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

Brian Harris-Kojetin, Director, CNSTAT
COPAFS Meeting
Washington, DC • June 1, 2018

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), Georgetown University
- **Michael E. Chernew,** Harvard University
- **Piet Daas,** Statistics Netherlands
- **Cynthia Dwork**, Harvard University
- **Ophir Frieder,** Georgetown University
- **Hosagrahar V. Jagadish**, University of Michigan
- **Frauke Kreuter**, University of Maryland
- **Sharon Lohr**, Westat, Inc.
- **James P. Lynch**, University of Maryland
- **Colm O'Muircheartaigh**, University of Chicago
- **Trivellore Raghunathan**, University of Michigan
- **Roberto Rigobon**, MIT
- **Marc Rotenberg,** Electronic Privacy Information Center

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.**

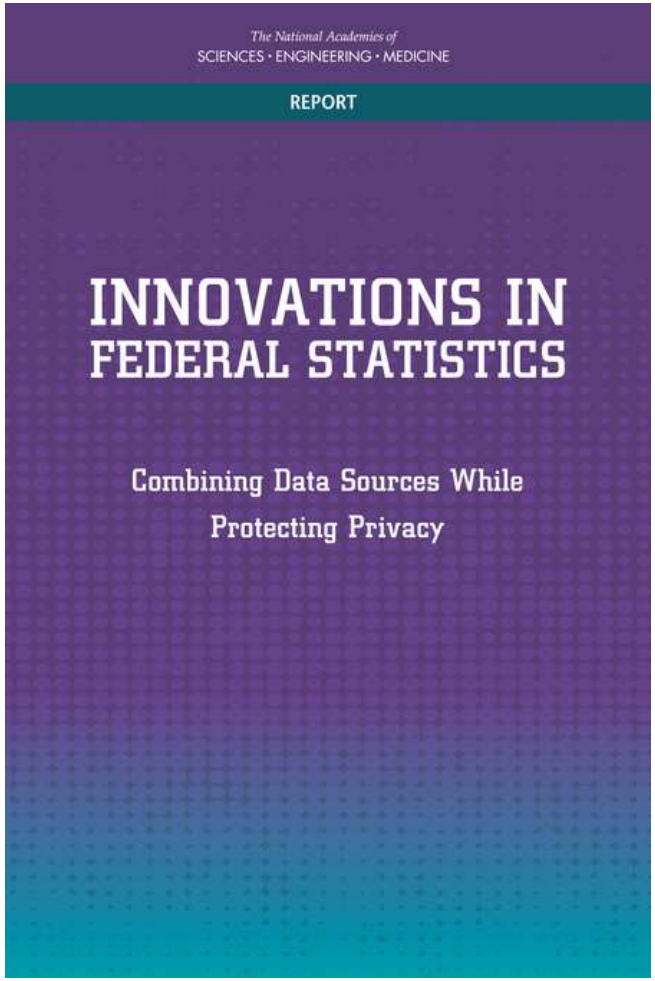
Acknowledgements

Funding for the panel was provided by

The Laura and John Arnold Foundation,

with additional support from the National Academy of Sciences Kellogg Fund.

The Panel's First Report:



Contents

- Chapter 1: Introduction
- Chapter 2: Current Challenges and Opportunities in Federal Statistics
- Chapter 3: Using Government Administrative and Other Data for Federal Statistics
- Chapter 4: Using Private-Sector Data For Federal Statistics
- Chapter 5: Protecting Privacy and Confidentiality While Providing Access to Data for Research Use
- Chapter 6: Advancing the Paradigm of Combining Data Sources

Current Challenges and Opportunities in Federal Statistics

Conclusion 2-3: The way that statistics are currently produced by Federal statistical agencies faces threats from declining participation rates and increasing costs.

- Although generally higher than other surveys, federal statistical surveys face increasing nonresponse and increased costs of data collection to maintain response rates
- Agency budgets have decreased or remained flat
- Agencies face increasing demands for more timely and more geographically detailed information
- Increasingly alternative data sources are available that offer the potential of faster and more detailed information

Current Barriers to Use of Alternative Data Sources

- Conclusion 3-4: **Legal and administrative barriers** limit statistical use of administrative datasets by federal statistical agencies.

Advancing the Paradigm of Combining Data Sources

RECOMMENDATION 6-1: A new entity or an existing entity should be designed to facilitate secure access to data for statistical purposes to enhance the quality of federal statistics

RECOMMENDATION 6-2: The proposed new entity should maximize the utility of the data for which it is responsible while protecting privacy by using modern database, cryptography, privacy-preserving, and privacy-enhancing technologies.

The Panel's Second Report:

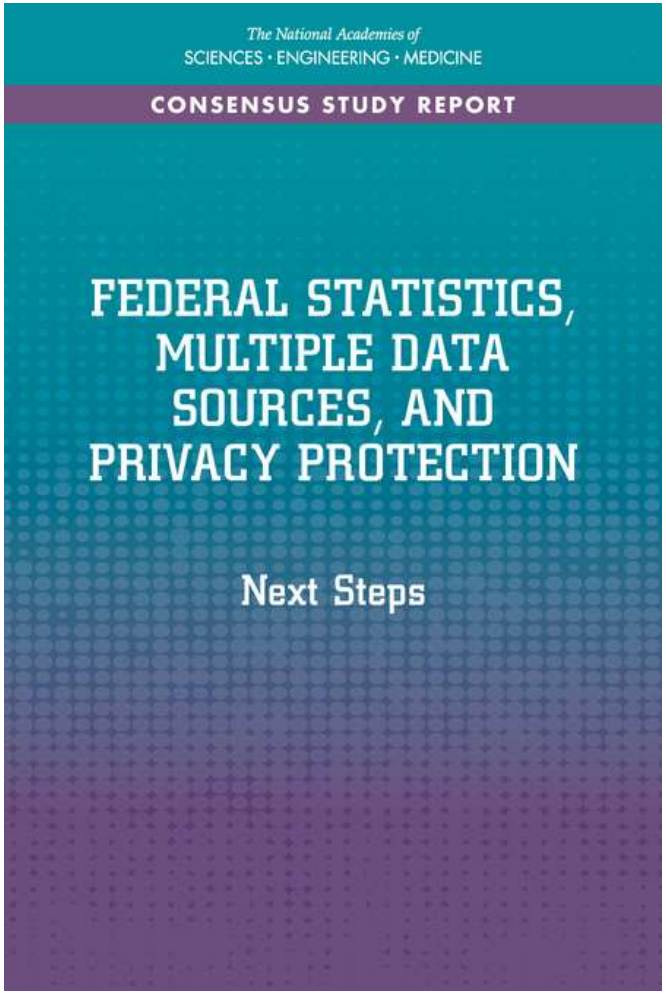


Table of Contents

2. Statistical Methods for Combining Multiple Data Sources

- Demands for More Granular Statistics
- Statistical Methods for Combining Data
- Next Steps for Combining Data Sources

Demands for More Granular 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.

Statistical Methods for Combining Data

- Record Linkage Approaches
- Multiple Frame Methods
- Imputation–Based Methods
- Small Area Estimation Models, with a variety of hierarchical structures

Statistical Methods for Combining Data

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

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.

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

- Issues for Federal Statistical Agency IT Systems
- System Architecture
- Data Processing Issues
- System Migration
- Personnel Staffing and Skills

System Architecture

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 a system, as well as the implications for protecting privacy and addressing data providers' concerns regarding privacy.

Data Processing Issues

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.

Personnel Staffing and Skills

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

- Personally Identifiable Information and Privacy Law
- Legal View of Privacy in the Context of Statistical Data Analysis
- The Scope of PII
- Examples Elucidating the PII/non-PII Issue
- Synthesis: A Proposed Liability Rule for PII
- Implications for Federal Statistical Agencies

The Scope of PII

- Currently legal experts and computer scientists have differing views of PII
- Combining data sources increases privacy risks through auxiliary data
- Data that previously could not identify an individual may now contain auxiliary information that could identify an individual

Implications for Federal Statistical Agencies

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.

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

- Security Threats
  - Securing Data
  - Securing Computation
- Inference Control Techniques
  - Statistical Disclosure Limitation
  - Data Enclaves
  - Differential Privacy
- Implications for Federal Statistical Agencies

Implications for Federal Statistical Agencies

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.

6. Quality Frameworks for Statistics Using Multiple Data Sources

- A Quality Framework for Survey Research
- Broader Frameworks for Assessing Quality
- Assessing the Quality of Administrative and Private-Sector Data
- The Quality of Alternative Data Sources: Two Illustrations

A Quality Framework for Survey Research

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 used developed careful protocols for understanding and reporting potential errors in their 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

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

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.

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.

A New Entity to Provide Vital Information through Enhanced Federal Statistics

- Attributes of the New Entity
  - Organizational Location
  - Functions
  - Technological Environment for Data Access
  - Access by Outside Researchers
  - Privacy
  - Transparency
  - Financing
  - Governance
- Implementation

Core Requirements for New Entity

- 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.
- 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 Privacy Act), but it may also require new legislation.
- 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 phenomenon. The authority needs to delimit what uses are forbidden as well as what uses are encouraged.
- It needs to be **staffed with personnel whose skills fit the needs of the recommended entity**, including advance IT architectures, data transmission, record linkage, statistical computing, cryptography, data curation, cybersecurity, and privacy regulations.

Organizational Location

The report discusses different options for possible location of the new entity including:

- A New Federal Statistical Agency
- Within an Existing Federal Statistical Agency
- A Federally Funded Research and Development Center
- A University-Based Public-Private Research Center

The report also discusses the advantages and disadvantages to each of these locations.

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 only for statistical purposes.

Functions

- The report discusses different approaches to allowing users to access data for statistical purposes within the entity.
- The entity could serve as a location to access and combine data and the staff would not do any data analysis.
- The entity could also be a “full-service” research institution who would also staff economists, statisticians, and other experts who can analyze data.

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.

Access by Outside Researchers

- Currently, researchers are subject to a lengthy variety of approaches to be approved to access statistical data.
- The panel believes that the entity could be useful to streamlining the research approval process.

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.

Transparency

RECOMMENDATION 7-5 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.

RECOMMENDATION 7-6 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.

Governance

- Governance of the entity will be driven by the location of the organization and the authorizing legislation
- Given the mission and nature of the recommended new entity, consideration should be given to additional structures and mechanisms for governance of the entity.

RECOMMENDATION 7-7 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.

Governance

RECOMMENDATION 7-8 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.

RECOMMENDATION 7-9 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 in producing and disseminating statistical information.

Implementation

- A strategic plan will be needed 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 points.
- The first phase might cover 5 years, at which time it would be useful to have a comprehensive review.
- The first phase needs to include expanded access to federal administrative and operational data that could be useful for federal statistics.
- Private-sector data could be included as part of the new entity in a later phase.
- 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.

THANK YOU!

For further information contact:
Brian Harris-Kojetin (bkojetin@nas.edu)