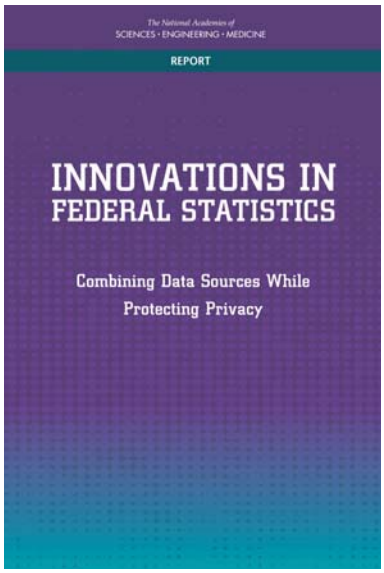
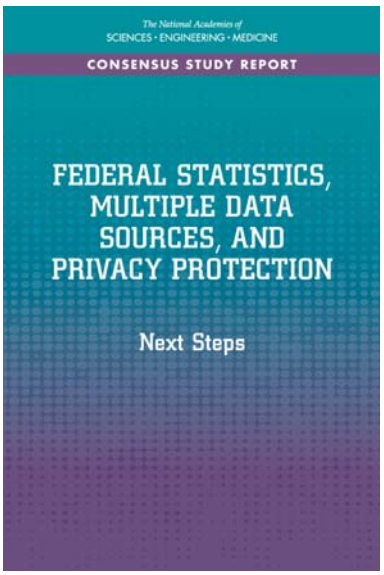
Consensus Study Reports
HIGHLIGHTS
October 2017

Two Reports Examine How to Improve Federal Statistics by Combining Multiple Data Sources



U.S. Federal government statistics provide critical information to the country and serve a key role in a democracy. Statistical indicators—for example, the unemployment rate or the rate of violent crime—guide the choices of policy makers and other decision makers. For decades, one of the primary methods for collecting data for federal statistics has been sample surveys designed for particular data needs. But the costs of such surveys have been increasing while response rates have been declining, and many surveys are not able to fulfill growing demands for more timely and detailed information.

The National Academies of Sciences, Engineering, and Medicine received funding from the Laura and John Arnold Foundation to examine how federal statistical agencies can shift toward the use of combinations of diverse data from government and private-sector sources instead of relying solely on a single survey or administrative records source. A committee of statisticians, social and computer scientists, and privacy experts were appointed to conduct the study.



The study yielded two reports, the first recommends that a new entity be created to enable statistical agencies to access and combine data from multiple sources, in order to offer policy makers and other decision makers more relevant, timely, and detailed statistics. The second report examines in depth how to implement this new approach—including statistical models for combining data from multiple sources, approaches for privacy protection, frameworks for assessing quality, and options for structuring and operating the new entity.

REPORT 1: *INNOVATIONS IN FEDERAL STATISTICS: COMBINING DATA SOURCES WHILE PROTECTING PRIVACY* (2017)

This report describes new sources of data that might be combined with survey data to yield more precise, timely, and relevant statistics. One new potential source of data is government administrative records—that is, data collected by government entities for program administration, regulatory, or law enforcement purposes. Another potential new source of information is data generated by the private sector for commercial use, such as credit card transactions, cell phone data, and Internet searches. However, not enough is known about the quality of these new data sources for federal statistics.

Some of the report's recommendations urge federal agencies to:

Evaluate the potential benefits and risks of using administrative data combined with federal survey data. 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.

Evaluate the potential benefits and risks of using private-sector data combined with federal survey data. Federal statistical agencies should systematically review their statistical portfolios and evaluate the potential benefits of using private-sector data sources, as well as the risks of using these data.

Use new methods to ensure that the privacy of individuals' and businesses' data is protected. Any consideration of expanding the use of data must have privacy as a core value. Federal statistical agencies should adopt modern database, cryptography, privacy-preserving, and privacy-enhancing technologies. 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.

The report notes that there are legal, regulatory, and policy barriers for statistical agencies to access these data sources, and while there are 13 agencies whose mission is primarily the creation and dissemination of federal statistics, there is no agency charged with facilitating access to and use of multiple data sources for the benefit of the entire statistical system.

Thus, the report's major recommendation is that **a new entity or an existing entity be designated to facilitate secure access to data to enhance the quality of federal statistics.** Privacy protections would have to be fundamental to the entity's mission. The data for which it has responsibility would need to have legal protections for confidentiality, and the strongest privacy protocols would need to be used for personally identifiable information while permitting statistical use.

REPORT 2: *FEDERAL STATISTICS, MULTIPLE DATA SOURCES, AND PRIVACY PROTECTION: NEXT STEPS* (2017)

This report, building on the previous one, offers detailed recommendations to guide agencies in creating the new entity to facilitate secure access to data from multiple sources. The report stresses that privacy protections should be at the forefront of the entity's design and identifies technological approaches that can minimize privacy risks. For example, secure multiparty computing could in some situations enable a statistical agency to compute a desired aggregate result without ever actually learning all the detailed data from the different data sources.

Among the report's recommendations:

Multiple data sources should be used to redesign current data collection efforts and estimation tasks to improve utility, timeliness, and cost efficiency of federal statistics. The report notes that more research is needed to understand these new approaches and that agencies will need to be careful and deliberative in making changes.

Federal statistical agencies should adopt a broader quality framework for statisti-

cal information that goes beyond the traditional quality measure of total survey error. Additional dimensions of quality that better capture user needs, such as timeliness, relevance, accuracy, accessibility, coherence, integrity, privacy, transparency, and interpretability should be considered.

The new entity should follow the principles and practices for federal statistical agencies and permit information accessed through it to be used only for statistical purposes. The entity needs strong legal authority to protect the confidentiality of data accessed through the entity and to ensure that the data are used only for statistical purposes.

The new entity should have an advisory committee to guide policies and best practices in privacy protection. 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.

Statistical agencies should ensure that their technical staff receive appropriate training in modern computing technologies such as secure multiparty computing, cryptography, and privacy-preserving and privacy-enhancing technologies, and for the new statistical skills needed for combining data from different sources. Because technology changes continuously and understanding those changes is critical for the statistical agencies' products, federal statistical agencies should ensure their IT staff receive continuous training to keep pace with those changes.

The report discusses the advantages and disadvantages of various options for locating the new entity, such as in a federal statistical agency, a federally funded R&D center, or a university-based public-private research center. Regardless of where the entity is established, federal statistical agencies should create partnerships with academia and external research organizations to develop the new methods needed for design and analysis using multiple data sources.

The report stresses the importance of transparency about the statistical activities conducted through the entity and the benefits of those activities. The report also offers recommendations about governance of the entity, noting that the director of the entity should report to a board of directors that includes representatives of the federal statistical agencies, experts on privacy, holders of data used by the entity, and users of statistical data.

THE COMMISSION ON EVIDENCE-BASED POLICYMAKING

At the same time the National Academies' panel was developing its second report on how to improve federal statistics by using multiple data sources, the congressionally established Commission on Evidence-Based Policymaking was looking at a related issue—how best to use administrative and survey data to evaluate government programs and alternative policy options. The National Academies panel's efforts in this domain are intended to complement and inform those of the Commission. Its second report, *Federal Statistics, Multiple Data Sources, and Privacy Protection: Next Steps*, is intended to help initiate a more detailed discussion among stakeholders to identify the best path forward for the federal statistical system.

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*.

For More Information . . . This Consensus Study Report Highlights was prepared by the Committee on National Statistics based on two Consensus Study Reports: *Innovations in Federal Statistics: Combining Data Sources While Protecting Privacy* (2017) and *Federal Statistics, Multiple Data Sources, and Privacy Protection: Next Steps* (2017). The studies were sponsored by the Laura and John Arnold Foundation with additional support from the National Academy of Sciences Kellogg Fund. 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. Copies of the Consensus Study Reports are available from the National Academies Press, (800) 624-6242; http://www.nap.edu.

Division of Behavioral and Social
Sciences and Education

The National Academies of
SCIENCES • ENGINEERING • MEDICINE

The nation turns to the National Academies of Sciences, Engineering, and Medicine for independent, objective advice on issues that affect people's lives worldwide.
www.national-academies.org