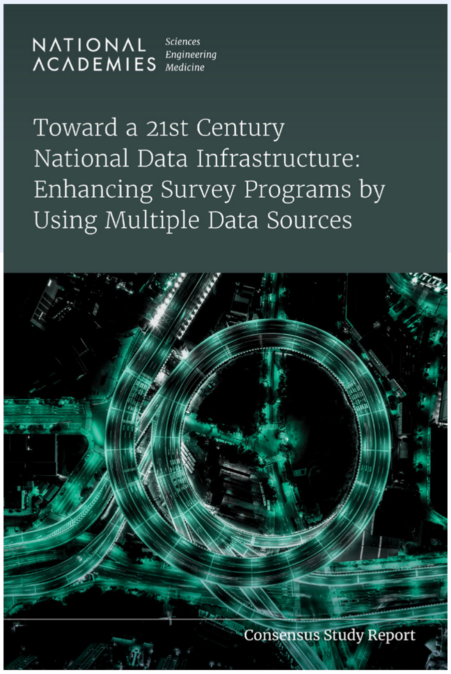
NATIONAL ACADEMIES *Sciences Engineering Medicine*

Consensus Study Report
Highlights

Toward a 21st Century National Data Infrastructure



NATIONAL ACADEMIES *Sciences Engineering Medicine*

Toward a 21st Century National Data Infrastructure: Enhancing Survey Programs by Using Multiple Data Sources

Consensus Study Report

Enhancing Survey Programs by Using Multiple Data Sources

Much of the statistical information produced by federal statistical agencies since the 1950s—information about economic, social, and physical well-being that is essential for the functioning of modern society—has come from sample surveys. Data from these surveys have been used to inform economic, social, and health policies; evaluate the effects of those policies; and monitor the health and economic circumstances of the population. They have also been used to inform decisionmaking by businesses and individuals, as well as produce vast quantities of economic, health, and social research that informs the public and can lead to societal benefits.

At the time they were established, many sample surveys represented the only way to obtain reliable, accurate, and regularly updated information about the population and businesses of the United States. But surveys have faced challenges in recent years that include decreasing response rates, increasing costs, and user demand for more timely and more granular data and statistics. At the same time, there has been a proliferation of data from other sources, including data collected by government agencies while administering programs (administrative records), satellite and sensor data, private-sector data such as electronic health records and credit card transaction data, and massive amounts of data available on the internet. How can these new data sources be used to supplement or replace some of the information currently collected on surveys, and to provide new frontiers for producing information and statistics to benefit American society?

To answer those questions, the National Academies, with funding from the National Science Foundation, appointed three consensus panels to develop a vision for a new data infrastructure for national statistics and social and economic research. *Toward a 21st Century National Data Infrastructure: Enhancing Survey Programs by Using Multiple Data Sources* is the second report in a series of three reports focused on separate aspects of the new data infrastructure, and it was produced by a diverse panel with expertise spanning areas of statistics, survey methodology, economics, sociology, psychology, public policy, equity analytics, public health, geography, and demography. The report discusses how survey programs might be affected by the use of alternative data sources. It also anchors key themes through examples—from the areas of income, health, crime, and agriculture—that represent different ways in which multiple data sources are, or could be leveraged, and that illustrate the types of challenges to be faced (see Box for some of the examples discussed in the report).

ROLE OF MULTIPLE DATA SOURCES

The report identifies four main ways that multiple data sources could improve national statistics, provide new resources for social and economic research, and promote data equity. These include:

- Providing information to evaluate and improve quality of data sources
- Giving additional information about survey respondents
- Producing statistics for small populations
- Creating data products directly from administrative data

IMPROVING DATA QUALITY

Use of multiple data sources can add value for official statistics and for research. However, combining information across data sources must be done carefully, with deep understanding of the properties of each component dataset and the statistics resulting from their combination. The process begins by evaluating the quality of each data source through assessing how well each

MULTIPLE DATA SOURCE EXAMPLES DISCUSSED IN THE REPORT

CREATING DATA RESOURCES FROM ADMINISTRATIVE RECORDS

- Linkage of geospatial, business, job, and demographic databases at the U.S. Census Bureau.
- Statistics Canada's *Disaggregated Data Action Plan.*

INCOME AND HEALTH

- Linking records to study whether income measurements are consistent across data sources.
- U.S. Census Bureau National Experimental Well-being Statistics project.
- U.S. National Vital Statistics System.
- Adding information about mortality, medical expenses, and housing to health survey records through data linkage.

CRIME

- Uniform Crime Reporting Program use of data from state and local agencies.
- Potential for increased use of administrative records and data integration to study crime.

AGRICULTURE

- Using statistical models to estimate crop yields for counties (U.S.) and provinces (Canada).
- Webscraping to augment agricultural sampling frames.

source meets the needs it is asked to address (fitness for use). Additional evaluations are needed of the quality of the data resources and of the statistics generated from combined datasets.

Numerous data sources, including probability samples, administrative records, and private-sector data, could be used to produce official statistics if they meet standards for quality. Each data source has specific tradeoffs in terms of timeliness, population coverage, amount of geographic or subgroup detail, concepts measured, accuracy, and continuing availability. Relying on multiple sources can take advantage of the strengths of each source while compensating for its weaknesses. (Conclusion 2-2)

The quality of statistics produced from multiple data sources depends on properties of the individual sources as well as the methods used to combine them. A new framework of quality standards and guidelines is needed for evaluating such data sources' fitness for use. (Conclusion 9-1)

ENHANCING DATA EQUITY WITH MULTIPLE DATA SOURCES

The use of multiple data sources can benefit data equity—promoting the collection and use of data in which all populations, and especially those that have been historically underrepresented or misrepresented in the data record, are visible and accurately portrayed.

Many data sources include or represent only part of the population of interest. Multiple data sources can be used to assess and improve the coverage of underrepresented groups, and to enable the production of disaggregated statistics. It is important to examine the representativeness and coverage of combined data sources to ensure data equity. (Conclusion 3-1)

Record linkage can merge information from separate data sources and add variables that are needed to produce disaggregated statistics. But linkage procedures may also introduce biases because linkage errors can disproportionately affect members of some population subgroups. It is important to assess data-equity implications of record-linkage methods. (Conclusion 3-2)

Data equity is an essential aspect of any data system. Documentation of equity aspects, including a discussion of the decisions to include or exclude population subgroup information and an evaluation of data quality for subpopulations of interest, will promote transparency. Development of standards for data equity, and procedures for regularly reviewing equity implications of statistical programs, would enhance efforts to improve data equity across the federal statistical system. (Conclusion 3-3)

IMPORTANT CONSIDERATIONS

Transparency and documentation of component datasets and of methods used to combine datasets are essential for producing trust in information created from multiple data sources, particularly as new types of data are used. (Conclusion 9-2)

Creating useful statistics and data products from combined data sources requires new skills. A new data infrastructure requires investment not only in data sources but also in the people who can work with those data. Beyond the technical challenges of developing new statistical methods, there are challenges for promoting data equity and public trust in integrated data. To take advantage of new data resources, it will be important for statistical agencies to invest in personnel, training, and cyberinfrastructure.

Use of multiple data sources is expected to play a major role in the future production of statistical information in the United States, but additional technical expertise and resources are needed to address the challenges involved in producing and assessing the quality of integrated data and statistics. (Conclusion 9-3)

CONTINUING WORK

Probability surveys have provided the nation with useful statistics on numerous topics for more than 80 years, and the panel that authored the report anticipates that they will continue to be used for producing statistics in many topic areas. Some statistics, such as the percentage of persons who were looking for work last week or the percentage of criminal victimizations that are reported to the police, rely on information that can only be provided by individuals in the population—a probability survey may still be the best method for collecting information on such topics. But there are many opportunities for enhancing survey information with data from other sources, or for reducing burden on survey respondents by obtaining information elsewhere. For some topics and for some parts of the population, administrative records or other data sources can provide more timely, accurate, or granular information than surveys, and at reduced cost.

For all individual data sources that feed into combined data sets and ultimately a new data infrastructure, continued investments in improving the quality of the underlying data are essential for ensuring that the resulting statistics are valid and reliable. This is particularly important given that data-quality concerns do not affect all population groups, geographic areas, or administrative units equally. A new data infrastructure, and ultimately data users, would benefit from changes to the underlying data sources that would facilitate data linkages.

PANEL ON THE IMPLICATIONS OF USING MULTIPLE DATA SOURCES FOR MAJOR SURVEY PROGRAMS **SHARON L. LOHR** (*Chair*), Arizona State University (*Emerita*); **JEAN-FRANÇOIS BEAUMONT,** Statistics Canada; **LAWRENCE D. BOBO,** Harvard University; **MICK P. COUPER,** University of Michigan; **HILARY HOYNES,** University of California, Berkeley; **KIMBERLYN LEARY,** Harvard T.H. Chan School of Public Health; **DAVID MANCUSO,** Washington State Department of Social and Health Services; **JUDITH A. SELTZER,** University of California, Los Angeles; **ELIZABETH A. STUART,** Johns Hopkins Bloomberg School of Public Health; **SHAOWEN WANG,** University of Illinois at Urbana-Champaign

STAFF **DANIEL H. WEINBERG,** Study Director (until December 2022); **KRISZTINA MARTON,** Study Director (from December 2022); **JOSHUA LANG,** Senior Program Assistant

FOR MORE INFORMATION
The Report Highlights was prepared based on the Consensus Study Report *Toward a 21st Century National Data Infrastructure: Enhancing Survey Programs by Using Multiple Data Sources* (2023). Copies of the Consensus Study Report are available from the National Academies Press at https://nap.nationalacademies.org/catalog/26804.

The study was sponsored by the National Science Foundation. 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.

Division of Behavioral and Social Sciences and Education

NATIONAL ACADEMIES *Sciences Engineering Medicine*