Combining Multiple Data Sources to Enhance U.S. Federal Statistics

Brian Harris-Kojetin, Director, CNSTAT
BigSurv18
Barcelona, Spain • October 26, 2018

The National Academies of Sciences, Engineering, and Medicine

- Non-profit, 501 c(3) organization
- Provide independent, objective advice from experts
- Academies members and distinguished **volunteers** serve on ad hoc panels and standing bodies such as CNSTAT

Committee on National Statistics
The National Academies of SCIENCES · ENGINEERING · MEDICINE

The Committee on National Statistics (CNSTAT)

- Established in 1972 as a standing unit of the National Academies on the recommendation of the President's Commission on Federal Statistics
- CNSTAT's mission is to improve the statistical methods and information on which public policy decisions are based.
- CNSTAT conducts and oversees consensus panel studies, workshops, expert meetings, and other activities
- Over its 46-year history, CNSTAT has produced over 270 reports.

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

Committee on National Statistics
The National Academies of SCIENCES · ENGINEERING · MEDICINE

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

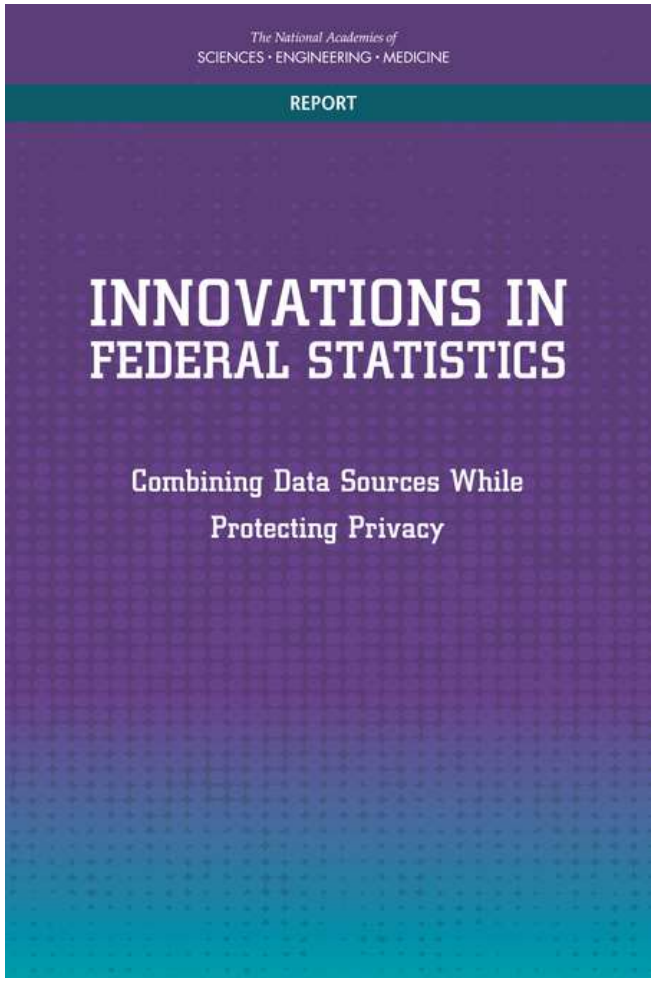
Funding for the panel was provided by

The Laura and John Arnold Foundation,

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

Committee on National Statistics
The National Academies of SCIENCES · ENGINEERING · MEDICINE

The Panel's First Report:
(available at www.nap.edu)



Contents

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

Using Government Administrative and Other Data for Federal Statistics

- **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**.

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.

Using Private Sector Data for Federal Statistics

- **Recommendation 4-1** Federal statistical agencies should systematically review their statistical portfolios and **evaluate the potential benefits of using private-sector data sources.**

Protecting Privacy and Confidentiality While Providing Access to Data

- **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.

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

The Panel's Second Report:

(available at www.nap.edu)

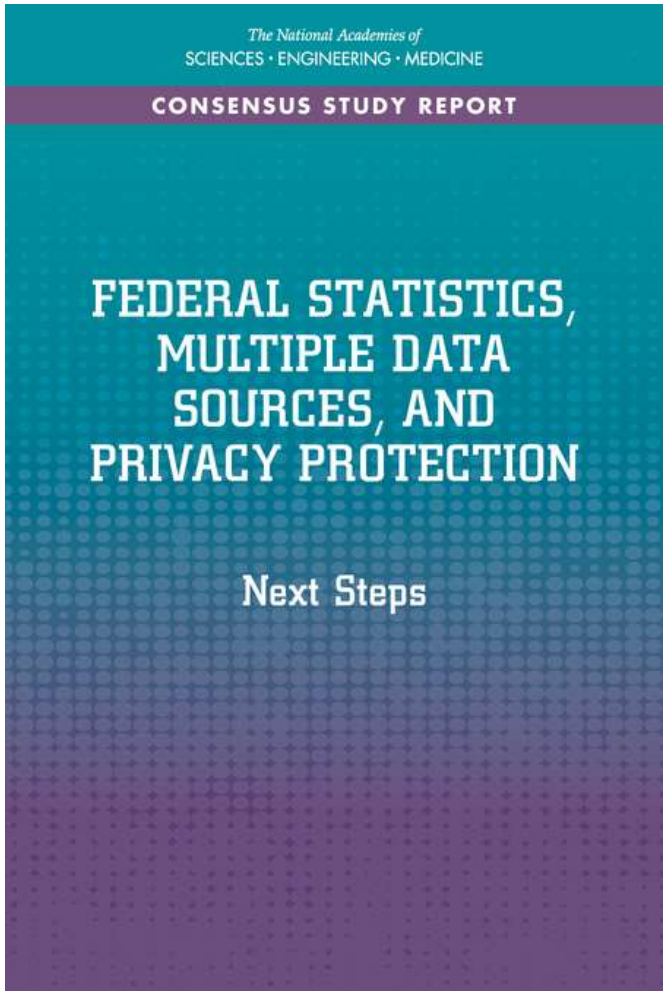


Table of Contents

Statistical Methods 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.

IT Infrastructure

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.**

Privacy 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.

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.

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.
- It has to have **strong authority to protect the privacy of data** that are accessed and prevent misuse.
- 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 phenomenon.
- 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.

Implementation

- A **strategic plan** will be needed for expanding the data sources accessible through the entity.
- The first phase needs to include **expanded access to federal administrative and operational data** that could be useful for federal statistics.
- 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)