NATIONAL ACADEMIES Sciences Engineering Medicine

Toward a 21st Century National Data Infrastructure:

Enhancing Survey Programs by Using Multiple Data Sources

Elizabeth A. Stuart

October 26, 2023

COMMITTEE ON NATIONAL STATISTICS

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 at 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 Urbana-Champaign

Study Directors: Daniel H. Weinberg and Krisztina Marton

NATIONAL ACADEMIES Sciences Engineering Medicine

Background

In spring 2021, the Committee on National Statistics received funding from National Science Foundation to convene three studies on a vision for a new national data infrastructure for federal statistics and for social and economic research.
Three independent consensus panels were appointed:

- Report 1: The components and key characteristics of a 21st century data infrastructure.
- **Report 2: The implications of using multiple data sources for major survey programs.**
- Report 3: The technology, tools, and capabilities needed for data sharing, use, and analysis

Statement of Task

The implications of using multiple data sources for major survey programs, including:

- Addressing changes in measurement with new data sources;
- Approaches for linking alternative data sources to universe frames to assess and enhance representativeness; and
- Implications of new data sources for population subgroup coverage, and life course longitudinal data

Interpreting the Statement of Task: Panel Decisions

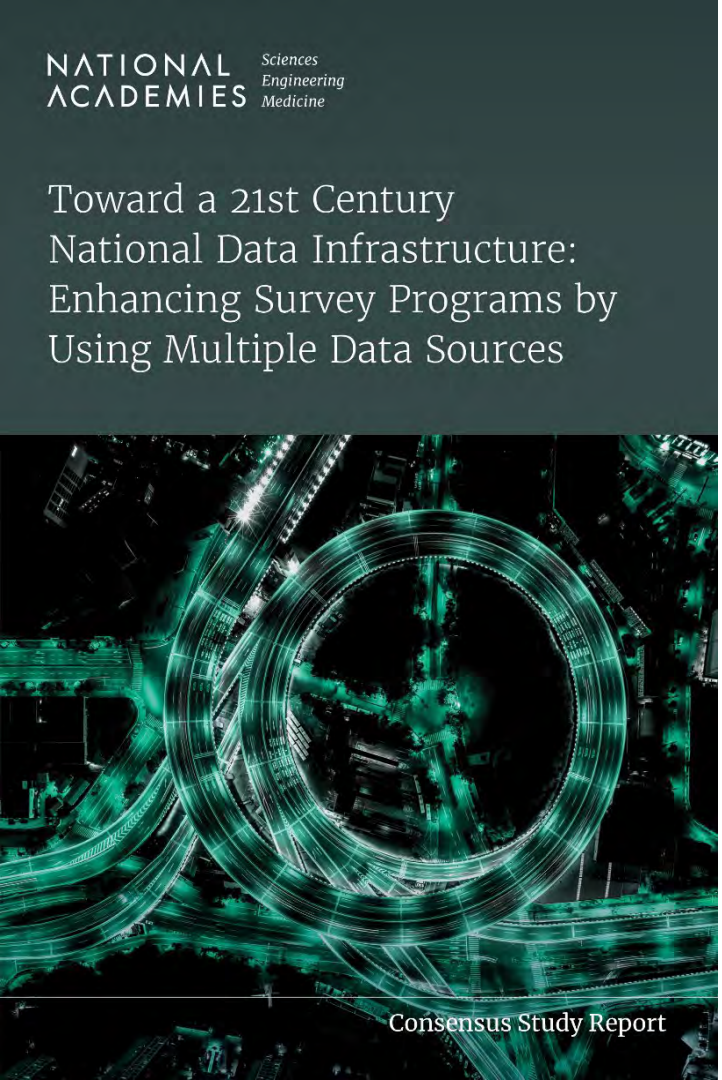
- Focus on use cases that represent different ways to take advantage of multiple data sources
  - Income and health: data linkage, measurement, life-course longitudinal data
  - Crime: data obtained from state programs and local law-enforcement agencies
  - Agriculture: small area models for crop estimates

Interpreting the Statement of Task: Panel Decisions

- Address issues of representativeness and population subgroup coverage through data equity lens
- De-emphasize topics tasked to Reports 1 & 3:
  - Acquiring data; legal agreements for data sharing
  - Computer infrastructure for integrated data
  - Public access to data
  - Privacy and confidentiality
- Our focus:
  - Information gains & challenges from different methods of combining data
  - Anchored through examples of state-of-the art practice + suggestions

Virtual Public Workshop

- May 16 & 18, 2022
- Five sessions:
  1. Opportunities for Using Multiple Data Sources to Enhance Major Survey Programs
  2. Measuring Crime in the 21st Century: A Panel Discussion
  3. Improving Agriculture Statistics with New Data Sources
  4. Data Linkage for Income and Health Statistics
  5. Issues in Data Equity



Chapters

Role of Multiple Data Sources

- **Enhance** information currently coming probability surveys
  - Improve national statistics
  - Provide new resources for social and economic research
  - Promote data equity
- Provide information to evaluate, improve quality of data sources
- Give additional information about survey respondents
- Produce statistics for small populations
- Create data products directly from administrative data

Improving Data Quality

CONCLUSION 2-1: Probability surveys still have an important role to play in the production of official statistics but face challenges from nonresponse and high costs. Probability surveys by themselves may not be able to meet increasing societal demands for timely and granular data. For these reasons, alternative data sources are increasingly important to complement surveys.

CONCLUSION 2-2: 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.

Utility	Relevance	Relevance refers to whether the data product is targeted to meet current and prospective user needs.
	Accessibility	Accessibility relates to the ease with which data users can obtain an agency's products and documentation in forms and formats that are understandable to data users.
	Timeliness	Timeliness is the length of time between the event or phenomenon the data describe and their availability.
	Punctuality	Punctuality is measured as the time lag between the actual release of the data and the planned target date for data release.
	Granularity	Granularity refers to the amount of disaggregation available for key data elements. Granularity can be expressed in units of time, level of geographic detail available, or the amount of detail available on any of a number of characteristics (*e.g.* (demographic, socio-economic).
Objectivity	Accuracy and reliability	Accuracy measures the closeness of an estimate from a data product to its true value. Reliability, a related concept, characterizes the consistency of results when the same phenomenon is measured or estimated more than once under similar conditions.
	Coherence	Coherence is defined as the ability of the data product to maintain common definitions, classification, and methodological processes, to align with external statistical standards, and to maintain consistency and comparability with other relevant data.
Integrity	Scientific integrity	Scientific integrity refers to an environment that ensures adherence to scientific standards and use of established scientific methods to produce and disseminate objective data products and one that shields these products from inappropriate political influence.
	Credibility	Credibility characterizes the confidence that users place in data products based simply on the qualifications and past performance of the data producer.
	Computer and physical security	Computer and physical security of data refers to the protection of information throughout the collection, production, analysis, and development process from unauthorized access or revision to ensure that the information is not compromised through corruption or falsification.
	Confidentiality	Confidentiality refers to a quality or condition of information as an obligation not to disclose that information to an unauthorized party.

Federal Committee on Statistical Methodology (2020, p. 4)

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

Equitable Data Working Group (2022) Definition

> “Equitable data are those that allow for rigorous assessment of the extent to which government programs and policies yield consistently fair, just, and impartial treatment of all individuals. Equitable data illuminate opportunities for targeted actions that will result in demonstrably improved outcomes for underserved communities.”

Enhancing Data Equity with Multiple Data Sources

- Investigate or improve coverage
  - Webscraped lists of small urban agricultural operations
  - Training data sets for artificial intelligence algorithms
- Enable finer data disaggregation
  - Statistics Canada *Disaggregated Data Action Plan*
- Produce model-based estimates for small subpopulations
  - Small Area Income and Poverty Estimates
- Assess and reduce measurement error
  - Race/ethnicity measurement: link surveys (self-report) to other sources
  - Income and program participation (link surveys to administrative data)

Enhancing Data Equity with Multiple Data Sources

- Add features to the data through data linkage
  - Akee (2022): link IRS data at Census Bureau to study income inequality for small race/ethnic groups
- Add features to the data through imputation
  - Impute race using surnames, block group demographic statistics
- But care is needed
  - Linkage quality differs across subpopulations
  - Equity implications of small area, imputation models
  - Confidentiality considerations (next report)
- Three conclusions on data equity

Enhancing Data Equity with Multiple Data Sources

CONCLUSION 3-1: 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 3-2: 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 3-3: 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.

Creating New Data Resources with Administrative Records

- Longitudinal databases from existing records
  - Longitudinal Business Database, Longitudinal Employer-Household Dynamics, Decennial Census Linkage Project
  - **CONCLUSION 4-1:** Longitudinally linked administrative records datasets provide a cost-efficient opportunity to study long-term outcomes, and they may have large sample sizes for key population subgroups that have low representation in other data sources. Careful curation and attention to linkage errors and data equity enhance the value of these datasets.
- Census Bureau *Frames* project
- National Vital Statistics System
  - Model for assembling state-administered data systems
- State and regional initiatives

Data Linkage to Improve Income Measurement

- Link survey and administrative data records (SSA, IRS) to:
  - Compare survey respondents, nonrespondents
  - Compare income, program participation reported on surveys with administrative records
- Two Census Bureau projects for improving income measures with linked data
  - Comprehensive Income Dataset Project
  - National Experimental Well-being Statistics Project

Data Linkage to Supplement Health Surveys

- Link survey and administrative data records to add variables
  - Subsequent life-course outcomes, including mortality
  - Medical expenses, housing assistance, pension information
- Linkage and data equity
  - Records with deficient linkage information are less likely to be matched
  - Best practices for investigating and documenting linked data quality
- Longitudinal linkage
  - Health and Retirement Study

Combining Multiple Data Sources to Measure Crime

- National Academies (2016, 2018)
- National Crime Victimization Survey
- Uniform Crime Reporting Program
  - 2021: switched to National Incident-Based Reporting System (NIBRS)
  - More details of incidents, but less agency participation

CONCLUSION 7-1: The National Incident-Based Reporting System (NIBRS) provides details about each crime incident that were not available in the previous Summary Reporting System of the Uniform Crime Reports. NIBRS represents an important step in the production of detailed and accurate crime statistics. But the transition to NIBRS is still underway and variations in measurement and data reporting across jurisdictions need further study.

Combining Multiple Data Sources to Measure Crime

- Other data sources
  - Police department websites
  - Webscraped and crowdsourced data
  - Data from regulatory, public health agencies
- Potential to improve crime and population coverage, accuracy, granularity
  - National Violent Death Reporting System

CONCLUSION 7-2: Improving crime statistics will require coordination of the National Crime Victimization Survey and Uniform Crime Reporting Program with new data sources that can provide timely and detailed information about crimes, including those measured in the current classification systems and those that are currently unmeasured. This will entail increased investment in research on directly using data collected by police departments and on developing new data resources.

Using Multiple Data Sources for County-Level Crop Estimates

- National Academies (2017)
- National Agricultural Statistics Service: small area models to produce objective estimates with measures of uncertainty
- Statistics Canada: replaced 2 of 6 annual field surveys with modeled estimates
- Potential for increased use of remotely sensed and private-sector data
- Data equity considerations and potential:
  - Small operations less likely to use precision agriculture
  - Linkage to study measurement, append demographic characteristics

Combining Data Sources for National Statistics: Next Steps

CONCLUSION 9-1: 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 9-2: 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 9-3: 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.

Continuing Work

- Report available for download from https://nap.nationalacademies.org/
- This report documented ongoing innovations for combining data
- Benefits of combining data for social and economic research
- Research needed:
  - Continuing to improve individual data sources
  - Combined data: quality, equity aspects, statistical methodology, communication
- Workshop 3: Approaches to Sharing Blended Data in a 21st Century Data Infrastructure: in progress!