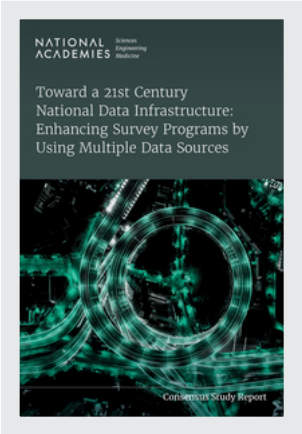


Toward a 21st Century National Data Infrastructure: Enhancing Survey Programs by Using Multiple Data Sources (2023)

DETAILS

278 pages | 6 x 9 | PAPERBACK

ISBN 978-0-309-69675-3 | DOI 10.17226/26804

CONTRIBUTORS

Sharon L. Lohr, Daniel H. Weinberg, and Krisztina Marton, Editors; Panel on the Implications of Using Multiple Data Sources for Major Survey Programs; Committee on National Statistics; Division of Behavioral and Social Sciences and Education; National Academies of Sciences, Engineering, and Medicine

SUGGESTED CITATION

National Academies of Sciences, Engineering, and Medicine. 2023. *Toward a 21st Century National Data Infrastructure: Enhancing Survey Programs by Using Multiple Data Sources*. Washington, DC: The National Academies Press. https://doi.org/10.17226/26804.

NATIONAL ACADEMIES *Sciences Engineering Medicine*

NATIONAL ACADEMIES PRESS
Washington, DC

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

Sharon L. Lohr, Daniel H. Weinberg, and Krisztina Marton, *Editors*

Panel on the Implications of Using Multiple Data Sources for Major Survey Programs

Committee on National Statistics

Division of Behavioral and Social Sciences and Education

Consensus Study Report

Prepublication copy, uncorrected proofs

NATIONAL ACADEMIES PRESS 500 Fifth Street, NW Washington, DC 20001

This activity was supported by a grant from the National Science Foundation to the National Academy of Sciences. 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.

International Standard Book Number-13: 978-0-309-XXXXX-X
International Standard Book Number-10: 0-309-XXXXX-X
Digital Object Identifier: https://doi.org/10.17226/26804

This publication is available from the National Academies Press, 500 Fifth Street, NW, Keck 360, Washington, DC 20001; (800) 624-6242 or (202) 334-3313; http://www.nap.edu.

Printed in the United States of America.

Suggested citation: National Academies of Sciences, Engineering, and Medicine. 2023. *Toward a 21st Century National Data Infrastructure: Enhancing Survey Programs by Using Multiple Data Sources*. Washington, DC: The National Academies Press. https://doi.org/10.17226/26804.

The **National Academy of Sciences** was established in 1863 by an Act of Congress, signed by President Lincoln, as a private, nongovernmental institution to advise the nation on issues related to science and technology. Members are elected by their peers for outstanding contributions to research. Dr. Marcia McNutt is president.

The **National Academy of Engineering** was established in 1964 under the charter of the National Academy of Sciences to bring the practices of engineering to advising the nation. Members are elected by their peers for extraordinary contributions to engineering. Dr. John L. Anderson is president.

The **National Academy of Medicine** (formerly the Institute of Medicine) was established in 1970 under the charter of the National Academy of Sciences to advise the nation on medical and health issues. Members are elected by their peers for distinguished contributions to medicine and health. Dr. Victor J. Dzau is president.

The three Academies work together as the **National Academies of Sciences, Engineering, and Medicine** to provide independent, objective analysis and advice to the nation and conduct other activities to solve complex problems and inform public policy decisions. The National Academies also encourage education and research, recognize outstanding contributions to knowledge, and increase public understanding in matters of science, engineering, and medicine.

Learn more about the National Academies of Sciences, Engineering, and Medicine at **www.nationalacademies.org**.

Consensus Study Reports published by the National Academies of Sciences, Engineering, and Medicine document the evidence-based consensus on the study's statement of task by an authoring committee of experts. Reports typically include findings, conclusions, and recommendations based on information gathered by the committee and the committee's deliberations. Each report has been subjected to a rigorous and independent peer-review process and it represents the position of the National Academies on the statement of task.

Proceedings published by the National Academies of Sciences, Engineering, and Medicine chronicle the presentations and discussions at a workshop, symposium, or other event convened by the National Academies. The statements and opinions contained in proceedings are those of the participants and are not endorsed by other participants, the planning committee, or the National Academies.

Rapid Expert Consultations published by the National Academies of Sciences, Engineering, and Medicine are authored by subject-matter experts on narrowly focused topics that can be supported by a body of evidence. The discussions contained in rapid expert consultations are considered those of the authors and do not contain policy recommendations. Rapid expert consultations are reviewed by the institution before release.

For information about other products and activities of the National Academies, please visit www.nationalacademies.org/about/whatwedo.

PANEL ON THE IMPLICATIONS OF USING MULTIPLE DATA SOURCES FOR MAJOR SURVEY PROGRAMS

SHARON L. LOHR (*Chair*), School of Mathematical and Statistical Sciences, Arizona State University (Emerita)
JEAN-FRANÇOIS BEAUMONT, Statistics Canada
LAWRENCE D. BOBO, Office of the Dean of Social Science, Harvard University
MICK P. COUPER, Institute for Social Research, University of Michigan
HILARY HOYNES, Goldman School of Public Policy at the University of California at Berkeley
KIMBERLYN LEARY, Harvard Medical School/McLean Hospital and Department of Health Policy and Management, Harvard T.H. Chan School of Public Health
DAVID MANCUSO, Washington State Department of Social and Health Services
JUDITH A. SELTZER, Department of Sociology, University of California, Los Angeles
ELIZABETH A. STUART, Department of Mental Health, Johns Hopkins Bloomberg School of Public Health
SHAOWEN WANG, Department of Geography and Geographic Information Science, University of Illinois Urbana-Champaign

Study Staff

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

COMMITTEE ON NATIONAL STATISTICS

REVIEWERS

This Consensus Study Report was reviewed in draft form by individuals chosen for their diverse perspectives and technical expertise. The purpose of this independent review is to provide candid and critical comments that will assist the National Academies of Sciences, Engineering, and Medicine in making each published report as sound as possible and to ensure that it meets the institutional standards for quality, objectivity, evidence, and responsiveness to the study charge. The review comments and draft manuscript remain confidential to protect the integrity of the deliberative process.

We thank the following individuals for their review of this report:

KATHARINE G. ABRAHAM, Department of Economics and Joint Program in Survey Methodology, University of Maryland
RICHARD V. BURKHAUSER, Department of Policy Analysis and Management, Cornell University
THOMAS A. LOUIS, Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health
BRUCE D. MEYER, Harris School of Public Policy, University of Chicago
SALLY OBENSKI, U.S. Census Bureau, Retired
RONALD PREVOST, Massive Data Institute, McCourt School of Public Policy, Georgetown University
TRIVELLORE RAGHUNATHAN, Department of Biostatistics, School of Public Health, University of Michigan

Although the reviewers listed above provided many constructive comments and suggestions, they were not asked to endorse the conclusions or recommendations of this report nor did they see the final draft before its release. The review of this report was overseen by Cynthia Clark, independent consultant, and Kathleen Mullan Harris, Department of Sociology, University of North Carolina. They were responsible for making certain that an independent examination of this report was carried out in accordance with the standards of the National Academies and that all review comments were carefully considered. Responsibility for the final content rests entirely with the authoring committee and the National Academies.

ACKNOWLEDGMENTS

This report of the Panel on The Implications of Using Multiple Data Sources for Major Survey Programs is the product of contributions from many colleagues, whom we thank for sharing their time and expertise. The panel was funded by the National Science Foundation, which has been a true partner in this endeavor, and we are especially indebted to Alan Tomkins, Daniel Goroff, Rayvon Fouché, and Cheryl Eavey for their support and for valuable discussions about the panel's goals and activities. Cheryl Eavey opened the workshop with comments about how the panel's activities complement other efforts at the National Science Foundation on enhancing data for social and economic research.

The panel benefitted greatly from the presentations provided during the virtual public workshop held on May 16 and 18, 2022. The experts the panel heard from can be clustered into the following perspectives and areas of expertise (see Appendix A for the workshop agenda and Appendix B for biographies of the workshop presenters):

- *Keynote speakers and discussants*: Robert Santos, Director, U.S. Census Bureau; Anil Arora, Chief Statistician of Canada; Joseph Salvo, University of Virginia; and Haoyi Chen, United Nations.
- *Experts on crime statistics*: Janet Lauritsen, University of Missouri-St. Louis; Ramiro Martinez, Jr., Northeastern University; Erica Smith, U.S. Bureau of Justice Statistics; and Derek Veitenheimer, State of Wisconsin.
- *Experts on agricultural statistics*: Linda Young, U.S. National Agricultural Statistics Service; Herbert Nkwimi-Tchahou, Statistics Canada; Martin Mendez-Costabel, Bayer Crop Science; and Michael Goodchild, University of California-Santa Barbara.
- *Experts on income and health statistics*: Jonathan Rothbaum, U.S. Census Bureau; Lisa Mirel, U.S. National Center for Health Statistics; Jessica Faul, University of Michigan; and Helen Levy, University of Michigan.
- *Experts on data-equity issues*: Steven Brown, Urban Institute; Randall Akee, University of California-Los Angeles; Frauke Kreuter, Ludwig-Maximilians-University of Munich and University of Maryland; Clarence Wardell, Chief Data and Equitable Delivery Officer, Executive Office of the President; and Margaret Levenstein, University of Michigan.

We would also like to thank the Chair of the Committee on National Statistics, Robert M. Groves, for his leadership and his insightful comments about a new vision for national statistics in the final workshop session.

The panel could not have conducted its work without the capable staff at the National Academies of Sciences, Engineering, and Medicine. Brian Harris-Kojetin, Director of the Committee on National Statistics, and Melissa Chiu, Deputy Director, provided invaluable support throughout the panel's activities, and their insightful comments improved the workshop and report. Joshua Lang did a magnificent job of organizing the panel meetings, ensuring the smooth operation of the workshop and other panel activities, and assisting with the report. Neeti Pokhriyal (Mirzayan Science Technology Policy Fellow) helped with literature reviews, and

Constance Citro, Daniel Cork, David Johnson, and Nancy Kirkendall provided helpful input for the report. Kirsten Sampson-Snyder organized the review process, and Susan Debad's thorough editing improved the readability and accessibility of the report. We are grateful to all of them for their contributions and help.

The crew at Spark Street Digital ensured that the technological aspects of the virtual workshop worked flawlessly and produced the video of the event. We appreciate their help in familiarizing participants with the webcast features and their behind-the-scenes support during the workshop.

Finally, we thank the members of the Panel on The Implications of Using Multiple Data Sources with Major Survey programs, listed on page (vi). As can be seen from the biographies in Appendix B, the panel members brought an impressive array of expertise and they generously volunteered their time to organize the workshop, gather evidence, and work on the report. The final report reflects the commitment and expertise of all panel members.

Sharon L. Lohr (*Chair*)
Daniel H. Weinberg (*Study Director*)
Krisztina Marton (*Study Director*)

CONTENTS

BOXES, FIGURES, and TABLES

BOXES

FIGURES

TABLES

LIST OF ACRONYMS USED IN THE REPORT

ACS	American Community Survey
AIAN	American Indian or Alaska Native
ASEC	Annual Social and Economic Supplement [of the Current Population Survey]
BJS	Bureau of Justice Statistics
CAPS	County Agricultural Production Survey
CDC	Centers for Disease Control and Prevention
CID	Comprehensive Income Dataset Project
CNSTAT	Committee on National Statistics
CPS	Current Population Survey
FBI	Federal Bureau of Investigation
FSA	Farm Service Agency
HRS	Health and Retirement Study
HUD	Department of Housing and Urban Development
ICDR	Integrated Client Data Repository [State of Washington]
IRS	Internal Revenue Service
JAS	June Area Survey
LEHD	Longitudinal Employer-Household Dynamics
MAF	Master Address File
NASEM	National Academies of Sciences, Engineering, and Medicine
NASS	National Agricultural Statistics Service
NCHS	National Center for Health Statistics
NCVS	National Crime Victimization Survey
NDI	National Death Index
NEWS	National Experimental Well-being Statistics
NHANES	National Health and Nutrition Examination Survey
NHIS	National Health Interview Survey
NIBRS	National Incident-Based Reporting System [of Uniform Crime Reports]
NVSS	National Vital Statistics System
OMB	Office of Management and Budget
PIK	Protected Identification Key
RMA	Risk Management Agency
SAIPE	Small Area Income and Poverty Estimates
SIPP	Survey of Income and Program Participation
SNAP	Supplemental Nutrition Assistance Program
SRS	Summary Reporting System [of Uniform Crime Reports]
SSA	Social Security Administration
SSN	Social Security Number
TIGER	Topologically Integrated Geographic Encoding and Referencing
UCR	Uniform Crime Reports/Reporting
USDA	U.S. Department of Agriculture

Summary

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; monitor the health and economic circumstances of the population; inform decisionmaking by businesses and individuals; and produce vast quantities of economic, health, and social research that informs the public and can lead to societal benefits. As the National Academies of Sciences, Engineering and Medicine report *Principles and Practices for a Federal Statistical Agency* stated: "It is impossible to capture the full economic and societal value of having reliable data on economic, social, health, agricultural, industrial, and environmental characteristics of the country" (NASEM, 2021b, p. 14).

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 a number of challenges in recent years, including 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 in the 21st century. Each panel was asked to examine a separate aspect of the new data infrastructure. The first panel's report (NASEM, 2023) discussed legal, privacy, and access issues related to using alternative data sources for official statistics, and it identified seven key attributes for a new data infrastructure.

The Statement of Task for this second panel, the Panel on the Implications of Using Multiple Data Sources for Major Survey Programs, directed the panel to examine how survey programs might be affected by the use of alternative data sources, 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.

A diverse panel—with expertise spanning areas of statistics, survey methodology, economics, sociology, psychology, public policy, equity analytics, public health, geography, and

demography—was formed to study these issues. The panel convened a 1.5-day virtual public workshop to seek input from external experts about survey programs that might benefit from use of non-survey data sources, and about how these data sources might be used to produce more accurate, detailed, and timely information. Realizing that no single workshop or report could possibly cover the implications of using multiple data sources for each of the thousands of federal data collections, the panel decided to focus on a small set of "use cases"—from the areas of income, health, crime, and agricultural statistics—that represent different ways in which multiple data sources are, or could be, exploited and that illustrate the types of challenges to be faced. Examples from these areas anchor the discussion of the report's themes.

Use of multiple data sources can add value for the production of official statistics as well as 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 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. Frameworks exist for evaluating the quality of data from probability samples; standards for the quality of integrated data and statistics would promote sound practices and help federal statistical agencies and data users understand these new data products.

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.

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.

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. Alternative data sources can advance data equity by identifying data gaps or misrepresentations, providing information about population members underrepresented in surveys (for example, persons experiencing homelessness or in institutions such as nursing homes), and producing statistics that are disaggregated by race, ethnicity, education, disability status, and other characteristics of interest.

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 3-3: 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.

This report discusses four main ways that multiple data sources could improve national statistics, provide new resources for social and economic research, and promote data equity. These improvements range from providing information for improving current surveys to having the option of replacing surveys altogether. Use of multiple data sources could:

- *Provide information for evaluating and improving quality of data sources.* Administrative and privately held data sources can identify subpopulations that are underrepresented in a sampling frame (a population list from which the sample is drawn) or that are especially prone to nonresponse. Standard survey practice involves comparing estimates of subpopulation sizes calculated from the survey with estimates from an external data source. If records can be linked across sources such that it is possible to identify which (if any) record in source B belongs to the same entity as a record in source A, the linkage can be used to identify, and add, records missing from the frame. This report discusses examples in which non-survey data sources are used to investigate demographic characteristics and socioeconomic status of nonrespondents to income and health surveys (see Chapters 5, 6), to obtain estimates of the number of people killed by law enforcement actions (see Chapter 2), and to identify small urban agricultural operations that are missing from the sampling frame of farms (see Chapters 3, 8). In some cases, information from an administrative source can be used to impute (fill in values using information from a statistical model or similar data records) data items that are missing in a survey.

  Linking records can also identify differences in the measurement of concepts across data sources. Chapter 5 discusses studies that compare income items self-reported on surveys with the same categories from linked tax or earnings records. Such studies are an important prelude to greater use of administrative data to supplement or replace information from surveys.

- *Obtain additional information about survey respondents.* Linking survey records with administrative data sources can provide information not measured in the survey, such as earnings histories and participation in food- or housing-assistance programs (see Chapters 1, 5, 6). Linkage can also provide information about life-course outcomes that occur after the survey, such as subsequent medical expenditures or mortality (see Chapter 6).

- *Produce statistics for small populations.* Survey sample sizes are typically insufficient to produce statistics for small demographic groups or geographic areas

with small populations. Administrative datasets may have large sample sizes but lack information (such as race or ethnicity) that would allow the production of statistics for those groups. Linking records across sources allows statistics to be produced from the administrative records information for groups whose membership is defined in survey or decennial census data. In other situations, information about relationships between race and ethnicity and other variables can be used to impute group membership for administrative data records (see Chapter 3).

Multiple data sources can also be used to produce statistics for small groups without the need to link individual records. This report discusses examples in which statistical models, relying on summary statistics computed from surveys, administrative data, and other sources, are used to produce statistics about income, poverty, health insurance, crime, and agriculture for counties or small demographic groups (see Chapters 2, 3, 7, 8).

- *Create data products and produce statistics directly from administrative data.* In some cases, after thorough research, surveys can be bypassed and statistics produced directly from administrative data sources. Chapter 4 discusses examples of U.S. Census Bureau and state-level projects that link records from various administrative data sources to create new data products.

  Some data sources used to produce statistics have relied on administrative data supplied by state and local governments since their inception. These include the National Vital Statistics System, which tracks births and deaths (see Chapter 4) and the Uniform Crime Reporting Program, which provides estimates of crimes known to the police (see Chapter 7). This report describes the federal-state cooperation that enables creating these datasets, as well as possible modifications that could lead to more timely statistics.

These methods show promise for enhancing data products of the federal statistical system, but care is needed to ensure that the resulting datasets and statistics are of high quality. Administrative and private data sources used to produce statistics should be dependable and continuing sources of accurate information, with consistent measurement of concepts, to ensure that statistics can be compared across times and locations.

CONCLUSION 4-4: Administrative records are a valuable source of information for official statistics and social and economic research. Each administrative records dataset considered for use in creating national statistics needs to be understood in terms of both its original and its proposed uses. This includes assessing the dataset's fitness for use, timeliness, continuing availability, population coverage, measurement of key concepts, and equity aspects.

Statistical methods used to combine information can provide new insights from data, but each method also has the potential to introduce errors. Models used to produce statistics for small geographic areas or to impute missing data values rely on assumptions about relationships

among variables that might not apply uniformly across population subgroups. These assumptions need to be carefully investigated and documented for data users.

Accurate record linkage provides additional information about populations and individual entities. However, when a record from Source A is mistakenly linked to a record from Source B that belongs to a different entity, the linked dataset record has erroneous information. Some data records contain insufficient identifying information to enable linkage across datasets and some subpopulations are more likely to have missed links than others (see Chapters 2, 3, 6). While record linkage can promote data equity by allowing calculation of statistics for small population groups, the method must be rigorously evaluated to identify unintended consequences for measurement and for the communities being measured.

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.

The first report in this series concluded that "[t]rust in a new data infrastructure requires transparency of operations and accountability of the operators, with ongoing engagement of stakeholders" (NASEM, 2023, p. 8). Many of the data products discussed in the current report are new, and the methods used to produce them may be new or unfamiliar. Documentation of all steps in the data-collection and production processes is needed to ensure that data users understand the properties and limitations of the statistics produced.

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.

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.

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.

Probability surveys have provided the nation with useful statistics on numerous topics for more than 80 years, and the panel 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, as discussed above, 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. These changes could include revised consent forms or the addition of new data items.

There is much work to be done. The first report in this series (NASEM, 2023) discussed challenges related to data infrastructure governance and data sharing, and the work needed to overcome those challenges. Many challenges to creating and sustaining a new data infrastructure have not yet been addressed by this or the previous report, and they will be studied in future reports in this series. These include the crucial issues of establishing cyberinfrastructure tailored to integrated data, sharing the benefits of enhanced data resources with researchers and the public while protecting the confidentiality of information contained in the data, investigating issues of data ownership, involving data users and community members in data decisions, and ensuring transparency. The panel believes that these challenges can be met and that a new data infrastructure can be developed to produce improved statistical information for the public good.

1. The Promise of Integrated Data

Probability surveys have been a cornerstone of federal statistics since the 1940s. Back then, almost any kind of data collection was expensive, and probability survey samples provided a way to produce accurate statistics without having to measure everyone. Probability surveys still serve that role, but they have faced a number of challenges in recent years, including declining response rates, increasing costs, and user demand for timelier and more granular data and statistics. Meanwhile, there has been a proliferation of other data 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.

There is increasing interest in using non-survey data sources together with probability surveys to improve official statistics and create new data resources for social and economic research. Data and statistics from the federal government "provide the foundation for policymakers, businesses, and individuals to make informed decisions regarding the economy, society, and their lives. An improved national data infrastructure would provide many societal benefits, including improved decisionmaking and more informed public policy."[1]

The Committee on National Statistics (CNSTAT) in the Division of Behavioral and Social Science and Education of the National Academies of Sciences, Engineering, and Medicine received funding from the National Science Foundation to convene three panels of experts in statistics, economics, social science research, survey methodology, privacy, public policy, and computer science, under the collective title *Toward a Vision for a New Data Infrastructure for Federal Statistics and Social and Economic Research in the 21st Century*.

Box 1-1 gives the Statement of Task for the three panels. Each panel was charged with convening a 1.5-day workshop on particular aspects of a vision for a new data infrastructure and writing a consensus panel report on those aspects. The first panel's workshop, *The Scope, Components, and Characteristics of a 21st Century Data Infrastructure*, was held on December 9 and 16, 2021.[2] This workshop explored recent data infrastructure initiatives in the federal government; presented examples of using private-sector data for statistical purposes; and discussed legal, privacy, and access issues in using alternative data sources for official statistics. Box 1-2 reproduces the seven key attributes for a new data infrastructure from the report on the first workshop. The third scheduled workshop, and possible additional future workshops, will delve more deeply into practical and legal considerations for obtaining access to data, information technology aspects of an infrastructure that draws on multiple data sources, and protecting the privacy of entities supplying data and the confidentiality of the data that are supplied.

[BOXES 1-1, 1-2 about here]

---

[1] https://www.nationalacademies.org/our-work/toward-a-vision-for-a-new-data-infrastructure-for-federal-statistics-and-social-and-economic-research-in-the-21st-century

[2] Video and presentations from the first workshop are available at https://www.nationalacademies.org/event/12-09-2021/the-scope-components-and-key-characteristics-of-a-21st-century-data-infrastructure-workshop-1a and https://www.nationalacademies.org/event/12-16-2021/the-scope-components-and-key-characteristics-of-a-21st-century-data-infrastructure-workshop-1b

The panel for this, the second of the three reports, was specifically directed to concentrate on issues relating to *The Implications of Using Multiple Data Sources for Major Survey Programs.* Which programs might benefit from the use of alternative data sources? How might non-survey data—data such as administrative records that are collected for purposes other than creating official statistics—supplement survey and census data to provide a more accurate, complete, and timely picture of U.S. residents, households, and businesses?

This report builds on previous CNSTAT reports about using multiple data sources to produce statistics and enhance research, including:

- *Modernizing Crime Statistics* (NASEM, 2016a, 2018);
- *Innovations in Federal Statistics: Combining Data Sources While Protecting Privacy* (NASEM, 2017c);
- *Federal Statistics, Multiple Data Sources, and Privacy Protection: Next Steps* (NASEM, 2017a);
- *Improving Crop Estimates by Integrating Multiple Data Sources* (NASEM, 2017b);
- *A Satellite Account to Measure the Retail Transformation: Organizational, Conceptual, and Data Foundations* (NASEM, 2021a);
- *A Vision and Roadmap for Education Statistics* (NASEM, 2022a);
- *Transparency in Statistical Information for the National Center for Science and Engineering Statistics and All Federal Statistical Agencies* (NASEM, 2022e);
- *Modernizing the Consumer Price Index for the 21st Century* (NASEM, 2022d); and
- *Toward a 21st Century National Data Infrastructure: Mobilizing Data for the Common Good* (NASEM, 2023), the report of the first panel in the project "Toward a Vision for a New Data Infrastructure for Federal Statistics and Social and Economic Research in the 21st Century" (Report 1 in Box 1-1).

This report examines current practice and potential for using data originating from administrative records, private-sector organizations, sensors and satellites, and other sources to enhance the timeliness, detail, and accuracy of information currently collected through surveys. The use of multiple data sources can promote data equity, through providing more accurate representation of population subgroups that have historically been underrepresented or misrepresented in the data ecosystem, as discussed in Chapter 3.

The chapter begins with an example that sets the context for the report and a brief discussion of what makes data fit for use. Section 1.1 describes the potential of combined data sources to improve evidence-based policymaking and gives an example in which using multiple data sources to investigate childhood lead exposure resulted in new information that was used to change policy. Section 1.2 discusses frameworks for evaluating the quality of statistics calculated from single and multiple data sources. Section 1.3 describes panel activities and approach to gathering information, and Section 1.4 provides a roadmap to the rest of the report.

1.1 AN EXAMPLE OF ENHANCING SURVEY DATA FOR POLICYMAKING

The U.S. Commission on Evidence-Based Policymaking was formed as the result of bipartisan legislation, the Evidence-Based Policymaking Commission Act of 2016 (U.S. Congress, 2016). One of the explicit charges to the Commission was to "[d]etermine the optimal arrangement for which administrative data, survey data, and related statistical data series may be

integrated and made available for evidence building while protecting privacy and confidentiality" (U.S. Commission on Evidence-Based Policymaking, 2017, p. 7). The Commission's final report stated: "There are many barriers to the effective use of government data to generate evidence. Better access to these data holds the potential for substantial gains for society. The Commission's recommendations recognize that the country's laws and practices are not currently optimized to support the use of data for evidence building, nor in a manner that best protects privacy" (p. 1). It also noted: "The strategy outlined in the Commission's report simultaneously improves privacy protections and makes better use of data the government already collects to support policymaking" (p. 3).

The ensuing Foundations for Evidence-Based Policymaking Act of 2018 (U.S. Congress, 2019) has become the cornerstone for many projected improvements in U.S. statistics. The Commission's report included a number of examples that demonstrated "the promise of evidence-based policymaking," specifically noting that "administrative data, collected in the first instance to serve routine program operation purposes, also can be used to assess how well programs are achieving their intended goals" (U.S. Commission on Evidence-Based Policymaking, 2017, p. 9). Examples included using administrative records to study permanent supportive housing for chronic homelessness, substance abuse education, and workforce investment. The Commission also pointed to the value of reducing the burden on survey respondents:

> Respondents have become less willing to participate in surveys and are increasingly reluctant to respond to questions about income. When they do answer questions about income, they are providing less accurate responses. The burden on respondents could be reduced and the accuracy of the data improved if statistical agencies were able to rely more on the income data the government already maintains to administer tax, income support, and social insurance programs (U.S. Commission on Evidence-Based Policymaking, 2017, p. 25).

Mirel (2022) reported on an example that showed how using multiple data sources could promote evidence-based policymaking for improving public health. In children, even small amounts of lead exposure can cause serious and irreversible mental and physical health problems; high levels can be fatal. But childhood lead poisoning can be prevented. Large declines "in blood lead levels occurred from the 1970s to the 1990s following the elimination of lead in motor-vehicle gasoline, the ban on lead paint for residential use, removal of lead from solder in food cans, bans on the use of lead pipes and plumbing fixtures and other limitations on the uses of lead" (President's Task Force on Environmental Health Risks and Safety Risks to Children, 2016, p. 5).

These declines are known to have occurred because the National Health and Nutrition Examination Survey (NHANES), a nationally representative survey initiated in 1960 to assess the health and nutrition status of adults and children in the United States, began measuring blood lead levels in 1976. According to NHANES data, the median blood lead level in children aged 1–5 dropped from 15 micrograms per deciliter in 1976–1980 to 0.6 micrograms per deciliter in 2017–2018, with most of the reduction occurring before 1990 (U.S. Environmental Protection Agency, 2022).

Despite this progress, "lead exposure remains an important public health problem among children particularly for those in high-risk groups" (Egan et al., 2021, p. 10).[3] A major source of childhood lead exposure in the United States "is lead-based paint and lead-contaminated dust found in buildings built before 1978."[4] Using data from the American Healthy Homes Survey, the U.S. Department of Housing and Urban Development (HUD) estimated that, in 2019, approximately 35 million housing units contained lead-based paint somewhere in the building, with about 90 percent of those units built before 1978 (HUD, 2021). Households receiving government housing assistance had statistically significantly lower levels of lead-based paint hazards than those not receiving assistance (11% versus 20%).[5]

While there were indications that HUD-assisted housing units had lower levels of lead hazards, no single dataset included both designations of HUD-assisted housing and information on children's blood lead levels, which would enable evaluating associations between children's health and living in HUD-assisted housing. The NHANES data contained blood lead levels and other health information about respondents, but no information on whether respondents lived in assisted housing. HUD's annual data about participants in housing-assistance programs (administrative records collected through the local housing authorities that administer the programs) had no information on tenants' health.

To study health characteristics (including blood lead levels) of children and adults living in HUD-assisted housing, HUD collaborated with the U.S. National Center for Health Statistics (NCHS), which administers the NHANES (Mirel et al., 2019a). Data from the 1999–2012 NHANES were linked to records for the same households in the HUD tenant data (with strict controls over access to those linked data).[6]

Researchers analyzing the linked data found that children living in HUD-assisted housing from 2005–2012 had lower blood lead levels than comparable children who did not receive housing assistance (see Ahrens et al., 2016). HUD used evidence from this and other observational research conducted on the linked NCHS-HUD data "to support the continued removal of lead-based paint hazards in HUD homes" and "cited this evidence in a proposed rule to lower the threshold for elevated blood lead level determination to align with CDC [Centers for Disease Control and Prevention] standards" (Mirel, 2022, slide 6).

By linking administrative records from HUD with survey data from NHANES, investigators could identify children in the NHANES dataset who lived in federally assisted

---

[3]Egan et al. (2021), analyzing NHANES data between 1976–2016, found that higher childhood blood lead levels were associated with non-Hispanic Black race/ethnicity, having family income below 130 percent of the poverty level, and living in older housing. See Rabin (1989) for a history of childhood lead poisoning in the United States.

[4]https://www.cdc.gov/nceh/tracking/topics/ChildhoodLeadPoisoning.htm

[5]The Residential Lead-Based Paint Hazard Reduction Act of 1992 (U.S. Congress, 1992) and other legislation instituted requirements for lead-based paint notification, evaluation, and reduction for housing receiving federal assistance.

[6]Lloyd et al. (2017) described the linkage process (also see NCHS, 2022c). To be eligible for linkage to HUD data, a NHANES participant must have consented for their data to be linked and provided sufficient data elements (including full or partial Social Security Number, full name, and month and year of birth) for the linkage to be attempted. About 65 percent of the 1999–2012 NHANES medical examination participants were eligible for linkage, and about 13 percent of those were linked to the HUD data (Lloyd et al., 2017, p. 14). In analyses using the linked data, NHANES participants who were matched with a record in the HUD data were considered to be receiving housing assistance, and linkage-eligible NHANES participants who could not be matched with a record in the HUD data were considered to be not receiving housing assistance. See Chapters 2, 3, and 6.

housing. Linking the two datasets produced information not available from either source by itself, without requiring additional data collection. The editors of the volume *Evidence Works* concluded: "Combining data can produce valuable insights" (Hart and Yohannes, 2019, p. 121).

1.2 PRODUCING STATISTICS THAT ARE FIT FOR USE

The U.S. Office of Management and Budget's (OMB) Statistical Policy Directive No. 1 states: "It is the responsibility of Federal statistical agencies and recognized statistical units to produce and disseminate relevant and timely information; conduct credible, accurate, and objective statistical activities; and protect the trust of information providers by ensuring confidentiality and exclusive statistical use of their responses…" (OMB, 2014, p. 71614). In 2021, OMB also issued guidance for implementing the Foundations for Evidence-Based Policymaking Act of 2018, which is part of a large collection of laws and regulations governing data sharing within the federal statistical system and with the public.[7] The guidance specified that the data be "fit for use" or "fit for purpose":

> Underlying all of the methodological approaches outlined here are the data collected and used in Federal evidence-building activities. Ensuring that those data are reliable, high-quality, and fit for their intended purpose is essential to restoring trust in Government (OMB, 2021, p. 11).

OMB's *Federal Data Strategy* was designed to create "a framework of operational principles and best practices that help agencies deliver on the promise of data in the 21st century" (OMB, 2019a, p. 1). In addition to desiring that agencies implement ethical governance and create a learning culture, the strategy specifically addressed four elements of "conscious design":

- *Ensure Relevance*: Protect the quality and integrity of the data. Validate that data are appropriate, accurate, objective, accessible, useful, understandable, and timely.
- *Harness Existing Data*: Identify data needs to inform priority research and policy questions; reuse data if possible and acquire additional data if needed.
- *Anticipate Future Uses*: Create data framework thoughtfully, considering fitness for use by others; plan for reuse and build in interoperability from the start.
- *Demonstrate Responsiveness*: Improve data collection, analysis, and dissemination with ongoing input from users and stakeholders. The feedback process is cyclical; establish a baseline, gain support, collaborate, and refine continuously (OMB, 2019a, pp. 2–3).

These OMB guidelines emphasize the importance of validating the quality of data and ensuring that they are fit for use—not just for the immediate purpose but also for possible future reuse. Groves and Lyberg (2010, p. 873) noted: "Because statistics are of little importance without their use being specified, 'fitness for use' is of growing importance as a quality concept.

[7] *Principles and Practices for a Federal Statistical Agency* (NASEM, 2021b, Appendix A) lists laws and standards that govern federal data collection and sharing.

… It is relatively common for national statistical agencies to refer to their quality frameworks as a means to achieve fitness for use."

Traditionally, statistics from probability surveys have been accompanied by margins of error or confidence intervals that provide a measure of their accuracy. Modern data-quality frameworks, however, argue that quality is multidimensional:

> Quality is defined as "fitness for use" in terms of user needs. This definition is broader than has been customary [sic] used in the past when quality was equated with accuracy. It is now generally recognised that there are other important dimensions. Even if data is accurate, they cannot be said to be of good quality if they are produced too late to be useful, or cannot be easily accessed, or appear to conflict with other data. Thus, quality is viewed as a multi-faceted concept. The quality characteristics of most importance depend on user perspectives, needs and priorities, which vary across groups of users…. [T]he OECD views quality in terms of seven dimensions: relevance; accuracy; credibility; timeliness; accessibility; interpretability; and coherence (Organisation for Economic Co-operation and Development, 2012, p. 7).

More recent statements on data quality have kept the same seven basic dimensions of quality but have added guidelines for assessing the quality of integrated data sources (Federal Committee on Statistical Methodology, 2018, 2020; Statistics Canada, 2019, 2022; Eurostat, 2021; see also the review of international quality standards in Czajka and Stange, 2018).

The Federal Committee on Statistical Methodology (2020, p. 2) also added a dimension of public trust to earlier ideas of "fitness for use," defining data quality as "the degree to which data capture the desired information using appropriate methodology in a manner that sustains public trust." Their data-quality framework, reproduced in Figure 1-1, encompasses 11 dimensions, categorized within the broader headings of utility, objectivity, and integrity. [FIGURE 1-1 about here]

As Brackstone (1999, p. 140) noted, dimensions of quality "are not independent of each other…. Accuracy and timeliness often have to be traded off against each other. Coherence and relevance can sometimes be in conflict as the needs of current relevance and historical consistency compete. Information provided to ensure information is interpretable will also serve to define its coherence." For example, making efforts to obtain complete data, studying measurement properties, cleaning data, and evaluating sources of uncertainty in individual and combined data sources all contribute to increased accuracy but also increase the amount of time needed to produce statistics.

The use of alternative data sources such as administrative records complicates the assessment of quality because of the many types of data sources and the many paths that can be taken to integrate data and statistics (see Chapter 2). Each individual data source has its own quality profile with respect to the dimensions in Figure 1-1 When multiple data sources are combined, quality assessments must consider the quality of each source as well as the quality of the combined data.

Paths for using multiple data sources, and possible implications for data quality, include:

- Using administrative records directly to give a picture of the population found in the administrative records system (see Chapter 4). In some situations, an administrative

data source may replace a survey; in such cases, it is important to ensure that statistics produced by the administrative data can be compared with previous statistics produced by the survey.

- Using administrative records or other data sources as input to statistical models developed to estimate population characteristics, as in the U.S. Census Bureau's Small Area Income and Poverty Estimates program (see Box 2-2) or the National Agricultural Statistics Service's Crops County Estimates Program (see Chapter 8). Quality assessment involves evaluating the performance of both the statistical models and the individual data sources.
- Linking administrative records or private-sector data records with records from a survey or the decennial census, to extend the number of attributes known about the entities in the survey or census. When individual records from a survey are linked with those from an administrative records dataset (as in the blood lead example discussed in Section 1.1) the accuracy of statistics calculated from the linked data depends on the quality of each individual data source, the accuracy of the data linkage, and the characteristics of the linked dataset.
- Merging datasets or integrating statistics calculated from separate datasets to compensate for the underrepresentation of certain population subgroups in some of the data sources. For example, information from a survey of the civilian noninstitutional population might be combined with information collected from institutions such as prisons and nursing homes. The quality of estimates depends on that of each source and on the alignment of the data sources (sometimes the same entity appears in multiple sources and duplication must be identified when producing population statistics). In addition, the concepts might be measured differently in the data sources, and possible consequences of the measurement differences need to be investigated.

For all of these paths, the resulting integrated datasets and statistics must be of sufficient quality to meet user needs. The National Academies (NASEM, 2017a, p. 109) emphasized that "the quality of administrative and private-sector data sources needs careful examination before being used for federal statistics," because of "the relatively recent novelty of the simultaneous use of multiple data sources and the fact that some potential new sources of data present new issues of data quality." The United Nations Inter-Secretariat Working Group on Household Surveys (2022) and Chen (2022) emphasized the importance of establishing a "total quality framework" for data integration.

One important aspect of fitness for use involves regularly produced statistics that are used to monitor aspects of society. Consistent measurement of statistics such as monthly unemployment rates or annual crime rates facilitates comparisons across time periods and geographic locations. Switching to administrative records or combined data sources may affect the time series for these indicators, and these potential effects need to be thoroughly investigated.

The use of multiple data sources can help improve the quality of data collected in surveys, even if the data are not combined. For instance, linking records for two sources that each measure wage income can provide information that can be used to improve income measurement. Non-survey data can also improve the quality of probability surveys by augmenting the sampling frame or providing information that can be used to adjust for nonresponse.

1.3 STUDY APPROACH AND INFORMATION GATHERING

Between December, 2021 and September, 2022, the panel held 9 closed virtual meetings to organize the 1.5-day workshop, decide on the study conclusions, and discuss drafts of the report.

Three early panel decisions defined the scope of the project:

- The federal government collects data on thousands of topics every year, from seat belt use to welfare of veterans to household energy consumption to adult literacy.[8] No report of reasonable length could possibly cover the implications of using multiple data sources for each of these surveys. The panel decided to focus on a small set of "use cases" that represent different ways that multiple data sources are, or could be, exploited and that illustrate the types of challenges to be faced.
    - Statistical agencies and researchers in the areas of income and health statistics have done extensive work on methods for linking survey and administrative records datasets. The panel decided to devote a workshop session to recent data-linkage projects involving income and health data that illustrate the current "state of the art" and show the potential for data linkage in other subject areas. These projects involved both cross-sectional datasets, which contain information for one time point, and longitudinal datasets, which follow individuals or businesses over time.
    - Crime statistics published by the Federal Bureau of Investigation are compiled from information submitted by individual law enforcement agencies (data submission is usually coordinated through state programs). The data collection is intended to be a census of the more than 18,000 law enforcement agencies in the United States. Challenges include missing data and ensuring consistency in the measurement of crime across agencies and across time.
    - Survey data about agriculture can be enhanced using information from administrative records, satellites, and sensors. In this application, survey data are collected on farm operations, as opposed to individual persons, and some of the issues faced are similar to those in other establishment surveys. Challenges include aligning geographic units in the data sources and developing models to produce crop estimates for small geographic areas.

[8]Many surveys are collected by the 13 principal U.S. statistical agencies (see NASEM, 2021b, Appendix B): the Bureau of Economic Analysis (U.S. Department of Commerce), Bureau of Justice Statistics (U.S. Department of Justice), Bureau of Labor Statistics (U.S. Department of Labor), Bureau of Transportation Statistics (U.S. Department of Transportation), Census Bureau (U.S. Department of Commerce), Economic Research Service (U.S. Department of Agriculture), Energy Information Administration (U.S. Department of Energy), National Agricultural Statistics Service (U.S. Department of Agriculture), National Center for Education Statistics (U.S. Department of Education), National Center for Health Statistics (U.S. Department of Health and Human Services), National Center for Science and Engineering Statistics (National Science Foundation), Office of Research, Evaluation, and Statistics (Social Security Administration), and Statistics of Income (U.S. Department of the Treasury). Other federal agencies also collect data; for example, the National Highway Traffic and Safety Administration (U.S. Department of Transportation) collects data on traffic crashes and seat belt use.

- The panel was tasked with examining the implications of using multiple data sources "to assess and enhance representativeness" and "for population subgroup coverage" (see Box 1-1). The panel decided to address these issues through the lens of data equity, examining how multiple data sources might affect the representation of population subgroups that have historically been underrepresented or misrepresented in the data record.
- The panel decided to exclude or de-emphasize topics that, while essential for the development of a new data infrastructure that uses multiple sources of data, were delineated in the Statement of Task (Box 1-1) as primary focuses of Reports 1 and 3. Thus, this workshop and report do not include extensive discussions of:
    - Legal agreements needed for data sharing;
    - Computer infrastructure for blended data;
    - Methods for providing public access to data; and
    - Methods for protecting the privacy and confidentiality of people, businesses, and other entities whose data are used.

  The panel recognizes, however, that these issues are crucial considerations and that the work ahead must integrate them into the vision for a new data infrastructure.

The public virtual workshop on "Implications of Using Multiple Data Sources for Major Survey Programs" was held on May 16 and 18, 2022. The five sessions of the workshop were organized according to decisions outlined above, with an overview session followed by the use cases and a final session on data equity:

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

The full agenda for the workshop is provided in Appendix A, and video and presentation slides are available online.[9] The panel asked workshop participants to explore how using alternative data sources such as administrative records, health records, satellite and sensor data, and private-sector data can improve the quality, granularity, timeliness, and equity of data in major survey programs.

This report relies on information presented by experts from federal and state governments, academic institutions, and international statistical organizations who participated in the workshop; public comments made during the workshop; and comments from the report's reviewers. In addition, panel members reviewed more than 800 books, research articles, technical reports, and informational websites to provide additional examples and background for the discussion. The report reflects the information available as of the fall of 2022, when the panel completed the bulk of the work on this report.

[9] https://www.nationalacademies.org/event/05-16-2022/the-implications-of-using-multiple-data-sources-for-major-survey-programs-workshop

1.4 ORGANIZATION OF THE REPORT

The remainder of the report proceeds as follows. Chapter 2 discusses various data types and their sources: probability samples; administrative records; private-sector data; satellite, sensor, and location data; convenience samples; and data obtained from social media, webscraping, and crowdsourcing. It also outlines some of the methods that can be used to combine data from multiple sources, such as linking data records, combining statistics from multiple sources, and using statistical models to predict values for missing data and to merge information from separate data sources.

Chapter 3 introduces the key theme of data equity. The chapter starts by defining aspects of data equity and then looks at examples of how using multiple data sources can improve the representation of population groups that have historically been underrepresented, unmeasured, or mismeasured in the data record. It also explores how misuses of available data sources might exacerbate data inequity.

Chapter 4 focuses on examples in which administrative records are used directly to produce statistics, largely bypassing surveys. The chapter begins with a description of three longitudinal databases assembled by the U.S. Census Bureau to study economic activity and population dynamics. The chapter then describes the *Frames* project, underway at the U.S. Census Bureau, is intended to link information from the Bureau's various databases to improve accuracy and inclusiveness of population and business listings maintained for drawing probability samples and other purposes. The National Vital Statistics System, coordinated by NCHS, is a model for cooperation in building an administrative data system based on data submissions by states. State-level systems of linked administrative records demonstrate both the promise of integrated data and the challenges of harmonizing data concepts across sources.

Chapters 5–8 concentrate on four subject areas—income, health, crime, and agriculture—each with a different experience in their use of administrative records and other non-survey data. Chapters 5 and 6 focus on the extensive programs of data linkage that have been implemented or are in progress for improving income and health statistics, respectively. Chapter 5 emphasizes the use of administrative data to study properties of income measurement, while Chapter 6 focuses on the ability to add data about health outcomes and expenditures to the records of survey participants. Chapter 7 discusses challenges in measuring crime as the Uniform Crime Reporting Program, which collects data on criminal offenses from law enforcement agencies, has migrated from a system that measured only counts of offenses to a system that records detailed information about the victims, offenders, and characteristics of incidents—but with fewer law enforcement agencies providing data to the federal government. The chapter discusses the potential for using statistical modeling and linkage to provide increased geographic and subpopulation detail and more timely statistics. Chapter 8 focuses on agricultural statistics, where external data sources including administrative and satellite data are already being used to improve crop estimates.

Chapter 9 concludes the report with a discussion of common themes for the case studies and opportunities and challenges for moving forward.

BOX 1-1 Statement of Task

The National Academies of Sciences, Engineering, and Medicine will appoint an ad hoc committee to produce three complementary reports on topics that will help guide the development of a vision for a new data infrastructure for federal statistics and social and economic research in the 21st century. The topics the committee will explore include the following:

Report 1: The components and key characteristics of a 21st Century Data Infrastructure including:

- The challenges and opportunities related to data infrastructure governance;
- The skills, capabilities, techniques, and methods required by the new data infrastructure; and
- Issues related to sharing non-traditional data assets, including state and local government, institutional, private sector, and sensor data;

Report 2: 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;*

Report 3: The technology, tools, and capabilities needed for data sharing, use, and analysis, including:

- Alternative approaches and techniques for protecting privacy and confidentiality;
- Alternative sustainable organizational models for data sharing; and
- Approaches to ensure transparency of the datasets, the use of the data, and the resulting products.

The committee for each report will convene a 1.5-day virtual public workshop for each topic to seek input from key stakeholders and external experts relevant to the specific charge. Each committee will issue a report that summarizes the committee's findings and conclusions from the workshop and other information gathered relevant to the charge, as appropriate. These reports will help inform a vision for a new data infrastructure and will not include recommendations. The three reports will follow institutional guidelines and be subject to the National Academies review procedures prior to release.

[END Box 1-1]

BOX 1-2 Seven Attributes of a 21st Century National Data Infrastructure Vision

1. Safeguards and advanced privacy-enhancing practices to minimize possible individual harm.
2. Statistical uses only, for common-good information, with statistical aggregates freely shared with all.
3. Mobilization of relevant digital data assets, blended in statistical aggregates to providing benefits to data holders, with societal benefits proportionate to possible costs and risks.
4. Reformed legal authorities protecting all parties' interests.
5. Governance framework and standards effectively supporting operations.
6. Transparency to the public regarding analytical operations using the infrastructure.
7. State-of-the-art practices for access, statistical, coordination, and computational activities; continuously improved to efficiently create increasingly secure and useful information.

SOURCE: National Academies of Sciences, Engineering, and Medicine Report *Toward a 21st Century National Data Infrastructure: Mobilizing Data for the Common Good* (NASEM, 2023, p. 4).

[END Box 1-2]

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.

FIGURE 1-1 Dimensions of data quality.

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

[END Figure 1-1]

2. Types of Data and Methods for Combining Them

This chapter briefly discusses types of data, including government surveys and data collected by government agencies while administering programs (administrative records), as well as non-governmental data from the private sector, sensors and satellites, the internet, and other sources. The chapter also discusses some of the methods that have been proposed for combining these data, providing background information for subsequent chapters discussing ways that federal statistical programs use, or could use, alternative data sources. These issues are described in lay terms, and readers interested in technical details are referred to the cited literature. The panel's approach complements that of the first National Academies of Sciences, Engineering, and Medicine report in this series (NASEM, 2023), which discusses issues of acquiring data, data governance, and key desired attributes of a new data infrastructure. Many of the data sources considered contain records on individual persons or businesses, raising concerns about informed consent, privacy, and confidentiality. Privacy issues will be studied in depth in subsequent reports in this series, but it is important to note that finding workable solutions to address these concerns is essential for data sharing and linkages.

2.1 TYPES OF DATA SOURCES

In the 19th and early 20th centuries, when many of the federal statistical agencies were formed, data were scarce and difficult to obtain. Today, data are plentiful. However, some data sources are more suitable than others for providing statistics that are fit for a particular use (see Chapter 1). Citro (2014) and the National Academies (NASEM, 2017c) detailed advantages and disadvantages of various types of data for producing official statistics. This section briefly describes the major types of data discussed in this report. Table 2-1 summarizes features of these data types that are related to their fitness for use.
[TABLE 2-1 about here]

Probability Samples

The 1930s were a period of devastating poverty and unemployment in the United States. Newspapers showed pictures of people standing in long bread lines; many people suffered the effects of unemployment personally. How many? No one knew. Estimates of unemployment varied widely, depending on the data sources and calculation techniques. There was no definitive source of information about conditions of the labor force and no way to track changes over time.

Throughout the 1930s, economists and statisticians explored methods for obtaining reliable, objective measures of unemployment and month-to-month changes. These efforts culminated in an April 1940 memorandum titled *Monthly Report for Unemployment*, which estimated the national unemployment rate to be 15 percent. This statistic was based on a survey of 8,000 households taken in March 1940, and it marked the first unemployment rate estimate from the survey now known as the Current Population Survey (CPS)—the longest-running probability survey in the United States.

Two major characteristics set the CPS apart from previous methods of measuring unemployment.

1. The CPS is a *probability sample.* Housing units, and non-institutional group quarters such as rooming houses and college dormitories, are randomly selected from the U.S. Census Bureau's Master Address File, a continually updated list of addresses in the United States. Each month's sample consists of about 60,000 eligible housing units.[10] The key characteristic of a probability sample, under ideal implementation, is that each subset of the population has a known, nonzero probability of being included in the sample. These probabilities can then be used to give an accurate assessment of the precision of each statistic calculated from the sample, typically through a margin of error or a confidence interval.[11]
2. The data are gathered for the express purpose of calculating statistics about the labor force. The statistical agency thus controls what information is collected and how it is collected. For the CPS, this means that unemployment is measured using questions specifically designed and tested for that purpose. Month-to-month and year-to-year changes in the unemployment rate can be calculated because unemployment is measured the same way every month.[12]

Following the success of the CPS, probability samples became widely used throughout the federal government. They allowed for faster and more frequent data collection than a population census, because accurate statistics for the nation as a whole could be calculated from a relatively small sample.[13] In 2022, probability samples still form the foundation for statistics in areas ranging from health to crime to agriculture to economic activity. The U.S. Census Bureau alone conducts more than 100 surveys of households and businesses each year.[14]

---

[10]See https://www.bls.gov/opub/hom/cps/design.htm and U.S. Census Bureau (2019) for descriptions of how housing units are selected for the CPS. The CPS has used a full probability sampling design since October 1943. See Bregger (1984) and Dunn, Haugen, and Kang (2018) for the history of the CPS.

[11]The following references, ordered from least to most technical, describe how probability sampling works: *Methods 101* videos from Pew Research (2017), which illustrate the basic ideas of random sampling (other videos in the series discuss issues such as question wording and mode effects); Appendix B of Federal Committee for Statistical Methodology (2020), which describes sources of error in various types of data, including probability samples; Kalton (2020), a short book describing the basics of probability sampling; Lohr (2022), a longer book describing how to design and analyze data from probability samples; and Skinner and Wakefield (2017), a compact and comprehensive description of survey methods at a high technical level.

The selection probabilities for a probability sample can be unequal, as long as they are known. For example, CPS sample sizes for states are determined so that each state-level estimate of unemployment, and the national estimate of unemployment, will attain a prespecified level of accuracy. States with smaller populations need higher sampling ratios to achieve that accuracy. The differing designed selection probabilities are accounted for in the estimation process (U.S. Census Bureau, 2019; https://www.bls.gov/opub/hom/cps/design.htm).

[12]The CPS questions have been revised at various points in its history, but revisions have been implemented such that changes over time can still be calculated. When the CPS questionnaire was revised in 1994, for example, the old and new questionnaires were run in parallel for 18 months so that statisticians could assess how changes in the questionnaire would affect labor force estimates (U.S. Census Bureau, 2019, p. 112).

[13]Of course, a probability sample does not have to be small. Technically, the U.S. Decennial Census of Population and Housing and other censuses can be considered to be probability samples because they are designed so that each subset of the population has a known probability (set equal to one) of being included in the census.

[14]https://www.census.gov/programs-surveys/surveyhelp/list-of-surveys.html

Despite their widespread use, however, probability surveys face challenges that diminish their usefulness as a sole source of information. These include:

- *Undercoverage and nonresponse.* As mentioned earlier, when population units are selected according to the probabilities specified by the survey design, those probabilities can be used to compute confidence intervals or margins of error that accurately quantify the accuracy of statistics calculated from the sample; in other words, the sample is designed to be *representative* of the population. But when the data are collected, the actual probability with which a unit is selected for the sample can end up differing from the designed probability, and this may result in a sample that is no longer representative of the population. The two main causes of discrepancies between the designed and actual probabilities (which may be unknown) are undercoverage and nonresponse.

  Units in a probability sample are drawn from a *sampling frame,* which is typically a list or map of units in the target population. *Undercoverage* occurs when some target population units are missing from the sampling frame—these missing units then have a zero probability of being included in the sample. The CPS target population is the civilian noninstitutional population aged 16 and over, so some parts of the U.S. population are excluded by design. Since the CPS's sampling frame is based on housing units, there is undercoverage of groups that are not in that frame, such as people experiencing temporary homelessness.

  Nonresponse occurs when units selected to be in the sample fail to provide data. Figure 2-1 displays response rates from 2000–2022 for the major surveys discussed in this report. Response rates for both face-to-face surveys and telephone surveys have decreased since the early 1990s, with the rate of decrease accelerating since 2010 (Williams and Brick, 2018; Dutwin and Buskirk, 2021).
  [FIGURE 2-1 about here]

  Undercoverage and low response rates are of concern because entities that cannot be selected for the sample or that fail to respond to a survey can differ systematically from entities that participate, causing survey estimates to be biased. Survey producers attempt to adjust for nonresponse through statistical modeling techniques involving weighting and imputation, but these methods are not guaranteed to remove nonresponse bias, especially that due to unobserved factors.[15]

- *Timeliness and frequency.* The timeliness of a data product is the length of time between the events described in the data and the data product's availability.

[15]The amount of bias in a survey estimate depends on: (1) the rates of undercoverage and nonresponse; and (2) how much survey participants differ from nonrespondents and persons who are not in the sampling frame. If respondents are similar to nonrespondents, then estimates from a survey will be approximately "on target" even with a low response rate (Groves, 2006). Conversely, a survey with a 90 percent response rate may have high bias if the nonrespondents have large systematic differences from the respondents. Survey weighting adjustments attempt to reduce the bias by using information known about responding and nonresponding units to obtain better estimates of the probabilities that each responding unit will appear in the final sample. See Mercer, Lau, and Kennedy (2018) for a nontechnical description of how weighting methods work.

Frequency relates to how often survey estimates are published. Timeliness and frequency vary among surveys—for example, the CPS produces monthly estimates of unemployment within four weeks of data collection; but the National Crime Victimization Survey (NCVS; see Chapter 7) publishes annual estimates of the number of violent and property crimes for a calendar year in September or October of the following year. Many statistics about health insurance coverage, transportation usage, education, agriculture, capital expenditures by businesses, and other topics, are published once a year. Some surveys, such as those used in the Centers for Disease Control and Prevention's (CDC) Youth Risk Behavior Surveillance System,[16] are conducted every 2 or 3 years, and other data collections are conducted even less frequently.

Some surveys may have a tradeoff between timeliness and other aspects of quality, such as accuracy and coherence (see Figure 1-1). For example, the Household Pulse Survey, launched by the U.S. Census Bureau in April 2020, was designed to obtain low-cost, rapid information about social and economic effects of the COVID-19 pandemic, including such topics as child care, education, food security, health, and household spending. Because the survey was collected online with only a short period for following up with nonrespondents, statistics from the Household Pulse Survey could be published within a few weeks of data collection. But the speed and cost savings came at the expense of response rate and coverage. The sample was drawn from the set of Master Address File units for which at least one e-mail address or telephone number was known, thus excluding from the survey those housing units without an associated telephone number or e-mail address. Response rates for the Household Pulse Surveys conducted between April and July 2020 were less than 5 percent.[17]

- *Granularity.* A high-quality probability survey of 60,000 households will give accurate estimates for the nation as a whole and for some population subgroups. But the sample size is too small for the survey, by itself, to give accurate estimates for smaller units of geography such as counties, or for demographic groups that account for a small percentage of the U.S. population. For cost and logistical reasons, surveys are often unable to meet the data demands for the high level of granularity needed to understand key population subgroups, which can result in data inequities (see Chapter 3).

  For example, NCVS annual data files contain information from about 150,000 household interviews and 240,000 person interviews each year but, because most persons are not robbed during the year, the dataset contains only a small number of robbery victims. A larger probability survey can provide more accurate statistics for

[16] https://www.cdc.gov/healthyyouth/data/yrbs/index.htm

[17] Information on the Household Pulse Survey can be found at https://www.census.gov/programs-surveys/household-pulse-survey.html. Technical documentation (see, for example, U.S. Census Bureau, 2022c and Peterson et al., 2021) described the weighting procedures used to adjust for undercoverage and nonresponse and provided a nonresponse bias assessment. Bradley et al. (2021) performed additional examinations of nonresponse bias by comparing Household Pulse Survey estimates with statistics from other data sources.

small population subgroups but can also be more expensive to conduct. In a typical year, a sample of about 3.5 million households is selected for the American Community Survey (ACS) and the large sample size allows annual estimates to be produced for geographic areas with populations of 65,000 or more. For areas with smaller populations, however, the survey sample size is still too small to produce accurate annual estimates. The U.S. Census Bureau takes two approaches to produce estimates for small-population areas. The first approach accumulates ACS data for a 5-year period, thereby approximately quintupling the sample size in the area but at the cost of having older data. The second approach incorporates data from other sources such as tax records to estimate statistics for the current year (see Section 2.2).

Statistics computed from probability surveys typically rely on data provided by respondents. If respondents provide inaccurate information—for example, if a respondent to the ACS misreports a household member's age, race, or income—those measurement errors affect the accuracy of statistics computed from the survey. Measurement errors, discussed in Groves et al. (2011), affect all the data sources discussed in this section.

Probability surveys have provided the nation with useful statistics on numerous topics for more than 80 years, and the panel anticipates that these surveys will continue to be used to produce statistics in many topic areas, particularly at the national level. Some statistics, such as the percentage of persons looking for work last week or the percentage of criminal victimizations reported to the police, rely on information that can only be provided by individuals in the population—a probability survey is often still the best method for collecting information on such topics. The main advantages of probability surveys are their ability to represent the entire population (subject to undercoverage and nonresponse) and the statistical agency's control over what information is collected and how concepts are measured. Probability surveys can also provide an independent check on other sources of information, such as data from private-sector sources or convenience samples, because the random selection process in a probability survey controls biases in sample selection.

Probability surveys are often costly, however, and provide data on a relatively small sample of the population. Decreasing response rates in recent years raise concern that estimates calculated from surveys might not represent characteristics of nonrespondents. In general, even a relatively large national probability survey will not be able, by itself, to provide accurate information about small subpopulations such as counties, school districts, or demographic groups that form small parts of the U.S. population. Alternative data sources may be able to improve the accuracy, timeliness, and granularity of statistics while reducing costs.

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.

Administrative Records Collected by Government Agencies

Administrative records are data collected by a federal, state, local, or tribal government agency for administrative purposes such as operating a program. Examples include income tax

data, Social Security Administration (SSA) benefit records, health care claims data from the Centers for Medicare and Medicaid Services, student data collected by state Departments of Education, data on planted acreage from federal crop insurance programs, and data from food assistance and transfer programs such as the Supplemental Nutrition Assistance Program (SNAP).

Administrative records datasets are typically large and contain detailed information. Many are longstanding and collect data on a regular basis. This allows calculation of annual (or more frequent) statistics describing the population included in the administrative records. In addition, when the same units reappear at multiple time periods, it may be possible to create longitudinal datasets that follow the units over time. For example, longitudinal datasets created from corporate tax records can be used to study business formation and growth (see Section 4.1).

Due to their scope, administrative records datasets often contain large sample sizes for subpopulations that might be represented by only a few persons in a probability sample. For example, the NCVS public-use dataset typically contains fewer than 300 robbery victims each year. The Federal Bureau of Investigation's (FBI) National Incident-Based Reporting System (NIBRS), which compiles data collected by law enforcement agencies, contains details on more than 100,000 robbery incidents that occurred in 2020.[18] This sample size allows a researcher to study characteristics of robberies for population subgroups (for example, characteristics of robberies occurring at night with female victims) in a way that is not possible with one year of data from the NCVS.

However, NIBRS records include only crimes that are known to law enforcement agencies that submit data to the FBI (see Chapter 7). If statistics are desired on all robberies—those unreported to the police as well as those appearing in NIBRS—then an additional source of information is needed. Similarly, a researcher studying families experiencing food insecurity will find much valuable information in SNAP data. But many needy families do not participate in the program, and a separate source of information is needed to study nonparticipants.[19]

The information collected in administrative records is determined by the entity administering the program. In probability surveys, data collection is tailored to information needed to calculate statistics. Administrative records information might be fit for its intended purpose (administering a program) but might not contain the "right" variables to meet the statistical needs.

Another challenge is that the agency collecting the administrative records can change the information that is collected. For example, the 2017 Tax Cuts and Jobs Act removed (through 2026) the U.S. federal income tax deduction for civilians' moving expenses, and subsequent individual tax returns collect moving-expense information only for members of the armed forces.

Moreover, as with surveys, information in administrative records may be inaccurate, inconsistent across sources, or subject to changes in definitions or processing procedures over time. A tax filer may misreport income or property values; a health care claim may contain incorrect treatment codes; a law enforcement agency may fail to record a crime or may misclassify the type of crime. Some administrative records information is verified from other sources, but some information, particularly information that is not needed for program administration, may be collected without verification. Judson and Popoff (2005, p. 20) commented: "There is considerable evidence that, if a particular field is 'important' to the

[18]U.S. Bureau of Justice Statistics (2021b); FBI Crime Data Explorer, https://cde.ucr.cjis.gov

[19]Evidence that many needy families do not participate comes from probability surveys (see Chapters 3 and 5).

administrative agency (either because it ties directly to the agency's funding or to its provision of services), then that field will likely be recorded with some care and some quality control. But if the field is an 'add on' or is otherwise superfluous to the agency's mission, it should be used with caution."

Some administrative datasets have continual data collection. That does not mean, however, that the dataset is in a form that is ready for statistical uses. It can take a lot of time to process and transfer data for statistical calculations. Some datasets may have duplicated records, with poor-quality information for identifying such duplicates. Sometimes administrative data records are updated as new information comes in, making it challenging to identify data from a fixed time point or to replicate statistics or analyses. Benzeval et al. (2020, p. 18) emphasized that "administrative data, like survey data, have data generating processes, errors can enter through those processes, and indeed, steps in those processes often include human actions."

The federal government collects some administrative records datasets directly, such as federal income tax records. Other administrative records are compiled from information submitted by state and local governments—examples include NIBRS, SNAP, and the National Vital Statistics System, which collects information from the states on births and deaths. When records are collected from states, care must be taken to ensure that states collect the same information in the same way (see Section 4.3). In addition, some state-level administrative datasets (such as SNAP data) are available only for certain states, not for the entire country.

Administrative datasets may be accompanied by documentation that is relevant to their programmatic use. For use in producing official statistics, however, documentation is needed that describes how each data element is measured, which population units are included (and excluded) from the data, and the limitations of the dataset. This documentation could be produced by the agency collecting the data or by the data user in consultation with the agency.

Records Collected by Private-Sector Organizations

Many private-sector firms and organizations generate large amounts of data that may be relevant to major survey programs (NASEM, 2017c, 2023). Some of these data sources are similar in structure to administrative records: examples include credit card transactions, electronic health records, grocery store scanner data, point-of-sale retail sales data, and stock market transactions.

Studds (2021) and the U.S. Census Bureau (2021d) described the use of private-sector data for the U.S. Census Bureau's Monthly State Retail Sales estimates, launched in September 2020.[20] Data users had been requesting more timely state-level data on retail sales; to meet that need, the U.S. Census Bureau combined data from three main sources: the Monthly Retail Trade Survey (a probability sample of retail businesses with paid employees), administrative data on gross payroll for retailers, and retailer point-of-sale data purchased from a private-sector firm. Studds (2021, slide 8) outlined challenges involved in working with private-sector data: although private-sector data products can be useful, they often do "not align to the federal statistical system and standards. A heavy lift is required to fully understand methodology, quality, and fitness for use."

[20] https://www.census.gov/retail/state_retail_sales.html

The U.S. Department of Agriculture uses grocery store scanner data and other private-sector data sources to study food access, health, and security.[21] The U.S. Bureau of Labor Statistics has been exploring the use of scanner and other private-sector data to supplement or replace information for the Consumer Price Index that is currently collected through surveys (Konny, Williams, and Friedman, 2022; NASEM, 2022d). The first National Academies report in this series (NASEM, 2023) discussed additional examples of the use of private-sector data by federal statistical agencies.

As with administrative records, private-sector data are usually collected for a purpose other than producing official statistics, and they may have substantial undercoverage of the population of interest. Electronic health records contain information on medical conditions for persons who engage with the health care system, but persons without insurance or ready access to medical care will be underrepresented in the dataset. Similarly, persons who have no credit cards or loans might not appear in credit bureau data or in credit card companies' transaction data. In addition, data may be available from only some insurers or some credit card providers, thus missing health claims or transactions processed by other companies.

If data are collected from multiple companies (for example, claims data from several health insurance companies), each company may measure different items, use different categories of measurement, and have its own data format. A company that initially participates in a data-sharing agreement may decide to end that participation at a later date or may change the data items that it collects. As with administrative records, data processing and standardization can be challenging.

Private-sector organizations may be reluctant to share data about individual persons or businesses out of concerns for privacy or competitive advantage. However, data combination does not necessarily require access to individual records (see Section 2.2). Summary statistics supplied by third-party organizations can also be combined with data from probability surveys and censuses.

Satellite, Sensor, and Location Data

Satellite and earth observation data are a rich source of information for official statistics, particularly for agricultural and environmental statistics (Global Strategy to improve Agricultural and Rural Statistics, 2017; United Nations Economic and Social Council, 2019). The National Agricultural Statistics Service's *Cropland Data Layer*, a map of land cover in the continental United States, is created primarily from satellite imagery. These data provide preliminary acreage estimates for major agricultural commodities and are used to improve agricultural sampling frames (Boryan and Yang, 2021; see Chapter 8).

Sensor and location data can provide additional information without increasing respondent burden. Traffic sensors collect information about the amount of traffic on selected roads. Data from weather and pollution sensors can be linked with other spatially identified data sources. Cell phone location data were used to study the association between county-level travel patterns and infection rates during the early parts of the COVID-19 pandemic (e.g., see Sehra et al., 2020).

Wearable fitness trackers generate data about wearers' heart-rate readings, sleep time, menstrual cycles, step counts, locations, and more. *The Economist* (2022) reported on the

[21] https://www.ers.usda.gov/topics/food-markets-prices/food-prices-expenditures-and-establishments/using-proprietary-data/

increasing use of data from wearable fitness trackers for disease surveillance and medical research. Traditional disease-surveillance methods rely on reports of symptoms from doctors and hospitals (e.g., see the influenza-surveillance system described by CDC, 2021b); by the time people seek medical care and the data are reported, an epidemic may already have progressed. Fitness trackers, by contrast, collect nearly continuous data from large numbers of people and may provide quicker warning of an outbreak.[22] Fitness trackers also collect data from people in their natural surroundings—for example, they measure sleep at home instead of in a researcher's sleep laboratory.

But *The Economist* (2022, p. 11) also warned that "disease-surveillance algorithms based on wearable devices might systematically miss what is happening with some types of people … [f]or example, algorithms might unwittingly be optimised for spotting outbreaks in wealthy areas where people are more likely to have been using high-end wearables for longer." Thus, these devices may be less likely to detect outbreaks in lower-income areas. Fitness trackers also produce data only when people wear them—there will be no sleep data from persons who charge their devices at night, for example.

As with other data sources, population coverage is a primary concern with sensor data. Cell phone location data exclude persons without phones, those who leave their phones at home, or those who turn off location tracking. Data from wearable fitness trackers are limited to persons who have them and remember to wear them—a group that may be more affluent or health conscious than the population at large.[23] Environmental data might be available for only a limited number of locations or pollutants, and the placement of measurement devices might be driven by nonstatistical considerations. Traffic sensors might be placed only on major arteries. Placement of gunshot sensors might be driven by past criminal activity and therefore unable to track new developments, biasing a local jurisdiction's view of "dangerous areas."

Measurement error is also of concern. While a wearable fitness tracker may accurately measure steps, it may be less accurate for other types of activity. As with other private-sector data, companies may use different measurement protocols or algorithms for calculating outputs such as heart rate. Algorithms may be proprietary or may change in response to new technologies and market forces.

Datasets from sensors are typically enormous and usually require substantial cleaning, editing, and transformation to be useful for analyses (Leroux et al., 2019). Choices about how to process raw data can affect the statistics produced. If companies make only processed data available, other users may be unable to validate, verify, or replicate the resulting statistics.

[22]Aggregating data from a convenience sample of about 50,000 Fitbit wearers in five states, Radin et al. (2020) reported that weekly measures of the proportions of persons with elevated resting heart rates and increased sleep durations (which can be signs of infection) correlated with rates of influenza-like illnesses from CDC's influenza surveillance system—but fitness tracker data were available about 10 days earlier and might provide faster warning of an influenza outbreak.

[23]One way to obtain fitness-device data that can be generalized to the population is to ask persons in a probability survey to wear the devices. If everyone agrees, then one has a probability sample of persons wearing fitness trackers. In some years, the National Health and Nutrition Examination Survey has asked a subsample of participants to wear accelerometers, to provide an objective measure of physical activity. Leroux et al. (2019) outlined challenges for working with the accelerometer data, but also noted the potential of these data for improving prediction of health outcomes.

Nonprobability or Convenience Samples

Technically speaking, any data collection that does not meet the criterion of being a probability sample can be classified as a nonprobability sample. For a convenience sample (a type of nonprobability sample), the primary consideration for inclusion of units in the sample is how easily those units can be recruited or located. Examples of convenience samples include crowdsourced data (discussed in the next section), surveys in which participants are recruited through a website advertisement, and samples that consist of the investigator's friends and neighbors.

Most low-quality convenience samples are unsuitable for the production of statistics that describe population characteristics because the respondents are not representative of the population of interest. In many cases, as in a "click on the link to participate" survey, the population of potential survey-takers is unknown. Kennedy et al. (2021, p. 1050) outlined reasons data from opt-in online surveys may be inaccurate, including "[r]espondents completing the same survey multiple times from different IP addresses, overseas workers posing as Americans, and algorithms designed to complete surveys." Respondents may have a different race, gender, or age than they claim, or may give "bogus" responses to questions (Kennedy, 2022).

Kohler, Kreuter, and Stuart (2019) laid out the separate assumptions needed to use a convenience sample to (1) estimate population characteristics (e.g., unemployment rate) and (2) estimate relationships among variables (e.g., comparing age-adjusted heart attack rates for persons with differing activity levels).[24] Researchers who estimate population characteristics from a convenience sample typically use statistical methods akin to nonresponse-adjustment methods for probability surveys, and they make the strong assumption that the methods remove any bias resulting from the volunteer nature of the sample (see Wu, 2022, and the literature referenced therein). But a convenience sample may still be useful for exploring relationships among variables even if it produces biased estimates of population characteristics. For example, an unrepresentative convenience sample might show an association between untreated hypertension and coronary artery disease, even though estimates of the population percentages having hypertension and coronary artery disease (and, perhaps, the degree of the association) are biased. Kohler, Kreuter, and Stuart (2019, p. 151) argued that the convenience nature of the sample is irrelevant "as long as the size of the causal effect is assumed to be some kind of a law of nature and is therefore the same for all research units, in all places, at all times."

Some convenience samples may provide valuable information because of their sheer size and their inclusion of population groups that have small sample sizes in other datasets. For example, the National Institutes of Health's *All of Us* research program is inviting more than one million people across the United States to "help build one of the most diverse health databases in history."[25] Participants sign up on the program website and are then asked to participate in surveys and to share their electronic health records. They may also give blood, saliva, and urine

[24]These considerations also apply to low-response-rate probability surveys, which can have bias remaining after nonresponse adjustment methods. Mercer et al. (2017) and Meng (2018) discussed frameworks for evaluating bias in nonprobability surveys.

[25]https://allofus.nih.gov/. Publications, statistics, and information on data collected are available from https://www.researchallofus.org/. As of June 16, 2022, the program had 501,000 participants, with 302,000 associated electronic health records and 368,000 biosamples.

samples for laboratory and DNA tests, and may share data from wearable fitness-tracking devices.

The *All of Us* program provides a large database of detailed longitudinal medical information, with emphasis on obtaining data from persons who have been historically underrepresented in biomedical research (Mapes et al., 2020). But it is not a probability sample—instead, participants volunteer to submit data, and participants may have health characteristics different from those of persons with similar demographic profiles who do not participate.

Data from Social Media, Webscraping, and Crowdsourcing

Vast amounts of data are available on websites and from social media companies. Crowdsourcing involves soliciting data from a large group of persons (usually online). Webscraping software applications are programmed to search for web pages that are relevant for a topic of interest and to extract information from those pages. These data sources are usually not under the control of a government agency or single private company, and they may be of poor quality. Self-reports of activities or views from social media can be unreliable—persons or organizations may portray themselves as they would like to be seen, not as they actually are—and the population of participants is not well defined. In many social media datasets, a small number of users account for the majority of postings; most account holders have little activity. See Couper (2013) for discussions of these and other issues with social media data.

These data sources, like many other convenience samples, can also be manipulated by an outside actor, which would be of particular concern for data used to produce official statistics. A computer or automated system can generate a massive number of social media posts to reflect almost any viewpoint desired. As one example, Himelein-Wachowiak et al. (2021) discussed the role of bots (short for software robots) in spreading misinformation about COVID-19.

Crowdsourced or webscraped data may be useful, however, when joined with other data sources. One potential use involves identifying population members that are missed by other sources. For example, *The Guardian* investigated how many persons in the United States were killed by law enforcement officers in the line of duty during 2015 and 2016. They performed web searches for news reports, gathered data from organizations that track law enforcement-related deaths, and asked readers to send information about fatalities they knew of. Recognizing that fatalities identified through crowdsourcing or web searches might not be caused by law enforcement (or might not even be real fatalities), *The Guardian* verified each fatality with law enforcement agencies and medical examiners' offices. *The Guardian's* database contained more than 1,000 fatalities resulting from encounters with law enforcement personnel for each of 2015 and 2016—about twice as many as reported in official statistics from the FBI (Swaine and McCarthy, 2016, 2017).

It may also be possible to make use of social media data together with a probability survey. For example, Hughes et al. (2021) asked participants in a probability survey to provide their Twitter handles, thus giving a representative sample of Twitter users (subject to nonresponse and nonconsent).

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.

2.2 METHODS FOR COMBINING DATA

The data sources described in Section 2.1 have distinct strengths and weaknesses. Judicious use of statistical methods for combining data sources can exploit the strengths and overcome the weaknesses. This section outlines some of the methods that can be used to combine information from data sources, with special reference to combining information from a probability sample with that from another source. For detailed descriptions of a variety of possible approaches, see Citro (2014); the National Academies (2017a); Lohr and Raghunathan (2017); and Elliott, Raghunathan, and Schenker (2018).

Linking Records

Record linkage, sometimes called entity resolution, involves identifying records from two or more data sources that have information about the same entity. If the linkage procedure is accurate, then the information from the sources can be merged, allowing researchers to study relationships among variables measured in the individual sources. One of the earliest large-scale linkage projects for income statistics linked survey records from the March 1973 CPS to administrative records from the SSA and tax records from the Internal Revenue Service. The linkage augmented the CPS information for each survey respondent with the respondent's earnings and benefits information (where available) from the SSA and information on the respondent's income subject to taxation from the tax records (Kilss and Scheuren, 1978).

Many methods exist for linking records and assessing the quality of the linkages. Historical overviews and descriptions of record-linkage methods are available in Harron, Goldstein, and Dibben (2016); Christen (2019); Asher et al. (2020); and Binette and Steorts (2022). Box 2-1 summarizes two popular methods: deterministic and probabilistic record linkage.

[BOX 2-1 about here]

Record-linkage techniques can be used to:

- Add variables measured in other data sources to the variables measured in a primary data source. This was the goal of the 1973 CPS record-linkage project mentioned above, and linkage allows researchers to study relationships among variables measured on the same individuals in separate data sources.

  Records of individuals in surveys can also be linked to characteristics of the areas where survey respondents live or groups to which they belong. For example, measures of environmental pollution can be linked to housing data, or a variable from the decennial census giving the percentage of Black Americans in the survey respondent's census tract might be added to the respondent's data record. See Chen (2015) for a discussion of issues to consider when linking "neighborhood" or "ecologic" variables.

- Merge two or more datasets that each contain part of the population, thereby augmenting the number of records in the dataset. Feldman et al. (2017) arranged for records from *The Guardian's* database of deaths caused by law enforcement actions (see Section 2.1) to be linked with records in the National Death Index (NDI).[26] By linking records, they could identify deaths that were present in both datasets, and thus remove duplicated records from their estimate of the total number of deaths caused by law enforcement actions.[27] They could also investigate characteristics of the records that were in *The Guardian's* database but were not attributed to law enforcement actions in the NDI. They found that law enforcement-related deaths were more likely to be misclassified in the NDI if the death was not due to a gunshot wound or if it occurred in a county with lower median income, but there were no significant differences in misclassification by race or ethnicity of the decedent.

- Create longitudinal datasets by linking records belonging to the same person over time, for example, merging high school records with information on college completion.

- Check the accuracy of information in a data source by comparing it to other sources.

Two types of errors can occur when linking records (Doidge and Harron, 2019). First, an algorithm might declare a link between two records that in fact belong to different entities, which is called a *false link.* For example, a procedure that matches by name and location might link a record for Philadelphia resident Michael Smith from data source A with a record for Philadelphia resident Michael Smith from data source B—but, in reality, these records belong to two distinct persons who share the name Michael Smith. The second type of error, a *missed link*, occurs when a match exists in the data sources but is not found by the procedure—Michael Smith in source A is actually the same person as J. M. Smith in source B, but the procedure does not link the records.

Linkage errors may affect some population subgroups more than others and may lead to flawed conclusions from linked datasets. This issue is central to considerations of data equity (see Sections 3.6 and 6.4). Reviewing studies that compared characteristics of linked versus unlinked records, Bohensky et al. (2010) identified several characteristics associated with lower linkage rates, including lower socioeconomic status, lower educational attainment, and

[26]The NDI is a comprehensive set of administrative records on deaths occurring in the United States since 1979. As of 2022, the NDI contains more than 100 million records, with information compiled by state registration areas from death certificates about day, location, and cause of death; age, sex, race, ethnicity, and marital status of the decedent; and additional information relating to the circumstances of the death. Its purpose is to "[p]rovide the public health and medical research community with an opportunity to obtain mortality follow-up information on their study participants" (NCHS, 2022a).

[27]Feldman et al. (2017) identified 599 deaths reported in *The Guardian's* database alone, 36 reported in the NDI alone, and 487 reported in both lists. By assuming that being listed in the NDI as a death caused by "legal intervention" was independent of presence in *The Guardian's* database, they estimated that 44 deaths (with 95 percent confidence interval [31, 62]) were in neither source, and that 1,166 law enforcement-related deaths occurred in the United States in 2015. Banks et al. (2019) described a pilot program conducted by the U.S. Bureau of Justice Statistics to improve its census of arrest-related deaths by including deaths found from web searches of news outlets, law enforcement agency documents, and other publicly available sources.

membership in a minority race or ethnicity group. Lower linkage rates for race or ethnicity groups may have systemic roots:

> Accurate data linkage relies on accurately recorded identifying information and well designed linkage algorithms. However, ethnic minorities are more likely to have missing or incorrect information in their health records, which might reflect structural biases in health systems (eg, ethnic minorities are more likely to be treated at health facilities with poorer overall data quality). Data capture systems are also typically designed around Western name standards (ie, a first, middle, and last name) and do not account for cultural differences in name structures (eg, Hispanic groups can have multiple first or middle names, and often two surnames, and Asian names can follow different ordering norms) (McGrath-Lone et al., 2021, p. e339).

CONCLUSION 2-3: Linking survey data with administrative records requires substantial expertise and investment. Decisions need to be made among reasonable alternative methods, and then periodically re-examined as data sources change or new linkage methods are developed. Documentation that assesses the quality of the linkages allows data users to evaluate the possible impact of linkage errors on analyses and to account for uncertainties in the linkage process.

Combining Statistics Calculated from Independent Data Sources

Many data sources will not contain sufficient information to allow individual records to be linked, but statistics from the sources can be combined to provide a more comprehensive picture than can be derived from any single data source. For example, Oronce et al. (2020) investigated the relationship between states' numbers of deaths from COVID-19, obtained from Johns Hopkins University's COVID-19 dashboard, and state-level measures of income inequality obtained from the ACS. This can be thought of as linking datasets at the state level. There is no linkage error for this example, but other statistical issues can arise from measurement errors, varying precision of state-level estimates, or misclassification.

Combining statistics can improve population coverage when data sources include information on different subsets of the population. For example, income tax records have detailed information about filing units (often, but not necessarily, households) that filed tax returns, but they exclude non-filers. Income statistics from a probability survey with information on non-tax-filers could be combined with income statistics from tax records to generate statistics for the entire population.

Multiple-frame probability surveys use this approach to improve population coverage and reduce data-collection costs (Lohr, 2021). Independent probability samples are taken from two or more sampling frames that together are assumed to include the entire population. For example, the quarterly agricultural surveys conducted by the U.S. National Agricultural Statistics Service (see Chapter 8 and NASEM, 2017b, p. 45) select a sample of farm operators from a *list frame*—a list of known U.S. farm and ranch operations containing information about each operation's size and the commodities it has produced in the past. Using the list frame allows the survey to be conducted in a cost-effective manner, but results in undercoverage because some farms are not on the list. Some of the quarterly agricultural surveys supplement the sample from the list frame

with another sample from an *area frame*—a list of all parcels of land areas (Davies, 2009). A sample of land segments is drawn from the area frame, and farm operators with land in those segments are included in the area sample. The area sample thus includes farm operators who are not in the list frame and gives more complete coverage.

A multiple-frame approach does not require the ability to link individual records from data source A to specific records in source B, but it does require knowledge of whether a record in source A is also in source B, and vice versa. If supplementing statistics from tax records with information from a probability survey, one must know which persons in the survey are represented in the tax return records, and which tax records are from persons outside of the survey population. In addition, multiple-frame estimates may be biased if the sources measure the concepts of interest differently. In a multiple-frame probability survey, the same questionnaire is used for the respondents to each survey, to promote consistency of measurements. When statistics are combined from other data sources, however, measurement differences may affect the results—for example, a tax form may exclude some types of income that are included in a survey question. Even if the income definitions are the same, persons may report different amounts on a survey than they would report on a tax return.[28] In these situations, it is important to define the population characteristics being estimated and how measurements from the various data sources relate to those characteristics.

Using Statistical Models to Combine Information

Record linkage and multiple-frame survey approaches use data that have been collected directly from survey participants or in administrative or private-sector data. If a survey participant is accurately linked with an administrative record, the merged data can be analyzed as if all variables came from a single source. In a dual-frame agricultural survey, one sample is selected from a list frame and a second sample is selected from an area frame, but all the information about planted acreage comes from survey respondents.

In other situations, a data source may contain information related to a variable of interest, but not the variable itself—for example, health care claims data do not report whether a person has ever had a heart attack but will have information about some of the medical treatments received in the past year. Alternatively, the only available data source may have information on a variable of interest, but only for part of the population. In these situations, statistical models may be used to estimate the quantities of interest from the available information, but the accuracy of such estimates depends on how well the model describes the population. This section briefly describes two types of models that will be referenced later in this report. Many other types of statistical models have been developed and this is an area of ongoing research. Elliott and Valliant (2017); Lohr and Raghunathan (2017); Kohler, Kreuter, and Stuart (2019); Beaumont (2020); Rao (2021); and Wu (2022) reviewed methods that use statistical models to combine information from multiple data sources.

Small Area Estimation. Consider a situation in which a probability sample measures a variable of interest such as household income, acres planted to corn, victimization by crime, or presence of a certain disease. The probability sample gives accurate estimates for the nation as a whole, but for some subpopulations (for example, states, counties, school districts, or

[28]Measurement errors can affect multiple-frame probability surveys, too. If list frame respondents provide data over the internet, and area frame respondents are interviewed in person, the differing modes of data collection may affect responses.

demographic groups) the survey sample size is too small to give an accurate estimate using the survey data alone—because of their small sample sizes, these subpopulations are called *small areas* or *small domains.*

Small area estimation methods, described by Rao and Molina (2015), combine information from the probability sample with information from other data sources (such as administrative records and private-sector data) to estimate characteristics of interest for small areas. These methods supplement the small amount of data available in some of the areas with predictions based on model assumptions about relationships between the survey data and data from other sources. The application of small area estimation methods requires careful model validation because the accuracy of the statistics produced depends on the validity of the model assumptions, particularly for the smallest areas or subpopulations.

Box 2-2 describes the U.S. Census Bureau's Small Area Income and Poverty Estimates program, which combines information from the ACS with data from administrative records to estimate the percentage of all persons, and the percentage of children, living in poverty for counties and school districts. Chapter 7 discusses small area models for estimating crime, and Chapter 8 discusses use of administrative and satellite data to calculate crop estimates.
[BOX 2-2 about here]

Imputation. Imputation methods predict values of missing data from other information. For example, some persons left questions blank or gave inconsistent information on the 2020 Census form. The U.S. Census Bureau filled in values for those missing items using other information available for that person or household if possible—for example, the respondent's first name might be used to fill in a missing value for sex, and information from administrative records or tax assessor records might be used to fill in a missing value for the question on whether the home is owned or rented. If the missing information could not be determined from available information, then the U.S. Census Bureau assigned values based on information from similar nearby households (Ramirez and Borman, 2021).

Haziza (2009), van Buuren (2018), and Chen and Haziza (2019) described some of the statistical methods that can be used to impute missing data. These methods can be used to impute individual items (such as missing values for race or sex in a survey) or can be used to impute entire missing records (as for households that refuse to participate in the survey). The accuracy of imputed values depends on how well the statistical model describes the relationships between the observed data and the missing data.

Imputation can also combine information from multiple data sources. For example, Raghunathan et al. (2021) used imputation to estimate the prevalence of each of 107 health conditions in the population of Medicare recipients. The Medicare Current Beneficiary Survey, conducted by the Centers for Medicare and Medicaid Services, contains a nationally representative sample of Medicare recipients (including persons residing in institutions, who are often excluded from other surveys). Although the survey collects self-reported information on only a few health conditions, it provides information on all of the health conditions of interest through linkage with Medicare claims data. But the claims data contain information only on persons who sought care during the billing year (and thus lack information on health conditions that occurred in the distant past, such a heart attack 6 years ago); in addition, care for specific conditions may be grouped together and not reported separately, claims codes can be incorrect, and claims data lack information for persons enrolled in a Medicare Advantage program. Thus, prevalence estimates calculated solely based on claims data would likely be lower than the actual prevalence.

Raghunathan et al. (2021) derived a statistical model using data from the National Health and Nutrition Examination Survey (which asks for self-reports and conducts medical examinations for a wide range of conditions) to correct for the underreporting of chronic health conditions in Medicare claims data. They used this model to impute indicator variables for each of the 107 conditions for Medicare Current Beneficiary Survey respondents.

CONCLUSION 2-4: Statistical methods such as small area estimation, imputation, and combining statistics for subpopulations can integrate information from multiple data sources without requiring individual records to be linked.

2.3 OPPORTUNITIES AND CHALLENGES FOR COMBINING DATA FROM MULTIPLE SOURCES

As will be demonstrated in Chapters 4–8, multiple data sources can improve the timeliness, granularity, and usefulness of data for estimating statistics currently calculated from probability surveys. At the same time, using alternative data sources also present challenges, as documented by the National Academies (2017a, 2017c) and Bee and Rothbaum (2019). These challenges, and ways of addressing them, are discussed in the remainder of this report. Challenges include:

- *Linkage errors*. Record linkage, the most commonly used method for combining data, can greatly augment information from a single data source by incorporating variables from other sources. But linkage may have errors, and uncertainty about linkages propagates to statistics calculated from linked data. When linking with the NDI, for example, are unmatched records from persons who are still alive or are they missed links?
- *Measurement errors*. A survey respondent may provide incorrect information, such as an incorrect value for earnings or utility payments. Administrative records and private-sector data may have measurement errors, too. For example, voter registration files may have incorrect demographic information and medical claims data may miscode diagnoses or medical procedures.
- *Missing data*. Administrative records may be available only for some locations, some population subgroups, or some years. Information such as sociodemographic characteristics may be missing entirely if not needed for the purposes of administering the program.
- *Concept alignment*. Data sources may define concepts of interest differently. For example, the CPS Annual Social and Economic Supplement asks about income received during the previous calendar year from sources including earnings, unemployment benefits, Social Security benefits, veterans' payments, and alimony (see Chapter 5).[29] Administrative data may capture only some of these sources of income, may define them differently, or may include additional sources.
- *Geographic alignment*. Satellite and sensor data may be collected for geographic units that are difficult to align with survey data or administrative records. Geographic

[29] https://www.census.gov/programs-surveys/cps/technical-documentation/subject-definitions.html#incomemeasurement

information may be inaccurate or unavailable in social media data or other data sources.

- *Entity alignment.* Data sources may measure different units. For example, one data source may measure household income, while another may measure income for individual persons. Or, one data source may look at sales by store, and another at sales by commodity. Sources that measure different units or entities can present challenges for linking records or combining statistics.
- *Stability of data sources*. The types of information collected in any data source can change, or data sources may disappear. An outside actor might even be able to manipulate some types of data if it became known these were being used to produce official statistics.
- *Population coverage*. The population intended to be represented by a probability survey does not necessarily match the population represented by a set of administrative records. In some cases, it may be possible to identify which units in the survey are missing from the administrative records population, but in other situations the overlap may be unknown.
- *Underrepresented population subgroups*. Some population subgroups may be underrepresented in all data sources (see Chapter 3).
- *Technical expertise*. Many of the new data sources and methods for data combination require skill sets beyond those needed for conducting traditional probability surveys. These include expertise in record linkage, statistical modeling for combining data, machine learning, data quality assessment, computer science, information systems management, and remote sensing technology.

Additional challenges with using multiple data sources include obtaining access to the data (which can include paying for data), establishing an information technology infrastructure for processing and storing data, and protecting the privacy and confidentiality of people and businesses whose information is in the datasets. These issues will be addressed in later reports of this series.

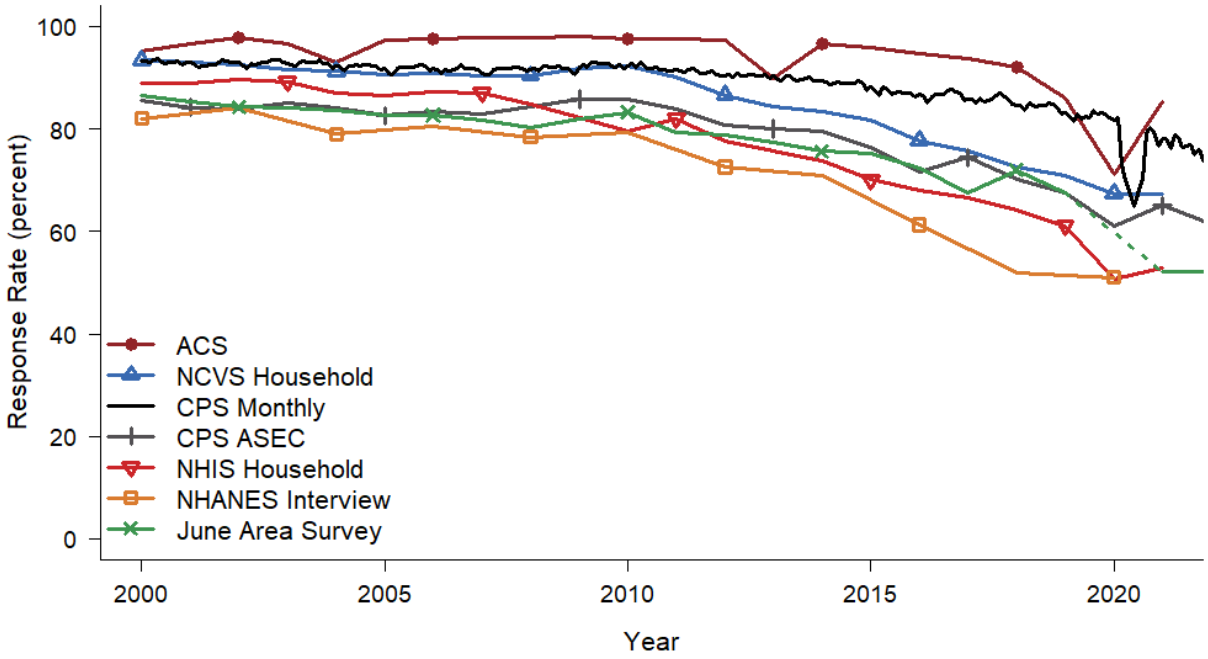


FIGURE 2-1 Response rates for selected surveys, 2000–2022.
NOTES: ACS = American Community Survey (Chapters 2, 3, 5)
NCVS = National Crime Victimization Survey (Chapters 2, 7)
CPS Monthly = Current Population Survey Monthly (Chapters 2, 5)
CPS ASEC = Current Population Survey Annual Social and Economic Supplement (Chapter 5)
NHIS = National Health Interview Survey (Chapter 6)
NHANES = National Health and Nutrition Examination Survey (Chapters 1, 6)—NHANES data are released in two-year cycles
JAS = June Area Survey (Chapter 8)
SOURCE: Panel generated.

For NCVS, NHIS, and NHANES, the graph displays household-level response rates. Additional nonresponse occurs when individual persons within households fail to provide data. NHANES response rates are for the interview component; response rates for the examination component are lower.[30] The dips in response rates in 2020 reflect the disruptions of in-person data collection from the COVID-19 pandemic. The JAS was not conducted in 2020.

[30]Sources for response rates (including the response rate definitions used) are https://www.census.gov/acs/www/methodology/sample-size-and-data-quality/response-rates/ (ACS); https://bjs.ojp.gov/sites/g/files/xyckuh236/files/media/document/samplesize.xlsx (NCVS); https://beta.bls.gov/dataViewer/view/timeseries/LNU09300000 (CPS monthly); Czajka and Beyler (2016); and individual-year documents on "Source and Accuracy of Estimates for *Income and Poverty in the United States*," for example, https://www2.census.gov/library/publications/2016/demo/p60-256sa.pdf (CPS ASEC); *Survey Description Documents* for individual years, available at https://www.cdc.gov/nchs/ (NHIS); https://wwwn.cdc.gov/nchs/nhanes/responserates.aspx (NHANES); https://www.nass.usda.gov/Publications/Methodology_and_Data_Quality/Land_Values/; Tran et al. (2010); and Linda Young, National Agricultural Statistics Service, personal communication (JAS).

BOX 2-1 Deterministic and Probabilistic Record Linkage

Record linkage methods are used to merge the information from entities in dataset A that are also in dataset B, or to determine that an entity in dataset A has no corresponding record in dataset B. The success of a linkage method depends on how well the available information in the two sources can identify individual persons, households, or businesses. Typical person-level identification variables include items such as Social Security Number (SSN), name, state and date of birth, state of residence, sex, race, and marital status.

In *deterministic record linkage*, a record pair is declared a match if the dataset A record and the dataset B record agree exactly on the specified identification variables (and there are no other pairings with the same identification variable values). For example, two records with the same SSN and first and last name might be declared a match. Not all records, however, have detailed and accurate identification information, and a deterministic system may miss valid matches.

Probabilistic record linkage methods compute a match score, based on the identification variables, for each possible pairing of a dataset A record with a dataset B record. Typically, "blocking" is used to reduce the number of potential candidate pairs that must be examined, for example by restricting comparisons to records that have the same state of residence or have similar last names. The score is the sum of the weights assigned to each identifying item used in the matching process. If the A record and the B record agree for the item, the weight is positive; if the records disagree outside of a prescribed tolerance, the weight is negative; if the item is missing in either record, the weight is zero. The weight for an item depends in part on how well that item identifies a person. For example, two records that agree on an uncommon first name such as "Quetzal" are more likely to belong to the same person than two records that agree on a common first name such as "Elizabeth," so the item weight for a match with "Quetzal" will be higher than the item weight for a match with "Elizabeth."

After the scoring, each record pair may be classified as a match, a nonmatch, or indeterminate. Pairs with scores exceeding a predetermined cutoff value are declared to be matches—the two records agree or nearly agree on many of the identification variables. Pairs with scores below another predetermined cutoff value are declared to be nonmatches—these disagree on enough of the identification variables to be thought to belong to different entities. If the highest score for a dataset A record is below the nonmatch cutoff, that record is considered to have no corresponding record in dataset B. Pairs with scores between the two cutoff values may undergo further review before a determination is made.

When a probabilistic linkage method is used, uncertainty about linkages can also be incorporated into standard errors of statistics. Reiter (2021) reviewed statistical methods that can account for uncertainties about linkage that occur when there are several plausible matches in dataset B for a record from dataset A, or when an administrative source used for linkage does not contain all individuals in the survey.
[END Box 2-1]

BOX 2-2 The Small Area Income and Poverty Estimates Program

The American Community Survey (ACS) produces annual estimates for geographic areas containing 65,000 or more persons. Even with an annual sample size of 3.5 million households, however, the survey by itself does not have sufficient information to produce reliable estimates of income and poverty for areas with populations less than 65,000, such as small counties or small school districts—in a geographic area containing 10,000 persons, the survey sample size is too small to allow accurate calculation of annual statistics about income and poverty. Some areas might even have no sample representation for a particular year.

The U.S. Census Bureau's Small Area Income and Poverty Estimates (SAIPE) program uses statistical models to compute estimates for small areas (areas in which the survey sample size is too small to calculate a reliable estimate using the survey data alone). The county-level estimation program computes estimates of the number of people in poverty as follows:[31]

1. Estimate the number of persons in poverty directly from ACS data for every county with survey data. The precision of each county's direct estimate depends on the survey sample size for that county—the estimate for a county with a large sample size, such as Los Angeles, will have a relatively small standard error; and the estimate for a county with a small sample size will have a relatively large standard error.

2. Collect data items that are related to county-level poverty from administrative records and previous censuses. For SAIPE county-level estimates, these data items include the estimated county population, the number of persons in the county receiving benefits from the Supplemental Nutrition Assistance Program (SNAP), and aggregated data calculated from federal tax information such as the median adjusted gross income of the state.

3. Develop a regression model predicting the logarithm of the direct estimate of the number of people in poverty (from Step 1) from the logarithms of the variables collected in Step 2.[32] Use the model to predict the number of people in poverty for every county in the United States. Note that the models use only summary statistics from the administrative data, not individual records.

4. Combine the direct (Step 1) and model-predicted (Step 3) estimates to obtain the small area estimate of the logarithm of the number of people in poverty for each county. The estimate depends more heavily on the direct estimate when the standard error of the direct estimate is low,

[31]Descriptions of the procedures used to produce SAIPE estimates, along with lists of input variables for the models, are given in Bell, Basel, and Maples (2016) and U.S. Census Bureau (2021a). The first SAIPE estimates, published in 1993, were developed to estimate the number of school-aged children in poverty in each school district for the purpose of allocating federal funds to school districts; see https://www.census.gov/programs-surveys/saipe/about/origins.html and National Research Council (2000) for the history of the SAIPE program. Before the launch of the ACS in 2005, the direct estimates in Step 1 were computed from the CPS Annual Social and Economic Supplement.

[32]Logarithms are used because the distribution of the number of persons in poverty for all counties is highly skewed. A small county may have only a handful of people in poverty, while a large county may have more than 100,000. The regression model is fit using only counties with nonzero estimated poverty from the ACS, and accounts for the varying magnitudes of standard errors associated with survey estimates.

and more heavily on the model-predicted estimate when the direct estimate has high standard error.

5. Transform each county's small area estimate from log scale to obtain an estimate of the number of persons in poverty, and ratio-adjust county-level estimates of the number of persons in poverty to ensure that they sum to an independently derived state estimate.

The regression model allows poverty to be estimated in counties whose ACS sample size is too small to produce a direct estimate. Standard errors for SAIPE estimates are calculated using the model and are typically smaller than standard errors of direct estimates.

[END Box 2-2]

TABLE 2-1 Characteristics of Data Sources

Characteristic	Probability Survey or Census Conducted by Statistical Agency	Administrative Records	Private-Sector Records	Sensor or Satellite Data	Convenience Samples, Social Media, or Crowdsourcing
Well-Defined Population	Yes. The agency defines the population of interest.	Yes. The population is usually defined by program requirements. For example, tax records include most persons and businesses that are required to file.	Depends on source.	Usually, since there is control over which data are measured	Usually not. Population members choose whether to be in the dataset. Sometimes posts are made by automated systems, not real persons.
Population Coverage	Usually high, but some parts of population may be underrepresented because of undercoverage or nonresponse (see Chapter 3).	High if population of interest equals the set of records, but administrative records often have undercoverage of the population of interest.	Depends on source and definition of population. Transaction records from a credit card company contain all records from that company's customers but exclude transactions with other credit cards or cash.	Depends on source. Satellite data may have high coverage; traffic sensors may cover only specific highways; cell phone location data are limited to cell phone users.	Usually poor because sample is self-selected.
Statistical Agency Control over Information Collected	High	Low; administrative records are collected for the purpose of the program, but cooperative	Low, but cooperative agreements may be possible.	Low unless data collection is contracted by agency.	Usually low

		agreements may be possible whereby additional variables are collected.			
Geographic or Subpopulation Detail	Limited for most probability surveys because of sample size; high for censuses.	High for large datasets if variables defining the subpopulations are measured.	Can be high for large datasets, but private-sector organizations might not be willing to share geographic locations or details about subpopulations.	High geographic detail	Usually low
Timeliness of Data Availability	Low for most surveys and censuses. Some surveys (such as CPS) provide monthly estimates; most other surveys provide annual or less frequent estimates.	Depends on how frequently data are processed and released. Data collection is often continuous, but data release may take time.	Depends on how frequently data are processed and released. Data collection is often continuous, but it may take time to release to statistical agency.	Depends on how frequently data are processed and released. Data collection is often continuous, and data may be released in "real time" or soon after collection.	Depends on how frequently data are processed and released. Data collection is often continuous.
Potential for Record Linkage	High if identifying information is available.	High if identifying information is available.	Depends on availability of linkage information.	High for geographical linkage.	Low, unless survey respondent provides social media account information.
Potential for Combining Statistics	High	High	High if population and subgroups are well defined.	High	Low, unless information can be verified from another source.

SOURCE: Panel generated. This table was inspired by Table 5.1 of Citro (2014), which evaluates additional dimensions.

3. Using Multiple Data Sources to Enhance Data Equity

Chapter 1 discusses the commonly adopted definitions of data quality as "fitness for use" and "fitness for purpose." But whose use, and what purposes? Guidelines leave these questions unanswered, to be determined for each particular data collection.

Every federal data collection is the result of conscious decisions—what parts of the population to include, how to locate and collect information about population members, what concepts to measure and how to measure them—and unintended consequences from undercoverage, nonresponse, measurement error, and events that disrupt data collection (such as government shutdowns or pandemics).

No single survey or other data collection can possibly gather all information that might be needed for all potential purposes. This report discusses the promise of using multiple data sources to augment information collected in federal surveys with data from government administrative records, private-sector data, and other data sources. But, as Giest and Samuels (2020, p. 560) argued, "the data used can have hidden data gaps that differ depending on how data was collected and analyzed as well as the kind of questions being asked." Hand (2020), Naudé and Vinuesa (2021), Akee (2022), Brown (2022), and Kreuter (2022) have described how gaps or omissions in data sources affect the ability to measure differences among subpopulations.

This chapter examines how multiple data sources might be used to identify or correct overt or hidden gaps or misrepresentations. Data-equity issues were discussed in all the workshop sessions; this chapter draws on material from the entire workshop and in particular on presentations in the session *Issues in Data Equity*. Section 3.1 defines equitable data and identifies aspects that are affected by the data-collection infrastructure. Sections 3.2–3.7 examine ways that using multiple data sources can promote data equity: by increasing the representation of population groups that, historically, have been underrepresented in the data record (Sections 3.2 and 3.3); by producing model-based estimates for small populations (Section 3.4); by providing information that can be used to improve measurement of concepts of interest (Section 3.5); and by enhancing the amount of information about individual records in a dataset (Sections 3.6 and 3.7). These sections also discuss possible harms that might arise from integrating data. Section 3.8 contains a discussion and outlines some steps for promoting data equity.

3.1 WHAT IS DATA EQUITY?

The 2021 Presidential Executive Order on *Advancing Racial Equity and Support for Underserved Communities Through the Federal Government* outlined the need for better data to measure equity:

> Many Federal datasets are not disaggregated by race, ethnicity, gender, disability, income, veteran status, or other key demographic variables. This lack of data has cascading effects and impedes efforts to measure and advance equity. A first step to promoting equity in Government action is to gather the data necessary to inform that effort. (Executive Order 13985, 2021, p. 7011).

The Equitable Data Working Group, formed pursuant to the Executive Order, defined equitable data as follows:

> 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 (Equitable Data Working Group, 2022, p. 3).

The Equitable Data Working Group noted that disaggregated data offer "more precise statistical indicators of population well-being, as well as insight into who can and cannot access government programs and whether benefits and services are reaching underserved and underrepresented communities" (Equitable Data Working Group, 2022, p. 2).

The set of communities considered to be underserved in data sources, or for which disaggregated statistics are desired, depends on the purpose of the analysis and may change over time. Examples of the need to consider subgroup variation include studies of location of services for persons experiencing homelessness, access to health care, and proximity to toxic waste sites. Identifying which characteristics to consider for evaluating data equity in representation requires attention to potential dimensions of inequality.

One way to identify characteristics on which data equity should be evaluated is to engage stakeholders in the data collection or analyses. Dimensions to be evaluated include stakeholders' "race, gender, gender reassignment, sexual orientation, religion or belief, age, disability, marriage and civil partnership status, pregnancy" that "could increase their vulnerability to abuse, adverse impact, or discrimination" (Leslie et al., 2022, p. 40). The relevance of each dimension depends on the specific research questions and the context in which research questions are posed. For instance, studying whether subgroup differences in exposure to poor air quality are related to health disparities requires information about the ages of people exposed (because of developmental differences in vulnerability), underlying conditions, and socioeconomic and racial/ethnic characteristics. A concern with access to reproductive health care may require information about the locations of reproductive-care clinics as well as personal information about those seeking care, including marital and cohabiting partner status, sex and gender, and age. In these examples, definitions of race, ethnicity, and marital and cohabiting partner status may vary over time as societal norms change.

Jagadish, Stoyanovich, and Howe (2021b, p. 1) noted that "the manner in which data systems are built and used can compound and exacerbate inequities we have in society. It can also introduce inequities where there previously were none. Avoiding these harms results in *data equity.*" Jagadish, Stoyanovich, and Howe (2021a, 2021b) considered four aspects of data equity:

1. *Representation equity* focuses on increasing the visibility of underrepresented groups in the data. A dataset that contains few members of an underserved community, or an unrepresentative sample from that community, will not be able to produce reliable statistics about that community.
2. *Feature equity* focuses on the availability of variables needed to identify population subgroups or measure characteristics of interest. Federal income tax Form 1040 does not ask about race, ethnicity, disabilities (other than blindness), or gender—that information is not relevant for processing tax returns—and thus cannot be used to

produce disaggregated statistics of income for those groups without additional information.[33]

3. *Access equity* focuses on equitable access to data and data products. The *Federal Data Strategy* lays out guidelines for making data produced by federal government agencies accessible to users: "Promote equitable and appropriate access to data in open, machine-readable form and through multiple mechanisms, including through both federal and nonfederal providers, to meet stakeholder needs while protecting privacy, confidentiality, and proprietary interests" (U.S. Office of Management and Budget, 2019a, p. 7). Accessibility includes presenting statistics and other data products in a form that is easy to understand and interpret.
4. *Outcome equity* focuses on the consequences for population groups affected by the data-collection and dissemination system. This includes *algorithmic equity* (Sikstrom et al., 2022), ensuring that automated decision-making tools used in health care, the criminal justice system, and other settings are fair (see Box 3-1). Outcome equity also includes effects of potential data disclosure, which may result in disproportionate harm to certain communities.

[BOX 3-1 about here]

Aspects of data equity are inherent in several of the data-quality dimensions displayed in Figure 1-1. A data product with representation equity will produce *accurate* and *reliable* estimates for subpopulations of interest. Producing accurate estimates for a subpopulation requires having a representative sample from the subpopulation with adequate sample size (*granularity*), and ensuring that the data measurements reflect user needs and consistently measure the concepts of interest across subpopulations (*relevance* and *coherence*). Feature equity requires the ability to identify subpopulation members in the data (*granularity*) and measure appropriate variables (*relevance*). A data product with access equity will be *accessible* and easily used by subpopulation members, and will be available in a *timely* fashion. Outcome equity is related to the quality dimensions of *scientific integrity, credibility, computer and physical security, and confidentiality*—ensuring that individuals and communities are not harmed through the collection and dissemination of their data.

The use of combined data sources has consequences for all four aspects of data equity. This chapter concentrates on representation equity and feature equity because combining data sources can improve coverage, augment sample sizes, and add variables. As will be discussed in future reports, combining information from multiple sources raises new concerns about protecting privacy, which may affect decisions about data access. Linking or otherwise combining datasets may also have consequences for outcome equity.

Levenstein (2022) asked: "How can using non-survey data—administrative data, commercial data, social media data—increase both representation (what we know about people and who they are) and representativeness (the ability of a dataset to reflect the distributions of our population)?" She argued that there is an increasing lack of representativeness from survey data, as surveys obtain high response rates only from people who trust government data collectors; therefore, administrative records or commercial data may have better representativeness of the parts of the population they collect data about because they include people who refuse to answer surveys. However, administrative records and other sources may

[33] https://www.irs.gov/pub/irs-pdf/f1040.pdf

also include only part of the population of interest. Multiple data sources may be able to present a more complete picture than any single data source.

Sections 3.2–3.7 examine some of the ways that combining data sources can improve representation or feature equity. However, it is important to note that integrated data are not necessarily "better" data for all purposes, and upcoming sections also discuss areas in which caution is needed.

3.2 INVESTIGATE OR IMPROVE COVERAGE OF A SURVEY

The coverage of a probability survey's sampling frame is the set of population members that could be selected for the sample. Units not in the frame have no chance of being included in the sample, so it is desirable for the sampling frame to have high coverage. External data sources are needed to investigate the coverage of a sampling frame. It has long been standard practice for federal agencies administering probability surveys to investigate potential undercoverage and nonresponse bias using other available data sources. U.S. Office of Management and Budget standards call for federal agencies to conduct a nonresponse bias analysis on surveys when the response rate is below 80 percent (OMB, 2006); one component of such an analysis involves comparing estimates calculated from survey respondents to known characteristics of the population obtained from an external source, such as administrative records or the decennial census.

Box 3-2 describes the use of an independent data source (the post-enumeration survey) to study characteristics of undercoverage in the 2020 Census. The representativeness of administrative records and other data sources can also be investigated through comparing summary statistics with estimates from high-quality data sources (when they exist).
[BOX 3-2 about here]

Information from multiple sources can be used not only to diagnose undercoverage but also to improve coverage. Datasets that contain units missing from the main sampling frame can be used to improve or supplement that frame. One data source may contain population members that are not found in a different source, so that the combined data cover more of the population than each data source by itself. The U.S. Census Bureau's *Frames* project (see Section 4.2) links records from several internal sampling frames and other sources, and the linkage can be used to identify areas of undercoverage and improve the frames.

Nontraditional data sources can also be used to improve coverage of sampling frames. For example, many agricultural surveys are conducted by taking a sample from a list of known agricultural operations, but these lists fail to include some smaller or transient operations. Coverage of list frames may be particularly low for sectors such as urban agriculture or local food farms (farms that distribute their products locally). Traditionally, the National Agricultural Statistics Service (NASS) has taken samples from an area frame to obtain full coverage, but using an area frame can be cost prohibitive in urban areas, where small agricultural operations may be scattered throughout the city (see Chapters 2, 8). Hyman, Sartore, and Young (2022) used webscraping to compile a list of operations that might be local food farms, and they linked the operations in the webscraped list with those in the NASS list frame. This gave three groups of potential local food farms: those in the webscraped list alone, those in the NASS frame alone, and those in both lists. After taking samples from both lists (in which respondents were asked questions to establish their status as farms), Hyman, Sartore, and Young (2022) were able to assess the NASS list frame's coverage of local food farms. They found that only about eight

percent of the local food farms in the webscraped sample were missing from the NASS list frame, but that those were more likely to be small operations.[34]

Hyman, Sartore, and Young (2022) linked the records to assess the coverage of the two lists and increase the coverage of the combined samples, but multiple-frame surveys can also be used to improve coverage without explicitly linking data, as long as there is some way to identify entities that could appear in more than one of the samples (for example, telephone surveys that select landline and cell phone samples ask respondents about their landline and cell phone usage, thereby identifying the respondents who are in both frames).

Multiple-frame surveys can improve the coverage of population groups that are difficult to sample in traditional probability surveys. For example, Iachan and Dennis (1993) combined samples from shelters, soup kitchens, encampments, and street locations to increase coverage of persons experiencing homelessness (see also Peressini, McDonald, and Hulchanski, 2010). Bird and King (2018) discussed uses of multiple-frame surveys for obtaining larger and more diverse samples of people who inject heroin and of victims of human trafficking.

Beals et al. (2021) conducted a survey of clients who received services from social or health agencies in the state of Washington. They drew separate random samples of clients from each of 10 program areas, ensuring a minimum number of completed interviews from each program area. Clients were asked about the program they were sampled from, and about services from any other programs they used in the previous year, thus including broader representation of service experiences than a simple random sample. To account for clients who might be served by multiple programs, each client's record in this multiple-frame sample was weighted according to the size of the population receiving that particular combination of program services.

Most of the examples in this section concern improving sampling frames used for selecting population members for probability surveys. However, linking survey data with data from an administrative data source can also be used to identify characteristics of people who might be eligible for a government program such as the Supplemental Nutrition Assistance Program (SNAP) but do not participate, and can provide insight into the fitness of the dataset for producing official statistics or for conducting social or economic research. For example, Newman and Scherpf (2013) linked records from SNAP data in Texas to records from the American Community Survey (ACS). ACS income questions were used to estimate eligibility for the program and the SNAP records, which have full coverage of the population participating in the program, were used to determine participation. The authors estimated that overall, about 63 percent of eligible state residents participated in the program. Participation was lower for eligible residents who were aged 60 or older and living alone or with another person aged 60 or older, living in limited-English-speaking households, or non-U.S. citizens. These findings identified population subgroups that might be targeted by information campaigns to increase awareness of SNAP benefits, and helped define characteristics of the SNAP population for social science researchers analyzing the data.

3.3 ENABLE FINER DATA DISAGGREGATION

One of the challenges of producing disaggregated statistics from surveys—even high-coverage and high-response-rate surveys—is that the sample size often limits the population

[34]The study by Feldman et al. (2017) discussed in Chapter 2, which linked records from a webscraped list of deaths caused by law enforcement personnel with records from the National Death Index, used similar methods for estimating the number of people killed by law enforcement.

groups for which reliable statistics can be produced. Administrative data sources, by contrast, are often much larger (even though they might not contain everyone in the population of interest). By combining survey data with administrative records, or combining multiple administrative data sources, researchers can form a dataset that has larger sample sizes from small population subgroups.

Chen (2022) described activities of the United Nations Inter-Secretariat Working Group on Household Surveys toward promoting data integration. One focus is on using multiple data sources to measure progress toward the United Nations Sustainable Development Goals—work that is closely related to data equity.[35] Key to the efforts for obtaining information at finer levels of aggregation is "[u]sing a common list of administrative units across censuses and surveys, and including identical census questions in subsequent household surveys" (United Nations Inter-Secretariat Working Group on Household Surveys, 2022, p. 7).

Arora (2022a, 2022b) discussed Statistics Canada's *Disaggregated Data Action Plan*, an initiative funded by the Canadian government in 2021 to produce timely, disaggregated data for population subgroups that have historically been less visible in probability samples, while maintaining accuracy and protecting confidentiality (see Figure 3-1). A large portion of the funding has been used to create new probability surveys or to increase the size of current probability surveys, to allow production of statistics at finer levels of aggregation. Questions aimed at identifying membership in specific subpopulations have also been added to existing probability surveys, allowing computation of statistics for those groups.[36]

[FIGURE 3-1 about here]

The use of probability surveys alone is not sufficient to respond to information needs about the Canadian population. Indeed, in 2022, 40 percent of Statistics Canada's programs relied, at least in part, on data from non-survey sources such as administrative records. Some programs, such as the Canadian Housing Statistics Program, have relied almost exclusively on administrative data. This program links data from existing administrative sources and the Canadian Census of Population to provide a comprehensive portrait of Canada's housing market, with the goal of including all residential properties in Canada and their owners (Arora, 2022b, p. 25). Property-level data are obtained from land registries and property assessment files. Owner-level information is also derived from land registries and property assessment files, and a variety of owner characteristics are linked from tax data, the Business Register, the Canadian Census of Population, and the Longitudinal Immigration Database. Owner information is supplemented with indicators of residency in the economic territory of Canada, which are obtained by linkage to various data sources, including tax and census data. The linkage is done through a variety of deterministic and probabilistic linkage methods to maximize coverage of the target population.

Consequently, the Canadian Housing Statistics Program allows Statistics Canada to produce estimates for small subpopulations, which was not possible using probability surveys alone. Estimates have been produced for subgroups formed by cross-classifying region and

[35]The 17 United Nations Sustainable Development Goals provide "a shared blueprint for peace and prosperity for people and the planet, now and into the future." They include goals "to end poverty and hunger everywhere; to combat inequalities within and among countries; to build peaceful, just and inclusive societies; to protect human rights and promote gender equality and the empowerment of women and girls; and to ensure the lasting protection of the planet and its natural resources." See https://sdgs.un.org/goals, https://sdgs.un.org/2030agenda, and https://sdgintegration.undp.org/human-rights-approach-data

[36]For example, a new question on gender was included in the 2021 Census of Population that allows all persons living in Canada to self-identify through the Census. See https://www.statcan.gc.ca/en/statistical-programs/instrument/3901_Q1_V7 for the census instrument.

property or owner characteristics, such as period of construction, residential property type, and immigration status.[37] Moreover, those estimates can be produced faster and with less cost. For example, it could be difficult and costly to reach and obtain timely responses from foreign owners through a probability survey, but information about this subpopulation is readily available from the database.

Although they are not affected by sampling errors, statistics computed from the Canadian Housing Statistics Program are not error free. For example, new construction may be missing from the database, and assessment values may fail to capture improvements performed without building permits. Linkage errors also exist, since some records have more accurate information for linkage than others. To evaluate the quality of the linkages, samples of linked records are manually reviewed and estimates of linkage error rates are calculated to ensure that linkages are of high quality.

Disaggregation can be temporal as well as spatial or demographic. Shapiro (2021) mentioned the lag in data availability from key data sources such as the ACS, and provided examples in which other data sources, though imperfect, provided more timely and more frequent data. For example, the Bureau of Transportation Statistics used cell phone location data to study travel patterns early in the COVID-19 pandemic.[38]

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.

3.4 PRODUCE MODEL-BASED ESTIMATES FOR SMALL SUBPOPULATIONS

Box 2-2 describes the U.S. Census Bureau's Small Area Income and Poverty Estimates (SAIPE) program, which uses summary statistics from administrative data as inputs to a regression model that predicts income and poverty for each county and school district in the United States. Small area estimation can also produce estimates for subpopulations other than geographic areas. The U.S. Census Bureau's Small Area Health Insurance Estimates program, which uses a model similar to that of SAIPE to estimate county-level health insurance coverage from the ACS and administrative records, also provides further breakdowns by demographic characteristics. Robinson and Willyard (2021) used a small-area model to generate heat maps of county-level estimates of the uninsured rate for working-age adults living in poverty and children under the age of 19, and state-level estimates for Hispanic, non-Hispanic White, and non-Hispanic Black populations.

Subpopulation estimates produced by small area modeling programs are predictions based on statistical relationships between the predictor variables and the quantities of interest. Those relationships might not hold for all subpopulations being predicted (e.g., a county predicted by a small area model to have a high uninsured rate may actually have a low uninsured rate because the county's predominant employers provide health insurance). The only way to detect such outliers is to obtain direct data from another source.

[37] For example, https://www150.statcan.gc.ca/t1/tbl1/en/tv.action?pid=4610005201
[38] https://www.bts.gov/daily-travel

As with the algorithms discussed in Box 3-1, the estimates from small area estimation procedures are predictions from a model fit to the areas that have survey data. A survey may be nationally representative (that is, accurate national estimates may be calculated from it) and yet have only a few data points for some subpopulations (for example, residents of small rural counties). If the relationship between the model inputs and the outcome of interest is different for certain small rural counties than for the counties that primarily determine the regression parameters, the small counties will be poorly predicted by the model. In general, though, having predictor variables in administrative data that are highly correlated with outcome variables will produce small area estimates that are more accurate, on average, than estimates calculated using the survey data alone.

3.5 ASSESS AND REDUCE MEASUREMENT ERROR

Record linkage can provide a cross-check on measurements of the same concept across data sources. Section 5.4 reviews studies in which records from surveys containing questions about income or transfer programs have been linked to administrative records data. By comparing the amount of wage earnings reported to surveys with that reported to the Internal Revenue Service (IRS) or Social Security Administration (SSA), or by comparing the response to the ACS question about receipt of SNAP benefits with SNAP administrative records, researchers can study patterns of measurement error in data sources and identify subpopulations for which better measurement methods are needed.

Whether and how individuals' characteristics are measured varies across datasets. At the most basic level, a data source may have no information about a characteristic—for example, a health survey or administrative data source may contain no questions about sexual orientation. A recent National Academies of Sciences, Engineering, and Medicine report (NASEM, 2022c) noted the wide variety of survey questions used to ask about sex, gender identity, and sexual orientation, and provided guidance for asking about these characteristics.

Similarly, measurement of race and ethnicity varies across data sources. Box 3-3 describes the race and ethnicity categories specified for federally sponsored data collections in the United States, and Figure 3-2 shows how these categories were implemented in the 2020 Census. Discrepancies across sources can be investigated by studying differences in race and ethnicity categories for records belonging to the same person, and such studies can point the way to methods that can improve measurement of those characteristics.

[BOX 3-3 about here]

[FIGURE 3-2 about here]

For example, participants in surveys and the decennial census are asked to select the race and ethnicity categories that best describe them. Information about age, sex, race, and ethnicity on death certificates, however, is not self-determined—it is usually entered by a funeral director based on observation or on information provided by an informant (often a relative).

Arias, Heron, and Hakes (2016) investigated how well race and ethnicity classifications from death certificates agreed with those from self-responses. They could not, of course, ask the decedents to report their race and ethnicity. However, they could examine Current Population Survey (CPS) records that were linked to records for the same persons who subsequently appeared in the National Death Index (NDI), and compare self-reported race and ethnicity from the CPS to the race and ethnicity in the NDI.

Arias, Heron, and Hakes (2016) found that agreement was close to 100 percent for persons who self-identified on the CPS as White or as Black/African American for the three time periods studied (1979–1989, 1990–1998, and 1999–2011). But, over the three decades studied, only 51–55 percent of decedents who self-identified as American Indian or Alaska Native (AIAN) on the CPS had that same classification on the death certificate, and only 72–80 percent of decedents identified as AIAN on the death certificate had self-identified as such on CPS. For persons self-identifying as Hispanic/Latino(a) on the CPS, about 90 percent had the same identification on the death certificate; of decedents listed as Hispanic on the death certificate, 91–96 percent had self-identified as Hispanic on the CPS. A follow-up study (Arias et al., 2021), linking records from the 2010 Census to the NDI, found similar misclassification rates for non-Hispanic AIAN decedents. If misclassification were not corrected, mortality rates for the AIAN population would be underestimated in statistics calculated from death certificates.

Race and ethnicity are fluid concepts, and they are also used fluidly. Record linkage can be used to study consistency across data collections in which information is self-reported. Using linked decennial census data, Liebler et al. (2017) found that 9.8 million people listed a different race and/or ethnicity in 2010 than in 2000; the most common changes were from "Hispanic some other race" to "Hispanic White" and vice versa.[39]

3.6 ADD FEATURES TO THE DATA THROUGH DATA LINKAGE

In addition to providing information on how the measurement of concepts varies across data sources, record linkage can be used to merge information from different datasets. This use of data linkage is discussed further in Chapter 6.

Adding Variables to a Dataset from Records Linked in Another Source

Brown (2022, slide 5) emphasized the importance of having disaggregated data by race for the purpose of measuring racial disparities. One challenge is that "disaggregated data are strong in some areas (employment, education) and lacking in others (health, wealth)." One way to calculate disaggregated statistics from data sources that do not measure the disaggregation variables is to link the records with a source that has those variables.

Akee (2022) described the value of linking data from the IRS and the U.S. Census Bureau to study income inequality. Income tax data contain a great deal of information on various types of income, as well as adjustments used in calculating adjusted gross income, but the individual income tax form (Form 1040) does not collect information on race and ethnicity. The ACS and decennial censuses do collect information on race and ethnicity; by combining tax information with ACS and decennial census information for the matched records, Akee, Jones,

[39]There may also be discrepancies across self-reported data sources based on who fills out the information on race and ethnicity. For the decennial census, a household respondent ("Person 1" in Figure 3-2) usually supplies demographic information for all household members. If no one in the household supplies information, a proxy reporter (for example, a neighbor or building manager) may be asked. But race and ethnicity information supplied by a household respondent or proxy reporter may differ from information that household members might supply if they were filling out the form. In the 2020 Census, proxy responses were obtained for about seven million households (U.S. Government Accountability Office, 2020). Comparing linked census records in which information was provided by a household member in one census and a proxy reporter in the other, Porter, Liebler, and Noon (2016) found high agreement between household and proxy respondents for White and Black residents, but more disagreements for multiracial and "some other race" categories.

and Porter (2019) could study trends in income inequality from 2000–2014 for relatively small race and ethnicity subgroups.[40] The authors were thus able to look at income and within-group income inequality over time for population groups such as Pacific Islanders, whose sample size is small in a typical survey. One of the challenges for the study was that the decennial census and ACS ask the race and ethnicity of each person, while Form 1040 is for tax-filing units (usually households). Individuals within tax-filing units may have different races or ethnicities. Akee, Jones, and Porter (2019) linked ACS race information to the primary and secondary filers in the tax records. They then used income equivalency weights to assign shares of the tax-return income to the primary and secondary filers, and analyzed the data at the person level. Akee et al. (2020) used similar linkage procedures to compare income inequality over time for recent immigrants who identify as Asian, Hispanic, and non-Hispanic White.

Data linkage at the area level can attach state-, county-, or neighborhood-level variables to a dataset. The Urban Institute's Spatial Equity Data Tool (Narayanan, Stern, and Macdonald, 2021; Urban Institute, 2021b; Brown, 2022) allows data users to upload their own data (e.g., locations of playgrounds or grocery stores) for comparison with estimates of community characteristics (e.g., percentage of residents who are under the poverty threshold) from the ACS. The Urban Institute (2021a) provided an example studying the characteristics of neighborhoods with electric vehicle charging stations.

Linkage Errors and Data Equity

Data linkage provides additional variables when the linkage is accurate. When a record from one data source is mistakenly linked to a record from another source that belongs to a different entity, however, the linked dataset record has erroneous information. Moreover, some records contain insufficient identifying information to enable linkage across datasets and, as discussed in Chapter 2, some subpopulations are more likely than others to have missed links. Akee, Jones, and Porter (2019, p. 1003) commented that "[t]he nonmatches between the IRS and the census race and ethnicity data likely occur among low-income individuals and minorities."

Bond et al. (2014) described the Person Identification Validation System used by the U.S. Census Bureau to assign a Protected Identification Key (PIK) to each record. The PIK is then used to link records across sources.[41] For some records, however, a PIK cannot be reliably assigned because there is insufficient identifying information or because the record does not uniquely match any of the administrative records (mostly from the SSA, but other administrative records are also used). Bond et al. (2014) used ACS data to study characteristics associated with lower linkage scores (and hence lower likelihood of receiving a PIK); for 2010, these characteristics included recent movers, Hispanic persons and persons of "some other race," non-U.S. citizens, immigrants, people who speak a language other than English at home, and people with low incomes or who are unemployed.

Reviewing studies on variation in linkage quality, Randall et al. (2018) found that most studies have been conducted on nested datasets, in which every record in dataset A was also expected to be in dataset B. In these studies, it could thus be assumed that any failure to link a record in dataset A with a record in dataset B is a missed link. When dataset A is not nested in dataset B, though, it may be difficult to determine whether failure to link an individual in A occurs because there is no record for that individual in B, or because a corresponding record

[40]Records were linked at the U.S. Census Bureau under strict confidentiality protections; see Section 5.2.
[41]https://www.census.gov/topics/research/stat-research/expertise/record-linkage.html

exists but the link was missed. For example, health surveys or medical datasets can be linked with the NDI to study subsequent mortality of study participants. But a failure to find a link does not necessarily mean that the person is still alive; it could be the result of low-quality information in the variables used for the linkage. Section 6.4 discusses data-equity issues for linkages of health surveys.

Linkage decisions can affect results. For example, Parrish et al. (2017) studied how linkage decisions affected estimates of child maltreatment in Alaska. They linked records from a probability sample of live births in 2009 with administrative data sources such as death records, records from child protective services agencies, and records from the Anchorage Police Department. Estimates of the incidence of child maltreatment (defined as having at least one report of maltreatment from the multiple sources in the six-year follow-up period) were up to 43 percent lower when a more restrictive linkage (without manual review) was used. Accounting for out-of-state emigration in the longitudinal linkage also affected estimates.

There is often a tradeoff between coverage and linkage errors. For linkages using methods similar to those in Box 2-1, analyses could be restricted to records with high match scores. But that would leave many records unlinked, and the resulting dataset would not be representative of the population, with undercoverage of population subgroups with lower linkage scores. On the other hand, using a low match score cutoff can result in false links. For example, it would be possible for the Canadian Housing Statistics Program (see Section 3.3) to restrict analyses to the set of records thought highly likely to be true matches. However, this would leave many records unlinked, and the resulting database would not be representative of the population, raising equity issues. To avoid these representativity issues, attempts are made to maximize the coverage of the target population, but this results in linkage errors.

Additional Equity Considerations for Data Linkage

Randall, Stern, and Su (2021) noted that variables added during data linkage may increase the risk that individuals could be identified from the data. Box 3-4 discusses how record linkage may affect privacy and confidentiality. Zook et al. (2017, p. 3) commented that "[p]rivacy also goes beyond single individuals and extends to groups. This is particularly resonant for communities who have been on the receiving end of discriminatory data-driven policies historically, such as the practice of redlining."

[BOX 3-4 about here]

In some administrative data collections, information about race, ethnicity, and other characteristics that might be desirable for data disaggregation is omitted by design. For example, because the Equal Credit Opportunity Act prohibits using characteristics such as race, religion, national origin, sex, and marital status in credit-scoring models, credit reports "generally may not include information on items such as race or ethnicity, religious or political preference, or medical history" (Cooper and Getter, 2020, p. 3).

The U.S. individual income tax form (Form 1040) does not ask about race, gender, or ethnicity. Adeyemo and Batchelder (2021) stated that current laws prohibit the IRS from acquiring that information from other agencies and that, in lieu of record linkage, the U.S. Department of the Treasury has been studying methods for imputing race and ethnicity onto tax data to study the "racial/ethnic equity implications of tax policy and tax administration questions,

which could ultimately enable a better understanding of the effectiveness and equity of a variety of tax provisions."[42]

Record linkage illustrates how a method that is promising for promoting data equity must be rigorously evaluated to identify unintended consequences both for measurement and for the communities being measured. Interventions to advance equity may reveal inequities or new challenges, requiring continual efforts to improve data equity involving researchers and the communities themselves.

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.

3.7 ADD FEATURES TO THE DATA THROUGH IMPUTATION

As mentioned in Chapter 2, federal statistical agencies often impute (fill in) values for missing items on censuses or surveys. They may use a deductive rule for some missing values (e.g., imputing a sex of female for a person who listed giving birth in the last year), substitute values from linked administrative data, replace the missing values with values from another record in the dataset that has similar characteristics on the nonmissing items, or use a statistical model to predict the missing items. In some applications, the model for performing imputations is developed on a separate dataset that contains both the variables to be imputed and variables that can be used to predict the items of interest. Imputation can enhance feature equity by providing estimates of missing items (for example, if a survey respondent skips an income question) or information needed to define subgroup membership.

Imputing Information Needed for Disaggregation

The Institute of Medicine (2009, p. 7) argued that the first step toward addressing health care disparities is to collect information on the quality of health care that is disaggregated by race, ethnicity, and language, and recommended: "Where directly collected race and ethnicity data are not available, entities should use indirect estimation [imputation] to aid in the analysis of racial and ethnic disparities and in the development of targeted quality improvement strategies, recognizing the probabilistic and fallible nature of such indirectly estimated identifications."

Administrative data sources can lack information that can be used to distinguish group membership. For example, one impediment to producing disaggregated data for Medicare beneficiaries has been the quality of race and ethnicity information in the administrative files, derived primarily from SSA records. The majority of current Medicare beneficiaries applied for a Social Security Number before 1980, when the application had only three response options for race and ethnicity: "White," "Black," and "Other."[43]

[42]Bearer-Friend (2019) discussed possible harms that might result if race and ethnicity information were collected on federal individual income tax forms, but also noted harms that can result when studies on implications of tax policies do not produce statistics that are disaggregated by race.

[43]The categories were expanded in 1980, and Social Security application forms in 2022 (https://www.ssa.gov/forms/ss-5.pdf) have two ethnicity options and seven race options in accordance with U.S.

Haas et al. (2019) investigated the performance of an imputation method that relies on surname and residential address data to refine race and ethnicity information for Medicare beneficiaries.[44] Each person was assigned a set of six initial probabilities of being in each of six race/ethnicity groups (White, Black, Hispanic, Asian or Pacific Islander, American Indian or Alaska Native, or multiracial) based on U.S. Census Bureau data about the race and ethnicity distributions for that person's surname. These probabilities were then refined with additional information about the race and ethnicity distribution for the census block group containing the person's address, and with information from the administrative files, such as the person's first name or preference for receiving materials in Spanish. Haas et al. (2019) checked the accuracy of the final set of probabilities by linking the administrative files with records from a survey of Medicare beneficiaries' health care experiences, which contained self-reported race and ethnicity. Correlations between the predicted probabilities and the self-reported race were high for persons identifying as White (0.90), Black (0.95), Asian/Pacific Islander (0.92), and Hispanic (0.88); for persons identifying as American Indian or Alaska Native, however, the correlation was only 0.54; and for persons identifying as multiracial, the correlation was 0.12.[45] Haas et al. (2019) recommended using the imputed probabilities to investigate health disparities among White, Black, Asian/Pacific Islander, and Hispanic Medicare beneficiaries. But because the probabilities were much less accurate for American Indian, Alaska Native, and multiracial beneficiaries, "the resulting probabilities for these groups are still not recommended for general use" (p. 21).

Equity Considerations for Imputation

Imputation methods fill in values for missing data using statistical models. But, as seen in the study by Haas et al. (2019) described in the preceding section, imputations can be less accurate for some population groups than for others. As with the prediction algorithms discussed in Box 3-1, imputation models are more accurate when developed on data for which the relationships among variables are similar to those in the data to be imputed.

Randall, Stern, and Su (2021) and Brown et al. (2021) considered potential harms that could result when race or ethnicity is imputed in a dataset. A flawed imputation can result in inaccurate estimates and conclusions, and in the misattribution of characteristics of one population subgroup to another. Jagadish, Stoyanovich, and Howe (2021b, p. 3) observed that "imputation of missing attribute values may involve an algorithm that depends on some model, which may itself be biased. For instance, zip code can be used to 'determine' race. Obviously, this cannot work at the individual level, because not everyone in a zip code is of the same race."

Brown (2022, slide 20) advised weighing the benefits of imputing race and ethnicity against the risks, by looking at the opportunity cost of imputation. Would it be better to use resources devoted to imputation to instead improve the data collection, or to use the "next-best available data?" Brown also recommended examining whether the imputed data would be fit for purpose.

Office of Management and Budget (1997) standards; however, the forms state that providing race and ethnicity is voluntary.

[44]The Bayesian Improved Surname Geocoding method is described by Elliott et al. (2009). Comenetz (2016) described the production of the 2010 Census Surname Table.

[45]Other researchers have found similar patterns of accuracy, validating the Bayesian Improved Surname Geocoding method (e.g., LeRoy et al., 2013).

Surveys commonly rely on imputations to fill in values for questions that respondents did not answer. As discussed in Chapter 5, many respondents leave income questions blank, and the missing income data may be imputed from model predictions, from another record in the data with similar characteristics, or from a separate data source. However, respondents may have left items blank because they did not want to share that information, and they might not have consented to participate in the survey had they known the information would be obtained from other sources. Box 3-5 discusses issues of informed consent regarding linkage and imputation. [BOX 3-5 about here]

3.8 DISCUSSION

As illustrated in this chapter, concepts of data equity, and the benefits that combining data sources can yield for equity, depend on context. The chapters that follow provide additional examples from the panel's workshop and related literature in which combining data sources has advanced data equity.

Record linkage or imputation may be used to add information about characteristics such as race, ethnicity, sexual orientation, or disability to datasets that do not contain that information. These techniques can also add variables to surveys that can enhance insight into survey data, including information that the survey respondent might not know (for example, the respondent might not know the value of health care claims paid by Medicare). But both procedures can also introduce errors into the data. Those errors may propagate or amplify existing biases, and the data-equity implications of various linkage and imputation methods need more study. Small area estimation combines information from separate sources to produce estimates for small geographic units such as counties, but these procedures also have implications for equity because some subpopulations might be poorly fit by the prediction models.

The panel views data equity as a core value to be considered when designing data-collection or data-integration systems and when evaluating quality of data products. Linkage errors can be reduced by including variables with high identification potential in data sources. The need for linkage or imputation can be obviated by data collections that include all variables of interest, but that may be impractical because of high respondent burden or privacy concerns. Similarly, small area estimation methods will not be needed if more data are collected (or an alternative data source is found that measures the concepts of interest). Changing major data collections, however, may not be desirable or economically feasible.

Many of the methods used for combining datasets are fairly new, and their equity implications have not been thoroughly studied. In the panel's view, much more research is needed to better understand the effects of data-combination methods on equity, and documentation for data-linkage projects and other data-combination methods should include implications for equity.

Nelson et al. (2020) stressed the importance of transparency when integrating data and provided a toolkit for embedding questions of data equity throughout the data lifecycle—in planning, data collection, data access, development of algorithms or statistical methods, data analysis, and dissemination. They observed: "Building data infrastructure without a racial equity lens and understanding of historical context will exacerbate existing inequalities along the lines of race, gender, class, and ability" (p. i), but responsible use of data integration, together with community knowledge and skills, can "advance government transparency and accountability in

data use, which is critical to building trust, community well-being, and improved outcomes" (p. 33).

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.

BOX 3-1 Artificial Intelligence and Data Equity

Artificial intelligence systems are increasingly used across society to guide decisions in applications such as voice-activated digital assistants (e.g., Siri and Alexa), personalized marketing, self-driving vehicles, facial-recognition algorithms, surveillance systems, financial fraud detection, signature verification, job-search applications, criminal sentencing, and medical diagnostics (Zhang and Lu, 2021).

An artificial intelligence system is typically trained on one or more datasets that contain an outcome variable as well as other variables that can be used to predict the outcome; during the training, a model is developed that predicts the outcome variable from the predictor variables. The model's performance is fine-tuned and evaluated using validation and test datasets (separate from the data used in the initial training). Often, the model continues to be refined as new data accrue.

The model's predictions, however, are only as good as the training data that feed it. Williams, Brooks, and Shmargad (2018) argued that algorithms discriminate based on data they lack—algorithms trained on data that underrepresent or misrepresent a population subgroup can have flawed predictions for members of that subgroup. Kreuter (2022) commented that "misrepresentation of subpopulations affects accuracy and can have considerable real-world consequences," and described examples in which artificial intelligence systems informed decisions about hiring, medical treatment, and criminal justice. In these applications, biased training data can lead to discriminatory outcomes because the training data are of "unequal quality for demographic subgroups."

As an example, Buolamwini and Gebru (2018) studied the accuracy of three commercial systems that classify a facial image as male or female. They assembled an evaluation dataset of facial images from 1,270 persons from three African countries (Rwanda, Senegal, and South Africa) and three European countries (Iceland, Finland, and Sweden). Images in the dataset were categorized as lighter-skinned if their Fitzpatrick (1988) skin type classification was I, II, or III, and as darker-skinned if they had Fitzpatrick skin type IV, V, or VI. The authors then used each classifier to predict the gender of the person associated with each image and found that the accuracy of gender classification differed greatly by gender and skin type. For all three systems, the error rate for lighter-skinned males was less than one percent. But for darker-skinned females, the error rates for the three systems were 20.8, 34.5, and 34.7 percent, respectively—for two of the classification systems, more than one in three darker-skinned females were assigned the wrong gender. Error rates for darker-skinned males and lighter-skinned females were in the middle, between 0.7 percent and 12 percent (Buolamwini and Gebru, 2018, Table 4).

None of the commercial classification systems mentioned which training datasets were used, but Buolamwini and Gebru (2018) found that more than 80 percent of the people in two standard facial analysis benchmark datasets had lighter skin; fewer than 8 percent were darker-skinned females. They recommended greater transparency about the training and benchmark datasets used in facial recognition algorithms. Kreuter (2022) commented that humans performing classification can make errors too, but "with the [artificial intelligence] system, whatever error happens is something that quickly scales."

Algorithmic bias may arise from decisions made in the model-fitting process (for example, omitting variables that might give better predictions), because the training data exhibit patterns of discrimination and the algorithms propagate those patterns,[46] or because the training data do not contain sufficient representation from particular subpopulations to enable accurate predictions for those subpopulations. Combining multiple datasets to obtain better representation can help with the last problem.

[END BOX 3-1]

[46]For example, Angwin et al. (2016) investigated the accuracy of 7,000 risk assessments for persons arrested in Florida. These assessments, which predicted the risk that a defendant would commit a future crime, were used in sentencing and parole decisions. The authors checked how many of the 7,000 persons were charged with crimes in the next 2 years and discovered that overall, only 20 percent of those who had been predicted to commit a future violent crime had done so. Black defendants were mislabeled as high risk at more than twice the rate as White defendants, and White defendants were mislabeled as low risk more often than Black defendants. Harcourt (2015) observed that, even if race is excluded as a possible predictor in the model, other predictor variables associated with race can lead to discriminatory predictions. He argued that prior criminal history was a proxy for race in risk algorithms.

BOX 3-2 Measuring Coverage of the 2020 Census

The U.S. Census Bureau's mission is to "count everyone once, only once, and in the right place" and for the 2020 Census, the Bureau had plans in place for enumerating "hard-to-count" people such as residents in remote villages, residents in group quarters such as college dormitories or nursing homes, people impacted by natural disasters such as hurricanes, and people experiencing homelessness (U.S. Census Bureau, 2020b).[47] But 2020 Census planning did not anticipate the COVID-19 pandemic, which started at the same time as operations for data collection and disproportionately affected efforts to reach hard-to-count parts of the population.

Since 1950, the U.S. Census Bureau has estimated undercoverage and overcoverage using a post-enumeration survey (a probability sample conducted independently of the census). People in the post-enumeration survey are linked with the census enumerations to determine who was missed or counted in error. The post-enumeration survey is not used to adjust census counts, but merely to assess accuracy and provide information for improving coverage of future censuses and surveys (Kennel, 2021; Marra and Kennel, 2022; Hill et al., 2022).[48]

Using the post-enumeration survey data, Khubba, Heim, and Hong (2022, Table 4) examined undercoverage and overcoverage of the U.S. household population in 2020 by race and ethnicity. Undercounts of Black (–3.3%), American Indian or Alaska Native (–0.91%; the estimated undercount for American Indian or Alaska Native residents of reservations was –5.64%), Hispanic or Latino (–4.99%), and some other race (–4.34%) populations were statistically significantly different from zero. Overcounts of the White (0.66%) and Asian (2.62%) populations were also significantly different from zero.

The post-enumeration survey was designed to produce national- and state-level estimates. Its sample size is not large enough to provide accurate estimates of coverage for small demographic groups or small geographic areas. In addition, the limited number of race categories on the census questionnaire (see Box 3-3) hindered the ability to evaluate coverage for finer subpopulations. Although Asian Americans were overcounted in the 2020 Census, that does not mean that all groups of Asian Americans were overcounted—Hmong Americans may have had a different level of coverage than Korean Americans. Khubba, Heim, and Hong (2022) also reported on 2020 Census undercoverage by age group (children aged 0–4 had the largest undercoverage, at about 3%), sex (male and female), and owner/renter status.

[END BOX 3-2]

[47]As one example of the research done to improve coverage, Fernandez, Shattuck, and Noon (2018) linked administrative records such as Medicaid enrollment data and tenant rental assistance program records with data from the 2010 Census and ACS to study characteristics of children who were missed in the 2010 Census and to inform strategies for achieving a better count in 2020.

[48]The U.S. Census Bureau also uses demographic analysis to assess census coverage, comparing census counts with those obtained from estimates produced using birth and death records, data on international migration, and other administrative records.

BOX 3-3 Measuring Race and Ethnicity in the United States

As of October, 2022, race and ethnicity classifications for all federally sponsored data collections are specified by the 1997 revision of Statistical Policy Directive Number 15 from the U.S. Office of Management and Budget (OMB). The *Revisions to the Standards for the Classification of Federal Data on Race and Ethnicity* state:

> This classification provides a minimum standard for maintaining, collecting, and presenting data on race and ethnicity for all Federal reporting purposes…. The standards have been developed to provide a common language for uniformity and comparability in the collection and use of data on race and ethnicity by Federal agencies.
> The standards have five categories for data on race: American Indian or Alaska Native, Asian, Black or African American, Native Hawaiian or Other Pacific Islander, and White. There are two categories for data on ethnicity: ‘‘Hispanic or Latino,’’ and ‘‘Not Hispanic or Latino’’….
> To provide flexibility and ensure data quality, separate questions shall be used wherever feasible for reporting race and ethnicity. When race and ethnicity are collected separately, ethnicity shall be collected first. (OMB, 1997, pp. 58788–58789).

The 1997 OMB standards encourage finer disaggregation when desired or useful, but all federal data collections involving race and ethnicity must include the five minimum categories for race and the two minimum categories for ethnicity. For example, the ethnicity and race questions for the 2020 Census, shown in Figure 3-2, expanded on the minimum categories with check boxes for 5 ethnicity and 15 race categories (including “some other race”).[49] Respondents could check multiple boxes for race and were also asked to write in more detailed information.

Before the original version of Directive 15 was issued in 1977, each federal agency could use its own categories for race and ethnicity, making it difficult to compare statistics across datasets. For example, death rates for subpopulations are calculated by dividing the number of deaths from administrative records by the subpopulation size from the decennial census or intercensal population estimates; the categories must be defined the same way for these rates to be meaningful.[50] Uniform standards allow statistics to be compared and combined across datasets.

[49]“Some other race” was not one of the categories listed in the OMB (1997) standards, but the 2005 Omnibus Appropriations Bill, which funded the census, stated that “none of the funds provided in this or any other Act for any fiscal year may be used for the collection of census data on race identification that does not include ‘some other race’ as a category” (Humes and Hogan, 2009, p. 13). Over time, “some other race,” which was originally intended to be a small residual category, has been selected by an increasing number of people who do not identify with the listed categories. In 2020 “some other race,” alone or in combination, became the second-largest race category, surpassing the Black population in size. More than 90 percent of the people who were classified as “some other race” were of Hispanic or Latino origin (Jones et al., 2021).

[50]Race and ethnicity have been determined by self-response in the census since 1960. Before then, the information was recorded by a census enumerator. Humes and Hogan (2009) reviewed the history of race and ethnicity measurement in the U.S. decennial censuses (also see OMB, 1997; 2016 for the history of standards for measuring race and ethnicity, beginning with the activities of the Federal Interagency Commission on Education in 1964).

Accurate and consistent measurement of race and ethnicity is crucial for data equity, and in fact equity concerns were a primary impetus for the development of Directive 15:

> Development of the data standards stemmed in large measure from new responsibilities to enforce civil rights laws. Data were needed to monitor equal access in housing, education, employment, and other areas, for populations that historically had experienced discrimination and differential treatment because of their race or ethnicity (OMB, 1997, p. 58782).

But the revised standards in Directive 15 (OMB, 1997) reflect the time when they were developed. Announcing a review of the standards, OMB (2016, p. 67399) wrote: "Since the 1997 revision, the U.S. population has continued to become more racially and ethnically diverse. Additionally, much has been learned about the implementation of these standards since they were issued approximately two decades ago."

The review will examine the format of the race and ethnicity questions (whether these should be separate questions, as in Figure 3-2, or combined) and will explore including additional categories (OMB, 2016). Both issues have implications for data equity. Mathews et al. (2017, p. ix) observed that "a growing number of people find the current race and ethnicity categories confusing, or they wish to see their own specific group reflected on the census questionnaire." The Institute of Medicine (2009, p. 62) noted that the Directive 15 categories (OMB, 1997) are often "too broad for effectively identifying and targeting disparities in health and health care." Additional race and ethnicity categories would permit data disaggregation for smaller population groups than is possible under the 1997 Directive. OMB has begun a formal review of the 1997 standards, with a goal of issuing revised standards in 2024 (OMB 2016; U.S. Census Bureau, 2022b; Orvis, 2022).

[END BOX 3-3]

BOX 3-4 Privacy, Confidentiality, and Data Equity

Issues of privacy and confidentiality are in the Statement of Task for the third report in this series (see Box 1-1) and will be discussed in detail in future workshops. However, privacy and confidentiality are an important part of data equity, and this box outlines some of the ways in which combining datasets might affect these issues.

Confidentiality refers to personal information shared with another person or organization that generally cannot be divulged to third parties without the express consent of the individual. Privacy refers to freedom from intrusion into one's personal matters and personal information. These terms are often used interchangeably in everyday life, but they mean different things from a legal and ethical standpoint. Generally, privacy applies to individuals and confidentiality applies to their information. Thus, discussions about "privacy-preserving" methods or privacy protection are really about confidentiality (third-party access to the data).

Federal statistical agencies typically acquire data under a pledge of confidentiality, promising survey respondents that the information they provide will be used for statistical purposes only and that the information will not be disclosed in identifiable form without respondents' consent (see Box 3-5; Appendix A of NASEM, 2021b describes federal laws protecting privacy and confidentiality of information).[51] Traditionally, this pledge has been honored through stripping information that could be used to identify individual respondents from publicly released datasets ("anonymizing" the data) and using other confidentiality-protection mechanisms such as swapping records, perturbing the data, or suppressing statistics whose release might enable individuals to be identified. Other mechanisms for protecting confidentiality of data include restricting access to the data. The Federal Statistical Research Data Centers, for example, strictly control who can access sensitive data and what types of analyses can be done.[52]

In general, there is a tradeoff between data utility and confidentiality: the more details that are published, the greater the disclosure risk. The only way to maintain complete confidentiality of information is to forego data collection. Otherwise, any dissemination of data carries a risk, however small, that individuals might be identified from the data. The President's Council of Advisors on Science and Technology (2014, pp. 38–39) stated: "Anonymization of a data record might seem easy to implement" but "as the size and diversity of available data grows, the likelihood of being able to re-identify individuals (that is, re-associate their records with their names) grows substantially."

Combining data sources, and particularly combining data through record linkage, can add information to datasets that could potentially be used to identify individuals in the data even if the original data are anonymized. Even if individuals are not identified, Randall, Stern, and Su (2021, p. 7) noted that information that is disclosed might cause harm to individuals and

[51]Examples of pledges of confidentiality can be found at https://www.bls.gov/bls/confidentiality.htm, https://www.census.gov/programs-surveys/ncvs.html, and https://www.cdc.gov/os/integrity/confidentiality/index.htm

[52]The Federal Committee on Statistical Methodology (2005) and the National Academies (2017a, 2017c) described approaches to protecting confidentiality of data. The 2020 Census used a new method called "differential privacy" to protect confidentiality, in which noise was added to statistical tables and artificial microdata were generated from the noise-infused statistics in the tables (U.S. Census Bureau, 2021c).

communities, thereby affecting outcome equity: “Linked credit bureau data, for example, could be used punitively to reinforce racially discriminatory lending practices or target predatory products.” Wardell (2022) mentioned possible unintended consequences of adding items to data collection instruments, particularly if that information might be used to threaten the rights, safety, and security of people in specific communities (such as LGBTQIA+ individuals). Potential unintended negative consequences increase the need to protect confidentiality.

Confidentiality-protection methods attempt to minimize those risks while still allowing useful statistics to be produced. Pujol and Machanavajjhala (2021) observed that there is also a tradeoff between confidentiality and data equity, and that measures intended to protect individual privacy may end up erasing properties of small groups.

Combining data sources requires a careful balance between competing needs. On the one hand, record linkage may provide better coverage of smaller population groups, improving representation equity and feature equity for those groups. On the other hand, creating more granular information through data integration may increase concerns about privacy and confidentiality. Addressing privacy concerns through confidentiality-protection methods that add noise to data may distort statistics for small subpopulations (see NASEM, 2020).

[END BOX 3-4]

BOX 3-5 Informed Consent and Data Ownership

Increased use of multiple data sources raises ethical and legal questions about data ownership and consent for linking data, with implications for data equity.

A key question is whether informed consent should be obtained from participants before linking data from separate sources. Currently, there are a variety of practices within the federal statistical system, guided by the legislation described in a report from the National Academies of Sciences, Engineering, and Medicine (NASEM, 2021b, Appendix A). For example, administrative records can often be linked without requiring consent; the U.S. National Center for Health Statistics requires participants to verbally consent to linkage; some data holders (such as the Social Security Administration or the Center for Medicare and Medicaid Services) have specific consent requirements that must be met before they provide linked data, but these in turn may differ by the agency or organization requesting the data (see NASEM, 2017a for further discussion of this issue).

If informed consent is required, what form should it take? Is opt-out consent (informing participants of the intended linkage and giving them an opportunity to opt out) sufficient? Is the provision of a signature or Social Security Number to document consent necessary? Again, there is wide variation in the ways informed consents are administered.

Data linkage raises ethical questions about the tradeoff between an individual's privacy and the desire for more—and more detailed—information about persons, households, farms, and businesses. For example, if a person chooses not to provide certain information (such as race, ethnicity, or income) in response to a voluntary survey, do statistical agencies and researchers have the right to get this information from another data source or to impute the information?

Should data providers (i.e., the individuals providing the survey or administrative data, whether voluntary or required by law) be informed of the secondary use of data collected for program administration or the repurposing of such data for statistical uses? If so, should consent be obtained for such uses?

If informed consent is required for data linkages, it creates a potential additional source of selection error. When asked, not all survey respondents consent to link their data to other sources (Jäckle et al., 2021a, 2021b). For example, consent rates to administrative data linkage have been declining in the Health and Retirement Study, in parallel with declining survey response rates (see Section 6.5 and NASEM, 2022b, p. 19). Furthermore, African American respondents are less likely to agree to have their data linked to Medicare records, and there is a tendency for the most highly educated to consent at slightly higher rates. These trends and differences may undermine the goal of data equity.

Questions of data ownership also arise with data obtained from multiple sources. Who owns the data—the agency that collects it, the agency that acquires it and repurposes it by linking it to other data, the third-party vendor that acquires or purchases the data from an original source, or the individual providing the data? If data providers own their own data, what rights does this imply regarding use of their data? For example, the European Union's General Data Protection

Regulation has a "right to be forgotten" or right to erasure.[53] Similar legislation is under consideration in the United States. How is this right upheld when information comes from multiple data sources?

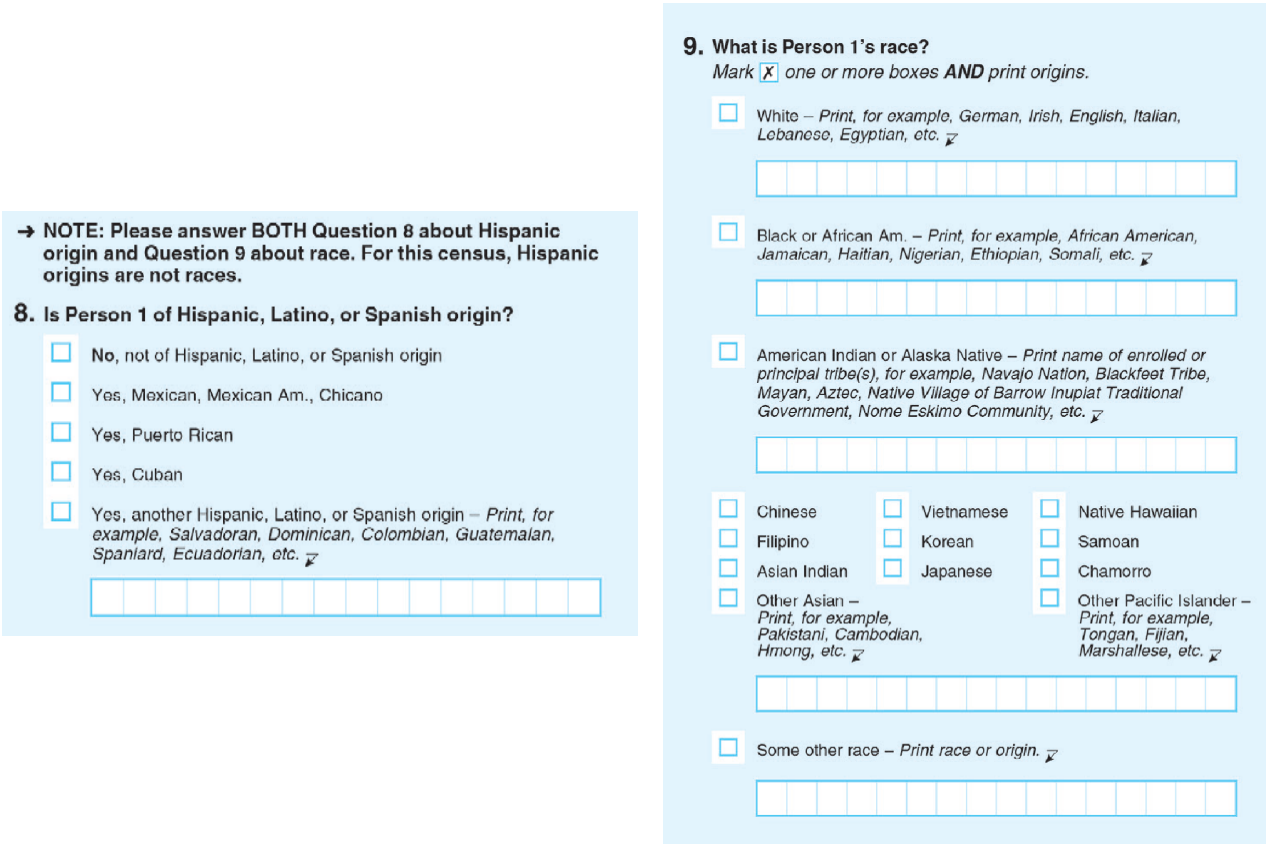
In summary, data equity and informed consent need to be balanced. On the one hand, the increased collection and use of data from or about minority individuals or communities (defined by race/ethnicity, sexual/gender orientation, disability status, or any other characteristic) may advance data equity. On the other hand, it is important to respect the rights of such individuals and communities to privacy (i.e., choice about the provision of data, implying voluntariness) and autonomy (i.e., control over the use of data, implying informed consent).

[END BOX 3-5]

[53] https://gdpr.eu/right-to-be-forgotten/

Figure 3-1 Statistics Canada *Disaggregated Data Action Plan*.

SOURCE: Statistics Canada, Catalogue no. 11-627-M, December 8, 2021 (https://www150.statcan.gc.ca/n1/pub/11-627-m/11-627-m2021092-eng.htm). Reproduced and distributed on an "as is" basis with the permission of Statistics Canada.



→ **NOTE: Please answer BOTH Question 8 about Hispanic origin and Question 9 about race. For this census, Hispanic origins are not races.**

8. Is Person 1 of Hispanic, Latino, or Spanish origin?

- ☐ **No**, not of Hispanic, Latino, or Spanish origin
- ☐ Yes, Mexican, Mexican Am., Chicano
- ☐ Yes, Puerto Rican
- ☐ Yes, Cuban
- ☐ Yes, another Hispanic, Latino, or Spanish origin – *Print, for example, Salvadoran, Dominican, Colombian, Guatemalan, Spaniard, Ecuadorian, etc.*

9. What is Person 1's race?
*Mark ☒ one or more boxes **AND** print origins.*

- ☐ White – *Print, for example, German, Irish, English, Italian, Lebanese, Egyptian, etc.*
- ☐ Black or African Am. – *Print, for example, African American, Jamaican, Haitian, Nigerian, Ethiopian, Somali, etc.*
- ☐ American Indian or Alaska Native – *Print name of enrolled or principal tribe(s), for example, Navajo Nation, Blackfeet Tribe, Mayan, Aztec, Native Village of Barrow Inupiat Traditional Government, Nome Eskimo Community, etc.*
- ☐ Chinese
- ☐ Filipino
- ☐ Asian Indian
- ☐ Other Asian – *Print, for example, Pakistani, Cambodian, Hmong, etc.*
- ☐ Vietnamese
- ☐ Korean
- ☐ Japanese
- ☐ Native Hawaiian
- ☐ Samoan
- ☐ Chamorro
- ☐ Other Pacific Islander – *Print, for example, Tongan, Fijian, Marshallese, etc.*
- ☐ Some other race – *Print race or origin.*

Figure 3-2 Ethnicity and race questions in the 2020 Census.

SOURCE: U.S. Census Bureau (2020a).

4. Creating New Data Resources with Administrative Records

The Foundations for Evidence-Based Policymaking Act of 2018 identified the potential for reducing burden on survey respondents by making use of administrative records already being collected for other purposes. The United States began using administrative data for statistical purposes even before the first decennial census in 1790; Box 4-1 lists some historical developments related to the use cases in this report. This chapter describes innovative projects that have used (or are planning to use) decennial census and administrative records to provide information that would otherwise be measured in a household survey, or not measured at all. The chapter relies in part on the workshop session *Opportunities for Using Multiple Data Sources to Enhance Major Survey Programs*.

[BOX 4-1 about here]

Section 4.1 describes the potential for creating longitudinal datasets from administrative records, illustrating the concept with examples of three databases created by the U.S. Census Bureau, and outlines additional linkage challenges involved when using data from multiple time points. Section 4.2 discusses a proposed U.S. Census Bureau project to link together four databases (demographic, geographic, jobs, and businesses) to provide the basis for a more integrated research program. Section 4.3 describes how a culture of innovation in one of the oldest U.S. data-combination programs, the National Vital Statistics System, is producing new and improved data products. Section 4.4 provides examples of linkage at state and regional levels and identifies challenges involved in combining data collected using different standards and protocols. Section 4.5 concludes the chapter with a summary of common themes and research needed to assess and document the quality dimensions of administrative records.

4.1 CREATING LONGITUDINAL DATABASES FROM EXISTING RECORDS

Surveys that interview new samples each year provide cross-sectional information—a snapshot of a particular moment in time. Repeated cross-sectional surveys can be used to estimate changes in aggregate statistics over time, such as year-to-year changes in the percentage of people who smoke cigarettes. Longitudinal studies, which measure the same set of persons or businesses at multiple time periods, can additionally provide information on individual trajectories and statistics such as the percentage of people who were smokers in 2020 but were nonsmokers in 2021. Longitudinal surveys collect data from the same persons or businesses at repeated points in time, but these are often expensive, and low initial response rates as well as survey drop-out over time can bias results.

An alternative is to link records belonging to the same person or business across existing datasets. As discussed in Chapters 2 and 3, linking records requires sufficient identifying information to determine that two records belong to the same entity. Additional challenges arise with longitudinal linkage of administrative datasets because recordkeeping standards and identification variables can change over time. Measurement standards and practices can also change—for example, the income variable from an administrative data source in 1990 might measure a different concept than the variable used for income in 2020, or new treatment codes in health care claims data may replace or supplement previous codes.

This section describes three U.S. Census Bureau projects that create longitudinal databases from administrative records and decennial censuses. Wagner and Layne (2014) described the U.S. Census Bureau's Person Identification Validation System, which relies on probabilistic linkage methods (see Box 2-1). When the U.S. Census Bureau acquires a dataset, the Bureau compares personally identifying information in that dataset with information in the Census Numident file.[54] Each record in the Numident file is assigned a unique, anonymous identifier called a Protected Identification Key (PIK). When a match is determined, the PIK from the Numident record is attached to the record in the dataset, allowing the new dataset to be linked with other U.S. Census Bureau data resources.[55]

Longitudinal Business Database

The Longitudinal Business Database is built from Internal Revenue Service (IRS) corporate and self-employment tax records that are used to maintain the U.S. Census Bureau's Business Register of nonfarm business establishments.[56] The Business Register, a regularly updated census of U.S. business establishments and firms with paid employees, contains information including business name and address, industry classification, size, employment, payroll, and receipts. It is the primary source for the annual County Business Patterns reports,[57] and it serves as a sampling frame for business surveys and censuses.

The Longitudinal Business Database includes linked Business Register records belonging to the same establishment across time, going back as far as 1976, and it also incorporates information from other sources such as economic censuses. This allows researchers to study year-to-year changes in private employment and the entrance or exit of establishments across industry types, locations, and size classifications.[58]

Chow et al. (2021) described how the U.S. Census Bureau addressed challenges involved in this longitudinal linkage. For example, an employer identification number present in one year but not the next might belong to a business that discontinued operation, but this situation might also arise because the business received a new identification number. The linkage procedure thus also considered name and address to reconcile broken links. Industry classification systems and

---

[54]The Census Numident file is derived primarily from the Social Security Administration Numerical Identification (Numident) file, which contains all transactions recorded for each Social Security Number.

[55]See https://www.census.gov/about/policies/quality/standards/standardc4.html for a description of statistical quality standards and confidentiality protections for linking data. The U.S. Census Bureau's administrative data inventory can be found at https://www2.census.gov/about/linkage/data-file-inventory.pdf

[56]See https://www.census.gov/econ/overview/mu0600.html and https://www.census.gov/programs-surveys/ces/data/restricted-use-data/longitudinal-business-database.html for descriptions of the Business Register and Longitudinal Business Database. Jarmin and Miranda (2002) described the development of the Longitudinal Business Database along with the history of earlier longitudinally linked establishment datasets such as the Longitudinal Research Database, which linked plant-level data from the Census of Manufactures and the Annual Survey of Manufactures (Davis, Haltiwanger, and Schuh, 1996).

[57]https://www.census.gov/programs-surveys/cbp.html

[58]Because the datasets involve IRS tax records, microdata can be accessed only by qualified researchers for approved projects in secure Federal Statistical Research Data Centers. Examples of research studies conducted using these datasets include Benedetto et al. (2007); Akee, Mykerezi, and Todd (2020); Cunningham et al. (2021); Goetz and Stinson (2021); Handley, Kamal, and Ouyang (2021); and Mahajan (2021). Kinney et al. (2011) described the creation of the Synthetic Longitudinal Business Database, intended for exploratory studies by a wider user base, in which data are simulated from statistical models intended to reproduce the structure of the real data.

other definitions of data elements have changed over time, requiring harmonization of the many versions of Business Register data across more than 40 years.

Data-equity considerations arise because less information is available about small, single-establishment firms than about large firms. An imputation model was developed for establishments with unknown beginning and end dates. But Chow et al. (2021, p. 30) noted that "the training data for the imputation model consist almost entirely of establishments born to large multi-unit firms whereas the set of establishments with missing data comes almost entirely from small multi-unit or single-unit firms" and advised researchers to "exercise caution" when using these imputed data.

Longitudinal Employer-Household Dynamics Database

The Longitudinal Employer-Household Dynamics (LEHD) program integrates data from federal censuses, surveys, and administrative records with administrative records collected by states to create a longitudinal database about employment. Each state collects quarterly earnings and employment data to manage its unemployment insurance program. Under the Local Employment Dynamics Partnership, states agree to share unemployment insurance system wage records and other administrative data with the U.S. Census Bureau.

Abowd et al. (2009) described the datasets that are linked in the LEHD and the procedures used to create the database. State data files contain information about economic activity, but little information about the individuals in the wage records. Each individual is assigned a PIK through the U.S. Census Bureau's linkage procedure, which allows demographic information to be appended. Additional information is linked from surveys (including the Current Population Survey [CPS] and the American Community Survey [ACS]), the Business Register, and other sources. These data can also be linked with other datasets containing PIKs; all research must be conducted in a restricted environment such as a Federal Statistical Research Data Center.

The resulting dataset contains longitudinal information about employers and employees that could not be obtained from a single probability survey. A household survey would not have information about the business characteristics of household members' employers, and a survey of businesses would not have detailed information about the employees, but the LEHD has both. This allows the LEHD to produce statistics about employment, earnings, job creation, and job-to-job flows (including characteristics of the origin and destination jobs and earnings changes resulting from job transitions) for detailed levels of geography and industry classification. Data can also be disaggregated by worker characteristics such as sex, age, education, race, and ethnicity.[59]

As with any linked dataset, statistics are affected by the coverage of the data and linkage errors. LEHD data are limited to employees of businesses required to file unemployment insurance system wage reports in participating states, and thus do not present a full picture of employment in the United States.[60]

---

[59] See https://lehd.ces.census.gov/data for other data and statistics available from the program.

[60] https://lehd.ces.census.gov/state_partners/ shows the set of states that participate in the partnership. Data are also obtained from the U.S. Office of Personnel Management to include federal employees. The previous National Academies of Sciences, Engineering, and Medicine report in this series (NASEM, 2023, p. 50) described the "time-consuming and daunting" process for negotiating data-sharing agreements with states to acquire data for the LEHD. Abowd and Vilhuber (2005) studied effects of linkage errors on individual job histories and aggregated statistics.

Decennial Census Digitization and Linkage Project

The Decennial Census Digitization and Linkage project, currently underway at the U.S. Census Bureau, will create a longitudinal database of records from decennial censuses from 1940 through 2020.[61] Records from 1940, 2000, 2010, and 2020 have already been linked. Combining records from other censuses is challenging because the digitized microdata for 1960 through 1990 contain all variables from the censuses except the one piece of information crucial for linking the records—the respondent names. Genadek and Alexander (2019) outlined a plan to scan and digitize the names from microfilmed census records, so that the remaining years can be linked.

Genadek and Alexander (2019, p. 3) stated that the "resulting data resource will expand our understanding of population dynamics in the U.S. far beyond what is currently possible, providing transformational opportunities for research, education, and evidence-building across the social, behavioral, and economic sciences." The linked data will, of course, have limitations. Because census data are available only for years ending in "0" and require a long processing time, the linked files lack both temporal granularity and timeliness. The decennial censuses have long undercounted certain populations, and undercounts before the U.S. Census Bureau used post-enumeration surveys to evaluate coverage were not as well studied (see Box 3-2). And, as discussed in Chapter 3, the quality of linkage information varies across population subgroups and across years, resulting in an inequitable distribution of linkage failures.[62]

Of course, many of the concepts and questions on the decennial census have changed over the decades, but full documentation detailing those changes is available.[63] The interactive infographic provided by the U.S. Census Bureau (2021f) shows the race categories used by each census between 1790–2020 and maps them to the U.S. Office of Management and Budget (OMB) categories described in Box 3-3 (OMB, 1997).

As an example of the type of research that can be done with the linked census files, Leach, Van Hook, and Bachmeier (2018) followed immigrant parents, their children, and their grandchildren from 1940–2014. To include intercensal years, they also linked data from selected years of the CPS and the ACS. They noted, however, that linkage errors could lead to bias. About 70 percent of the children of immigrant parents observed in the 1940 Census were assigned a PIK, and their characteristics differed from those of children without PIKs. In addition, for individuals who cannot be linked, it may be unclear whether the linkage failure is because of insufficient linkage information or because the individual died or emigrated. Leach, Van Hook, and Bachmeier (2018) attempted to correct for these sources of potential bias by using weighting methods similar to those used to adjust for nonresponse.

CONCLUSION 4-1: Longitudinally linked administrative records datasets provide a cost-efficient opportunity to study long-term outcomes, and they

---

[61] https://www.census.gov/programs-surveys/dcdl.html. The files are available to researchers with approved projects through the Federal Statistics Research Data Centers.

[62] For example, the set of persons obtaining Social Security Numbers (SSNs) has changed over time. The Social Security Act of 1935 excluded domestic and agricultural workers from the program. Domestic and agricultural workers, a large percentage of whom were Black, were less likely to have SSNs for earlier censuses and their records were thus more likely to be subject to linkage errors.

[63] Bohme (1989) and U.S. Census Bureau (2002) provided historical context and listed the questions and instructions to enumerators for each census.

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.

4.2 THE *FRAMES* PROJECT

The U.S. Census Bureau provides data and information about the people and economy of the United States. Some of that information comes from the decennial census and surveys that the U.S. Census Bureau conducts; other information comes from administrative records, private-sector data, or other sources. These data products are crucial for the functioning of the democracy, by providing freely available data and statistics to inform decisions made by businesses, policymakers, researchers, and ordinary citizens (NASEM, 2021b).

Motivated by declining survey response rates, increased data-collection costs, and demand for more timely and granular data, the U.S. Census Bureau has proposed a new vision for an "enterprise approach" to statistical data. Santos (2022, slide 3) articulated the goals of the approach: "Improved collaboration with stakeholders and partners, improved data quality, stronger computing power, proliferation of alternative unofficial data products, and new technologies." The data ecosystem is designed to 1) "provide a cloud-centric data storage and computing platform for survey operations"; 2) provide sampling frames that are linkable to other sources and accessible for research purposes; 3) provide modernized data-collection and acquisition solutions that are cost effective, efficient, and scalable; and 4) broadly disseminate publicly available data products in a way that facilitates their use (Santos, 2022, slide 9). This is a long-term project, and, according to Santos, it will require "a decade or more of concerted effort to become sustainable and to achieve maturity."

Figure 4-1 shows a schematic of the *Frames* project. Initial steps involve linking four internal U.S. Census Bureau frames: geospatial, business, job, and demographic (Ratcliffe, 2021a, 2021b). These frames use information from the Master Address File/Topologically Integrated Geographic Encoding and Referencing system (MAF/TIGER, see Box 4-1), the Business Register, the inventory of jobs linked to businesses (underpinning the LEHD database), and administrative records databases used in conjunction with the 2020 Census. All the data sources can be used individually to produce statistics about society but linking them will allow for insights that span the individual topics. As seen in Figure 4-1, further linkages are planned with other U.S. Census Bureau data resources, public records, and data acquired from private-sector sources.

[FIGURE 4-1 about here]

Keller et al. (2022, p. 2) stated that linking multiple data sources at the U.S. Census Bureau "represent[s] a necessary evolution beyond the survey-only model that has reached scientific and practical limits in an era of increasing demand for more data, more often, and more urgently. It holds the promise of producing more timely, robust, and accurate findings and to more fully reflect the diversity of the nation's racial and ethnic composition."

Salvo (2022) commented that the *Frames* project will provide "the scaffold … for the capture and integration of massive amounts of information [leading to] a universal frame that could form the foundation for a transformative capability to integrate" data. He emphasized the importance of such data for local governments and mentioned, as one example, local planners' needs for timely and granular data during the COVID-19 pandemic. To make this project valuable to planners at all levels of government, once the frames are integrated, data must be "curated" to make them consistent, accessible, and "actionable at a local level." Salvo also

recommended developing "use cases to frame discussions with researchers for research agenda development."

One such use case involves challenges involved in improving researchers' understanding of nursing home residents: "Nursing homes are businesses, nursing homes are places where people live, nursing homes have workers" but the "different dimensions of the nursing home picture are not integrated" (Salvo, 2022). Obtaining a comprehensive picture of elder care requires data from many sources, including census and survey data about demographics, income, and health from federal statistical agencies; administrative data from agencies such as the Centers for Medicare and Medicaid Services; information about nursing homes and their employees from sources such as the Business Register and the LEHD; data from state Departments of Public Health and Social Services; and data from the private sector and nonprofit organizations such as the Kaiser Family Foundation.

Challenges in realizing the *Frames* vision include identifying data relevant to the particular problem to be addressed and the fitness for use of those data, as well as obtaining new, high-quality data. Harmonizing varying definitions of concepts and relevant geographies is also critical. Ratcliffe (2021b, slide 3) noted: "Frames exist in an uncoordinated and unintegrated environment" and "[n]o process exists that allows for the direct linkage of information contained in one frame with information in any other frame."

Santos (2022) highlighted the importance of using a data-equity lens to improve policies and practices. The data-equity goals of the *Frames* project include improving coverage of underrepresented groups (capturing individuals who may be in one data source but not others) and increasing sample sizes for small population subgroups, thus enabling production of statistics about those subgroups.

All the individual data sources are incomplete, however, and their union may be incomplete as well. The LEHD, for example, contains only people working for employers in participating states, and may miss self-employed persons. Business files based on tax records will underrepresent new businesses and overrepresent failed businesses. Address files might not capture all new construction or housing abandonment, particularly in sparsely settled or unincorporated locations. Administrative records may be available for only some locations, some population subgroups, or some years. As discussed in Box 3-2, the decennial census differentially undercounts certain race and ethnicity groups. An ongoing evaluation program is important for assessing data-equity impacts of the *Frames* project.

Discussing the potential use of administrative records in the decennial census, McClure, Santos, and Kooragayala (2017) noted:

> The Census Bureau has researched the use of administrative records in enumeration for decades, yet the full implications of such a methodology are still unclear. How accurate is the methodology for different subpopulations? What assumptions about accuracy have been made? What are the costs, risks, and benefits of this approach? Understanding the proposed methodology and the substantive consequences of incorporating it in the census is as critical as understanding the benefits. This is especially true for subpopulations that may have their civil rights affected as a consequence of this new approach. (p. viii).

McClure, Santos, and Kooragayala (2017, p. 12) also observed: "People who do not routinely interact with society's public institutions are less likely to be represented in administrative records (i.e., they are more 'off the grid') … The limited information about these

people that may still be found in these sources could be more likely to be incomplete or inaccurate (e.g., emergency room visits by undocumented immigrants or the homeless)."

One essential aspect of administrative data linkage projects is ensuring public trust, as was emphasized in the previous National Academies of Sciences, Engineering, and Medicine report in this series (NASEM, 2023). In a discussion of records-based alternatives to the decennial census, the National Research Council (1995, p. 62) noted that the "prospect of ongoing linkage of federal, state, and local government data would be opposed by many people." Linkage of administrative sources requires acquiring and processing the data, but typically does not require obtaining consent from persons or businesses whose records are found in the data. The previous National Academies report emphasized that "transparency is critical to building the trust essential to engendering widespread support for a new data infrastructure" and "must be a stated requisite in the legal basis of a new data infrastructure, as well as part of that infrastructure's data-governance framework" (NASEM, 2023, pp. 58–59). The Commission on Evidence-Based Policymaking (2017, p. 17) stressed that "[i]ndividual privacy and confidentiality must be respected in the generation and use of data and evidence" and "[t]hose engaged in generating and using data and evidence should operate transparently, providing meaningful channels for public input and comment and ensuring that evidence produced is made publicly available."

The linkages in the *Frames* project can facilitate study of population groups missed in each source and can point the way to improving coverage and representation—although they cannot help with populations missed by all sources. It may be possible to use data in one frame to update information in another—for example, using information in the Business Register to update MAF listings (Ratcliffe, 2021a). But there are many challenges ahead for this work, including assessing coverage and the impact of linkage errors (see Conclusion 3-2), and it requires cooperative research across the U.S. Census Bureau.

Santos (2022) emphasized the importance of continuing to develop innovative methods of using and combining datasets and of encouraging cooperation among the divisions that house data. "Baking innovation into Census Bureau operations" will require new skills and adaptability, and Santos stressed the need for "human capital strategies, so that [the U.S. Census Bureau] can better recruit, develop, and retain a dynamic and diverse workforce."

CONCLUSION 4-2: Linking administrative data and sampling frames can enable useful future data linkages for social science research and evidence-based policy analysis. However, combined data sources do not necessarily have either full population coverage for generating national statistics or sufficient sample sizes to investigate differences among population subgroups.

4.3 THE NATIONAL VITAL STATISTICS SYSTEM

Sections 4.1 and 4.2 described U.S. Census Bureau activities in linking records obtained from administrative records and censuses. Another model for bypassing surveys and using administrative records directly involves acquisition and standardization of administrative records directly from state and local governments. This is the approach taken by the National Vital Statistics System (NVSS), which keeps track of all births and deaths in the United States.

The NVSS is the oldest national example of cooperative data sharing in the United States, dating back to 1880 (see Box 4-1). It is coordinated by the U.S. National Center for Health Statistics (NCHS) within the Centers for Disease Control and Prevention (CDC). Data are provided through contracts between NCHS and vital registration systems operated by the 50 states, two cities (Washington, DC and New York City) and five territories. The legal requirements for registering births and deaths rest with states, but states work together with NCHS to build a uniform system that provides national data (NCHS, 2021a).

Uniformity of data collection is promoted through use of standard certificates of death, fetal death, and live birth. These are revised periodically in cooperation with state vital statistics offices. Additionally, "model procedures for the uniform registration of the events are developed and recommended for nationwide use through cooperative activities of the jurisdictions and NCHS."[64] These specify the duties of the state registrar, procedures for recording births and deaths, and regulations covering disclosure of information from vital records.

Consequently, the same minimal set of information is collected in every state (some states collect additional information). Death records include information on the decedent's residence, birthplace, surviving spouse, location of death, race, ethnicity, sex, educational attainment, marital status, and cause of death. The race and ethnicity categories accord with OMB standards discussed in Box 3-3 (OMB, 1997). The death certificate also contains an item asking for the decedent's Social Security Number (SSN), which facilitates linkage to other sources.

Several characteristics make the NVSS a model for cooperative data collections. First, it has extraordinarily high coverage of the population of births and deaths. Murphy et al. (2017, p. 3) stated that more than 99 percent of deaths are included in the system. This coverage was accomplished after long collaborative effort—it took 53 years to get all states to contribute data (see Hetzel, 1988).

NCHS also has an ongoing program for quality improvement in data collection, processing, and dissemination. It conducts regular investigations of measurement error in demographic and cause-of-death information (e.g., see Section 3.5 and Hedegaard and Warner, 2021).

Since the NVSS is a census of all vital events, it is highly granular and can be used to produce statistics about small population subgroups. But, at present, the data are not timely: there is a lag between the vital events and the release of the final data file. The final mortality report for deaths in 2019 was published in July 2021 (Xu et al., 2021), although provisional data were available earlier.

A modernization program is underway "to transform the National Vital Statistics System into a tool for real-time public health surveillance."[65] NCHS is working with states to improve the timeliness and quality of death data; the NCHS Modernization Tool Kit provides training materials, tools, and documentation to help jurisdictions establish and learn to use electronic death-reporting systems. These systems are expected to not only speed the production of data—one short-term goal is for NCHS to receive at least 80 percent of mortality records within 10 days of the event—but also promote more complete and more accurate information because data items can be validated as they are entered. The modernization is part of a larger effort within the

[64]Quoted from https://www.cdc.gov/nchs/nvss/about_nvss.htm, which also provides links to the standard forms, model procedures, and guidance for persons completing certificates.

[65]https://www.cdc.gov/nchs/nvss/modernization/goals-accomplishments.htm. See also NCHS (2021b).

CDC to collect more timely data and promote interoperability among data collections including vital records, electronic health records, and electronic laboratory reports (CDC, 2021a).

CONCLUSION 4-3: The National Vital Statistics System can serve as a model for assembling state-administered data programs into coordinated, standardized national databases of administrative records that can be linked to other data sources.

4.4 LINKING DATA AT THE STATE OR REGIONAL LEVEL

The NVSS is perhaps the most complete and successful example of federal coordination of data that are submitted by states. Standardized certificates are used to ensure that submitted information is consistent across locations. For other data collections, for which there are no national standards or federal coordination, each state designs its own data collections to meet program administration needs. This lack of uniformity makes it challenging to use these data for national statistics or research that is national in scope. Because these data are collected for program administration, they exclude individuals who might have been eligible for the programs but did not enter the system. Negotiating data-sharing agreements is also a challenge (NASEM, 2023).

To provide insights on local issues, several state and regional collaboratives have formed to link state administrative data. These collaborations focus on data harmonization in subnational areas, which can lead to greater consistency across data collections. This section focuses on three examples: an integration of data about children and families in Illinois, current work in the State of Washington, and the multistate Coleridge Initiative.

Illinois Integrated Database of Child and Family Programs

Chapin Hall at the University of Chicago began building the Integrated Database of Child and Family Programs in the mid-1980s, to study the children's services system in Illinois (Goerge, van Voorhis, and Lee, 1994; Kitzmiller, 2013). At the time, each agency serving children and families had separate datasets. The database integrates data from Illinois and Chicago agencies that administer the foster care system, investigate child abuse and neglect, and administer assistance and health insurance programs such as public housing, the Supplemental Nutrition Assistance Program (SNAP), and Medicaid. Additional information is obtained from Chicago Public Schools, Chicago Police Department, the juvenile court system, birth certificates, and other sources. The linked database contains longitudinal information about the experiences of all families and children receiving child protective services since 1990.

This database has been widely used to research important child-welfare issues. Examples of such analyses are Goerge, Harden, and Lee (2008), on the consequences of teen childbearing for child abuse, neglect, and foster care placement; Goerge et al. (2009), who analyzed child care subsidy participation and employment outcomes among low-income families in Illinois, Maryland, and Texas using state administrative data linked with the 2001 ACS; Gennetian et al. (2016), on the association between timing and frequency of SNAP program benefits and student outcomes in grades 5–8, as measured by school disciplinary records; and Herz et al. (2019), who studied youth who experience both the child welfare and juvenile justice systems. Most recently,

a study of families who experienced services in multiple public systems highlighted issues faced by these families (Goerge and Wiegand, 2019).

Washington State Department of Social and Health Services

The Research and Data Analysis Division of the Washington State Department of Social and Health Services integrates data from dozens of administrative systems to support research and other analytic use cases. Data are integrated at the individual level into a repository referred to as the Integrated Client Data Repository (ICDR), which is designed to protect privacy and confidentiality. Additional agreements with state agency data suppliers define the governance processes in place to authorize analytic activities that use ICDR data. Examples of the types of data for Washington State residents contained within the ICDR, some dating back to the 1990s, include:[66]

- Medicaid and Medicare claims data spanning domains of physical health, mental health, substance use disorder, long-term care, and developmental disabilities;
- Child welfare system data;
- Food and cash assistance data;
- Vocational and supported employment services;
- Housing program and homelessness data;
- Vital records, including births and deaths;
- Employment and earnings data from the unemployment insurance system; and
- Criminal justice data spanning domains of arrest, jail booking, adjudication, incarceration, and community supervision.

Most data sources are updated on at least a monthly basis. ICDR data are analyzed for a wide range of use cases, including:

- Quasi-experimental analysis of program and service impacts on client outcomes;
- Predictive modeling of populations at risk of adverse outcomes;
- Measurement of quality of services received according to defined standards of care;
- Analysis of disparities and differences in client experiences by race/ethnicity and other demographic characteristics;
- Clinical decision support for care management of high-risk Medicaid beneficiaries; and
- Ad hoc descriptive policy analysis.

Multistate Collaborations

The previous two examples combined data sources within a single state. The final example in this section describes a collaborative effort to establish a multistate data

[66] https://aisp.upenn.edu/network-site/washington-state; Mancuso and Huber (2021). An extensive library of State of Washington health and human services publications is found at https://www.dshs.wa.gov/ffa/research-and-data-analysis. Its research projects are supported by ad hoc funding from state agency program partners, typically with a federal grant as the underlying source of the funding.

infrastructure. Many metropolitan areas straddle state boundaries, but data sources from those states may be in separate enclaves and in incompatible formats. Cunningham et al. (2022) noted that the Foundations for Evidence-Based Policymaking Act of 2018 calls for changes in the way federal data are accessed and used, and that similar changes are needed for state and regional data.

The Coleridge Initiative has organized collaborations that allow for regional data sharing and access.[67] The Initiative does this by providing a secure cloud-based platform, the Administrative Data Research Facility, where confidential microdata can be accessed and linked. The provision of training programs to build the capacity of agency staff to work with the data is an important component of the initiative. Kreuter, Ghani, and Lane (2019) described a program that teaches government employees how to analyze confidential, individual-level data that originate from administrative datasets. The program includes modules on analytical design, database management, data visualization, record linkage, machine learning and text analysis, statistical inference, confidentiality, and data ethics.

Kuehn (2022b) identified the need for a multistate data infrastructure by focusing on the needs of Ohio, for which several metropolitan areas (in particular, Cincinnati, Toledo, and Youngstown) straddle state boundaries. Kuehn (2022b, pp. 8–9) noted: "On the technical side, Ohio and its regional partners needed a secure, usable data platform that could flexibly host data from several states without threatening the states' control of their own data…. Each [state] has different data governance practices that have resulted in different approaches to collaboration." An example of the data produced is the Multi-State Postsecondary Dashboard, which examined, among other topics, the percentage of graduates from each major who are employed, and their locations (in-state or out-of-state) and earnings.[68]

Fischer et al. (2019, p. 677) outlined challenges for developing and maintaining an integrated data system. Foremost is gaining access to a service provider's confidential records, which "requires the cultivation of trusting and mutually beneficial relationships." Additionally, they noted:

> Since all administrative data in an IDS [Integrated Data System] were originally collected for program purposes, not research, the attention to accuracy and reliability is not as high as would be expected for data collected in controlled research settings. As a secondary data source, the richness and quality of data in an IDS is dependent on the quality of underlying administrative records. Data quality standards are applied after the fact through examining aberrant patterns and addressing outliers, but adjustments are necessarily imperfect. Similarly, changes in technology used by data providers can result in changes to data already being supplied to an IDS. For example, data providers may have funders that have required them to change the type of information they collect or how they collect it. Ongoing communication with the data partner has been essential during these times of transition in order to guard against unintended data lapses or misinterpretation (Fischer et al., 2019, p. 679).

[67] https://coleridgeinitiative.org; Cunningham et al. (2022); and Kuehn et al. (2022a). As of June 2022, the Coleridge Initiative has worked with Arkansas, Connecticut, Illinois, Indiana, Kentucky, Maine, Michigan, Missouri, New Hampshire, New Jersey, Ohio, Rhode Island, Tennessee, Texas, and Vermont.

[68] https://coleridgeinitiative.org/projects-and-research/multi-state-post-secondary-dashboard/. See also Cunningham (2021).

The Coleridge Initiative approach, in particular, deals with issues of harmonizing data across states in ways that could scale to larger projects.

State-level linkages have demonstrated the value of administrative data both for research and for state-level program monitoring and evaluation. While states have developed useful research and data privacy-protecting practices, cross-border population mobility and differing legal, technical, financial, and practical considerations across states make these initiatives difficult to scale to the national level. Multistate initiatives such as the Coleridge Initiative provide ideas for harmonizing data concepts and promoting data sharing, and these initiatives have potential for scaling to larger regions.

4.5 USING ADMINISTRATIVE RECORDS TO PRODUCE STATISTICS

As the examples in this chapter demonstrate, using administrative records is not simply a matter of grabbing a convenient dataset off the shelf (or from the cloud) and popping it into a statistical software package. The data user needs to understand the quality and properties of the administrative records and often must do substantial data cleaning and processing before combining administrative data with other data sources. The Federal Committee on Statistical Methodology noted:

> Statistical agencies in many countries have extensive, well-established methods for identifying and reporting threats to quality in data collected and designed for statistical purposes, particularly sample surveys. Methods are less well-developed for dealing with threats to quality from sources other than surveys, such as administrative records and readings from sensors, and other data originally collected for nonstatistical purposes (Federal Committee on Statistical Methodology, 2020, p. 1).

Using administrative data for statistical purposes requires an understanding of the processes used to create, collect, and process the data (Singh et al., 2020). Several frameworks have been proposed for assessing administrative data quality, including Daas et al. (2009); Iwig et al. (2013); Seeskin, Ugarte, and Datta (2019); Statistics Canada (2019); United Kingdom Statistics Authority (2019); and United Nations (2019). These include assessments of the components of quality described in Figure 1-1 and checklists for reporting on quality. Rothbard (2013) provided practical advice on preparing administrative records for analytical use.

Goerge and Lee (2002) discussed the importance of cleaning administrative data prior to linkage and analysis. They noted that administrative data often lack documentation about measurement and quality, and that intensive research is needed to understand the processes behind collecting, processing, and storing the data. Sometimes the original architects of administrative data have moved on to other projects and the institutional history has been lost. Culhane et al. (2010, p. 6) wrote that "many agencies are often too busy with business processes to assess their own data quality on a regular basis" and "an external hosting partner who reviews the data can provide an opportunity for data improvement." Documentation on using administrative datasets for statistical purposes must be prepared by researchers if not supplied by the originating agency.

Boruch (2011) stressed the importance of evaluating and documenting sources of error in administrative records and gave a taxonomy of issues to consider. These include the meaning of key measures such as homelessness and disability, which may vary across programs or personal perspectives, or the distinction between urban and rural residence; the variation over time in definitions, such as changes in race categories; and the difficulty in collecting accurate information on topics such as income, for which the content of probing questions varies across data sources. Boruch (2011) also mentioned issues that might make linking more difficult, such as names versus nicknames, errors in reporting identification information such as SSNs, and coding errors that might occur during data entry.

Section 4.2 addresses one aspect of data equity for data integration efforts: the potential of improving statistics for historically underrepresented population subgroups by obtaining data from multiple sources. Addressing other data-equity aspects, Santos (2022) emphasized the importance of engaging with data consumers as well as with persons and businesses who provide data through surveys or indirectly through administrative records, to better understand their needs and to increase trust and confidence. This raises important questions about equitable approaches for public data access, confidentiality of linked information, data ownership, and the effect of data-combination programs on public trust—trust from all parts of the population. These issues will be explored in future workshops in this series.

This chapter focuses on the value of databases constructed solely from administrative records. Chapters 5–8 discuss examples of integrating administrative records and other data sources with surveys to improve statistics about income, health, crime, and agriculture.

CONCLUSION 4-4: Administrative records are a valuable source of information for official statistics and social and economic research. Each administrative records dataset considered for use in creating national statistics needs to be understood in terms of both its original and its proposed uses. This includes assessing the dataset's fitness for use, timeliness, continuing availability, population coverage, measurement of key concepts, and equity aspects.

BOX 4-1 Historical Uses of Administrative Records for Statistical Purposes: Selected Examples

Administrative records have been used to produce statistics in the United States since the nation's founding. Many of the early milestones involved producing statistics from newly established administrative data collections. The last 50 years have seen an increase in the use of administrative records in combination with other data sources. Here are some selected highlights:

- On July 31, 1789, the first U.S. Congress approved "An Act to regulate the Collection of the Duties imposed by law on the tonnage of ships or vessels, and on goods, wares and merchandises imported into the United States," which directed customs collectors "to receive all reports, manifests and documents made or exhibited to him by the master or commander of any ship or vessel, … *to make due entry and record in books* to be kept for that purpose, all such manifests and the packages, marks and numbers contained therein." (U.S. Congress, 1845, p. 36, emphasis added). Treasury secretary Alexander Hamilton transmitted the first statistical summaries of these data to Congress in January 1791, cross-classifying tonnage and duties by vessels' nationality and the state receiving the goods (Hamilton, 1791). Cummings (1918) reviewed the proliferation of statistics from administrative data that followed these initial statistics on foreign commerce.
- In 1880, the U.S. Census Office established a federal-state cooperative data system that still operates today: a national death-registration area consisting of states and cities providing death statistics deemed of sufficient quality to be tabulated. A national birth-registration area was established in 1915. In 1880, the death-registration system contained only two states (Massachusetts and New Jersey) and a few large cities, but by 1933, all 48 states and the District of Columbia had been admitted to the birth- and death-registration system and each was reporting at least 90 percent of deaths (Hetzel, 1988; National Research Council, 2009, Appendix B; Rothwell, Freedman, and Weed, 2014). In 1946, responsibility for statistics about births and deaths was transferred to the U.S. Public Health Service; today, the National Vital Statistics System is coordinated and guided by the U.S. National Center for Health Statistics (see Section 4.3).
- Following the authorization of an income tax in the 16th Amendment of the U.S. Constitution, the Revenue Act of 1916 called for "the preparation and publication of statistics reasonably available with respect to the operation of the income tax law and containing classifications of taxpayers and of income, the amounts allowed as deductions and exemptions, and any other facts deemed pertinent and valuable" (U.S. Congress, 1917, p. 776). The first statistical report was issued in 1918 (Dalton, 2007), and the Statistics of Income Division in the Internal Revenue Service (IRS) continues to publish aggregate tax information.
- The first volume of the Uniform Crime Reports, containing statistics voluntarily contributed by police departments using the uniform crime classifications specified by the International Association of Chiefs of Police (1929), was published in 1930 (see Chapter 7).
- The 1935 Social Security Act set up a system of continuous data reporting by employers, presenting an opportunity to develop a longitudinal dataset to study work history. The Continuous Work History Sample, initiated in 1941 by the predecessor of the Social Security Administration (SSA), is the oldest major longitudinal dataset in the United States, with data extending back to 1937. It is a probability sample of administrative records, consisting of one percent of all Social Security Numbers (SSNs) ever issued,

along with demographic and geographic information about the persons with those SSNs and annually updated information on earnings, benefits, and payroll-tax contributions (Perlman, 1951; Smith, 1989; Compson, 2022).

- In the 1940s, the U.S. Census Bureau began expanding the use of administrative data for estimating U.S. and state population sizes in noncensus years (county population estimates were added in the 1960s; see U.S. Bureau of the Census, 1947a, 1947b, 1967, 1968). Postcensal population estimates are calculated by adding births, subtracting deaths, and adding net migration to estimates from the most recent census. They currently rely on information about births and deaths from the administrative records in the National Vital Statistics System (see Section 4.3), and on information about domestic migration obtained from tax and Medicare records.[69]
- After World War II, the U.S. Census Bureau "started making extensive use of record files from the Internal Revenue Service and Social Security Administration to develop mailing lists for economic census and surveys, and, eventually, to provide aggregate data, as in the County Business Patterns program and for smaller establishments in the economic census" (Kilss and Alvey, 1984, p. 1). In the early 1970s, the Bureau constructed the Standard Statistical Establishment List (now called the Business Register) from economic census records and administrative data from the IRS and SSA; the register is continually updated using information from U.S. Census Bureau programs and administrative records (see Section 4.1; U.S. Bureau of the Census, 1979; Jarmin and Miranda, 2002). During the 1970s the U.S. Department of Agriculture, used record-linkage techniques with a variety of data sources to improve list frames for agricultural surveys (Allen, 2008).
- In 1973, one of the earliest large-scale survey linkage projects linked records from the Current Population Survey (CPS) with administrative records data from the IRS and SSA. This interagency collaboration allowed researchers to merge variables about earnings and benefits from the administrative data with variables obtained from survey respondents (see Section 2.2).
- The first estimates from the Small Area Income and Poverty Estimates (SAIPE) program, which uses administrative data as input to statistical models for predicting poverty rates in areas with small survey sample sizes, were published in 1993 (see Box 2-2).
- The 1994 Census Address Improvement Act (U.S. Congress, 1994) authorized the U.S. Postal Service to share its Delivery Sequence File, the list of all delivery point addresses served by postal carriers, with the U.S. Census Bureau. The Master Address File (MAF), the U.S. Census Bureau's inventory of all known living quarters, was created in the late 1990s by merging the Delivery Sequence File with the inventory of living quarters enumerated in the 1990 Census (Uhl, 2011). After the 2000 Census, the MAF was integrated with the Topologically Integrated Geographic Encoding and Referencing (TIGER) system, a spatial database developed for the 1990 Census that captures geographic features such as streets, rivers, lakes, and railroads, as well as boundaries of political and census units (National Research Council, 2003). The continually updated MAF/TIGER system is used in decennial census operations and as a sampling frame for surveys.

[69] https://www.census.gov/programs-surveys/popest/about.html. Population estimates are used for allocating federal funds, adjusting for survey nonresponse (since they provide independent population estimates for demographic groups), and as input to programs such as the SAIPE program.

- The U.S. Census Bureau developed the Statistical Administrative Records System in the late 1990s, combining seven national administrative records datasets to test the feasibility of an administrative records census (Prevost and Leggieri, 1999; Judson, 2000). The *Frames* project (see Section 4.2) builds on this work.
- In 2007, a consortium of agencies and research organizations published the first report from the Medicaid Undercount Project, established to study discrepancies in Medicaid enrollment counts between survey estimates and administrative records (SNACC, 2007).[70] The U.S. Medicaid program was established in 1965 to provide health insurance for people with limited income, but a number of studies found that survey estimates of the number of persons receiving Medicaid were substantially lower than the number of persons known to be receiving Medicaid from state-level administrative records (see, e.g., Lewis, Ellwood, and Czajka, 1998). By linking records from the Current Population Survey Annual Social and Economic Supplement (CPS ASEC) with administrative records (the Medicaid Statistical Information System), the Medicaid Undercount Project team identified reporting errors on the CPS ASEC as the main source of the discrepancies.

[END BOX 4-1]

[70]The acronym SNACC comes from the first letters of the collaborating agencies: the University of Minnesota's State Health Access Data Assistance Center, the National Center for Health Statistics, the Department of Health and Human Services Office for the Assistant Secretary for Planning and Evaluation, the Centers for Medicare and Medicaid Services, and the U.S. Census Bureau. See SNACC (2010); Davern et al. (2008, 2009); Noon, Fernandez, and Porter (2019); and Boudreaux et al. (2019) for summaries of the project's findings.

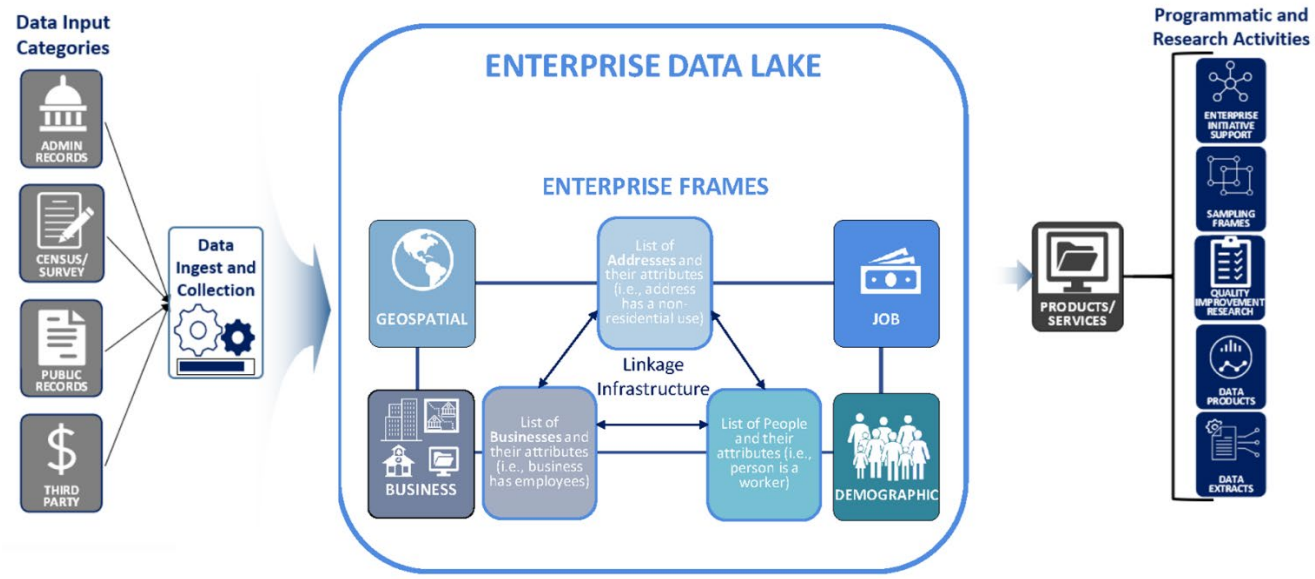


FIGURE 4-1 The U.S. Census Bureau's *Frames* Project.

SOURCE: Santos (2022, slide 11).

5. Data Linkage to Improve Income Measurement

Much of the work on linking survey data with administrative records in the United States has been in the area of income statistics. One of the earliest large-scale linkage projects linked records from the Current Population Survey (CPS) with administrative records data from the Internal Revenue Service (IRS) and Social Security Administration (SSA) (see Chapter 2). Numerous agencies and business organizations collect or assemble data related to income, and many of these data records contain identifying information such as Social Security Numbers (SSNs) that make it feasible to link records across sources.

As seen in Sections 3.6 and 4.2, data sources about income have been linked to increase the sample size for subpopulations, to add variables about subpopulation membership that permit calculation of disaggregated statistics, and to correct for measurement error in surveys. There is also a large literature in which researchers have used linked data to study issues of policy concern, such as poverty disparities, income inequality, earnings volatility, and effects of tax policies.

This chapter, relying in part on presentations in the workshop session *Data Linkage for Income and Health Statistics*, focuses on using record linkage to improve measurement and understanding of income and related concepts. When survey records have sufficient personally identifying information, it becomes possible to link them with tax records from the IRS, detailed earnings records from the SSA, and records from state-administered programs such as the Supplemental Nutrition Assistance Program (SNAP). Comparing income types and amounts reported to surveys with IRS or SSA records can identify areas for which better measurement methods may be needed (for example to address underreporting or overreporting in various sources), and can facilitate modeling efforts to calculate statistics about income and to impute missing income items in surveys and other data sources.

Section 5.1 reviews key federal income surveys and lays out some data-equity issues for income measurement, and Section 5.2 describes sources of administrative records data on income. Section 5.3 discusses how linking income surveys with administrative records can provide information about survey nonresponse bias and improve population coverage. Section 5.4 discusses examples of studies that have linked data sources to assess the accuracy of information about income and participation in food- and housing-assistance programs. Section 5.5 describes two major U.S. Census Bureau projects—the *Comprehensive Income Dataset* project, which examines the effect of improved income measures on poverty estimates, and the *National Experimental Well-being Statistics* project, which proposes to blend administrative and survey data to create more accurate income estimates.

5.1 INCOME DATA COLLECTION ON SURVEYS

Numerous federal statistical agency surveys collect data about income and poverty. This section describes three major ongoing data collections from the U.S. Census Bureau and the Bureau of Labor Statistics that have been used widely in record-linkage activities: the American Community Survey (ACS), the Current Population Survey Annual Social and Economic

Supplement (CPS ASEC), and the Survey of Income and Program Participation (SIPP).[71] It also discusses the strengths and limitations of the household surveys used to report official statistics on income, including issues of nonresponse.

American Community Survey

The ACS has been collecting household and personal income data continuously since 2005. It was designed to replace the decennial census "long form," in which a sample of the population received additional questions about numerous topics, including income. Goals when launching the ACS were to shorten the decennial census form and thereby encourage response to the census, and to provide more timely estimates than the every-ten-years statistics previously produced by the long form.

Because of the ongoing data collection and large sample size (about 3.5 million addresses are selected for the sample each year), the ACS can produce annual statistics about income for geographic areas containing 65,000 or more persons. Estimates for smaller areas are calculated by aggregating ACS data over 5 years, adjusting dollar amounts for inflation. Each year, a new set of 5-year estimates is produced, using data from the most recent 5 years, and these provide a rolling picture of income and poverty. The ACS thus provides more timely estimates than the census long form it replaced. One-year and five-year ACS estimates in tabular form are usually published in September and December of each year, respectively, with microdata issued one month later.

The income questions on the ACS are similar to those on the 1990 and 2000 Censuses (see Figure 5-1). These questions ask about each person's total income and income from each of eight sources. The ACS is a multipurpose survey, however, and can ask only a limited number of questions about income. The two surveys discussed next collect more detailed information on aspects of income.

[FIGURE 5-1 about here]

Current Population Survey Annual Social and Economic Supplement

As discussed in Chapter 2, the CPS originated in a program established in 1940 to provide direct measurement of national unemployment each month, on a sample basis. The monthly CPS sample still serves that purpose but is periodically supplemented by additional data collections that concentrate on specific aspects of the nation's social or economic well-being.[72]

---

[71]Many other federal surveys measure concepts related to income, including surveys that focus on other subjects. See Czajka and Denmead (2008) for a comparative study of income data collected by eight major household surveys.

Surveys that are not conducted by federal agencies also provide opportunities for linkages with administrative records. For example, the Panel Study of Income Dynamics, conducted by the University of Michigan's Institute for Social Research, is a longitudinal household survey that began in 1968 (https://psidonline.isr.umich.edu/). Since then, information on a subset of individuals in the original sample and their descendants has been collected at regular intervals (the sample has been occasionally refreshed to enhance representativeness). Thus, researchers have the potential to link survey observations and administrative records for multiple individuals from the same extended, multigenerational family. Chapter 6 discusses the Health and Retirement Study, an ongoing multicohort panel study of the U.S. population aged 50 and over.

[72]A list of CPS supplements from 2005–2021 is given at https://www.census.gov/programs-surveys/cps/about/supplemental-surveys.html. These include, among other topics, supplements on displaced workers, contingent workers, disability, tobacco use, computer and internet use, and food security.

The oldest of these supplements, established in 1947, is now known as the Annual Social and Economic Supplement (ASEC). The CPS ASEC is the source of official national estimates of income and poverty.[73]

ASEC data are collected once a year (in February through April) to "provide data concerning family characteristics, household composition, marital status, educational attainment, health insurance coverage, foreign-born population, prior year's income from all sources, work experience, receipt of noncash benefit, poverty, program participation, and geographic mobility" (U.S. Census Bureau, 2019, p. 15). While the ACS asks about eight major sources of income, the questions in the CPS ASEC are more detailed, providing information about more than 50 potential income sources.

The sample size of the CPS ASEC is smaller than that of the ACS; in 2021, the CPS ASEC sample consisted of about 91,000 addresses (Shrider et al., 2021, p. 23). CPS ASEC estimates are issued in September of each year. They are accompanied by reports and news releases on income and poverty, with microdata issued one month later.

Survey of Income and Program Participation

The ACS and CPS ASEC are both cross-sectional surveys. The SIPP, which began in 1984, is a longitudinal survey whose "mission is to provide a nationally representative sample for evaluating: 1) annual and sub-annual income dynamics; 2) movements into and out of government transfer programs; 3) family and social context of individuals and households; and 4) interactions among these areas" (U.S. Census Bureau, 2021b, p. 1). The SIPP captures aspects of income and program participation not measured by the ACS or the CPS ASEC, such as changes in household composition, periods of program participation, and detailed data on assets and liabilities (which play a role in determining program eligibility).

The SIPP is organized as a series of national panels, which are samples of households selected to be interviewed periodically over multiple years. The typical duration of a panel ranges from 2.5–4 years. Panels begun at different times overlap, permitting cross-sectional as well as longitudinal analyses (Nwaoha-Brown et al., 2021; U.S. Census Bureau, 2021b). SIPP estimates must be longitudinally processed to ensure consistency over time and to separate data from panels being interviewed at the same time, and they are published with a longer time lag than the CPS ASEC. For example, 2020 SIPP data covering 2019 were released in October 2021.

Strengths and Limitations of Survey Data on Income

The three major surveys providing data about income have different strengths and limitations. This section discusses information that can only be obtained through a survey, and addresses aspects that may be improved by linkage with administrative records.

Each of the three household surveys provide information on entire family and household units. This is critical for measuring poverty and financial well-being. The large sample size of the ACS allows publication of statistics for detailed levels of geography and small demographic groups for the eight income sources it measures (see Figure 5-1). The CPS ASEC and SIPP

[73] https://www.census.gov/data/developers/data-sets/Poverty-Statistics.html and Shrider et al. (2021).

contain detailed questions about income, noncash benefits, and program participation.[74] The questions in each of those surveys have been developed through research programs that include stakeholder input and extensive testing. The surveys also ask about topics that would not be available from administrative records on income. For example, the ACS asks about education and disability; the CPS ASEC asks about health insurance and child care expenses; the SIPP asks about unpaid time away from work and adult and child well-being. All three provide family relationships and demographic characteristics that are not available in administrative data.

Because these are probability samples, selection and representation issues are controlled by the survey designer. The surveys have high population coverage overall (although the surveys exclude some parts of the population by design and some residential addresses are missed by the sampling frame); by contrast, administrative datasets exclude subpopulations that are not part of the program being administered. The surveys also undergo regular quality evaluations.[75]

All three surveys provide national estimates of the income concepts they measure (and the ACS provides estimates for states and smaller geographic areas) as well as statistics for some demographic subpopulations. But not even the ACS has sufficient sample size to provide separate estimates for every subpopulation that might be of interest, and statistical models are needed to estimate income and poverty for small subpopulations (see Box 2-2). In addition, surveys can ask only a limited number of questions and are thus constrained in the amount of information they can collect about income and subpopulation characteristics.

Income estimates from all three surveys are affected by two types of nonresponse. Some households and group quarters residents who are selected for the sample do not participate in the survey, usually because they cannot be reached or refuse to participate (called unit nonresponse because no information is supplied by the sampled unit). Additionally, some households that participate in the survey fail to answer one or more survey questions (called item nonresponse).

Figure 2-1 shows unit response rates for the ACS and CPS ASEC. Participation in the ACS is mandated by law, and consequently the survey has high unit response rates. However, as seen in Figure 5-2, ACS item nonresponse rates for income items have increased over time. In 2005, 18.0 percent of respondents were missing data for at least one type of income; that percentage rose to 32.9 percent of respondents in 2021. The CPS ASEC and SIPP have higher unit nonresponse than the ACS,[76] and also have item nonresponse for income questions (Hokayem, Bollinger, and Ziliak, 2015).

[FIGURE 5-2 about here]

The ACS and other household surveys usually use weighting to adjust for unit nonresponse and imputation to adjust for item nonresponse (U.S. Census Bureau, 2014). There is no guarantee, however, that these methods eliminate potential bias from nonresponse. Weighting adjustments typically ensure that estimates from the survey agree with known population counts for housing units and persons in demographic categories, but nonrespondents might still differ systematically from respondents with respect to other characteristics. As discussed in Chapter 3,

[74]See https://www2.census.gov/programs-surveys/cps/techdocs/cpsmar22.pdf and https://www.census.gov/programs-surveys/sipp/tech-documentation/questionnaires.html for the CPS ASEC and SIPP questionnaires, respectively.

[75]See, for example, https://www.census.gov/acs/www/methodology/sample-size-and-data-quality/ and U.S. Census Bureau (2019).

[76]SIPP response rates are calculated separately by panel and interview and are not directly comparable with the other surveys, but SIPP response rates have also declined over the years. The response rate for the first interview of the 2018 panel was about 58 percent (Nwaoha-Brown et al., 2021).

imputation fills in missing items using data from similar individuals or prediction models, but imputed values can differ from actual values.

Item nonresponse tends to be higher for questions perceived as sensitive, such as those about income, than for other types of questions (Rässler and Riphahn, 2006). Moreover, item nonresponse is more prevalent among some subpopulations than others. Meitinger and Johnson (2020, p. 171) concluded that "in the U.S., African American, Asian, Hispanic, and Native American respondents have each demonstrated higher levels of [item nonresponse], compared to non-Hispanic whites, in social surveys." As noted above, when there is differential nonresponse, data inequities might arise from failing to account for unobserved characteristics in the imputation procedures.

Decreasing response rates for surveys and, in particular, for income items raise concern about potential bias that may be more pronounced in some subpopulations than others. The next section describes administrative data sources that might be used to supplement or study properties of income surveys.

5.2 ADMINISTRATIVE RECORDS SOURCES FOR INCOME DATA

Many administrative data sources collect information about income. For example, the SSA knows how much each recipient receives in Social Security benefits each year. Other sources, such as income tax records, can provide information on particular types of income. Data from these sources can be compared to survey responses or possibly substituted for survey information.

Data from the Internal Revenue Service

The IRS collects information returns (W-2s, Form 1099s) from employers and tax returns (Form 1040s) from individuals, as well as tax returns from corporations and other organizations. The U.S. Census Bureau has direct access to specified items of individual and corporate income tax information under its enabling legislation. The data are protected by limiting access to a small number of Census Bureau employees with "special sworn status" and by following procedures in Title 26 U.S.C. and IRS regulations implementing data sharing. The U.S. Census Bureau uses these data in several ways, including:

- To improve the sampling frame for periodic business surveys such as the Annual Survey of Manufactures;
- As economic census data for certain "nonsampled" single-establishment firms, typically small firms (to reduce small business owners' response burden from filling out forms);
- To produce statistics otherwise not available (such as for the Business Formation Statistics, Business Dynamics Statistics, and Nonemployer Statistics programs);
- As the basis for aggregation to small levels of geography for use in modeling programs (such as the Small Area Income and Poverty Estimates program, described in Box 2-2);

- To create models of federal income and payroll taxes for computing post-tax income (used in the Supplemental Poverty Measure program);[77]
- To evaluate accuracy of income reporting (see Section 5.4); and
- To improve nonresponse adjustments for surveys and censuses.[78]

IRS data have some drawbacks. Data are available only for individuals who have Forms 1040, 1099, or W-2; this excludes many persons with low income and those not in wage and salary employment. Even if submitted, tax data may not be fully accurate. Earnings data are mostly complete in tax records, as federal regulations govern their reporting by employers, and they are eventually used by SSA to determine Social Security benefits. But other types of income, such as self-employment or tips income, are subject to underreporting (IRS, 2019).

In addition, tax returns are linked to tax-filing units, not households. For example, spouses who are married and living together can file separately, and dependents of divorced parents living separately can be reported on the return of the parent with whom they do not reside. Additional challenges to linking tax units to households include typographical errors in addresses, inaccurate addresses (e.g., post office boxes, rural routes, lack of apartment numbers), and outdated addresses. Even if one can reconstruct households accurately (see Larrimore, Mortenson, and Splinter, 2021), tax forms contain little demographic information, which limits options for income tabulations (e.g., those with or without child dependents, those with or without a 65+ exemption). Inaccurate reporting can affect the use of tax data as benchmarks, and inconsistency in tax units versus households can affect linkage accuracy. In addition, some tax items are unavailable to the U.S. Census Bureau.

Data from the Social Security Administration

The SSA has data files that can provide earnings histories and Social Security benefit amounts. Those earnings histories are necessary to compute Social Security benefits once a program participant retires, and the data are provided to SSA by the IRS. SSA earnings records provide a longitudinal earnings history associated with each SSN. SSA data are also linked to longitudinal surveys such as the Health and Retirement Survey, to supplement the survey information (see Chapter 6).

Administrative Data from Other Government Agencies

Other federal agencies collect transfer program data that can be used to verify or enhance survey data. Examples include the Supplemental Security Income program administered by SSA, SNAP benefits for states reporting those benefits, and information about housing units receiving federal assistance from the U.S. Department of Housing and Urban Development (HUD).

---

[77]The Supplemental Poverty Measure "extends the official poverty measure by taking account of many of the government programs designed to assist low-income families and individuals that are not included in the official poverty measure" (Fox and Burns, 2021, p. 1).

[78]Mule (2021) described the use of administrative records for nonresponse in the 2020 Census; Rothbaum and Bee (2021) and Rothbaum et al. (2021) described their use to help reweight the CPS ASEC and the ACS for unit nonresponse.

Hokayem, Raghunathan, and Rothbaum (2022, p. 82), studying item nonresponse, found that "there are clear efficiency gains from using administrative data" in CPS ASEC imputation models. See Benedetto et al. (2013) for implementation of similar models for the SIPP.

5.3 USING ADMINISTRATIVE DATA WITH INCOME SURVEYS

Administrative records have the potential to address some of the shortcomings of survey data, including small sample sizes that limit the groups for which disaggregated statistics can be produced, inaccuracies in reporting income, and possible bias from misreporting and nonresponse. Administrative records also may contain information on some populations that are excluded from surveys (e.g., persons living in institutions). Administrative data can be used to reduce respondent burden in surveys by replacing survey questions or providing additional information without adding survey content. Combining earnings information from surveys with administrative data has the potential to provide more accurate estimates of income than could be calculated from either source by itself. [79]

The Interagency Technical Working Group on Evaluating Alternative Measures of Poverty (2021, pp. 38–39) discussed three main ways to correct survey data for misreporting: "(1) rules-based approaches; (2) statistical- or regression-based modeling; and (3) direct substitution of survey reports with administrative records. These three approaches could be used independently or in combination." Rules-based approaches impute participation using program rules: "[f]or example, in some states any person receiving public assistance is categorically eligible for Medicaid and SNAP" and thus Medicaid participation could be imputed for persons receiving public assistance. The SIPP uses the second approach.[80] The National Experimental Well-being Statistics project, discussed in Section 5.5, uses the third approach.

Linking administrative data to the sampling frame for a survey can provide information about characteristics of nonrespondents and the nature of potential nonresponse bias (see, e.g., Sakshaug and Antoni, 2019; Rothbaum and Bee, 2021). For example, Bee, Gathright, and Meyer (2015) linked the 2011 CPS ASEC sampling frame to IRS records and found that survey respondents and nonrespondents differed in demographic characteristics such as marital status but had similar income distributions.

Effects of item nonresponse in surveys can also be studied through linkage with administrative records. For example, Bollinger et al. (2019) studied item nonresponse to the earnings question in the 2006–2011 CPS ASEC microdata by linking the data with SSA's detailed earnings records. They found a U-shaped pattern, in which nonresponse was highest for extreme high- and low-earning individuals. Celhay, Meyer, and Mittag (2021) linked New York State SNAP and public-assistance data to the ACS, the CPS ASEC, and the SIPP, and found that nonrespondents to survey questions about SNAP and public assistance were more likely to be program recipients.

The linkage process, however, can also introduce errors and inequities, as discussed in Section 3.6, and linkage errors can interact with errors from nonresponse. For example, when studying linkages from the 1998–2009 CPS ASEC to SSA detailed earnings records, Hokayem, Bollinger, and Ziliak (2015) found that the characteristics of matched and unmatched individuals differed for survey respondents and nonrespondents. Survey respondents who could be matched

[79] See Bee and Rothbaum (2019) and Meyer and Mittag (2021). There has also been preliminary work on replacing survey questions (particularly on the ACS) with administrative records (e.g., O'Hara, Bee, and Mitchell, 2016; and NASEM, 2016b).

[80] https://www.census.gov/programs-surveys/sipp/methodology/data-editing-and-imputation.html and https://www.census.gov/programs-surveys/sipp/tech-documentation/user-notes/2020-usernotes/chngs-imputd-earngs.html

with a corresponding SSA record were 14 years younger on average than survey respondents who could not be matched, while matched nonrespondents were only 3 years younger on average than nonmatched nonrespondents. Moreover, "matched respondents are statistically much less likely to be a high-school dropout or to be living in poverty than nonmatched respondents. These gaps are relatively small among nonrespondents" (p. 939).

Survey respondents who can be linked can thus have systematic differences from respondents who cannot be linked, and both can differ from nonrespondents. Omitting unlinked records can potentially lead to biases in the resulting analysis, so it is important for analysts to understand the characteristics of unlinked entities from each data source. Many linkage studies referenced in this report use weighting methods to attempt to compensate for missing links, as discussed in Section 6.4. These methods essentially treat linkage failures as a form of nonresponse and distribute the weights of units that cannot be linked among a set of units with similar characteristics (e.g., race, ethnicity, age) who can be linked. An additional option is to incorporate linkage uncertainty into standard errors of the statistics (Reiter, 2021).

Finally, some people may be missed by both the survey data and the administrative records data. For example, low-income households were more likely to be nonrespondents to the CPS ASEC in 2020 and to also be missing from tax data because they are not required to file (Rothbaum and Bee, 2021). Thus, the linked data can end up underrepresenting people who did not respond to the survey, did not have an administrative record, or lacked a strong set of identifying information to enable data linkage.

Bee and Rothbaum (2019) suggested some methods for redesigning surveys to take advantage of administrative records:

> For example, we could assess from administrative records available prior to interview (through address linkage) the likelihood that members of a given housing unit will be PIKed [assigned a unique identifier used for linkage] and have particular administrative records. We could then use this information to adjust the probability that a given individual is asked particular income questions. This could reduce respondent burden on those that are more likely to have administrative data, maintain the questions for those likely not to have administrative data, and preserve a sample of each group with survey responses for modeling and imputation.
>
> Another possibility is to use administrative information to over-sample subsets of the population, such as those that more likely have income that is not covered by administrative data or those that are less likely to be assigned a PIK. Similarly, survey questionnaires could focus on capturing information to improve linkage and representativeness, and to cover topics that are difficult to capture in administrative records, such as subjective well-being and informal employment (pp. 35–36).

5.4 STUDYING MEASUREMENT OF INCOME AND PROGRAM PARTICIPATION

This section examines examples of research conducted on using administrative records to assess the accuracy of survey reports on key sources of cash income—earnings, retirement, and pension income. It also looks at comparisons of survey data with administrative records from noncash transfer programs such as SNAP, to study the accuracy of participation and income

reporting. This section does not include a comprehensive literature review (more extensive literature reviews can be found in Meyer and Mittag, 2021, and the Interagency Technical Working Group on Evaluating Alternative Measures of Poverty, 2021), but the examples show the potential of data linkage for improving accuracy in both administrative records and surveys.

Earnings typically account for approximately 80 percent of total income (Rothbaum, 2022, slide 12), and a number of studies have compared various types of earnings from the surveys discussed in this chapter with earnings for the linked records in administrative data. Examples include:

- Pedace and Bates (2000) analyzed income misreporting propensities and magnitudes using the 1992 SIPP linked to SSA earnings records for wage and salary and self-employment earnings, concluding that "the 1992 SIPP accurately estimates the net number of earnings recipients, but tends to underestimate the amounts received…. [R]espondents on the lowest end of the income distribution tend to overreport earnings, while those at the higher end of the earnings distribution are more likely to underreport earnings" (p. 173). They also looked at demographic characteristics associated with large discrepancies between SIPP and SSA record amounts, and found that large discrepancies were more common among males, persons reporting Hispanic ethnicity, persons reporting Black or Asian race, those who were married/divorced/separated (relative to never married), and persons in certain occupational categories.
- Using linked CPS ASEC/SSA records from 1998–2009, Hokayem, Bollinger, and Ziliak (2015) found that poverty rates among matched respondents were, on average, 1.7 percentage points lower using ASEC earnings than earnings from the SSA data. They speculated that "under-the-table" earnings may show up in the ASEC but are not reported to tax authorities.

Self-employment income, a component of earnings, is particularly susceptible to discrepancies between survey and administrative data and is likely underreported in both.

- Abraham et al. (2021) linked CPS ASEC records with tax information for the same individuals from the SSA detailed earnings record files to study self-employment income, which is more likely to be underreported in tax records than is wage income. They concluded: "The disagreement between these two data sources is both large and growing. Over the period from 1996 through 2015, 51.5% of those reporting CPS-ASEC self-employment income had no self-employment income for the same year on their tax returns. Even more striking, over the same period 66.7% of those with self-employment income on their tax returns did not report it in the CPS-ASEC" (p. 827).
- Eggleston, Klee, and Munk (2022) found that 32 percent of 2014 SIPP respondents who reported only unincorporated self-employment had no corresponding tax form in the SSA detailed earnings records. They noted: "The lack of tax forms for a self-employed worker may indicate low or negative profits—recall unincorporated self-employed workers are only required to file a 1040-SE if their earnings exceed $400—or it may indicate the self-employed worker did not report their income to the IRS" (p. 13).

Retirement and pension income are also amenable for study with SSA and IRS data. A number of studies have found that earned income from retirement and Social Security benefits is reported with high accuracy in surveys, but other forms of retirement income (for example, from pensions and individual retirement accounts) may be underreported.

- By linking 2013 CPS ASEC records for persons aged 65 or older to administrative data records supplied by the SSA, Bee and Mitchell (2017) were able to examine discrepancies at the individual record level (previous studies had compared aggregated statistics from separate sources). They found that the CPS ASEC underestimated retirement income overall, but "across most of the income distribution, we find that retirement income underreporting is mainly responsible for the overall income discrepancy, while self-reported earned income and Social Security benefits correspond well with administrative records" (pp. 2–3). Furthermore, most of the underreporting occurred because people who received retirement income failed to report any of it; when reported, retirement income amounts matched well.
- Dushi and Trenkamp (2021) used data from the 2016 CPS ASEC linked with IRS and SSA records to examine the extent to which administrative records could improve income estimates. For the population aged 65 or older, they found that "supplementing the CPS ASEC with IRS and Social Security administrative data results in a higher estimate of pension income's share of aggregate income, less estimated reliance on Social Security, and a lower estimated rate of poverty" (p. 3).

Food and housing assistance participation can be studied by comparing survey data to SNAP records and records from HUD. A number of studies have found that survey estimates of amounts received from SNAP benefits are lower than amounts from administrative records (see, e.g., Meyer, Mok, and Sullivan, 2015). Linking records allows researchers to explore discrepancies for individual households.

- Shantz and Fox (2018) linked records from SNAP data in seven states to 2009–2015 CPS ASEC data and found that more than 40 percent of SNAP recipients did not report receipt on the survey. Celhay, Meyer, and Mittag (2021), linking CPS ASEC, ACS, and SIPP data with SNAP records in New York State, confirmed the high rate of discrepancies found in earlier studies, and commented that differences were most pronounced for households with Hispanic or Black householders.
- Meyer and Mittag (2019) linked CPS ASEC records from New York State with HUD and state administrative records, and found that 36 percent of housing-assistance recipients did not report receipt in the survey.

These studies indicate that the patterns for discrepancies between survey data and administrative records differ by survey, type of income or program, and, in some cases, by population subgroups. This has implications for the use of administrative records to impute data, develop imputation models, or serve as a substitute for survey data collection, and for the use of administrative records data in promoting data equity. Imputing Social Security benefits, for which centralized, detailed records exist, is likely to result in more accurate data than

respondents' recollections of the benefits they received. Similarly, SNAP files have complete coverage of program participants. Self-employment income may be inaccurate in all sources.

5.5 USING LINKED INCOME DATA TO IMPROVE INCOME STATISTICS

While many of the studies using linked income data from surveys and administrative records have addressed undercoverage, nonresponse, and reporting differences, other studies have looked at the effects of adjusting survey data on outcomes of interest, such as poverty or income distributions, for the population as a whole and for subpopulations. This section describes two projects at the U.S. Census Bureau that have linked survey and administrative data, and explores their potential for improving understanding of income dynamics and poverty.

Comprehensive Income Dataset Project

The Comprehensive Income Dataset (CID) Project began at the U.S. Census Bureau as an internal project but is now transitioning to one whose product is available to outside researchers at a Federal Statistics Research Data Center. One of the motivating factors for creating the CID was that underreporting of various types of income has worsened over time (Meyer, Mok, and Sullivan, 2015).

The CID links each of four household surveys—the ACS, the CPS ASEC, the SIPP, and the Consumer Expenditure Interview Survey (collected by the U.S. Census Bureau for the Bureau of Labor Statistics)—to administrative records including:

- IRS-supplied tax return data;
- SSA Detailed Earnings Record and Master Beneficiary Record (Social Security and Supplemental Security Income);
- Federal housing assistance data from HUD (including the Public and Indian Housing Information Center and Tenant Rental Assistance Certification System files);
- Medicare and Medicaid enrollment data from the Centers for Medicare and Medicaid Services' Medicare Enrollment Database and Medicaid and Children's Health Insurance Program Statistical Information System; and
- Selected state data from SNAP; the Special Supplemental Nutrition Program for Women, Infants, and Children; Temporary Assistance to Needy Families; public assistance programs; and the Low Income Home Energy Assistance Program (see Medalia et al., 2019, pp. 4–5).

Medalia et al. (2019, p. 6) offered anticipated outcomes from the CID Project, "including improving the Census Bureau's household surveys, becoming a critical resource for policymakers to evaluate policies, programs and taxes, and offering better evidence for researchers investigating a diverse range of topics." Furthermore, linkage to Social Security and Supplemental Security Income files might permit actual benefit amounts to be substituted for the questions normally asked on income surveys, which could reduce respondent burden and increase accuracy. Such substitution could be expanded to programs administered by states, such

as SNAP, if complete and standardized data reporting from states to the federal government could be achieved.[81]

Data from the CID Project have been used in numerous research projects.[82] For example, Meyer et al. (2021b) concluded that incorporating administrative data has a larger impact near the bottom of the income distribution, and that estimates calculated without incorporating administrative data overestimate poverty and underestimate the anti-poverty effects of safety net programs. In addition, administrative data can shed light on populations not covered by surveys. For example, the sheltered homeless population is excluded or underrepresented in most surveys, and the unsheltered homeless population is excluded from all surveys except the decennial census; using linked data, Meyer et al. (2021a) found persistently low well-being for these populations.

National Experimental Well-being Statistics Project

A second comprehensive U.S. Census Bureau project, The National Experimental Well-being Statistics (NEWS) Project is underway. The NEWS Project is closely related to the CID Project and has the goal of developing better federal income statistics. Rothbaum (2022) discussed the potential of NEWS for expanding the set of income and resource statistics produced by the U.S. Census Bureau and for developing "best possible" estimates for income topics that make use of the range of data available. In addition to producing measures comparable to existing income, resource, and poverty statistics (including inequality), NEWS researchers hope to produce "mobility, opportunity, and volatility" statistics that focus on income and earnings dynamics. The systemic integration of multiple data sources will allow researchers to study and address potential biases from individual data sources (for example, from missing data or misreporting) through linkage with other data sources, and to produce new statistics that would not be possible from a single source.

According to Rothbaum (2022, slides 8–9), data sources for the NEWS Project include surveys such as the ACS and the CPS ASEC as well as the decennial census. The administrative data include IRS and SSA data, the U.S. Census Bureau's Master Address File, and the Longitudinal Business and Longitudinal Employer-Household Dynamics databases (see Section 4.1). Additional information is included from state and federal programs such as SNAP and programs administered by the HUD, the Department of Veterans Affairs, and the Centers for Medicare and Medicaid Services. The Census Bureau also integrates private-sector data on home values into NEWS.

Examples of methodological studies related to specific aspects of NEWS include Jones and Ziliak (2022) on the Earned Income Tax Credit; and Fox, Rothbaum, and Shantz (2022) on SNAP. In addition, NEWS researchers are using linked administrative data to adjust CPS ASEC weights for unit nonresponse (Rothbaum and Bee, 2021).

[81]This was recommended by the U.S. Commission on Evidence-Based Policymaking (2017, p. 2): "Where appropriate, states that administer programs with substantial Federal investment should in return provide the data necessary for evidence building."

[82]A partial bibliography can be found at https://cid.harris.uchicago.edu/, which also describes planned future linkages for the project.

Using Administrative Records to Improve Income Measures

Rothbaum (2019) discussed strengths and limitations of using administrative records to improve measures of income. He listed three options (slide 3): 1) direct replacement, assuming that the administrative records are correct and substituting their information for survey responses (but assuming administrative records are free, or nearly free, from error is a big assumption); 2) using the survey response alone, when administrative records are unavailable or survey results have been shown to be reasonably accurate; or 3) combining information from both sources, since both sources provide information about "true" underlying income but both also have errors.

Bee and Mitchell (2017) suggested that administrative records may improve measurement of earnings, self-employment income, and income for those aged 65 and older. Income for older Americans is often from benefit programs, particularly Social Security and Supplemental Security Income, so direct substitution is possible if timely data are available from the SSA (Rothbaum, 2019, slide 6).

Rothbaum (2019) noted that some earnings (e.g., tips) may be under- or unreported to the IRS but may be reported on surveys. Sources such as the Longitudinal Employer-Household Dynamics database (see Section 4.1) may have information on nontaxable income that is missing from tax records. Rothbaum (2019) advocated improving earnings measurement through first obtaining the "best" estimate of earnings from administrative records, then comparing that estimate with survey data. The process requires guidelines for deciding which estimate is "best" and for which individuals the survey (or administrative record) value is preferred.

Self-employment income is more difficult to adjust through modeling and imputation, and the studies cited in Section 5.4, as well as audit studies, suggest there is substantial underreporting on tax forms and surveys. Some survey respondents report self-employment as wage and salary earnings (or not at all), and for some self-employment income there are no third-party information returns (such as Form 1099). One option is to develop imputation models for self-employment, using relationships between self-employment income and other characteristics from audit studies.

If administrative data are to be used to replace or impute survey items, timeliness of the administrative data is a critical issue. For example, SSA data used for earnings analyses are generally not available until the following year. One approach is to use time-series modeling, assuming that relationships between survey and administrative data within demographic and socioeconomic groups hold across time. Revised annual estimates could be released when the full set of administrative data becomes available, as is done in the National Income and Product Accounts. But tax laws and program administrative rules can change, which can make models developed on earlier data invalid for the current year.

Other challenges include false links and linkage failures as noted in Sections 2.2, 3.6, and 5.3, including the inability to link about 10 percent of survey respondents with administrative records (Rothbaum, 2019, slide 19). Linkage failures may occur because of problems with identification information or because administrative records have limited geographic, time, or population coverage.

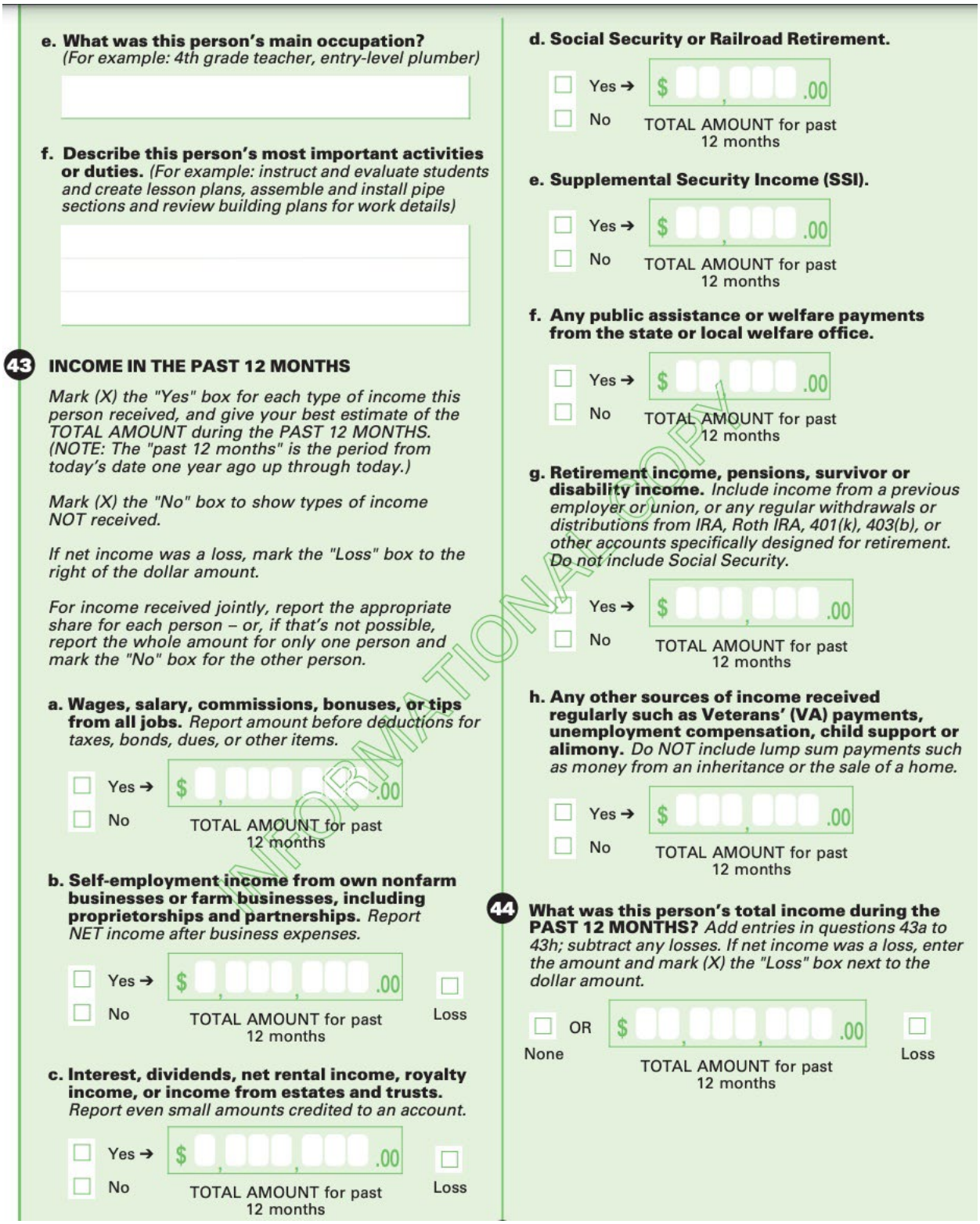
In their review of the U.S. Census Bureau *Frames* project discussed in Section 4.2, Keller et al. (2022, p. 28) commented that "measuring income accurately … would benefit from curating and integrating multiple data sources [particularly in] capturing income for the bottom

and top 10 percent of the income distribution where survey data is less useful due to more complex income payments."[83]

Income measurement remains an active area for survey and administrative data development. With the development of data linkages, there is great potential for improving federal statistics.

CONCLUSION 5-1: Comparison of survey data with linked administrative records can provide statistical agencies with valuable information on measurement quality as well as guidance for further investigations and improvements.

[83]Better understanding of income, consumption, and wealth is the focus of a 2022 National Academies consensus panel. See https://www.nationalacademies.org/our-work/an-integrated-system-of-us-household-income-wealth-and-consumption-statistics-to-inform-policy-and-research



e. **What was this person's main occupation?** *(For example: 4th grade teacher, entry-level plumber)*

f. **Describe this person's most important activities or duties.** *(For example: instruct and evaluate students and create lesson plans, assemble and install pipe sections and review building plans for work details)*

43 INCOME IN THE PAST 12 MONTHS

Mark (X) the "Yes" box for each type of income this person received, and give your best estimate of the TOTAL AMOUNT during the PAST 12 MONTHS. (NOTE: The "past 12 months" is the period from today's date one year ago up through today.)

Mark (X) the "No" box to show types of income NOT received.

If net income was a loss, mark the "Loss" box to the right of the dollar amount.

For income received jointly, report the appropriate share for each person – or, if that's not possible, report the whole amount for only one person and mark the "No" box for the other person.

a. **Wages, salary, commissions, bonuses, or tips from all jobs.** *Report amount before deductions for taxes, bonds, dues, or other items.*

Yes → $ .00 TOTAL AMOUNT for past 12 months
No

b. **Self-employment income from own nonfarm businesses or farm businesses, including proprietorships and partnerships.** *Report NET income after business expenses.*

Yes → $ .00 TOTAL AMOUNT for past 12 months Loss
No

c. **Interest, dividends, net rental income, royalty income, or income from estates and trusts.** *Report even small amounts credited to an account.*

Yes → $ .00 TOTAL AMOUNT for past 12 months Loss
No

d. **Social Security or Railroad Retirement.**

Yes → $ .00 TOTAL AMOUNT for past 12 months
No

e. **Supplemental Security Income (SSI).**

Yes → $ .00 TOTAL AMOUNT for past 12 months
No

f. **Any public assistance or welfare payments from the state or local welfare office.**

Yes → $ .00 TOTAL AMOUNT for past 12 months
No

g. **Retirement income, pensions, survivor or disability income.** *Include income from a previous employer or union, or any regular withdrawals or distributions from IRA, Roth IRA, 401(k), 403(b), or other accounts specifically designed for retirement. Do not include Social Security.*

Yes → $ .00 TOTAL AMOUNT for past 12 months
No

h. **Any other sources of income received regularly such as Veterans' (VA) payments, unemployment compensation, child support or alimony.** *Do NOT include lump sum payments such as money from an inheritance or the sale of a home.*

Yes → $ .00 TOTAL AMOUNT for past 12 months
No

44 What was this person's total income during the PAST 12 MONTHS? *Add entries in questions 43a to 43h; subtract any losses. If net income was a loss, enter the amount and mark (X) the "Loss" box next to the dollar amount.*

None OR $ .00 TOTAL AMOUNT for past 12 months Loss

FIGURE 5-1 American Community Survey income questions, 2022.
SOURCE: https://www2.census.gov/programs-surveys/acs/methodology/questionnaires/2022/quest22.pdf<census.gov/acs

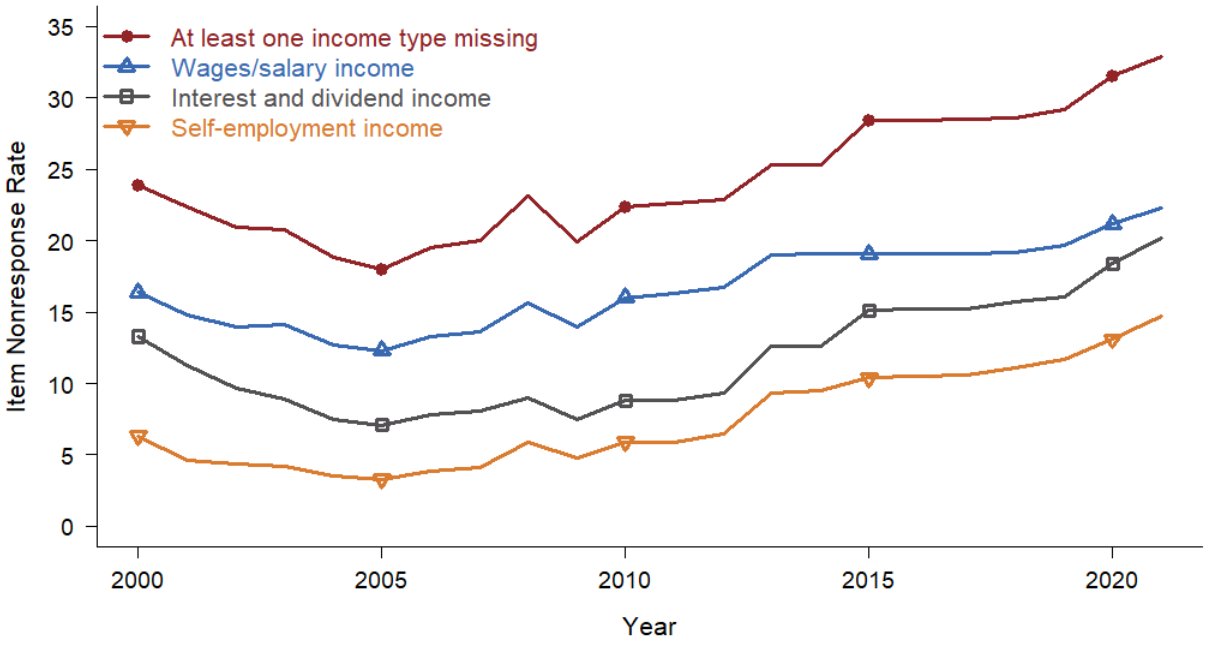


FIGURE 5-2 Item nonresponse for selected income types, American Community Survey, 2000–2021.

NOTE: Nonresponse rates before 2005 are from the experimental precursor surveys to the ACS. Income types not shown are Social Security or Railroad Retirement; Supplemental Security; public assistance; retirement; other. The patterns for these other income sources over time mirror those shown and fall between the lines for wages/salary and self-employment income.

SOURCE: Panel generated with data from https://www.census.gov/acs/www/methodology/sample-size-and-data-quality/item-allocation-rates

6. Data Linkage to Supplement Health Surveys

As with income, much work has been done on linking health survey data with administrative records. Many health data sources contain personally identifying information that permits record linkage; these include surveys about health, administrative data such as Medicare and Medicaid claims or databases of birth and death records, state data records, health claims submitted to private insurers, and electronic health records from government agencies (e.g., the Department of Veterans Affairs, the Indian Health Service, and municipal hospitals) and from private hospitals and doctors. These data have been used to study potential bias from missing data and to suggest improvements to measurement methods, as with the income studies discussed in Chapter 5.

This chapter focuses on the use of linked household survey and administrative data to enhance the study of health conditions and outcomes, as emphasized in the workshop session *Data Linkage for Income and Health Statistics*. For example, survey respondents might know they were hospitalized, but not the precise condition(s) treated, the results of all tests that were done, the actual medical procedures undertaken, or the total costs. By adding variables from administrative records on health claims or deaths to health survey data (which may provide information not available in administrative records such as demographic information, health attitudes and behaviors, and health conditions from self-reports or medical examinations), researchers can gain additional insights about health and diseases nationally and in population subgroups.

Sections 6.1 and 6.2 review key household surveys and administrative data sources used in linkage projects by the U.S. National Center for Health Statistics (NCHS). Section 6.3 highlights recent data-linkage activities by NCHS, and Section 6.4 discusses data-equity implications of the linkages. Section 6.5 examines challenges involved in linking data from longitudinal surveys, with a focus on data linkage with the Health and Retirement Study, a panel survey of Americans over the age of 50.

6.1 SURVEYS FROM THE U.S. NATIONAL CENTER FOR HEALTH STATISTICS

Many data linkages have involved two of the key household surveys administered by the NCHS: the National Health Interview Survey (NHIS) and the National Health and Nutrition Examination Survey (NHANES).[84]

National Health Interview Survey

The NHIS, the largest household survey conducted by NCHS, is a face-to-face, cross-sectional survey that monitors the health of the U.S. civilian noninstitutionalized population

[84]NCHS also conducts many other surveys and these are listed at https://www.cdc.gov/nchs/. Other federal agencies also conduct surveys about health topics. For example, the Current Population Survey regularly has a supplement on tobacco use, the U.S. Veterans Health Administration conducts surveys about veterans' health and use of health care, and the Substance Abuse and Mental Health Services Administration conducts the National Survey on Drug Use and Health.

through interviews with survey participants. The NHIS has been conducted continually since 1957, but the survey design and content have been updated periodically to take advantage of new developments in survey methodology, include new health topics, reduce respondent burden, and harmonize content with other health data sources. Some content is included every year, including demographic information, health insurance coverage, health care access and use, chronic conditions, health-related behaviors such as diet and physical activity, and functioning and disability. Other questions are asked on a rotating schedule.[85]

One "sample adult" aged 18 years or older and one "sample child" aged 17 years or younger (if applicable) are selected randomly from each respondent household. Sampled adults provide their own health information if able to do so (otherwise information is provided by a proxy); information about the sample child is collected from a parent or other knowledgeable adult. In 2021, there were 29,482 sample adult interviews and 8,261 sample child interviews (NCHS, 2022b, p. 10).

As with the Current Population Survey (see Chapters 2 and 5), the target population for the NHIS is the U.S. civilian noninstitutionalized population:

> The NHIS universe includes residents of households and noninstitutional group quarters (e.g., homeless shelters, rooming houses, and group homes). Persons residing temporarily in student dormitories or temporary housing are sampled within the households that they reside in permanently. Persons excluded from the universe are those with no fixed household address (e.g., homeless and/or transient persons not residing in shelters), active duty military personnel and civilians living on military bases, persons in long-term care institutions (e.g., nursing homes for the elderly, hospitals for the chronically ill or physically or intellectually disabled, and wards for abused or neglected children), persons in correctional facilities (e.g., prisons or jails, juvenile detention centers, and halfway houses), and U.S. nationals living in foreign countries (NCHS, 2022b, p. 11).

National Health and Nutrition Examination Survey

The NHANES began in the 1960s to assess the health and nutritional status of U.S. adults and children.[86] It provides information not available from other health surveys because it has both interview and examination components. The interview asks questions about demographic and socioeconomic characteristics as well as dietary and health-related questions. The examination component, conducted by trained medical personnel, includes laboratory tests and medical, dental, and physiological measurements. Because it measures aspects of health directly, data from the NHANES can be used to estimate the prevalence of major diseases and risk factors. NHANES findings are also the basis for national standards for such measurements as height, weight, and blood pressure.

Interviews are conducted in respondents' homes and medical examinations are performed in mobile examination centers that travel to the areas included in the sample. Because of the expense of conducting medical examinations of survey respondents, the sample size for the

[85]See https://www.cdc.gov/nchs/nhis/about_nhis.htm and NCHS (2020c) for overviews of the survey, and https://www.cdc.gov/nchs/nhis/2019_quest_redesign.htm for a description of content in any given year.

[86]See https://www.cdc.gov/nchs/nhanes/about_nhanes.htm and NCHS (2020b) for overviews of the survey.

NHANES is smaller than for the NHIS: about 5,000 adults and children each year. Data must typically be accumulated for multiyear periods to allow computation of estimates for population subgroups.

Like the NHIS, the NHANES is a sample of the civilian noninstitutionalized population. Persons experiencing homelessness, persons residing in institutions such as nursing homes and prisons, and persons in the military are excluded.

Strengths and Limitations of Health Survey Data

A survey is the only way to measure some health topics, and the NHIS and NHANES both ask a broad array of questions about health, nutrition, and physical activity that are unavailable from administrative records.[87] The NHANES examination component identifies health conditions that might be unknown to the survey participant—for example, some survey participants may be unaware that they have diabetes—and that information would not be found in any other data source.[88]

As with all household surveys, however, both the NHIS and NHANES have been subject to decreasing response rates, with accelerating declines since 2010. Figure 2-1 shows the response rates for the screener portion of the NHIS (which obtains the roster of household members for selecting the sample adult and child) and the interview portion of the NHANES. Additional nonresponse occurs because some sampled adults and children in the NHIS do not participate in interviews, some NHANES respondents do not participate in the medical examination, and participants may have missing data for survey items. In 2021, the NHIS response rates for the sample adult and sample child interview were each close to 50 percent (NCHS, 2022b, p. 10). Of the 27,066 persons sampled for the 2017–2020 NHANES, 51.0 percent were interviewed and 46.9 percent were examined.[89]

The low response rates in recent years raise concern about possible nonresponse bias that might remain in the survey data after weighting adjustments for nonresponse are performed. Administrative data sources can be used to investigate how well nonresponse adjustments remove bias (see Section 6.3), but administrative records datasets may also omit parts of the population.

6.2 SOURCES OF ADMINISTRATIVE DATA ON HEALTH

The NCHS has a robust program linking data from its surveys with administrative data, including the National Death Index (NDI), Social Security and Supplemental Security Income benefit records collected by the Social Security Administration (SSA), data on Medicare and Medicaid/State Children's Health Insurance Program from the Centers for Medicare and Medicaid Services, and administrative data for participants in the Department of Housing and Urban Development's (HUD) largest housing-assistance programs (the Housing Choice Voucher program, public housing, and privately owned subsidized multifamily housing).

---

[87]As discussed in Section 2.2, some information about these topics may be available from fitness-tracking devices, but data from these devices are typically available only through convenience samples.

[88]Survey participants receive a report on results of the thorough medical examination as one of the benefits of participation. A participant is notified of any urgent health problems immediately.

[89]https://wwwn.cdc.gov/nchs/data/nhanes3/ResponseRates/NHANES-2017-2020-Response%20Rates-2017-March2020-508.pdf. Data collection for this sample ended in March 2020 because of the COVID-19 pandemic.

The NDI contains records of nearly all deaths occurring since 1979 (see Section 2.2), and provides information that cannot be gathered from a household survey: date, location, causes, and circumstances of deaths.

Administrative records from the Centers for Medicare and Medicaid Services provide the opportunity to study changes in health status, health care utilization and costs, and prescription drug use among Medicare and Medicaid participants.[90] NCHS is provided with Medicare program enrollment and claims/encounters data for survey participants who are matched with Medicare administrative records. Using Medicare and Medicaid data together with survey data allows researchers to study some of the populations excluded from the NHIS and NHANES, such as people living in institutions. Health care needs and expenditures of nursing home residents differ from those of people of similar age who live in households or noninstitutional group quarters, and Medicare and Medicaid data, either alone or combined with other data sources, can provide information on the health trajectories of nursing home residents.

Medicare and Medicaid data are not available for everyone in the U.S. population, however, since both programs have eligibility requirements. Medicare federal health insurance is limited to people who are 65 or older, people under 65 with disabilities, and people with end-stage renal disease. Medicare eligibility and enrollment files, containing information on demographics, reason for Medicare eligibility, and type of Medicare enrollment (fee-for-service Original Medicare or Medicare Advantage), are available for everyone in the program. But Medicare claims data generally do not include information about beneficiaries enrolled in Medicare Advantage plans, which are operated by private companies that contract with Medicare (NCHS, 2016); in 2021, about 44 percent of beneficiaries were enrolled in such plans.[91]

Federal law specifies mandatory eligibility groups for state Medicaid programs, including low-income families and individuals receiving Supplemental Security Income; some states cover additional groups.[92] But Medicaid data do not have full coverage for studying health and expenditures of the low-income population because not all eligible people participate in the program. Certain population subgroups are particularly likely to be nonparticipants and thus be without health insurance. Using American Community Survey data, Lukens and Sharer (2021) estimated that Black and Hispanic adults accounted for nearly 60 percent of the 2019 "coverage gap"—adults with incomes below the poverty line but who do not have Medicaid or other insurance.[93]

Medicare and Medicaid files both lack information about people who do not participate in Medicare or Medicaid—including those with private, or no, health insurance. Moyer (2021) discussed NCHS initiatives for using private-sector data.

HUD data, too, cover only part of the population: people receiving housing assistance through HUD's three largest programs. HUD's administrative data, submitted by local public

[90] https://www.cdc.gov/nchs/data-linkage/CMS-Medicare-Restricted.htm and https://www.cdc.gov/nchs/data-linkage/medicaid.htm

[91] https://www.cms.gov/newsroom/news-alert/cms-releases-latest-enrollment-figures-medicare-medicaid-and-childrens-health-insurance-program-chip; https://data.cms.gov/collection/cms-program-statistics

[92] https://www.medicaid.gov/medicaid/eligibility/index.html describes Medicaid eligibility requirements.

[93] Being below the poverty line does not exactly coincide with Medicaid eligibility because eligibility criteria vary across states. Children from low-income families, however, are eligible for Medicaid in all states, and the percentage of eligible children enrolled in Medicaid across states ranged from 81–98 percent in 2018 (Schor and Johnson, 2021). Keisler-Starkey and Bunch (2022, Figure 4) estimated from CPS ASEC data that in 2021, 5 percent of all children under age 19 had no health insurance coverage, but the uninsurance rate was 8.6 percent for Hispanic children and 18.6 percent and 22.6 percent for foreign-born and noncitizen children, respectively.

housing authorities and contracted private owners or managers of apartment buildings, contain housing, income, and program information for participants.

6.3 DATA LINKAGE AT THE U.S. NATIONAL CENTER FOR HEALTH STATISTICS

The NCHS data linkage program "aims to maximize the scientific value of the Center's population-based surveys, by linking NCHS survey data with data collected from vital and other administrative records. Linked data files enable researchers to augment information for major diseases, risk factors, and health service utilization, by linking exposures to outcomes and in some cases introducing a longitudinal component to survey data" (NCHS, 2022b, p. 118).[94] Golden and Mirel (2021) and Mirel (2022) gave overviews of the program.[95]

In addition to linking individual records, NCHS also performs linkages at the area level. Addresses are geocoded to standard census geography units, which allows researchers to merge area-level statistics such as county poverty rate or air quality with the survey data.

The linked data have been used for two main purposes. First, as with the income studies discussed in Section 5.4, linked data have been used to study accuracy of items in survey and administrative datasets. Linked data have also been used to study questions about health and to provide information that can be used to promote evidence-based health policy.

Linkages to Examine Accuracy of Health Data

As with linked income data, researchers have used linked health data to study the concordance between survey reports and information in administrative records, or to assess effects of survey nonresponse. For example, Keyes et al. (2018) studied potential nonresponse bias in the NHIS by comparing age-adjusted mortality rates estimated from survey respondents (with mortality status determined by linkages with the NDI) with population mortality rates from the National Vital Statistics System. Other researchers have examined the concordance between survey and administrative data on topics including Medicare enrollment (Gindi and Cohen, 2012), Medicaid enrollment (Mirel et al., 2014), receipt of rental assistance or Social Security disability benefits (Boudreaux, Fenelon, and Slopen, 2018; Mirel et al., 2019b), and reports of childhood asthma (Zablotsky and Black, 2019).

For example, Day and Parker (2013) compared self-reported diabetes in the 2005 NHIS with information about diabetes in linked Medicare claims files, using a procedure typical of concordance studies. They linked NHIS participants aged 65 and over with their Medicare records, finding that 93 percent of survey respondents who reported they had diabetes had a diabetes indicator in the Medicare files, but only 67 percent of those with a diabetes indicator in the Medicare files self-reported the condition on the NHIS. Day and Parker (2013) suggested that the discrepancy may have occurred because respondents misunderstood the survey questions or their doctors' diagnoses.

---

[94] An inventory of NCHS survey data already linked with administrative records can be found at https://www.cdc.gov/nchs/data/datalinkage/LinkageTable.pdf. Linked data from NCHS can be accessed for approved research projects at the NCHS Research Data Center or through the Federal Statistics Research Data Centers.

[95] See also https://www.cdc.gov/nchs/data-linkage/index.htm for a general description. NCHS (2022d, 2021c) described the specific procedures used to link NCHS survey data to the NDI and Medicare/Medicaid records.

Linkages to Study Health Outcomes and Associations

Linking health survey data with administrative data can provide information on health outcomes and associations with other participant characteristics that can inform medical practice and health policy. Mirel (2022) listed areas in which linked data have been used in evidence-based policymaking: to study health insurance coverage and costs, to evaluate policies such as smoking-cessation programs, and to generate evidence that can be used to improve public health. She mentioned the following examples of studies that used linked NCHS data:

1. Excess deaths associated with underweight, overweight, and obesity (NHANES-NDI linked data; Flegal et al., 2005);
2. Air pollution exposure and heart disease mortality (NHIS-NDI; Parker, Kravets, and Vaidyanathan, 2018);
3. Differences in adult mortality by education level (NHANES-NDI; Rogers, Hummer, and Everett, 2013);
4. Comparing health characteristics of people who chose Medicare Advantage with those who chose Original (fee-for-service) Medicare (NHANES-Medicare enrollment; Mirel et al., 2012);
5. Use of health services among Medicare enrollees who were previously uninsured (NHIS-Medicare enrollment and claims; Decker et al., 2012);
6. Medical costs of chronic kidney disease in the Medicare population (NHANES-Medicare claims; Honeycutt et al., 2013);
7. Housing assistance and children's blood lead levels (NHANES-HUD; Ahrens et al., 2016; see Section 1.1);
8. Cigarette smoking and adverse health outcomes among adults receiving federal housing assistance (NHIS-HUD; Helms, King, and Ashley, 2017); and
9. Association between housing assistance, health insurance coverage, and unmet medical needs (NHIS-HUD; Simon et al., 2017).

The research in these studies could not have been done with the survey data alone or with the administrative data alone. In the first three studies, linkages between NCHS survey data and the NDI allowed researchers to examine the association between personal characteristics and risk factors (measured in the surveys) and mortality. Parker, Kravets, and Vaidyanathan (2018) also used the geocoding of NHIS data to link each survey participant with an annual estimate of fine particulate matter for the participant's census tract. The researchers were thus able to control for risk factors such as body mass index and smoking status (from NHIS) when examining the association between air pollution and heart disease mortality.

In studies 4 through 6, information about the type of Medicare plan, health care usage, and medical costs came from the Medicare data. The health surveys also do not ask about housing assistance; that information, for studies 7–9, came from the linked HUD data.

In concordance studies, comparisons are done using the set of records that can be linked, and conclusions typically apply only to those data. For studying health outcomes, however, it is desired to make inferences to the U.S. population or specific subpopulations. The NHIS and NHANES are designed to be representative of the U.S. civilian noninstitutional population at the time of the survey, but the set of records that can be linked is not necessarily a random subsample of respondents (Golden et al., 2015, p. 38). In addition, some administrative records

datasets include only part of the population of interest (for example, Medicare data do not have claims information on Medicare Advantage participants). The next section describes approaches for addressing potential differences between records that can, and cannot, be linked.

6.4 LINKAGE AND DATA EQUITY

This section looks at data-equity issues for linked datasets and possible steps for investigating and documenting them. The issues are described in the context of the NCHS linkages described in Section 6.3 but apply to other data-linkage programs as well.

Linkage Eligibility

Linkage with NHIS or NHANES records is performed only for "linkage-eligible" participants—those who have provided consent and have sufficient personally identifiable information to enable successful linkage. For NHIS, "[s]urvey participants are informed of NCHS' intent to conduct data linkage activities through a variety of procedures such as 'advance letters,' participant brochures, and during the interview when verbal consent is requested" (NCHS, 2022b, p. 118). Participants are asked to supply the last four digits of their Social Security Numbers (or, if unwilling to provide that, asked if they consent to linkage that uses other identifying information). Children are linkage eligible if consent is provided by their parent or guardian and they have enough identifying information to enable linkage, but that consent applies only to administrative data about events occurring before the child reaches the legal adult age of 18.

One approach for analyzing linked datasets is to treat ineligibility for linkage as an additional stage of nonresponse, and to perform weight adjustments similar to those used to adjust for nonresponse. NCHS (2022d) described the procedure used to produce survey weights for analyzing linked NHIS-NDI data, which involved adjusting the survey weights for linkage-eligible respondents so that they sum to known population counts for sex, age, race, and ethnicity subgroups. This procedure produces estimates similar to those that would be obtained from all NHIS respondents if, within each demographic subgroup, health characteristics of linkage-eligible persons are similar to those of non-linkage-eligible persons.

Many surveys conduct nonresponse bias analyses, and similar analyses can be carried out to investigate possible bias from differences in linkage eligibility across subpopulations. For example, Aram et al. (2021) found that about 88 percent of sample adults in the 2010–2013 NHIS were linkage-eligible regardless of age group, sex, and education. Linkage eligibility was slightly higher (about 90%) for adults with diabetes or obesity, and slightly lower for Hispanic and non-Hispanic Asian adults (85.5% and 85.6%, respectively).

Aram et al. (2021) also investigated possible linkage bias by comparing estimates of demographic and health characteristics (diabetes, hypertension, obesity, fair or poor self-rated health, having a doctor's office visit in the past year, and smoking) for the full NHIS sample (using the nonresponse-adjusted weights) with estimates calculated from the set of linkage-eligible records (using linkage-eligibility-adjusted weights). They found that, while there were large differences for some of these characteristics before 2007, estimates were similar for the 2010–2013 NHIS, indicating that restricting to linkage-eligible records did not increase bias for these characteristics for those years. Lloyd et al. (2017) investigated potential bias in linked NHIS-HUD and NHANES-HUD data by comparing estimates of housing characteristics and

demographic information computed from HUD administrative files with estimates calculated from the set of linked records.

Linkage Errors

Probabilistic linkage procedures compute a "match score" for pairs of records (see Box 2-1). Pairs with high match scores are thought likely to belong to the same person, and pairs with low scores are likely to belong to different persons. Record pairs with scores in the middle might or might not be a true match. There is evidence, however, that linkage uncertainties and errors affect some population groups more than others (see Chapters 2 and 3). Miller, McCarty, and Parker (2017, p. 83) wrote: "With data coming from multiple sources, there will be differences in availability, quality, and format of unique identifiers, which could disproportionately affect minority populations."

Lariscy (2017) studied data-equity issues related to linkage uncertainty by examining the distribution of match scores for Black and White men and women in the NCHS linkage of data from the 1986–2009 NHIS with the NDI (see NCHS, 2009 for the linkage procedures used for these data and NCHS, 2022d for current linkage procedures). Lariscy (2017) found that linkage quality was lower for Black adults than for White adults. Among the persons whom NCHS had determined to be deceased, 51 percent of Black women and 54 percent of Black men were in Class 1 (considered to have a high likelihood of being a true match), compared with 59 percent of White women and 66 percent of White men. Black decedents had lower mean scores than White decedents, indicating less certainty about the matches. Similarly, a higher percentage of White men and women who were deemed to be still living were placed in Class 5 (considered to have a high likelihood that there is no match in the NDI) than were Black men and women. In a similar study, Lariscy (2011) found more linkage uncertainty for Hispanic adults (and especially for foreign-born Hispanic adults) than for U.S.-born non-Hispanic White adults under the linkage consent rules and procedures used at that time.

Mortality rates estimated from linked data may be less accurate for population subgroups with more uncertainty about linkages. For example, Black et al. (2017) found that even small numbers of missed links between a survey and the NDI can result in large underestimation of mortality rates for older age groups, a phenomenon they dubbed the "Methuselah effect."[96]

Investigating and Documenting Properties of Linked Survey Data

NCHS has performed multiple investigations of the quality of linked datasets, and a perusal of their work suggests some "best practices" for investigating and documenting the quality of linked data.[97]

---

[96]The effect occurs because a survey respondent who died at age *a* but is not matched to the NDI inflates the denominator of the estimated mortality rate (the estimated number of persons still alive) for all ages greater than *a.* For each successive age group, as the number of "real" survivors in the denominator decreases, the number of "nonreal" survivors in the denominator increases (because of the cumulative missed links of all persons younger than that age group), resulting in a higher proportion of "nonreal" survivors in the denominator and a too-large estimate of the percentage of persons who live to an advanced age. See Arias (2021) for a discussion of how data quality affects comparisons of longevity across race and ethnicity groups.

[97]See also the guidance presented by Bohensky et al. (2011); Davern, Roemer, and Thomas (2014); and Gilbert et al. (2018).

- Identify the exact datasets that were linked, with an assessment of coverage, missing data, and measurement methods. Describe how the data were collected, maintained, cleaned, and processed for each source. Provide references to the full documentation of the individual data sources, including nonresponse bias analyses of the surveys being linked (or supply such documentation if it does not exist).
- Provide full documentation of the linkage method used, including descriptions of the data elements used for linkage, the accuracy of those elements for each data source, and the algorithm followed. Also provide documentation of weighting adjustments or other methods used for estimating population characteristics from the linked data.
- Report rates for linkage consent and eligibility, with disaggregated statistics by age, sex, race, ethnicity, and other subgroups. If probabilistic linkage is used, provide information about the distribution of match scores for population subgroups.
- Provide disaggregated estimates of linkage error rates, with a description of how these were estimated. How many missed links and false links were found in validation studies?
- Analyze additional bias that may occur when restricting analyses to the set of linkage-eligible individuals or linked records. As part of this analysis, compare estimates computed from the linkage-eligible respondents with estimates from the full set of survey respondents. If the administrative records form the population of interest, compare characteristics calculated from the set of linked records with characteristics calculated from the full set of administrative records.
- Investigate discrepancies in measurements between the survey and the administrative dataset, for example, differences in self-reports of disease and reports in claims data.
- Describe how linkage errors and uncertainties about linkage might affect analyses performed on the linked data. For some linkage methods, uncertainties about linkage can be a component of measures of uncertainty for statistics produced from the linked data.

Each step involves consideration of data-equity aspects. Discussions—within the agency and with data users and community members—of how a proposed linkage project might affect population subgroups can promote transparency and raise awareness of community concerns. What are the potential benefits and harms of the linkage, and should the effort even be undertaken? How does linkage quality vary by age, sex, race, ethnicity, disability, and other characteristics? What are the implications of those disparities for research performed on the data? Future reports in this series will address privacy and confidentiality concerns for data linkage.

CONCLUSION 6-1: The U.S. National Center for Health Statistics has linked many of its surveys with administrative records datasets, providing valuable resources for investigating long-term health outcomes and promoting evidence-based policy. These linkage procedures and documentation can serve as models for other partnerships between program-oriented and federal statistical agencies.

6.5 LINKAGE OF LONGITUDINAL HEALTH SURVEYS

Longitudinal datasets allow researchers to investigate the dynamics of human behavior, such as how participation in government transfer programs might relate to subsequent labor force behavior or utilization of health care services. Understanding these interactions in a dynamic environment is helped by linkage with administrative datasets. Chapter 4 described two longitudinal datasets formed by linking administrative records: the Longitudinal Business Database and the Longitudinal Employer-Household Dynamics database. Data from the longitudinal Survey of Income and Program Participation (see Chapter 5) have been used to study the dynamics of poverty over time.

Linkage of longitudinal surveys presents challenges additional to those for linking cross-sectional surveys (Calderwood and Lessof, 2009). As discussed in Section 4.1, population coverage of administrative datasets may change over time (for example, Medicaid coverage expanded after passage of the Affordable Care Act in 2010) or data-access rules may change. Characteristics measured in administrative data and definitions of those characteristics may also change over time, and variables used to link records may be missing or may use different categories across administrative datasets. Attrition in a longitudinal survey, when combined with missing administrative data and missed links, can cause the set of survey respondents having data across all time periods and for all variables of interest to be small. Issues of consent for longitudinal data linkage are also more complex (Jäckle et al., 2021a).

This section illustrates data linkages with the Health and Retirement Study (HRS), a nonfederal longitudinal survey (Faul and Levy, 2022).[98] The HRS started in 1992, with a nationally representative sample of about 12,000 people who were between the ages of 51 and 61 at the time of the initial face-to-face interview. Additional cohorts of persons over the age of 50 have been added every 6 years so that there are approximately 20,000 respondents at any point in time; more than 40,000 respondents have participated altogether. Both members of a couple are included in the sample for all cohorts, and participants are interviewed every 2 years.

The repeated interviews allow researchers to study changes in health and economic circumstances that are associated with aging. The study collects detailed information about demographic characteristics, cognition, health status and functional limitations, use of health care services, work history and employment, retirement plans, net worth, income, health and life insurance, family structure, and subjective well-being.

Each new cohort is selected through a probability sample of households. As participants age, however, some of them may move into nursing homes, and these respondents are retained and followed in the sample. Thus, although the HRS does not sample from nursing homes at the time of recruitment, the sample contains members of the U.S. nursing home population, and weights are constructed to allow researchers to study that population (Sonnega et al., 2014; Lee et al., 2021).

An important data-equity issue for the HRS is inclusion of people with cognitive impairments. Excluding people who are physically or mentally unable to answer survey

[98]See Sonnega (2017) and Fisher and Ryan (2018) for overviews of the HRS. Sonnega et al. (2014) described the sample design and weighting. The HRS is conducted by the University of Michigan Institute for Social Research as a cooperative agreement with the U.S. National Institute on Aging, with additional funding from the SSA. The National Institute on Aging also sponsors the Longitudinal Studies of Aging Network at the University of Michigan (https://micda.isr.umich.edu/networks/longitudinal-studies-of-aging/) to promote research related to data-collection procedures and measurement issues, and has been working to expand and facilitate linkages between aging studies that it funds and administrative records (Rose Li and Associates, 2016, 2019).

questions would create bias and result in underestimates of the prevalence of conditions such as dementia. The HRS asks a proxy respondent (usually a family member) to provide information about a participant who cannot or is unwilling to answer questions after the baseline interview, or when an interview started but the interviewer has concerns about the participant's ability to provide accurate information. About 9 percent of interviews overall, and 18 percent of those for persons aged 80 or older, are with proxy respondents (Sonnega et al., 2014).

HRS data are linked to sources of administrative information at the individual level. Respondents must consent to having their data linked. Faul and Levy (2022) reported that from 1996–2018, consent for linkage to Medicare records was obtained after three attempts for about 85–90 percent of respondents. Linkage to SSA records provides earnings histories, benefit histories, and application histories for disability and Supplemental Security Income of HRS participants.

One of the main goals of the HRS is to understand the relationship between medical history and financial status and how health care usage changes as people age. For respondents who consent to linkage, information about diagnoses and costs of treatment has been obtained from Medicare and Medicaid records. For HRS participants who served in the military, medical records have been obtained from the Department of Veterans Affairs. Linkage to the NDI tracks mortality. Information on employer-provided pension plans is obtained from businesses at which respondents are or have been employed.[99]

Researchers can access HRS data linked with other sources in a protected research environment. Faul and Levy (2022) mentioned the following recent studies that used linked data:[100]

- Studying potential bias from dropouts and proxy reporters in the HRS, using NDI data to identify respondents who died and Medicare claims data to identify the earliest reported diagnostic code for dementia (Weir, Faul, and Langa, 2011).
- Monetary cost of dementia, using self-reports on the HRS to estimate out-of-pocket spending and nursing home costs, and linked Medicare claims data to identify costs paid by Medicare (Hurd et al., 2013).
- Long-term consequences of sepsis for cognition and physical function, obtaining characteristics of hospitalizations for severe sepsis from Medicare claims data (Iwashyna et al., 2010).
- Knowledge about Social Security and pensions, comparing self-reported expected Social Security and pension income with benefit entitlements calculated from SSA earnings histories and employer pension plan descriptions (Gustman and Steinmeier, 2005).
- Impact of employer match on retirement contributions, linking with SSA data to obtain earnings histories (Engelhardt and Kumar, 2007).
- Delayed diagnoses of dementia for Black and Hispanic older adults, using HRS data on cognitive and daily function and linked Medicare/Medicaid claims data to identify the time of dementia diagnosis (Lin et al., 2021).

[99]See https://hrs.isr.umich.edu/data-products/restricted-data/available-products for a list of datasets linked to the HRS. In addition, the Census-Enhanced HRS project is linking HRS data to U.S. Census Bureau data on characteristics of respondents' employers (https://cenhrs.isr.umich.edu/).

[100]A bibliography of studies that have used the HRS is at https://hrs.isr.umich.edu/publications/biblio/. Fisher and Ryan (2018) gave an extensive description of the research areas involving the HRS.

CONCLUSION 6-2: Longitudinal surveys provide perspectives on individual and household behavior not available in cross-sectional surveys. Data from such longitudinal surveys can be enhanced through data linkages to create new opportunities for social science research.

7. Combining Multiple Data Sources to Measure Crime

The United States has two major collections of national statistics about crime. The first is the Uniform Crime Reporting (UCR) Program administered by the Federal Bureau of Investigation (FBI), which compiles data from law enforcement agencies. The second is the National Crime Victimization Survey (NCVS), an annual sample survey administered by the Bureau of Justice Statistics (BJS) that asks persons aged 12 and older in a randomly selected set of households about their experiences with crime. James and Council (2008); two National Academies of Sciences, Engineering, and Medicine reports (NASEM, 2016a, 2018); Lohr (2019); and Morgan and Thompson (2022) provided overviews of these two data collections.

Table 7-1 displays the types of crime included in the UCR and the NCVS. Because UCR statistics are compiled from law enforcement agency submissions, they include only crimes that are reported to the police and thus undercount the total numbers of crimes against residents. The NCVS asks persons about their victimization experiences, and thus has information about criminal victimizations that are not reported to the police as well as those that are reported. Because the NCVS is a household survey, though, it does not measure crimes against businesses and organizations (which are measured in the UCR if known to the police); it also does not measure crimes against persons living in institutions (such as nursing homes or prisons), persons experiencing homelessness, and children under age 12. Furthermore, NCVS respondents may forget or fail to mention some of the victimizations they experienced. And some crimes, such as corporate or environmental crime, are not measured in either data source.

[TABLE 7-1 about here]

There is therefore great potential for using multiple data sources to enhance statistics about crime. The NCVS, as a household probability survey, can be blended with other sources using methods such as data linkage and small area estimation. While some challenges are similar to challenges in the areas of income and health statistics (for example, coverage of only the noninstitutionalized population), others are unique to the NCVS because of the relative rarity of crime and the sensitive nature of the information collected.

The UCR Program presents a distinct set of data-combination challenges. It is, in essence, a cooperation between states and the federal government of the same type as the National Vital Statistics System (see Chapter 4). Individual law enforcement agencies submit data on crimes within their jurisdictions to state UCR programs, which, after data processing, forward them to the FBI.[101] The UCR is intended to be a census of incidents known to the more than 18,000 law enforcement agencies in the United States. It thus has the potential to produce detailed information about crime for small geographic and demographic subpopulations, but challenges include missing data (some agencies do not submit data or submit data for only part of the year), ensuring the quality of the information collected and reported, and aligning the information with data from other sources.

The National Academies provided a comprehensive review of and a vision for the future of crime statistics, with an emphasis on crime classification and measurement (NASEM, 2016a, 2018). Box 7-1 reproduces some conclusions and recommendations from those reports. This

[101]Some law enforcement agencies submit their data directly to the FBI instead of through state programs.

chapter examines developments that have occurred since those National Academies reports and, in particular, the potential of combining data sources for measuring crime, as discussed in the workshop session *Measuring Crime in the 21st Century*.

Sections 7.1 and 7.2 describe the UCR Program and the NCVS, respectively, and identify challenges that might be addressed through use of multiple data sources. Section 7.3 outlines other national data collections about crime, and Section 7.4 explores the potential for obtaining more timely crime statistics directly from police department databases and websites. Sections 7.5 and 7.6 describe some initiatives for combining data sources to study crime, and Section 7.7 discusses possible future directions for using multiple sources to improve the quality and equity of data about crime.

[BOX 7-1 about here]

7.1 THE UNIFORM CRIME REPORTING PROGRAM

The UCR Program combines data voluntarily submitted by states and law enforcement agencies and has thus relied on multiple data sources since its inception in 1930. From 1930–2020, UCR statistics were based on data collected in Summary Reporting System (SRS) format. Law enforcement agencies reporting to the SRS provided monthly counts of "Part I offenses" (homicide, rape, robbery, aggravated assault, burglary, larceny/theft, and motor vehicle theft) occurring in their jurisdictions and the number cleared by arrest.[102] In the 1960s, the FBI expanded the UCR data collection to encompass more detailed information about homicides, including age, sex, and race of the victim and offenders, circumstances of the crime, weapons used, and relationship between victim and offender. The Supplementary Homicide Report data were incident-based—instead of aggregate counts, details were collected separately for each homicide incident. For other crimes, however, the SRS collected only summary statistics.

In 2016, the FBI announced that the SRS would be retired on January 1, 2021, and that all UCR data submissions from 2021 onward would be through the National Incident-Based Reporting System (NIBRS). NIBRS began in the late 1980s with the goal of obtaining more detailed information about crime, and it extends incident-based data collection to all of the 52 types of offenses measured. Through 2020, the FBI encouraged NIBRS submission but allowed law enforcement agencies and state UCR programs to report UCR data in either SRS or NIBRS format; NIBRS data were converted to SRS format to compute national statistics for the amount of Part I crime. Beginning in 2021, SRS data were no longer accepted.

Table 7-2 outlines the differences between the SRS and NIBRS data collections and describes the types of detailed information measured in NIBRS. The additional variables measured in NIBRS allow its crime information for demographic groups to be combined or contrasted with information from other sources, enabling "the analysis of data in proper geographic, demographic, sociological, and economic context" as called for in Conclusion 2.1 of a National Academies 2018 report on *Modernizing Crime Statistics* (NASEM, 2018; see Box 7-

[102]See FBI (2013, p. 110) for the "Return A" form used to collect data for the SRS. In addition to crime counts, the form also collected monthly breakdowns of these statistics by characteristics such as weapons used. The original seven Part I offenses were defined by the Uniform Crime Reporting Committee, which established the form of the UCR (International Association of Chiefs of Police, 1929). Arson was added as a Part I offense in 1979, and two human trafficking offenses were added in 2013. Volumes of *Crime in the United States* through 2020, however, reported statistics only for the seven original Part I crimes (the only major modification of the original definitions occurred when the definition of rape was revised in 2013). For Part II offenses such as simple (non-aggravated) assaults, fraud, vandalism, and drug abuse violations, only arrest data were collected.

1). Jarvis (2015) and Hanson (2021) described other advantages of NIBRS data relative to SRS data.

[TABLE 7-2 about here.]

Lauritsen (2022a), Smith (2022), Martinez (2022), and Veitenheimer (2022) emphasized the advance represented by a national dataset of police-reported crime containing information beyond mere crime counts. The additional details about incidents allow tabulations by victim and offender demographics, relationship between victim and offender, and other characteristics described in Table 7-2. Veitenheimer (2022) commented that the transition to NIBRS allows more focus "on the contextual information and the characteristics of certain crimes like homicide or robbery or burglary or fraud or drug cases or sexual assaults, to look at things like victim and offender demographics, who's committing crimes against who, what are the circumstances behind some of the aggravated assaults that have occurred, what's the makeup of drug seizures that have happened or are happening for drug crime, what sorts of weapons or how often have weapons been involved in the commission of crimes."

As an example of the potential for exploring contextual information and characteristics of crime, Smith et al. (2018) highlighted how NIBRS data could be used to better understand sexual violence. Martin (2021) created an interactive report on sexual assault statistics for 15 states (those certified to report all of their 2019 crime data in NIBRS format): users can click on a state to view statistics about the percentage of violent victimizations that involved a sexual assault; incident characteristics such victim-offender relationships, demographics of victims and offenders, and weapons used; and rates of sexual assault by location type, time of day, and victim age, race, ethnicity, and sex. None of these statistics could have been calculated from the SRS data format used through 2020.

The NIBRS data present a tremendous opportunity for enhancing understanding about crime. But they also present new challenges for calculating and interpreting estimates of crime numbers and rates. The panel identified four main challenges:

1. *Missing Data.* In 2020, the last year in which SRS-format data were accepted, UCR estimates of crime were based on data contributed by 15,875 of the 18,623 law enforcement agencies in the country (85%). The agencies submitting at least three months of data served areas representing about 97 percent of the U.S. population—close to full coverage.[103]

   For the 2021 UCR estimates, the first to be computed entirely from NIBRS data, agency participation and population coverage were much lower. The 2021 UCR estimates were based on data submitted by 11,333 of 18,806 law enforcement agencies (60%), serving areas that represent about 65 percent of the U.S. population (FBI, 2022b, p. 19; Berzofsky et al., 2022, p. 4). In some states, all law enforcement agencies submitted 2021 NIBRS data; in other states, including California, Pennsylvania, and Florida, fewer than 3 percent of agencies submitted NIBRS data.

[103]Source: FBI Crime Data Explorer, https://cde.ucr.cjis.gov and Barnett-Ryan and Berzofsky (2022). Only agencies with at least three months of submitted data were used for estimating crime statistics. Note that statistics in the Crime Data Explorer are revised as new data come in and may differ slightly from statistics in this report, which were retrieved between April and October, 2022. The coverage statistics for the SRS include agencies that reported data for only part of the year; their data for missing months were imputed. NASEM (2016a, p. 47) commented that because of these partial reporters, the actual coverage of the SRS has been lower than claimed.

Only 62 of the 87 agencies serving populations of 250,000 or more participated in the 2021 NIBRS; nonparticipating agencies included the New York City and Los Angeles Police Departments (FBI, 2022a; BJS, 2022b).

Thus, unlike the SRS data used for the UCR in 2020 and previous years, NIBRS has large amounts of missing data. The agencies submitting NIBRS data in 2021 were essentially a convenience sample from the population of law enforcement agencies.[104] The FBI estimated national crime statistics for 2021 using data from the participating agencies (FBI, 2022b). Barnett-Ryan and Berzofsky (2022) and Berzofsky et al. (2022) gave nontechnical summaries of the estimation procedures used for 2021, which attempt to compensate for nonparticipating agencies' missing data through statistical modeling, weighting, and imputation.

2. *Uncertainty Estimates for National and State Crime Statistics.* One big change for 2021 UCR statistics is the addition of confidence intervals for the estimates. Through 2020, the FBI reported counts alone for the UCR, with no standard errors or other measures of uncertainty; Berzofsky et al. (2022, p. 6) stated that confidence intervals for SRS data were unnecessary because of the high coverage rate of the SRS.[105] For 2020, for example, the FBI reported 21,570 homicides and 921,505 aggravated assaults with no measures of uncertainty.[106] In 2021, however, because of the high number of nonreporting agencies for NIBRS, national estimates were accompanied by confidence intervals and state-level estimates were produced only for states with high NIBRS participation. The FBI estimated that there were 22,900 homicides in 2021, with 95 percent confidence interval [21,300, 24,600] (FBI, 2022b, p. 3). Because the agencies participating in NIBRS were not from a probability sample, the validity of these estimates and confidence intervals relies on how well the statistical model accounts for missing data. The panel could not evaluate the quality of the 2021 UCR estimates or the NIBRS estimation procedures because technical documentation, with details of the modeling process, had not yet been published as of October, 2022. Piquero et al. (2022) stated that technical documentation will be released at a later date.

[104]Note that, in 2013, the FBI and BJS attempted to obtain a representative sample of agencies submitting data to NIBRS (U.S. Bureau of Justice Statistics, 2021a). They selected a probability sample of 400 law enforcement agencies from the set of agencies that had submitted SRS data in 2011. The goal was to expedite the sampled agencies' transition to NIBRS; if all 400 sampled agencies submitted NIBRS data, then the FBI would be able to calculate unbiased estimates of crime: the agencies that were already submitting NIBRS data in 2011 would represent themselves, and the probability sample of 400 agencies would represent the rest of the population. As of August, 2021, however, only 210 of the 400 agencies in the probability sample were certified for the NIBRS program (https://bjs.ojp.gov/sites/g/files/xyckuh236/files/media/document/NCS-X_Sample_Agencies.pdf). Because of the high and nonrandom nonresponse, the 210 agencies do not form a representative sample of the agencies that were reporting data in SRS format in 2011.

[105]With almost all of the population served by an agency that reported SRS data for at least part of the year, the missing data would have little effect on estimates. If confidence intervals had been produced for SRS data, accounting for the imputation used for agencies reporting fewer than 12 months of data, they would likely have been narrow.

[106]Table 1, *Crime in the United States, 2020,* https://cde.ucr.cjis.gov (data reported on Sept. 30, 2021).

3. *Estimating Year-to-Year Changes in Crime.* The switch from SRS to NIBRS created a discontinuity in the time series for crime statistics from the UCR. The method used to estimate year-to-year changes in crime from 1930–2020, comparing annual crime totals from the SRS data, could not be used to estimate the change in crime from 2020 to 2021 because the SRS system was retired (Rosenfeld, 2022).
   Changes in crime from 2020 to 2021 were estimated by applying the NIBRS estimation method to the agencies that submitted NIBRS data in 2020 (FBI, 2022b). But 2020 had more missing NIBRS data than 2021 and, consequently, estimates from both years have low precision.[107] The FBI (2022b, p. 1) stated that most changes in crime between 2020 and 2021 were not statistically significantly different from zero, but "that the main contributor to that finding is the large amount of variation—both random and systematic—that is measured in the 2020 data due to low coverage of participating agencies." The estimates need higher precision to detect changes in crime.

4. *Evaluating Measurement Quality.* With the increased amount of information collected in NIBRS comes increased potential for missing data and measurement error on each item collected. Lynch (2018, p. 449) commented on the contrast in studies of data quality for survey and administrative data: "This contrast is apparent in the discussion of crime trends in the NCVS and the UCR where the survey had extensive discussions of sampling and measurement error, but almost nothing was said about the UCR except perhaps for missing data." Characteristics such as category of offense, race, ethnicity, relationship, circumstances, and possible bias motivation for the crime are provided by the law enforcement agency (in contrast to surveys, in which data elements are self-reported), and more research is needed on the uniformity and accuracy of these measures.

The panel anticipates that some of these challenges will be resolved as more law enforcement agencies submit NIBRS data. The 2021 crime estimates for states with high NIBRS coverage, such as South Dakota and Oregon, have narrow confidence intervals because little estimation is needed—almost all law enforcement agencies in those states submitted data. It will be important to continue monitoring NIBRS coverage in the coming years.

In addition, alternative data sources could be used to address some of the challenges in producing accurate crime estimates. Possible paths include using data obtained directly from police departments to impute crime statistics for nonparticipating law enforcement agencies or to study measurement quality (see Sections 7.4 and 7.6).

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.

[107] In 2020, only 9,993 law enforcement agencies, covering 53 percent of the population, submitted data in NIBRS format (https://cde.ucr.cjis.gov). The FBI (2022b, p. 3) estimated from the 2020 NIBRS data that there were 22,000 homicides in 2020, with 95 percent confidence interval [21,000, 23,000].

7.2 NATIONAL CRIME VICTIMIZATION SURVEY

The NCVS began in 1972 as an effort to measure crimes that were not reported to the police, and to learn about the details of crimes from victims' perspectives.[108] When the NCVS began, the UCR was collecting only counts of offenses known to the police, without details on characteristics of victims and offenders (except for homicide). The NCVS was designed to meet four primary objectives:

1. To develop detailed information about the victims and consequences of crime;
2. To estimate the numbers and types of crimes not reported to the police;
3. To provide uniform measures of selected types of crimes; and
4. To permit comparisons over time and types of areas (BJS, 2021b, p. 4).

The NCVS was launched in part to provide an independent measure of the crime statistics in the UCR, and the set of crimes measured by NCVS parallels, but does not exactly coincide with, those collected in the SRS of the UCR (NASEM, 2016a, p. 51). However, NCVS definitions of crimes differ from the definitions in both the SRS and NIBRS, and measurement of characteristics of victims, offenders, and incidents also differ in the sources (see Table 7-1).

The NCVS asks respondents about all victimization incidents in the measured categories that they have experienced, whether reported to the police or not. Follow-up questions ask details about victimizations such as relationship to offender, location of the incident, injuries, financial losses, and whether the victimization was reported to police. From the outset, NCVS data have shown that the UCR Program fails to capture substantial amounts of crime—in 2019 and 2020, about 40 percent of violent crimes and one-third of property crimes were reported to the police (Morgan and Thompson, 2021).[109]

In recent years, annual estimates from the NCVS have been based on about 240,000 interviews of persons aged 12 and older from a probability sample of households.[110] From its inception through the early 1990s, the NCVS regularly achieved household response rates exceeding 95 percent. In the first two decades of the 21st century, however, response rates have dropped (see Figure 2-1). In 2020, about two-thirds of households eligible for the survey completed interviews. There was additional nonresponse because only 83 percent of the persons within responding households agreed to participate in an interview, giving an overall person-level response rate of 56 percent (Morgan and Thompson, 2021; Peterson and Will, 2021). Moreover, population subgroups have differing within-household response rates, with lower rates for persons under age 25 (the age group most likely to be victimized by violent crime), for males, and for Hispanic persons and those of a race other than Black or White, raising questions about possible nonresponse bias, particularly for groups underrepresented among the respondents.

The NCVS sample is designed to give precise estimates of victimization for the nation as a whole. However, the sample size is not large enough to produce reliable estimates for all

[108]The name "National Crime Victimization Survey" was adopted in 1993. Before that, it was called the "National Crime Survey." In this report, the acronym NCVS is used to refer to the entire data collection.

[109]Violent crimes include rape and sexual assault, robbery, aggravated assault, and simple (non-aggravated) assault. Property crimes include burglary, motor vehicle theft, and theft.

[110]https://bjs.ojp.gov/programs/ncvs

subpopulations of interest, and often several years of data must be accumulated to compute estimates of violent and property crime for regions or states, as in Lauritsen (2022b). In 2016, the sample size was increased to allow production of state-level estimates for the most populous states using three years of aggregated data (BJS, 2022a).

State-level estimates have also been produced using small area models similar to those described for the Small Area Income and Poverty Estimates program (see Box 2-2). These models predicted a state's NCVS crime counts from state-level crime counts from the UCR Program, along with information from the American Community Survey and the decennial census (Fay, 2021; Liao, Zimmer, and Berzofsky, 2021). Small area estimates represent one promising arena for combining data sources to obtain more detailed information about crime (see Section 7.5 for other potential methods for combining statistics at the area level to enhance the value of crime statistics).

7.3 OTHER NATIONAL DATA SOURCES ABOUT CRIME

The UCR and NCVS are valuable sources of data, but both have limitations in population and crime coverage (see Table 7-1). Estimates from each data source are typically published in September or October of the following year, thus lacking timely availability for studying impacts of events such as the COVID-19 pandemic or changes in laws. This section and Section 7.4 discuss other data sources that might be used, either singly or in combination with the NCVS and UCR, to enhance knowledge about crime.

National Vital Statistics System

Information about homicide is also available through the National Vital Statistics System (NVSS) (see Section 4.3 and Regoeczi and Banks, 2014). Definitions of homicide differ slightly from those in the UCR, but homicide rates measured through the NVSS have closely tracked those from the UCR over time. The NVSS data allow calculation of disaggregated statistics by the victim's state of residence, age, race, ethnicity, sex, marital status, educational attainment, cause of death, and injuries sustained. As with the UCR and the NCVS, there is a lag in publishing mortality statistics (although the data-modernization system underway is expected to speed production of statistics). Unlike the 2021 NIBRS data, the NVSS has nearly full coverage of deaths.

Other Surveys About Crime

The NCVS is the only national survey that collects data on a wide range of crimes, but other surveys ask about specific types of crime. The National Intimate Partner and Sexual Violence Survey, for example, asks about past-year and lifetime experiences with sexual violence and about the health consequences of that violence (Black et al., 2011). Individual localities also collect their own surveys about crime and perceptions of safety.

Crime measurement in surveys is sensitive to the questions asked, how the survey is administered (by an interviewer or self-administered; in person or by telephone, mail, or internet; and who else is present when the respondent answers the survey questions), and nonresponse (Cook et al., 2011; Catalano, 2016). That sensitivity shows up in differences in crime estimates between the NCVS and other surveys, and can make it challenging to directly combine or compare estimates. For example, in 2011, the estimated number of rape and sexual assault

victimizations from the NCVS was 244,190; the estimated number of rape victims from the National Intimate Partner and Sexual Violence Survey was 1,929,000—nearly eight times larger (U.S. Government Accountability Office, 2016, p. 25).[111]

Data Collected by Regulatory Agencies

Several government agencies, including the U.S. Postal Inspection Service, Federal Trade Commission, Securities and Exchange Commission, and Environmental Protection Agency, collect data about specific types of crime as part of their regulatory missions. The Federal Trade Commission (2022), for example, publishes annual national and state statistics on fraud and identity theft reports it has received. Like the UCR, these data collections include only crimes that come to the attention of the agencies, which may be a small fraction of the total crimes committed.

Data from Crowdsourcing and Webscraping

Chapter 2 describes *The Guardian's* database of killings by police, assembled from reader reports and by webscraping of news stories. Other data sources include the Global Terrorism Database, a database of terrorist incidents from around the world from 1970 onward, and the Gun Violence Archive.[112] Lauritsen (2022a) commented that "crowdsourced data can be useful for these types of incidents because most terrorists seek publicity, and because many shootings become known to media." Crowdsourced and webscraped data typically require validation from external data sources but may be available earlier than data from the UCR and NCVS.

7.4 POLICE DEPARTMENT DATA

Individual law enforcement agencies are another potential source for data on crimes around the country. Most large police departments post crime statistics on their websites; some update the statistics daily. Websites of the New York City, Los Angeles, and Chicago Police Departments, for example, provide up-to-date maps and statistics on crimes in city neighborhoods.

---

[111]Krebs (2014) ascribed a large part of the difference to the questions asked about sexual assault. The National Intimate Partner and Sexual Violence Survey asked nine (for women) and eleven (for men) behaviorally specific questions describing acts that are considered to be rape, and detailed specific examples of nonconsent. The NCVS asked two general screening questions about being forced or coerced to engage in unwanted sexual activity. Lohr (2019) discussed other potential reasons for the differences in the two sets of survey estimates, including survey context, response rate, and mode of data collection. The redesigned NCVS questionnaire, to be phased in during 2024, contains revised, behaviorally specific questions about rape and sexual assault (Truman and Brotsos, 2022).

[112]The Global Terrorism Database (https://start.umd.edu/gtd) contains information about the date and location of the incident, weapons used, the number of casualties, and the identity of the perpetrator (when known). The Gun Violence Archive (https://www.gunviolencearchive.org) was established in Fall 2013 with the "goal to provide a database of incidents of gun violence and gun crime. To that end we utilize automated queries, manual research through over 7,500 sources from local and state police, media, data aggregates, government and other sources daily. Each incident is verified by both initial researchers and secondary validation processes."

Noting that many police departments post crime data online, Planty et al. (2018) explored the possibility of using data scraped from police department websites to supplement UCR data. They observed a number of challenges in doing so, however—primarily, a lack of uniformity in how statistics are compiled and presented. Police departments may use crime definitions for their websites that are different from those used by the UCR, with various data formats and frequencies of reporting. Police departments also have various practices for updating information. For example, a crime might be recorded as an aggravated assault on a police department website, but if the victim later dies from the injuries, UCR protocols call for the crime to be classified as a homicide. Data on a police department's website might not be updated after the initial posting.

Despite the lack of uniformity, data from comprehensive police department websites have the advantages of timeliness and granularity, which allow for real-time analyses of crime trends. Although the UCR Program releases some updates during the year, final statistics about crime in a particular year are typically not available until September of the following year. Websites and databases may also have more precise information about locations and characteristics of crimes. But not all police departments report data online, and the set of police departments with comprehensive websites is not representative of the nation as a whole, especially given the resources required to keep such data up to date and accurate. While such sources do not provide national coverage of crimes, they may contain sufficient temporal and geographic granularity to provide richer data for specific jurisdictions.

The time lag for producing statistics reduces the usefulness of UCR data for studying effects of crime-prevention programs or external events. Many researchers were concerned about the effect of the COVID-19 pandemic on crime rates. In the absence of timely UCR data, they used crime data published online from selected cities to compare estimates of homicides and certain other violent crimes within those cities before and during the pandemic (e.g., see Ashby, 2020; Boman and Gallupe, 2020; Kim and Phillips, 2021; Rosenfeld and Lopez, 2022; and Schleimer et al., 2022). As the authors acknowledged, however, these datasets are not nationally representative and crime definitions and measurements (particularly for crimes such as intimate partner violence) vary across cities.

Other researchers have created databases of offenses from publicly available crime data. Ashby (2019) described the Crime Open Database, containing 16 million offenses from ten U.S. cities over an 11-year period. Data were obtained from open-access crime databases of each city and converted to consistent formats for geolocation. Offense lists from each city were manually mapped to NIBRS categories.

Although data from police departments do not necessarily use the same crime categories and protocols as NIBRS, they may be useful for improving the accuracy of NIBRS estimates. Berzofsky et al. (2022) did not mention using crime data external to the NIBRS system in the 2021 estimation methods, but including data obtained from nonreporting law enforcement agencies' websites (when available, and particularly for larger agencies) in an imputation model may be helpful for improving accuracy of national NIBRS statistics.

7.5 COMBINING STATISTICS COMPUTED FROM MULTIPLE DATA SOURCES

Social science researchers have linked statistics calculated from UCR program data or from local police departments to area-level statistics calculated from the census or ACS to investigate factors associated with higher crime rates. For example, Stucky, Payton, and

Ottensmann (2016) linked geocoded UCR data from the Indianapolis police department with publicly available tract-level income statistics from the ACS, finding that lower levels of income, and higher within-tract income inequality, were associated with higher UCR violent and property crime rates. Martinez (2022) discussed the wealth of information available from local police departments and medical examiner offices for studying crime, which includes narratives that provide context for many of the crimes. Martinez (2015) linked data in homicide files from police investigative units and medical examiners' offices in five cities (Chicago, El Paso, Houston, Miami, and San Diego) with publicly available tract-level information from the decennial census to study homicide in Latino and immigrant communities.

The increased sample size for the NCVS, and the additional contextual and demographic information for NIBRS, provide new opportunities for combining subpopulation statistics from these sources with other data. For example, the NCVS small area estimation models (see Section 7.3) used crime counts from the pre-NIBRS-conversion UCR Program and state-level statistics from other administrative and survey data sources to predict the amount of crime at the state level. The new information collected in NIBRS could be used to improve predictions from these models and possibly allow small area estimates to be calculated for smaller geographic areas and for demographic subpopulations.

Past comparisons of UCR and NCVS data have involved national statistics—for example, Morgan and Thompson (2021) compared the rate of crime reported to the police in the UCR and the NCVS for rape and sexual assault, robbery, aggravated assault, burglary, and motor vehicle theft. This set of crimes was measured in both systems with similar, although not identical, definitions. Demographic subpopulations could not be compared because the UCR SRS Program collected only count data; smaller geographic areas could not be compared because the NCVS sample size limited production of these estimates. When estimates for the 22 most populous states are published from the NCVS (see Section 7.2), however, NCVS state-level estimates of crimes reported to the police will be able to be compared with state-level estimates from NIBRS. The contextual information collected by NIBRS will also allow comparison of the two sources for demographic subgroups.

There is also potential for combining NIBRS statistics calculated for small geographic areas (where there is complete reporting) with data from other sources. Fouch and Martin (2022) outlined plans for a "NIBRS Data Dashboard" that will provide context for crime by linking area-level statistics about crime with statistics from sources such as the U.S. Census Bureau's Community Resilience Estimates.[113] This Dashboard would allow researchers to study relationships between crime rates estimated from NIBRS (disaggregated by victim or offender characteristics, weapon use, and other characteristics if desired) and county-level information on characteristics such as health insurance coverage, poverty, and demographics.

Data from the NCVS and NIBRS (or other law enforcement data) could also be combined to obtain larger sample sizes of crime victims. Early discussions about the NCVS explored the idea of using a dual-frame survey (Turner, 1983), in which the NCVS sample of households would be supplemented by a probability sample of persons taken from police records. Although a dual-frame approach was not adopted for the original NCVS, it may be time to explore the idea anew, in light of the availability of detailed records from NIBRS. Chromy and Wilson (2013) discussed the potential of using multiple-frame surveys to obtain larger samples of sexual assault victims. These could also be used to explore measurement differences among data sources.

[113] https://www.census.gov/programs-surveys/community-resilience-estimates.html

There are several challenges in linking crime data at the area level. Data sources may use different definitions or measurements of crime. Classification errors (i.e., what constitutes a crime and what type of crime it is) may affect data sources differently. Resolving such differences and aligning definitions may help improve the quality of crime statistics.

A second challenge in linking NIBRS or police department data at the area level involves geographic alignment of areas covered by law enforcement agencies. Many areas are served by multiple law enforcement agencies. A crime occurring on a university campus, for example, may be investigated by one or more agencies: the city police department, the state police department, the county sheriff, or the university police department. The UCR Program has protocols for avoiding duplication when two or more agencies are involved in the investigation of the same offense, but duplication may be an issue if data are obtained directly from law enforcement agencies. In addition, crime locations may be recorded differently. NIBRS and police department data count a crime incident in the state and jurisdiction where it occurred; the NCVS and NVSS count it at the victim's residence. Discrepancies may arise for crime location when data are combined at small geographic levels.

Studying and combining statistics at the subpopulation level would allow more insights into crime and measurement of crime without linking individual records. Record linkage might be explored with data sources containing sufficient identifying information, but because of the sensitive nature of criminal victimization, it is important to prioritize confidentiality and consent issues (see Boxes 3-4 and 3-5).

7.6 LINKING INDIVIDUAL RECORDS ACROSS DATA SOURCES

There has been less linkage of survey data records with administrative records for crime than there has for income and health statistics. There have, however, been initiatives in which records from various administrative data sources are linked to provide a more comprehensive picture of some types of crime, to study the accuracy of crime data, or to provide detailed information to law enforcement agencies. This section provides a few examples.

Linkage to Add Variables about Crime Incidents, Victims, or Offenders

The National Violent Death Reporting System, a state-based surveillance system administered by the U.S. Centers for Disease Control and Prevention, links information about persons who died by suicide or homicide from death certificates, coroner or medical examiner reports, and law enforcement reports. Some states include information from additional sources such as NIBRS reports, state-level Child Fatality Review team data, hospital data, and court records. Each data source contains different information about the decedent and the circumstances of the crime, and the linked data present a more comprehensive picture of homicide victims—with more than 600 data elements—than can be compiled from any single source (U.S. Centers for Disease Control and Prevention, 2022).

Crosby, Mercy, and Houry (2016) outlined research made possible by the National Violent Death Reporting System that could not have been conducted using only a single source.[114] The system allows researchers to identify violent deaths that occur in the same event, thereby enabling study of topics such as characteristics of homicides followed by a suicide.

[114] https://www.cdc.gov/violenceprevention/datasources/nvdrs/index.html lists recent publications using the linked data.

The individual record linkage in the National Violent Death Reporting System is feasible because there are multiple data sources on violent deaths and records contain identifying information that can be used for linkage. For other crimes, however, information available from other sources may be more limited. Researchers may be able to link police records of aggravated assaults with hospital data, for example, but there may be little additional information for reports of fraud.

Datasets can be linked on several dimensions. Issues for linking crime data are similar to those for linking data about health, in which units vary across data sources (e.g., persons, doctors, hospitals, diagnoses, health care claims). The National Violent Death Reporting System links data related to each violent death. NIBRS and police department data could potentially be linked by location, incident, victim, or offender. The NCVS has some capacity for longitudinal linkage of persons or households through its panel survey design; survey participants could potentially be linked with other data sources such as the National Death Index.[115]

The focus of this chapter is on using multiple data sources to measure crime, but it is important to note that many researchers have linked data sources to study arrest and prosecution, correctional populations, and recidivism. For example, the Criminal Justice Administrative Records System links records across the criminal justice system to create a longitudinal dataset that follows individuals from arrest through discharge (Finlay, Mueller-Smith, and Papp, 2022). Other studies have explored linking BJS administrative datasets about persons in correctional institutions with other data sources (e.g., see Carson, 2015; Goerge and Wiegand, 2019; and Fernandez et al., 2022).

Linkage to Study Crime Measurement or Law Enforcement Procedures

Data sources measure crime in different ways, and more research is needed on their measurement properties. Record linkage can be used to compare measurement of crime concepts across data sources, and thereby suggest improvements to measurement methods. Early NCVS research on the feasibility of measuring crime through a survey examined how accurately survey participants recalled details about victimization incidents by comparing responses to the survey with linked police reports for the incident (Lehnen and Skogan, 1981).

A more recent example of data linkage to study measurement is from Pattavina, Hirschel, and Scearbo (2013), who requested paper copies of original local police department incident reports for a sample of incidents in NIBRS involving intimate partner violence. They compared the results "obtained from coding information directly from the jurisdiction's police reports with those from the same incidents that were submitted electronically to the FBI NIBRS data program" (p. 27). They found that while gender, location, offense, and injury variables were similar in the two sets of records, there was a large discrepancy in the substance use variable—NIBRS records reported substance use in 12 percent of the incidents, while the independent reviewers coded substance use in 26 percent of the incidents.

Wadsworth and Roberts (2008) linked data from the Supplementary Homicide Reports of the UCR with police department records, to study patterns of missing items and evaluate the accuracy of imputations for items missing from the Supplementary Homicide Reports but present in the police data. A similar approach could be used to study accuracy of data elements in NIBRS, perhaps for a probability sample of law enforcement agencies.

[115]The NCVS conducts seven interviews, at six-month intervals, at sampled addresses. But the residents at those addresses can change during the course of the interview series.

Record linkage can also provide information that can be used to inform law enforcement agency procedures. Veitenheimer (2022) described a project to inventory unsubmitted sexual assault kits in Wisconsin. The investigators developed "linkages between those kits that were inventoried and the sexual assault incidents that were reported or supposed to be reported" in NIBRS, and used that information to identify ways the state Department of Justice could "educate law enforcement agencies on […] reporting sexual assaults".

Some individual police departments link multiple data sources, sometimes in conjunction with artificial intelligence algorithms, to allocate law enforcement resources or predict where crime is likely to occur. These predictive policing (sometimes called "data-driven policing") programs use a variety of datasets that can include the police department's internal datasets on crime complaints and arrests; city and state agency data about foreclosures, vacant buildings, building code violations, and transit ridership; U.S. Census Bureau data about neighborhood characteristics; information gathered from online sources; data from automated license plate readers and surveillance cameras; and data purchased from brokers. In predictive policing, datasets are not combined to measure the amount of crime, but rather to develop strategies to prevent or detect it. Brayne (2017) and Ferguson (2017) described predictive-policing methods as well as equity concerns that can arise from algorithmic biases of the type described in Box 3-1. The National Academies noted that predictive policing information "has already been used by police to determine on-street activity, and may ultimately prove useful in refined statistical collection as well" (NASEM, 2016, p. 124).

7.7 IMPROVING THE QUALITY OF CRIME DATA

The previous sections give examples of data-combination activities that have been used or are anticipated for the near future. This section discusses potential longer-term projects in which combining data sources could enhance quality of data about crime.

Improve Population and Crime Coverage

Lauritsen (2022a) argued that any depiction of crime in the United States is incomplete without reference to *all* crime types. She stated that the UCR and NCVS both focus on "street crimes, which, we have learned from a century of criminological research, is disproportionately found in poor areas and sociologically disadvantaged communities," but "neglect of measuring many crime types beyond those available in the UCR and NCVS has produced an incomplete and biased picture of who commits offenses and who experiences the greatest harms from violations of the law." The National Academies proposed an alternative crime classification that encompasses not only the violent and property crimes measured in the UCR and NCVS, but also acts involving fraud, deception, and other types of crime (NASEM, 2016, Section 5.2; see Box 7-1).

Even for the crimes within the scope of the UCR and NCVS, both data-collection programs miss some crimes and some parts of the population. The UCR Program, of course, captures only crimes that are known to the police and forwarded to the FBI. The SRS did not allow researchers to study crime (other than homicide) for subpopulations because it did not collect information about the circumstances of the crimes. NIBRS data have more information on demographics and circumstances, but data from many law enforcement agencies are missing from the FBI statistics.

The NCVS excludes persons living in institutions such as nursing homes and prisons, as well as persons experiencing homelessness and children under age 12. Other subpopulations may be underrepresented in the survey because of nonresponse (see Section 7.2).

Using NIBRS and the NCVS together provides a fuller picture of crime than either source alone, with the NCVS providing information on crimes not reported to the police and NIBRS providing information on crimes against businesses and people who are out of scope for the NCVS (and potentially providing insight into possible nonresponse bias in the NCVS). But, as shown in Table 7-1, some crimes are missing from both data collections—for example, crimes against children aged 0–11 that are not reported to the police. Addington and Lauritsen (2021) described other surveys and administrative datasets that have information about intimate partner violence and violence against children but emphasized that all data sources are incomplete.[116]

Enable Production of Disaggregated Statistics

The Office for Victims of Crime (2013) emphasized the importance of obtaining more information about:

> …the incidence and prevalence of crime victimization in historically underserved populations, as well as the barriers they face in asserting their rights as victims and gaining access to services. These populations include persons with disabilities, boys and young men of color, adults and juveniles in detention settings, youth and women who are trafficked, LGBTQ victims, undocumented immigrants, Americans who are victimized while living in foreign countries, and American Indian/Alaska Native peoples (p. 3).

Crimes against some of these underserved populations cannot currently be studied with NIBRS data because the characteristics defining the subpopulations are not measured. For example, one of the data elements in NIBRS concerns bias motivation for the incident (FBI, 2021). But, for crimes without a known bias motivation, NIBRS does not collect information on the sexual orientation or disability status of the victim, so NIBRS data by themselves cannot give estimates of crimes against the LGBTQIA+ population or against persons with disabilities.

However, the NCVS can provide estimates of victimizations for persons in those groups, because it began asking all respondents about sexual orientation, gender identity, and disability status in 2016. Harrell (2021) found that, from 2017–2019, the rate of violent victimization against persons with disabilities was about four times the rate against persons without disabilities. Examining data from 2017 through 2020, Truman and Morgan (2022, p. 1) found that "rates of violent victimization were significantly higher for persons aged 16 or older who self-identified as lesbian, gay, or bisexual than for those who identified as straight" and that the rate of violent victimization against persons who self-identified as transgender was 2.5 times higher than the rate against persons who self-identified as cisgender. These differences in victimization rates could not be studied with NCVS data before 2016.

It may be possible to study victimization in other historically underrepresented or unidentified groups by linking data with other sources. For example, Nixon et al. (2017) linked records from an Australian registry of persons with intellectual disabilities with a statewide police database, to study criminal charges and victimizations against persons in the registry.

[116]The National Data Archive on Child Abuse and Neglect (https://www.ndacan.acf.hhs.gov/index.cfm) collects microdata from various sources that concern violence against children.

Even when attributes are collected, however, they may be missing or subject to measurement error. NIBRS collects data on race and ethnicity of victims and offenders, when available, but these are input by law enforcement personnel and thus may differ from the race or ethnicity that would be self-reported. Linking NIBRS data with other sources, as done by Arias, Heron, and Hakes (2016) for death certificate information (see Section 3.5) could provide information about the accuracy of race and ethnicity information, as well as other characteristics, in NIBRS data.

Improve Cooperation for Data Collection

Most of the examples of combining data sources in Chapters 5 and 6 focus on linking survey data with administrative records. For crime, however, one of the two major data sources is itself a blending of data contributed by states and individual law enforcement agencies. In that respect, the UCR Program resembles the NVSS (see Section 4.3). While the NVSS is also voluntary and in early years only a few states participated, the system now collects data in standardized form from every state.

The UCR Program, however, has historically lacked the level of personnel or financial resources that enabled the NVSS to achieve nearly complete population coverage, though funds were available to help law enforcement agencies convert to NIBRS. Despite its limited resources, the UCR Program managed to attain a high level of cooperation for the SRS used through 2020. But converting police department data systems to collect and report the more detailed information required by NIBRS is expensive and the program requires local personnel to have expertise in the data-collection and reporting protocols (Smith, 2017). Barnett-Ryan and Swanson (2017) identified lack of funding for state programs as a major contributor to the variability in data quality, and they recommended more research to assess the effects of state quality-control programs on the quality of NIBRS submissions.

As stated in Conclusion 7-1, the UCR Program is still in transition, with incomplete data reporting and variation in measurement methods across jurisdictions. Short-term priorities include improving coverage and consistency of measurement for NIBRS, as well as continuing to develop and refine statistical methods that produce accurate statistics, with valid measures of uncertainty, for the population of law enforcement agencies covered by NIBRS. Sections 7.4 and 7.6 mention that alternative data sources could contribute to these endeavors, through improving imputation models for missing data and through providing external sources of information for evaluating measurement accuracy.

Longer-term activities, however, which may include a more fundamental consideration of some of the alternative data sources described in Sections 7.3 and 7.4, will require a different approach. The National Academies reports on *Modernizing Crime Statistics* concluded that continued improvement in crime statistics will require "enhancements to and expansions of the current data collections, as well as new data collection systems" (NASEM, 2018, p. 6) and that there is "currently no entity responsible for reporting on the full range of crimes" (NASEM, 2018, p. 10). These reports recommended that the U.S. Office of Management and Budget establish a structure for the governance of "the complete U.S. crime statistics enterprise" (see Box 7-1). Such a structure could include crimes measured by regulatory agencies, such as information on fraud and identity theft collected by the Federal Trade Commission.

Acquiring and using data from other sources will require cooperation from data providers: "Data sharing is incentivized when all data holders enjoy tangible benefits valuable to

their missions, and when societal benefits are proportionate to possible costs and risks" (NASEM, 2023, p. 6). NIBRS requires law enforcement agencies to submit data in NIBRS format. This has advantages of standardizing data collection and data elements, but also has costs to law enforcement agencies and states.

An alternative model for data sharing, particularly if previously excluded types of crime and new data sources are to be included, is to "shift some burden of data standardization from respondents to the state and federal levels" (NASEM, 2018, Conclusion 3.1; see Box 7-1). As an example, in a panel discussion, Smith (2022) commented on the potential for using data directly from law enforcement agencies to obtain more timely national statistics:

> We could be looking experimentally at what it means to bring in crime incident data directly from the source—not as a bypass to how official statistics are captured by the FBI through the NIBRS system [but to enhance] the information that we get, collecting it in a much more timely way for a smaller subset of agencies and then being able to expand that picture with these national collections that are already in place. Some of that direct connection to crime incident data could involve narrative where possible, capitalizing on the [artificial intelligence] and [machine learning] tools that we have available to us now, to really try to understand more specifically what is the connection between what police see, what we see in the victimization data in the NCVS, and some of these other sources of information. There is a lot that technology can do for us.

The COVID-19 pandemic underlined the urgent need for nationally representative and timely data about crime: UCR and NCVS statistics for 2020, published in September 2021, could not help local jurisdictions decide how to deal with changes in crime patterns caused by lockdowns and changed activity patterns. Some researchers filled the void by assembling their own datasets from conveniently available online data (see Section 7.4), but these datasets were not nationally representative.

As suggested by Smith (2022), more timely data could be assembled for crimes known to the police by selecting a probability sample of law enforcement agencies to provide real-time crime incident data. The burden of providing such data could be substantially reduced by developing procedures that could take data in the format supplied by each agency and convert it to the format needed for the statistics. In this model, the sampled law enforcement agency would provide data that it already collects, along with documentation (perhaps developed jointly with BJS) about the data elements. BJS would then map the agency's data onto standardized crime, demographic, and circumstance categories. This would reduce the burden on individual data providers while providing timely national statistics about crime. Technology might similarly be able to help speed measurement of crimes not reported to the police by, for example, collecting real-time reports of incidents from a probability sample of people who were provided with smartphones for that purpose.

The panel holds that the area of crime statistics could benefit greatly from increased use of multiple data sources to improve coverage of crimes, improve coverage of populations and businesses affected by crime, and allow research about the differential impact of crime on subpopulations. Profiting from these data sources, however, will require investment in data infrastructure, personnel, and statistical methods for working with the data sources, as well as a structure for coordinating data collection across agencies.

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.

Table 7-1 Crimes Included in the Uniform Crime Reporting (UCR) Program and the National Crime Victimization Survey (NCVS)

	Represented in NCVS	Not Represented in NCVS
In UCR	Crimes against noninstitutionalized U.S. residents aged 12+ that are reported to police and are measured by both data sources[a]	Crime types measured in UCR but not measured in NCVS (e.g., homicide) Crimes reported to police and measured in UCR against: ● Businesses and organizations ● People out of scope for NCVS (e.g., children aged 0–11, people in institutions, people experiencing homelessness, non-U.S. residents) ● NCVS respondents who do not report the crime on the survey
Not in UCR	Crimes against noninstitutionalized U.S. residents aged 12+ that are measured in NCVS and not reported to police (or not measured in UCR)	Crime types measured in neither UCR nor NCVS (e.g., fraud against government agencies, environmental crimes) Unrecognized crimes (e.g., romance scams in which victims and law enforcement are unaware a crime has been committed) Crimes not reported to police against: ● Businesses and organizations ● People out of scope for NCVS ● NCVS respondents who do not report the crime on the survey

[a]Hanson (2021) and https://ucr.fbi.gov/nibrs-in-brief listed the 52 Group A Offenses and 10 Group B Offenses that are measured by NIBRS. The major categories for the Group A Offenses are: animal cruelty, arson, assault offenses, bribery, burglary, counterfeiting/forgery, destruction/vandalism of property, drug/narcotic offenses, embezzlement, extortion/blackmail, fraud offenses, gambling offenses, homicide offenses, human trafficking offenses, kidnapping/abduction, larceny/theft offenses, motor vehicle theft, pornography/obscene material, prostitution offenses, robbery, sex offenses, stolen property offenses, and weapon law violations. The NCVS measures rape and sexual assault, robbery, aggravated and simple assault, burglary, theft, motor vehicle theft, and pocket-picking/purse-snatching. In addition, NCVS supplements have measured fraud, identity theft, and school crime for various years. A new NCVS questionnaire is expected to be phased in during 2024 (Truman and Brotsos, 2022).

NOTE: This table gives the classification under the idealized UCR in which all law enforcement agencies submit data to the FBI.
SOURCE: Panel generated.

Table 7-2 Uniform Crime Reports Estimates under the Summary Reporting System and the National Incident-Based Reporting System

	Summary Reporting System (2020)	National Incident-Based Reporting System (2021)
Crimes reported	Numbers and rates for homicide, rape, robbery, aggravated assault, burglary, larceny/theft, and motor vehicle theft. Statistics included only the most serious offense for each incident.	Estimated numbers and rates for 52 types of crime. Up to 10 offenses are counted for each incident.
Law enforcement agency coverage	15,875 of the 18,623 agencies submitted data for at least three months of the year.	11,333 of the 18,806 agencies submitted data for at least three months of the year.
Population coverage	97 percent of U.S. population lived in an area served by at least one reporting agency.	65 percent of U.S. population lived in an area served by at least one reporting agency.
Geographic detail	Estimates reported for all states and metropolitan statistical areas, plus tabulations within states by type of community (metropolitan statistical area, other cities, rural) and counties.	Estimates for some states, metropolitan areas, and types of agencies suppressed because of insufficient data.
Incident detail	Incident-level details on victim and offender demographics and relationships, weapon used, and crime circumstances collected for homicides; counts alone for other crimes.	Includes date; time; location type (e,g., restaurant, home, cyberspace); age, race, ethnicity, sex of victims and offenders; relationships between victims and offenders (e.g., spouse, sibling, neighbor, employer, stranger); injuries; property loss; weapons; alcohol or drug involvement; bias motivation (if offense was recorded as being motivated by bias against race, religion, disability, ethnicity, or sexual orientation); clearance and arrest information.
Estimation procedure	Data reviewed for quality and outlier detection. Imputation of crime counts for agencies with fewer than 12 months of data.	National crime statistics estimated using statistical models. Details of the procedure had not yet been published as of October, 2022.
Standard errors	Not reported because of high population coverage.	Estimates accompanied by confidence intervals generated through estimation procedure.

SOURCE: Panel generated with information from Addington (2019); U.S. Bureau of Justice Statistics (2021a); FBI (2021, 2022b); Barnett-Ryan and Berzofsky (2022); Berzofsky et al. (2022); and National Archive of Criminal Justice Data (2022).

BOX 7-1 Selected Conclusions and Recommendations from the National Academies of Sciences, Engineering, and Medicine Reports on *Modernizing Crime Statistics*

Conclusion 2-1: The aim of modern crime statistics is the effective measurement and estimation of crime. Accurate counting of offenses and incidents is important, but the nation's crime statistics will remain inadequate unless they expand to include more than just simple tallies with no associated measure of uncertainty or capacity for disaggregation. Through the collection of associated attribute data, the crime statistics we suggest should—at minimum—enable the analysis of data in proper geographic, demographic, sociological, and economic context, and provide the raw material for important measures related to an offense (such as the harm it causes) in addition to its count (NASEM, 2018, p. 32).

Conclusion 2-3: Improvement in the nation's crime statistics will require enhancements to and expansions of the current data collections, as well as new data collection systems for the historically neglected crime types highlighted by the proposed crime classification (NASEM, 2018, p. 39).[117]

Conclusion 3-1: A stronger federal coordination role is needed in the production of the nation's crime statistics: providing resources for information systems development, working with software providers to implement standards, and shifting some burden of data standardization from respondents to the state and federal levels. The goal of this stronger role is to make crime data collection a product of routine operations (NASEM, 2018, p. 53).

Conclusion 3-2: Having an effective governance structure for the complete U.S. crime statistics enterprise is critical. There is currently no entity responsible for reporting on the full range of crimes in the proposed classification (most notably for top-level categories 6-11) (NASEM, 2018, p. 54).

Recommendation 3.1: The U.S. Office of Management and Budget (OMB) should explore the range of coordination and governance processes for the complete U.S. crime statistics enterprise—including the "new" crime categories—and then establish such a structure. The structure must ensure that all of the component functions of generating crime statistics are conducted in concordance with the sensibilities, principles, and practices of a statistical agency. It should provide for user and stakeholder involvement in the process of refining and updating the underlying classification of crime. The new governance process also needs to take

[117]The proposed crime categories are:

1. Acts leading to death or intending to cause death
2. Acts causing harm or intending to cause harm to the person
3. Injurious acts of a sexual nature
4. Acts of violence or threatened violence against a person that involve property
5. Acts against property only
6. Acts involving controlled substances
7. Acts involving fraud, deception, or corruption
8. Acts against public order and authority
9. Acts against public safety and national security
10. Acts against the natural environment or against animals
11. Other criminal acts not elsewhere classified (NASEM, 2016a, p. 126)

responsibility for the dissemination of data products, including the production of a new form of *Crime in the United States* that includes the "new" crime categories (NASEM, 2018, p. 61).

SOURCE: NASEM, 2016a, 2018.

[END BOX 7-1]

8. Using Multiple Data Sources for County-Level Crop Estimates

Previous chapters have focused on enhancing information about the U.S. population by using administrative records directly or linking them with surveys. This chapter looks at an example in the arena of business statistics—in particular, the use of surveys, administrative records, and remote sensing data to produce county-level estimates of agricultural production.

The U.S. Department of Agriculture (USDA) has been involved in producing county-level crop estimates since 1917 (Cruze et al., 2019).[118] A National Academies of Sciences, Engineering, and Medicine report described the importance of these estimates:

> Participants in agricultural markets rely on such information to make decisions: for producers, about what to grow and how to manage inventories; for processors and traders, about how to organize production and determine sales; and for retailers and consumers, about how to anticipate costs and assess the availability of food. When market participants share a common understanding of the fundamentals of supply and demand, market transactions accurately reflect the value of commodities to those along the supply chain and help ensure that food is grown, processed, and consumed at the lowest cost to the nation (NASEM, 2017b, pp. 6–7).

The National Academies (NASEM, 2017b) reviewed procedures then used by the USDA National Agricultural Statistics Service (NASS) to produce county-level estimates for crops (including planted acres, harvested acres, production, and yield by commodity) and recommended pursuing a model-based approach relying on multiple data sources (see Box 8-1). The approach that panel recommended would build on modeling research performed by NASS, the U.S. Census Bureau, Statistics Canada, and other agencies.

[BOX 8-1 about here]

This chapter describes the models that NASS has developed since the National Academies' 2017 review (NASEM, 2017b), considers challenges for integrating data from agricultural and other business surveys, and outlines additional ways that NASS might take advantage of multiple data sources. While Chapters 5 through 7 focus on linkage of income and health data and consolidation of crime data submitted by states, this chapter focuses on the use of small area models to integrate information from administrative and other data sources.

Section 8.1 briefly reviews data sources that might be used for producing crop estimates. Sections 8.2 and 8.3 discuss statistical modeling approaches taken by NASS and Statistics Canada, respectively, to incorporate data from non-survey sources into crop-estimation programs, relying in part on presentations from the workshop session on *Improving Agriculture Statistics with New Data Sources*. Section 8.4 explores opportunities for continued improvement of agricultural statistics.

[118]For the history of agricultural statistics in the United States, see U.S. Department of Agriculture (1969); Allen (2008); and https://www.nass.usda.gov/About_NASS/History_of_Ag_Statistics/

8.1 DATA SOURCES FOR CROP ESTIMATES

This section summarizes the main data sources that NASS has used to make county-level crop estimates in the United States, as well as other data sources with potential to improve model-based estimates: private-sector data and data obtained from social media, webscraping, and crowdsourcing.[119]

Probability Samples

NASS conducts hundreds of surveys every year.[120] A census and three probability surveys provide information for NASS's Crops County Estimates Program.

- The *Census of Agriculture* is taken every 5 years with the purpose of providing "a complete count of U.S. farms and ranches and the people who operate them."[121] It collects information on characteristics of farm operators, land use, production practices, income, and expenditures. Although the intent is to include every agricultural operation, the Census of Agriculture has undercoverage, nonresponse, and misclassification. Some farms, particularly smaller operations, are not on the mailing list that serves as the sampling frame for the census, and some operations are misclassified. Estimates of the total number of farms and acreage devoted to agriculture are adjusted for undercoverage, nonresponse, and misclassification using information from the June Area Survey (USDA, 2019).
- The *June Area Survey* (JAS) collects information on "crop acreage, grain stocks, cattle inventory, hog inventory, sheep and goat presence, land values, farm numbers, technology use, and value of sales data."[122] It is called an "area survey" because the sample is drawn from an area frame that identifies parcels of land (Davies, 2009). For the JAS, land segments of approximately one square mile are selected for the sample, and interviewers attempt to interview every farm operator within the boundaries of the sampled land segments. Because the sampling frame consists of parcels of land, the JAS has full coverage of all farm operators (although there is still nonresponse because some of the sampled operators cannot be reached or decline to participate in the survey).

---

[119] See also Stubbs (2016, p. 1), who categorized "big data" sources for agriculture as "public-level big data," which are "collected, maintained, and analyzed through publicly funded sources, specifically by federal agencies (e.g., farm program participant records, Soil Survey, and weather data)" and "private big data," which "represent records generated at the production level and originate with the farmer or rancher (e.g., yield, soil analysis, irrigation levels, livestock movement, and grazing rates)."

[120] See https://www.nass.usda.gov/Surveys/ and https://www.nass.usda.gov/Surveys/Guide_to_NASS_Surveys/index.php for listings and descriptions of NASS surveys and programs. Schnepf (2017, p.5) noted that NASS uses these surveys to publish "about 400 national agricultural statistical reports and thousands of additional state agricultural statistical reports covering more than 120 crops and 45 livestock items."

[121] See https://www.nass.usda.gov/AgCensus/ for a description of the Census of Agriculture and https://www.census.gov/history/www/programs/agriculture/census_of_agriculture.html for its history.

[122] https://www.nass.usda.gov/Surveys/GuidetoNASSSurveys/JuneArea/index.php

- The quarterly (March, June, September, and December) *Agricultural (Crops/Stocks) Surveys* are conducted in all states except Hawaii, and they provide national estimates and early-season predictions of acreages, yields, and production for major crops. Farm operators are asked about the total number of acres they operate and how much acreage is devoted to each commodity of interest.[123] The main samples are selected from list frames—lists of known farm operations—and thus do not include operations not on the list, but farms from the June Area Survey "that are not included in the list frame sampling population are subsampled for the March, September, and December surveys so that the target population is completely represented" (NASS, 2022, p. 1).
- The *County Agricultural Production Survey* (CAPS), conducted annually at the end of the harvest season, supplements the county-level sample sizes from the Agricultural Surveys. All counties in the 44 states in which CAPS is conducted must be represented in the sample, although the commodities studied are specific to each state. The survey is mainly conducted by mail and telephone.[124]

Figure 2-1 displays response rates for the JAS from 2000–2022. Response rates for the quarterly Agricultural Surveys dropped from about 85 percent in the early 1990s to the 60 percent range in 2016 (Johansson, Effland, and Coble, 2017). The December 2021 Agricultural Survey had a response rate of 50.1 percent, a decrease from the 55.7 percent response rate from the previous December (NASS, 2022, p. 6). In addition, there is item nonresponse to the questions about specific commodities.

Schnepf (2017, p. 16) wrote: "The potential bias related to nonresponse becomes increasingly important for more localized estimates. For example, NASS estimates remain most accurate at the national level, but low response rates become increasingly important for estimates at the state and especially county levels." Increasing nonresponse to agricultural surveys suggests that assessment of alternate data sources is an appropriate next step, as recommended by the National Academies report on *Improving Crop Estimates by Integrating Multiple Data Sources* (NASEM, 2017b; see Box 8-1).

Administrative Records

Several administrative data sources provide information related to crop estimates. Agencies that collect data through program administration include the USDA Farm Service Agency (FSA), which collects individual producers' farm record data, federal payments, and loan information used in administering various farm programs; and the USDA Risk Management Agency (RMA), which collects individual farm yield and loss information to administer the Federal Crop Insurance program.[125]

Farmers who elect to participate in FSA programs provide the agency with planted acreages and crop types. Because participation in FSA programs is voluntary, estimates of planted acreage from FSA data alone will usually underestimate the total amount of planted

[123] https://www.nass.usda.gov/Surveys/Guide_to_NASS_Surveys/Crops_Stocks/index.php

[124] https://www.nass.usda.gov/Surveys/Guide_to_NASS_Surveys/County_Agricultural_Production/

[125] An additional possible administrative records data source is the USDA National Resources Conservation Service, which collects data on conservation plans, geospatial data, and conservation program activities and payments to meet the USDA's responsibilities under the Soil and Water Resources Conservation Act of 1977. See https://www.nrcs.usda.gov/wps/portal/nrcs/main/national/about/

acreage for a crop (which will also include acreage from farmers who do not participate in FSA programs).[126] FSA planted acreage data can be considered as a lower bound for the true amount of planted acreage. The RMA, in its role as an underwriter of crop insurance policies, receives data from crop insurance providers about failed acreage (acreage that was planted but not harvested, perhaps because of local weather or flooding) and checks submissions for accuracy before making payments to farmers.[127] As with FSA data, there is undercoverage of the population because some farmers do not participate in a crop insurance program. In general, the FSA and RMA data have high coverage of planted acres for major commodities (NASEM, 2017b, p. 57). However, Cruze et al. (2019, p. 303) noted that some groups are particularly prone to undercoverage, for example "known Amish communities in Pennsylvania and other midwestern states may represent significant portions of local agricultural activity but tend not to participate in federal or commercial crop insurance programs."

One complication in combining these administrative sources with survey data is that NASS, FSA, and RMA use different definitions of farms. NASS defines a farm as "any establishment from which $1,000 or more of agricultural products were sold or would normally be sold during the year."[128] NASS associates one or more operators with each farm on its list frame. For the FSA, a farm "is made up of tracts that have the same owner and the same operator" (NASEM, 2017b, p. 48). The RMA does not define farms but collects information from entities that purchase crop insurance from approved insurance providers.

The National Academies report on *Improving Crop Estimates by Integrating Multiple Data Sources* (NASEM, 2017b) recommended that NASS adopt the FSA's Common Land Unit (similar in spirit to a farm field) as its basic spatial unit, to enhance interoperability and facilitate linkage of data sources (see Box 8-1).[129]

Satellite, Aerial Imagery, and Sensor Data

The Global Strategy to improve Agricultural and Rural Statistics (2017) provided an overview and guidelines for using remote sensing in agricultural statistics, with chapters on land cover mapping and monitoring, detailed crop mapping, and crop area and yield estimation. Important survey-related uses of remotely sensed data include improving coverage of list frames

[126]For overviews of FSA programs and the information collected by FSA, see https://www.fsa.usda.gov/Assets/USDA-FSA-Public/usdafiles/FactSheets/2016/farm_service_agency_programs.pdf, https://www.fsa.usda.gov/Assets/USDA-FSA-Public/usdafiles/FactSheets/2019/arc-plc_farm_bill_comparisons-fact_sheet-aug-2019.pdf, and https://www.fsa.usda.gov/Assets/USDA-FSA-Public/usdafiles/FactSheets/2022/fsa_cropacreagereporting_factsheet_22.pdf

[127]For more information on the RMA see https://www.rma.usda.gov/en/Fact-Sheets/National-Fact-Sheets/About-the-Risk-Management-Agency

[128]https://www.nass.usda.gov/About_NASS/History_of_Ag_Statistics/index.php

[129]FSA defined the Common Land Unit as an "individual, contiguous farming parcel," which is the smallest unit of land that has a permanent, contiguous boundary; common land cover and land management; and a common owner and/or common producer association. See the 2017 Common Land Unit information sheet at https://www.fsa.usda.gov/Assets/USDA-FSA-Public/usdafiles/APFO/support-documents/pdfs/cluinfosheet2017Final.pdf. Ali and Dahlhaus (2022) discussed interoperability in the agricultural data context.

and improving the efficiency of sampling designs; many agricultural surveys use information on land cover to stratify the sampling design (Carfagna and Carfagna, 2015).[130]

Various remote sensing sources could be used as inputs to crop models: "An increasing number of satellites, aircraft, drones, flux towers, and weather stations collect geospatially referenced data that may be useful for monitoring crop-growing conditions. These data may be available from other government agencies or for purchase from private companies" (NASEM, 2017b, p. 67).

Carletto, Dillon, and Zezza (2021, p. 4453) noted that:

> Remote sensing data are being used and adapted for countless purposes in farm management, agricultural programs, agricultural statistics, and empirical agricultural economics…. For empirical applications in agricultural economics, remote sensing data offer the promise of far greater accuracy, objectivity, temporal resolution, and coverage, than could be achieved through traditional survey methods relying on farmers' self-reporting. However, remote sensing datasets are not immune from measurement error…. Errors can be introduced through the measurement technology, the algorithm to convert the measurement into a variable for analytical use (e.g., rainfall), or the resolution of the data. Errors can also occur in linking remote sensing data to the household, plot, or farm on which the analysis is run, as well as by using variables that are not 'fit for purpose' from an agronomic perspective.

NASS uses data from satellite and aerial imagery to create the Cropland Data Layer, a detailed map of crops grown across the continental United States.[131] Historically, the Cropland Data Layer has had 85–95 percent accuracy for major crops (Young, 2022, slide 4). NASS adjusts for bias with a regression model that uses the observed acreages for a specific crop recorded during the June Area Survey.

Although the Cropland Data Layer is highly accurate overall, there are data-equity issues in that land classification based on satellite observation is less accurate for smaller fields, which may produce multiple crops or have land parcels smaller than one pixel. Smaller holdings are more likely to have the "mixed-pixel problem," meaning a pixel contains more than one type of ground cover and may be inaccurately classified.

Satellite imagery is a valuable resource for producing crop estimates, but comes with challenges described by workshop participants.[132] Nkwimi-Tchahou et al. (2022) mentioned the effects from clouds and other contaminants on data quality, the intensive information technology needs for processing satellite imagery data, potential comparability problems when satellites change (because of differences in resolution), and the rare possibility of satellite failure.

---

[130]When data from remote sensing are used for stratifying a survey design, misclassification errors do not affect the validity of the survey. More accurate data from remote sensing will improve the efficiency of the design, but any misclassification errors will be corrected during the ground survey.

[131]See Craig (2010) for a history of the Cropland Data Layer, and Boryan et al. (2011), https://data.nal.usda.gov/dataset/cropscape-cropland-data-layer, and https://www.nass.usda.gov/Research_and_Science/Cropland/sarsfaqs2.php for information on data sources and uses.

[132]See also Gallego, Carfagna, and Baruth (2010), who identified characteristics associated with quality of data from remote sensing systems, including accuracy, objectivity, and cost-efficiency. They identified the main characteristics for use in agricultural applications as spectral resolution, spatial resolution, and the ability to provide data for large areas of land for low cost. They also mentioned the need for "high temporal frequency" to be able to "follow crop growth during the season" (p. 204).

Goodchild (2022) emphasized the uncertainty inherent in using remote sensing data: "The pixels of remote sensing … are not sharp boundaries on the Earth's surface, but instead the contents of one pixel bleed quite substantially into the contents of a neighboring pixel." Goodchild also expressed concern about propagation of uncertainties through the estimation system: "We are combining datasets which have different, independent, uncertainties associated with them." He illustrated this with an example: Common Land Units are often defined by physical boundaries, such as roads, but farmers often do not plant to the edge of a road.[133]

Private-Sector Data

Private-sector entities (agricultural producers) provide data through NASS agricultural surveys and administrative records. But many farm operators collect much more detailed information than is submitted to surveys. Coble et al. (2018, p. 82) commented on "the remarkable growth in producers' ability to collect data pertaining only to their own operation through the growth of techniques and technologies such as grid soil sampling, telematics systems for farm equipment, Global Navigation Satellite Systems (GNSS), farm aerial imagery acquired via small unmanned aerial systems (sUAS), and the like."

These detailed data are used in precision agriculture, a field that emerged in the 1980s to take advantage of technological advances in global navigation satellite systems, geographic information systems, and computing to enable data-driven decisions about planting, fertilizer use, pest and disease management, and other aspects of agricultural production. The International Society of Precision Agriculture (2019) defined precision agriculture as "a management strategy that gathers, processes and analyzes temporal, spatial and individual data and combines it with other information to support management decisions according to estimated variability for improved resource use efficiency, productivity, quality, profitability and sustainability of agricultural production."

Early uses of precision agriculture involved adapting fertilizer distribution to soil conditions. Since then, uses have become more sophisticated, combining information from "sensors, information systems, enhanced machinery, and informed management to optimize production by accounting for variability and uncertainties within agricultural systems" (Gebbers and Adamchuk, 2010, p. 828). Stubbs (2016, p. 8) emphasized the dependence of data collection on "physical technology, such as sensors, imagery, drones, radar, and other technologies all working together to provide detailed information about soil content, weeds and pests, sunlight and shade, nutrient deficiencies, moisture, and other factors…. Data collection is an ever-expanding area of big data and includes a number of key players, including" equipment manufacturers, chemical companies and applicators, and developers of technologies such as radio frequency identification. Mendez-Costabel (2022) described uses of linked geospatial and other data sources to predict performance of seed varieties under various growing conditions (e.g., open fields, greenhouses, and small land holdings) and climates.

Individual farmers are the main beneficiaries of precision agriculture, as it provides them with better data for making decisions. But these data also hold promise for improving agricultural statistics through integration with surveys and administrative data. The National Academies report on *Improving Crop Estimates by Integrating Multiple Data Sources* (NASEM,

[133]With the advent of precision agriculture (see below), some farmers may use tractor-based data to report the actual acreage planted, which may differ from acreage that would be reported using Common Land Units.

2017b, p. 52) commented that these data might be used to reduce burden on survey respondents or to impute data for nonrespondents.

There are several challenges, however, in using private-sector data to improve agricultural statistics (see Chapter 2). Sourav and Emanuel (2020), reviewing recent trends of "big data" technology in the field of precision agriculture, argued that sensors and machinery help farmers track temperature, humidity, and soil conditions, but the data require processing. There may also be data gaps, measurement errors, lack of documentation, or proprietary data-manipulation methods that cannot be shared.

Undercoverage may occur because not all farms use precision agriculture: large corporate farming operations are more likely to have the resources to collect such data. This may lead to data inequities, in which more timely and accurate information is available for areas with large farm operations compared with areas consisting mainly of small farms.

Beyond those are the challenges of obtaining—and continuing to obtain—access to the data. Hurst (2016, p. 6) stated that many farmers using data and analytics "are reporting higher yields, fewer inputs, more efficiency, less strain on the environment, and higher profits. Yet many are also expressing concerns about privacy, security, portability, and transparency in how their data is used, and who exactly has access." Stock and Gardezi (2022, p. 6) also highlighted concerns about the ability of agricultural technology firms and data consolidators to protect the confidentiality of farmers' data: they may have "a royalty-free license over this data, giving them unrestricted permission to access."

Hurst (2016, p. 8) mentioned the issue of data ownership, noting that "the individual farmer's data has considerably more value than the average consumer's data." Ryan (2019) discussed the potential for a digital divide, in which farmers with data can prosper more than those without. Public data from NASS could mitigate some of that impact, but care is needed to ensure that the benefits of data are shared by all.

One issue with using private-sector data sources for agriculture is also shared with other types of business data. Private companies collect many types of data that give them a competitive advantage, and that advantage may be lessened if data are shared. The previous National Academies report in this series (NASEM, 2023) discussed possible benefits that could be offered to encourage data sharing, including the value of timely and granular data, confidentiality protection, and financial incentives. One benefit would be the development of standards for the collection and processing of data.

Data from Social Media, Webscraping, and Crowdsourcing

The unknown coverage of social media, webscraping, and crowdsourcing data makes it difficult to use these data as a single source to produce statistics, but they can provide valuable information when verified and combined with other sources. In agricultural statistics, webscraped and crowdsourced data have been used for expanding sampling frames and providing "ground truth" to verify data obtained from other sources such as satellite images. Hyman, Sartore, and Young (2022) described the use of webscraping to assess the coverage of local food farms in the NASS list frame (see Section 3.2). Webscraping has similarly been used to identify urban agricultural operations (Young, Hyman, and Rater, 2018) and farmers' markets (Young and Jacobsen, 2022). In these studies, the researchers created a list of terms that might be used on websites to identify operations (e.g., "urban farm" or "community garden") and verified that the operations were in the target population. These efforts advance data equity by improving

coverage of small farms that are missing from the list frame and expensive to capture in an area frame.

In ground-truthing applications, participants visit sites that correspond to the satellite images, to verify the crops grown.[134] Goodchild and Li (2012) offered three approaches to ensuring quality of "volunteered geographic information": crowdsourcing, referring to "the ability of a group to validate and correct the errors that an individual might make" (p. 112); social, relying on "a hierarchy of trusted individuals who act as moderators or gate-keepers" (p. 114); and geographic, relying on "a comparison of a purported geographic fact with the broad body of geographic knowledge" (p. 115).

Fritz et al. (2019) discussed the possibility of using data from smartphones and social media: "The increased amount of smartphones all over the world, even among low income farmers, usually the group responsible for the largest agricultural uncertainties, allows for increased opportunities to self-report geo-located crops and parcel practices, including planting dates, fertilizer application, irrigation and expected yields, through the use of purpose-designed mobile applications" (p. 270). They also suggested: "The food security and early warning community should also make greater use of the latent predictive capacity of social media and sources such as web search data" (p. 270), and gave examples of social media messages that could have given early warning of lower-than-average wheat yields.

8.2 MODELING CROPS COUNTY ESTIMATES IN THE UNITED STATES

Crops county estimates are used for many purposes. County yield data from surveys are used by USDA for various programs, including those administered by USDA's Farm Service Agency and Risk Management Agency. For example, when a natural disaster such as drought or flooding impacts crop production, these data are crucial to the agriculture industry. They are also used by government agencies, researchers, and organizations "to determine many production and economic values on a small area basis" (Schnepf, 2017, p. 17).

County-level crop estimates published before 2020 were the result of an expert review process directed by the USDA Agricultural Statistics Board, which considered survey data (in particular, the quarterly Agricultural Surveys and CAPS) and other sources of information (such as administrative records from FSA and RMA) when determining an official estimate for each county (NASS, 2012; NASEM, 2017b; Cruze et al., 2019). To ensure consistency across geographic units of varying sizes, the Agricultural Statistics Board first determined the final national and state estimates for crop yield, acreage, and production. They then set estimates for agricultural statistics districts (sets of contiguous counties) and counties, ensuring that county totals summed to district totals, and district totals summed to state totals. Once the official estimates were approved, they were subject to NASS production standards for confidentiality and consistency across different-sized geographic units (Cruze et al., 2019).

Although the historical process made use of administrative records information such as that from the FSA and RMA, that information was incorporated through the expert judgment of the Agricultural Statistics Board, not through a statistical model. The process of manual

[134]For example, Lesiv et al. (2019) discussed an effort to estimate field sizes across the globe using crowdsourcing to assess the contribution made by smallholder farms to food production. Saralioglu and Gungor (2020) provided a literature review of the use of crowdsourcing to validate remote sensing data. One example of crowdsourcing is the Geo-Wiki Project (https://www.geo-wiki.org/), "which enables volunteers from around the world to help make land cover maps more accurate" (p. 99).

assessment of separate inputs was time consuming and needed to be repeated for each state and commodity separately. Young (2022) noted that because of the subjective input from the Agricultural Statistics Board, there was a "lack of transparency and reproducibility." In addition, no measures of uncertainty (such as margins of error) were reported.

The National Academies' report *Improving Crop Estimates by Integrating Multiple Sources* recommended that NASS revise the county-level crop estimates program by using statistical models that rely on multiple data sources (see Box 8-1). They recommended development of small area statistical models as in the U.S. Census Bureau's Small Area Income and Poverty Estimates program (see Box 2-2). The vision for 2025 had three components:

> First, NASS prepares its county estimates using a transparent and well-documented process, publishing measures of uncertainty along with point estimates. Second, the NASS list frame is a georeferenced farm-level database, serving as a sampling frame for surveys and facilitating the use of farm data in statistical analysis. Third, NASS acquires all relevant georeferenced administrative and remotely sensed and ground-gathered information and uses this information to complement its traditional survey data (NASEM, 2017b, p. 17).

NASS has taken important steps toward realizing the vision in *Improving Crop Estimates by Integrating Multiple Sources*. Successive stages in the model-development process to include non-survey sources of data have been documented in a series of journal articles and conference presentations.[135] The current panel anticipates that NASS will issue an official methodology report that consolidates the information in the research reports and describes the current production models, as that will provide important documentation for data users.

The models that have been developed can be viewed as extensions of those used for the Small Area Income and Poverty Estimates program, with additional features to meet the special challenges of producing crops county estimates, such as the additional information from FSA and RMA that can be used to set a lower bound on planted acreage in each county. Separate models were needed for planted acres, harvested acres, and yield (or production) for each commodity.[136]

A model-based estimate for the number of acres planted to a particular crop in a county relies on a direct estimate for that county (computed from the survey data), auxiliary information from administrative data (which provide a lower limit for planted acreage) and other sources of covariates. For the model described in Erciulescu, Cruze, and Nandram (2019), the direct estimate for acreage came from CAPS and the auxiliary data considered as covariates included:

- Planted acreage totals reported to FSA;
- Insurance claims totals for failed acreage reported to RMA;
- Information on maximum planted acreage for each operator in the NASS list sampling frame;

[135]See, for example, Cruze et al. (2019); Erciulescu, Cruze, and Nandram (2018, 2019, 2020); Chen and Nandram (2022); Chen, Nandram, and Cruze (2022); and Nandram et al. (2022).

[136]Young and Chen (2022) noted that because production is the product of yield and harvested acres, only three models are needed for each commodity: planted acres, harvested acres, and either yield or production.

- Cropland Data Layer information on planted acreage, derived from satellite imagery; and
- Monthly weather variables from the National Oceanographic and Atmospheric Administration.

It was desired that estimates produced at differing levels of geography be consistent with each other, with county totals aggregating to agricultural statistics district totals, and district totals aggregating to state totals. This was done by estimating acreage for both agricultural statistical districts and counties in the same model, thereby ensuring that the estimates for counties within a district summed to the district estimate.[137] At the end of the estimation process, district and county estimates were multiplied by a common factor that ensured they summed to state-level estimates.

Logical constraints among the quantities measured—for example, the number of harvested acres for a county must be less than or equal to the number of planted acres—were also incorporated into the small area models. Participation in the FSA and RMA programs is voluntary (see Section 8.1), so total planted acres from those administrative datasets would miss the acreage from nonparticipating farm operators. Totals from the administrative records could, however, be viewed as "informative lower bounds" for the planted acreage in each county, and Chen, Nandram, and Cruze (2022) incorporated constraints into the planted acreage model by requiring the county-level estimate of planted acreage to be at least as large as the maximum acres planted to the crop, as determined from FSA and RMA values. Similar constraints were introduced for other models (for example, the RMA value for failed acreage provided a lower bound for that quantity).

Ensuring county-district-state agreement and including lower bounds from FSA and RMA into the estimation process "led to estimates that were consistent with the expert opinion used by the members of the Agricultural Statistics Board, which enabled the model to be considered for production" (Young and Chen, 2022, p. 890). According to Young (2022, slide 13), models for 13 crops were used for production estimates beginning in 2020, after rounding and review by state field office staff.

Moving to small area estimates based on statistical modeling has had several advantages. First, the models have increased the transparency and reproducibility of the estimate-production process. Second, the models allow calculation of measures of uncertainty, such as variances or coefficients of variation, about the estimates. And third, "the automation of modeling, rounding, and enforcing coherence across geospatial scales has led to a substantial savings in staff time" (Young and Chen, 2022, p. 895).

Young and Chen (2022) described the process of moving these estimates into a production mode:

> Transitioning to these models being the foundation for major survey programs including those associated with the principal federal economic indicators has

[137]Erciulescu, Cruze, and Nandram (2019) accomplished this with a hierarchical Bayesian subarea level model, in which the areas were agricultural statistics districts, and the subareas were counties. As described in Chen, Cruze, and Young (2021), the full process uses three univariate Bayesian subarea models working in concert: (1) A *planted area model* constrained by known minimum administrative totals from FSA and RMA; (2) A *harvested area model* that uses survey harvested-to-planted ratio and transformation to produce coherent harvested area totals; and (3) a *crop yield model* with geographic benchmarking to generate distributions and summaries for crop production totals.

> required substantial changes in the final stages of the NASS processes and a major cultural shift…. For the reviews within the state field offices and by the Agricultural Statistics Board, tools are available to facilitate the review process, but were not designed for the inclusion of modeled estimates or their measures of uncertainty. These tools had to be revised to integrate the modeled estimates into the review process. Following the 2020 growing season, small area models became the foundation for crop county estimates for the 13 nationally reported crops (p. 893).

CONCLUSION 8-1: The National Agricultural Statistics Service has made substantial progress in the difficult process of developing models to produce crop estimates at different levels of geography. Important advances include producing objective estimates with measures of uncertainty.

8.3 MODELING CROP ESTIMATES IN CANADA

NASS county-level crop estimates rely on survey data as the basis for the modeling. In Canada, models have been used to completely replace some surveys. Nkwimi-Tchahou et al. (2022) developed models for estimating mid-season field crop yields from alternative data sources that could potentially be used to replace survey estimates. Traditionally, Statistics Canada collected six field crop surveys each year, three of which asked about crop yields. The July and September surveys dealt with mid-season estimated crop yields, and the November survey asked about actual yields for the season. But mid-season estimates usually underestimated the final values for the actual yields. Nkwimi-Tchahou et al. (2022, slide 2) asked: "Can we make use of alternative data sources to reduce cost and response burden and produce a mid-season set of estimates of equal or better quality than the mid-season surveys provide?"

Data sources they considered as sources of predictor variables included:

- A weekly Normalized Difference Vegetation Index computed from satellite imagery (in general, higher values are associated with higher crop-yield potentials);
- Agroclimatic data about temperature, precipitation, hours of sunshine, soil moisture, and other characteristics; and
- Crop insurance data provided by provincial crop insurance corporations, with information on "the location of the land parcel, what crop is being grown, the acreage of each crop sown and, after the growing season, the resulting yield" (Nkwimi-Tchahou et al., 2022, slide 4).

Previous research reported in Brisbane and Mohl (2014), Reichert et al. (2016), and Statistics Canada (2020b) investigated models that could be used to produce mid-season estimates of crop yield and production that could, potentially, replace estimates from the September Farm Survey. Predictor variables included Normalized Difference Vegetation Index and agroclimatic data available in August, as well as information from the July Farm Survey; these models did not include crop insurance data. Accuracy of estimates was evaluated by comparing estimates to final yields from the November survey. Reichert et al. (2016, p. 11) found that "estimates produced by the yield model were comparable to those produced by the September Farm Survey in terms of relative difference from the November Farm Survey estimates for the 15 crops modelled." As a result, Statistics Canada decided to replace the

September Farm Survey with estimates of field crops from the model, resulting in less burden on survey respondents and reduced costs, as well as earlier publication of mid-season estimates. Reichert et al. (2016, p. 12) noted that this "replacement of a statistical field crop survey with a remote sensing model-based administrative approach is a first for any statistical agency worldwide."

Statistics Canada researchers then explored whether both the July and September surveys could be eliminated (Nkwimi-Tchahou et al., 2022). They dropped July yield as an explanatory variable and included crop insurance data. Crop insurance data presented additional challenges for model building. The first challenge was acquiring access to the data from data providers, and data were not available for all provinces. Moreover, the data structure varied across provinces, with some provinces having more information than others. Undercoverage was also a challenge (as with the USDA's RMA data; see Section 8.1) because not all crops are insured. Finally, when multiple crops were grown within a parcel, insurance data did not tell where, within the parcel, each type of crop was grown, which prevented associating its exact Normalized Difference Vegetation Index.

The model with crop insurance data was first studied with data from Manitoba. To simplify modeling, mixed-crop parcels were dropped from the model, and estimates were adjusted to compensate. It was also assumed that uninsured crops have a similar yield as insured crops. Technical details of the model, as well as estimates and their coefficients of variation, are given in Statistics Canada (2020a).

Nkwimi-Tchahou et al. (2022) reported that, despite the challenges in acquiring and standardizing data, the July estimates from the model were much closer to the November values than estimates from the July survey. Furthermore, using the models, they could publish estimates for additional, less-common, crops that could not be published from survey-based estimates. Nkwimi-Tchahou et al. (2022, slide 16) concluded that "[t]he methodology has shown to be a good replacement for mid-season surveys estimates," in particular because of the cost savings and reduced burden on survey respondents.

One challenge in relying entirely on model-based estimates is the assumption that the relationship between the predictor variables and the response variable is the same for the predicted years as it was for the dataset used for model development. Nkwimi-Tchahou et al. (2022, slide 16) noted that the model had more difficulty in extreme years such as 2021, a year of severe drought for Alberta, Saskatchewan, and Manitoba. In addition, the models "still generally underestimate the values from the end of season survey." They suggested that adding variables to the models, or using machine-learning methods, might further improve predictions.

The modeling efforts of Statistics Canada demonstrate the promise of using satellite imagery along with administrative records data for producing crop estimates that could replace estimates from surveys. As a result, Statistics Canada was able to reduce the number of Field Crop Surveys from six to four (March, June, November, and December), while relying on model-based estimates of yields and production based on satellite imagery in July and September.[138]

8.4 OPPORTUNITIES FOR IMPROVING AGRICULTURAL STATISTICS

The Data Foundation and AGree Initiative (2022) argued that farmers today face unprecedented challenges including supply chain disruptions and extreme weather events, and

[138]https://www23.statcan.gc.ca/imdb/p2SV.pl?Function=getSurvey&SDDS=3401

that timely and accurate data are essential for addressing critical issues related to food and agriculture:

> Modernizing the national data infrastructure for the agricultural sector is the linchpin to provide critical agricultural insights, improve the effectiveness of farm bill programs, and deliver better value for farmers and taxpayers. Harnessing existing data from government, industry, and individual sources has the potential for farmers to work in a more productive, streamlined manner and economically empower rural America (p. 5).

Many opportunities exist for continued improvement of the accuracy, timeliness, detail, and transparency of agricultural statistics through the use of multiple data sources. In the short term, continued research on small area models is likely to be fruitful in producing even more accurate estimators, through inclusion of additional predictor variables (perhaps acquired from new data sources) and new developments from statistical research. Other approaches to improving small area models could also be considered, including further investigations into the properties of the datasets used as model inputs, or exploring groupings of counties other than agricultural statistics districts (NASEM, 2017b, p. 95). Young (2022, slide 16) reported that NASS is investigating the use of drones and in-situ sensors to provide data, although establishing a nationwide system would be costly.

There are also opportunities for continuing to improve data equity for agriculture. Presenting an analysis of U.S. farm owners, operators, and workers by race, ethnicity, and gender, Horst and Marion (2018, p. 14) noted the importance of producing statistics that are disaggregated by these characteristics, and concluded: "Survey data should also enable intersectional analysis across race, ethnicity and gender, at national, regional state and county-levels…. We also urge collection of more detailed demographic data following emerging best practices."

Remotely sensed data can provide information about which crops are grown, but not about whose crops they are. Survey data or administrative records are needed to answer questions about demographic characteristics of farm owners and workers, and about impacts of USDA programs on small-scale, female, minority, and new farmers. Roberts and Hernandez (2021, p. 4) argued that it is important for population groups to have more than mere representation in the data and "that there is a compelling need to improve the participation of women, people living with disabilities, and other marginalized groups in all aspects of open data for agriculture and nutrition."

For county-level crop estimates, equity aspects could be explored by comparing measures of uncertainty about model inputs and outputs with county-level statistics about poverty, race, ethnicity, and other characteristics calculated from the decennial census or American Community Survey. An important part of data equity is identifying areas with less accurate estimates and taking steps to improve those estimates. As with the income studies in Section 5.4, it may be possible to use administrative records to study survey measurement properties and nonresponse bias.

In the longer term, improvements in data quality from non-survey sources may reduce dependence on surveys in the future. One promising area is exploring the potential for using private-sector precision-agriculture data. This would involve investing in an infrastructure for using such data that includes data standards, system interoperability, incentives for data holders to provide their data, cybersecurity, and consideration of data equity. The Data Foundation and

AGree Initiative (2022, p. 3) described eight attributes that would be key for this infrastructure: "farmer and public trust, privacy and confidentiality protections, independence, data acquisition, scalability, stable funding, oversight and accountability, and intergovernmental support." A pilot study, in which a probability sample of farm operators was selected to supplement or replace their survey data with data from internal operations, would provide information about ways of using these data to shift burden away from survey respondents.

With the increasing availability of satellite remote sensing, on-the-ground sensor networks, and social media, there is a great opportunity to improve agricultural statistics by combining these data sources at fine spatial and temporal scales. The spatial and temporal resolution of these data sources tend to become increasingly detailed as technology advances (Wang and Goodchild, 2019). While rapid change and the variety of such data sources often translate into higher uncertainty for analysis results and can affect scientific reproducibility, there are new opportunities for geospatial analysis and statistical approaches to support scalable data integration with adequate uncertainty quantification (Wang 2016). Recent advances in artificial intelligence and machine learning provide an opportunity to harness diverse data sources for improving prediction of crop types and yields at various spatial and temporal scales (Cai et al., 2018; Jiang et al., 2019). Integration of such advances with cyberinfrastructure and cyber-based geospatial information systems and science (cyberGIS) is important to the data-intensive transformation of national agricultural statistics leading to more intelligent, robust, and transparent outcomes (Lyu et al., 2022).

CONCLUSION 8-2: Remotely sensed data have great potential for improving agricultural production models. The resolution and quality of the data are important considerations when choosing appropriate geographic units for modeling and analysis. Private-sector data, such as data from precision agriculture, could also be of value if data-sharing mechanisms that protect privacy can be developed. Data sharing could be improved by cross-agency cooperation to develop and use interoperable geographic units, and by development of quality standards for non-survey data.

BOX 8-1 Selected Recommendations from the National Academies of Sciences, Engineering, and Medicine Report *Improving Crop Estimates by Integrating Multiple Data Sources*

RECOMMENDATION 2-1: The National Agricultural Statistics Service should evolve the Agricultural Statistics Board role from one of integrating multiple data sources to one of reviewing model-based predictions; macro-editing; and ensuring that models are continually reviewed, assessed, and validated.
RECOMMENDATION 2-2: The National Agricultural Statistics Service should achieve transparency and reproducibility by developing, evaluating, validating, documenting, and using model-based estimates that combine survey data with complementary data in accordance with Office of Management and Budget standards.
RECOMMENDATION 2-3: The National Agricultural Statistics Service (NASS) should adopt and use the following publication standard:

- County-level estimates may be withheld to protect confidentiality.
- County-level estimates may be withheld because NASS deems them unreliable for any use, based on its measure of uncertainty.
- All other county-level estimates will be published, along with their measures of uncertainty.

RECOMMENDATION 2-4: The National Agricultural Statistics Service should develop and publish uncertainty measures for county-level estimates.
RECOMMENDATION 2-8: The National Agricultural Statistics Service should adopt the Farm Services Agency's Common Land Unit as its basic spatial unit.
RECOMMENDATION 2-9: The National Agricultural Statistics Service should be prepared to maintain alternative geospatial field-level boundaries (e.g., resource land units and precision agriculture measurements) in its databases to facilitate completing the geospatially referenced farm-level database.
RECOMMENDATION 3-5: The National Agricultural Statistics Service should develop a precision agriculture reporting option for the County Agricultural Production Survey/Acreage, Production, and Stocks survey system. Farmers who reported relevant precision agriculture data would either not receive an additional survey form or receive one that was simplified and easy to use.
RECOMMENDATION 3-8: The National Agricultural Statistics Service should explore collaboration with other U.S. Department of Agriculture agencies that are actively involved in remote sensing applications to obtain access to data with finer spatial resolution and possibly also to share in the costs of processing those data.
RECOMMENDATION 3-9: The National Agricultural Statistics Service (NASS) should keep abreast of emerging data sources; how they are used; and how they might be used to improve county estimates, especially of yield. Based on a careful evaluation, NASS might consider purchasing data.

SOURCE: NASEM (2017b, pp. 3–4).

[END BOX 8-1]

9. Combining Data Sources for National Statistics: Next Steps

In this series of reports, the Committee on National Statistics (CNSTAT) is laying out a vision for a reimagined data infrastructure—one that relies on multiple data sources in addition to probability surveys—for generating official statistics in the United States. The first report, *Toward a 21st Century National Data Infrastructure: Mobilizing Information for the Common Good* (NASEM, 2023), articulated key attributes of the envisioned infrastructure (see Box 1-2).

This report explored implications of using multiple data sources for expanding or replacing information currently collected in major survey programs. The panel examined recent activities in building frames at the U.S. Census Bureau, and explored aspects of current and potential future practices for combining data in four areas: income, health, crime, and agriculture. These areas were chosen to illustrate diverse methods, challenges, and uses of data combination.

Household surveys have been a fundamental means of data collection about both income and health, and these topics have also been the subject of detailed administrative data collections. This report discusses how record linkage has been used to improve measurement and to increase the number of data attributes associated with respondents to income and health surveys. Linking income survey responses with Internal Revenue Service (IRS) tax data and transfer program benefits data has provided valuable insights about the accuracy of survey responses and alternative perspectives on key measures such as poverty and income distribution. Linking health survey records with the National Death Index has allowed researchers to evaluate mortality risks associated with health conditions. Record linkage and imputation have also enabled researchers to make use of large administrative databases such as Medicare claims data to produce statistics that are disaggregated by race and ethnicity. In these usages, the administrative data contain records for everyone participating in the program, but often do not contain accurate (or any) information about race, ethnicity, or other characteristics for which disaggregated statistics are desired. Linking with a source such as the decennial census or the American Community Survey attaches that information to the administrative records, and can identify characteristics of people who are eligible for the program but do not participate.

Crime statistics present additional challenges for integrating data. One of the major data sources, the National Crime Victimization Survey, is a household survey. The other major source, the Uniform Crime Reporting Program, compiles crime statistics from data submitted by states and individual law enforcement agencies. This program faces challenges similar to those of other programs that compile administrative records from states: missing data from states and agencies that do not make submissions, the need to assess and improve quality of data that are supplied, and the need to resolve measurement differences among data suppliers.

Obtaining accurate and timely statistics for agriculture exemplifies some of the challenges faced by establishment surveys. The nature of agricultural statistics also opens opportunities to rely more heavily on data from satellites and sensors in addition to administrative records. There is also potential to make use of the detailed data that many farm operators and agribusinesses collect in their precision agriculture programs. This report focuses on the use of small area models to combine data from various sources for producing county-level crop estimates.

Of course, statistical agencies in these and other areas have done a great deal of additional work on combining data sources. This report does not explore other topics or survey programs, but the examples in the report illustrate challenges and opportunities for other subject areas as well.

9.1 THEMES FOR COMBINING DATA

Each example studied in this report presents unique challenges and opportunities, but the examples share some common themes.

Multiple Data Sources Can Add Value for Official Statistics and Research

Some information, such as opinions and personal experiences, can be collected only through surveys. But there are growing demands for more timely, more granular, and more accurate data on an ever-increasing number of topics. Administrative records and other data sources, either combined with or in place of surveys, can help meet those demands.

Administrative records offer four main benefits for contributing to official statistics. The first benefit is the sheer size of many administrative datasets. Income tax records contain information about every tax filer; a state's dataset for the Supplemental Nutrition Assistance Program contains records for all residents participating in the program; the National Death Index contains information on almost all deaths occurring after 1979. Second, some administrative datasets may have information on population members not represented in surveys, such as persons in nursing homes or survey nonrespondents. Third, administrative data are already being collected for other purposes, so the only costs for their use involve acquiring them, studying and documenting their properties, and repurposing them for producing statistics. Fourth, administrative data provide alternative perspectives on concepts measured in surveys, and thus can contribute to improved understanding of the measures in both data sources.

The previous National Academies of Sciences, Engineering, and Medicine report in this series (NASEM, 2023) explored the potential of using private-sector data to produce official statistics. Challenges of using private-sector data for official statistics are greater than the challenges of using government-collected administrative records, in part because of the limited history of public-private data cooperation. However, private-sector data such as those collected through precision agriculture programs or private health insurance companies could potentially improve federal statistics and create new data resources for social and economic research—if these data can be shown to be reliably available, accurate, and cost-effective sources of information.

There are multiple ways to take advantage of alternative data sources (see Chapter 2). When data sources contain high-quality information for identifying individual entities, data records can be linked. Data linkage is not the only way to combine data sources, however. Statistical models can be used to combine data for individuals, or to combine statistics for geographic areas or population subgroups. Small area models (discussed in Chapters 2, 7, and 8) can be a cost-effective way of providing useful estimates for small geographic areas because such models can "borrow strength" from similar areas and make use of correlated data from administrative records and other sources. For all these methods, however, the quality of estimates depends on the quality of the individual input data sources and the statistical properties of methods used to combine them.

Probability surveys have important strengths that in many cases cannot be entirely replaced by administrative records or information from other existing sources. These strengths range from the probabilistic design itself to the collection of information that can only be obtained by asking a sample member directly. Additional research is needed to identify specific ways that data from other sources can add value to probability surveys, for example through providing information to improve survey design and measurement, augmenting survey information through data linkage, or reducing respondent burden.

Quality of Integrated Data and Statistics

Multiple data sources show great promise for improving official statistics and enhancing research, but using multiple sources is more complicated than producing statistics from a single source. It is challenging to evaluate the quality of data from a single source and to assess how various sources of uncertainty affect statistics. Evaluating the quality of statistics produced from combined data sources is even more challenging.

The Federal Committee on Statistical Methodology (2020) discussed factors that affect the accuracy of integrated data. These factors include the contributors to error from each source: sampling error (for surveys), undercoverage, missing data, measurement error, and processing error. Standard procedures exist for reporting sampling error for surveys, but assessing bias from undercoverage and nonresponse is much more challenging. Statistics from administrative records are usually reported without measures of uncertainty (as with Uniform Crime Reports through 2020), but they are affected by undercoverage, missing data, and measurement and processing error. For example, tax records from the IRS do not include everyone in the population, and certain types of income, such as self-employment income, may be underreported (see Chapter 5).

Additional factors affect accuracy of statistics computed from combined data sources. Linkage error can result, for example, in appending the wrong person's race to a data record, or in coding a person as living when in fact that person is in the National Death Index but the link was missed. Harmonization error, which arises when sources have different units or definitions for data elements (e.g., pixels in satellite data might not match up with farms or fields in another dataset; data sources might report information for nonsynchronous time periods, or sources may use different definitions for seemingly identical concepts) can lead to bias in estimates. Modeling error, which can occur when an imputation or small area estimation model is a poor fit for part of the population, can cause model predictions to be inaccurate.

Combining data sources also affects other dimensions of data quality (see Figure 1-1). An administrative data source might have information that is more granular than survey data, but the data might not be available to the statistical agency soon enough to produce timely statistics. Many survey programs produce public-use datasets that can be downloaded from the internet, but administrative records are often available only to approved researchers in restricted settings (if available at all). To assess the overall "fitness for use" of statistics computed from combined sources, one must understand the purposes for which each dataset was created, the populations covered, the quality and limitations of each data element, and the properties of the data-combination method used. Some of the elements in a data source may have limitations that make them unsuitable for use as outcome variables, but could be deemed useful for other purposes, such as nonresponse adjustment or small area estimation.

A previous National Academies report on combining data sources (NASEM, 2017c, p. 2) recommended: "Federal statistical agencies should systematically review their statistical

portfolios and evaluate the potential benefits and risks of using administrative data." As part of ongoing quality-improvement programs, such systematic reviews could also consider procedures to take advantage of new data sources as well as changes in existing data sources.[139]

The U.S. Office of Management and Budget (OMB, 2002, 2019b) provided guidance to federal agencies for implementing the Information Quality Act.[140] During pre-dissemination reviews of data products, "each agency should consider the appropriate level of quality for each of the products that it disseminates based on the likely use of that information" (OMB, 2019b, p. 2). Additionally, agencies should "provide the public with sufficient documentation about each dataset released to allow data users to determine the fitness of the data for the purpose for which third parties may consider using it" (OMB, 2019b, p. 4). As work on combining data sources progresses, it is important to continue to invest in improving the individual data sources—probability surveys, administrative records, and other data—that feed into a new data infrastructure.

In some cases, metrics and standards used to evaluate survey data may be adapted to apply to other data sources, but new methods and standards are needed to evaluate the quality of statistics produced from multiple sources. The first report in this series provided examples of standards that would be useful for a new data infrastructure (NASEM, 2023, Appendix 3B). The large volume of data from alternative sources could be further mined to build analytics that may provide additional insights into data-quality issues and that could be used to guide data collections that are consistent, reliable, and aligned with relevant fitness-of-use criteria.

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.

Transparency and Documentation

CNSTAT's *Principles and Practices for a Federal Statistical Agency* emphasized the importance of transparency and documentation of data products:

> Federal statistical agencies must have credibility with those who use their data and information. The value of a statistical agency rests fundamentally on the accuracy and credibility of its data products. Because few data users have the resources to verify the accuracy of statistical information, users rely on an agency's reputation to disseminate high quality, objective, and useful statistics in an impartial manner (NASEM, 2021b, p. 31).
>
> A statistical agency must be transparent about how it acquires data and produces statistics and be open about the strengths and limitations of its data…. Openness requires that statistical releases from an agency include a full description of the purpose of the program; the methods and assumptions used for data collection,

[139] *Principles and Practices for a Federal Statistical Agency* (NASEM, 2021b, p. 43) listed continual improvement and innovation as one of the five principles for federal statistical agencies: "Federal statistical agencies must continually seek to improve and innovate their processes, methods, and statistical products to better measure an ever changing world."

[140] Treasury and General Government Appropriations Act of 2001, p. 106–554, § 515(a) (2000) (as codified at 44 U.S.C. § 3516, note).

> processing, and estimation; information about the quality and relevance of the data; analysis methods used; and the results of research on the methods and data (NASEM, 2021b, p. 95).

Chapter 7 of the National Academies' report *Transparency in Statistical Information for the National Center for Science and Engineering Statistics and All Federal Statistical Agencies* (NASEM, 2022e) described best practices for documenting, retaining, releasing, and archiving data. Tables 7.1–7.6 in that report outlined that panel's recommendations regarding the information that a statistical program should retain or archive, the documentation that should be available internally for program staff, and the documentation that should be made available to the public. Table 7-4 of that same report focused on information that should, in that panel's opinion, be made available to the public when record linkage is used:

> A description of the specific data files that were matched should be provided routinely as part of the technical reports or versioned data documentation. Study-specific information and technical reports should be made available to the public. A description of the techniques used for record linkage, the variables used to match on, and a description of how the matching algorithm is implemented, including how uncertain matches are treated, should be made available to the public as part of technical reports on data quality. If available, the estimated error rates for the record linkage routine in this environment should be provided, and if not available, any information on the quality of such a match should be provided instead (NASEM, 2022e, p. 161).

When preparing the current report, this panel found that the amount of detail provided in documentation varies across data collections. Documentation about datasets and data integration from the U.S. Census Bureau, the U.S. Bureau of Labor Statistics, and the U.S. National Center for Health Statistics was relatively easy to find on agency websites. Methodology reports for surveys provided well-organized and detailed descriptions of data collection, processing, and estimation procedures. The panel found similar high-quality documentation for administrative records systems coordinated by these agencies, such as the National Vital Statistics System. Methodology reports from these agencies could serve as models for other agencies that are developing documentation for current programs, and such reports could be a starting point for developing documentation guidelines to assess fitness for use and to address data-equity concerns.

> **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.**

Data Equity

The use of multiple data sources to advance data equity is a major theme of this report. As Leary (2022) concluded from the workshop presentations: "It is clear that the future in many ways is equitable data science, and that equitable access to government programs and services

requires, as we've heard, data that are timely, appropriately granular, and as [Robert] Santos and his colleagues … put it, 'good enough and fit for purpose.'"

The introduction of probability surveys in the 1930s and 1940s was motivated in part by equity considerations, even though the term "equity" was not featured in the writings of the time. Hansen, Hurwitz, and Madow (1953, p. 9) wrote: "When the determination of the individuals to be included in a sample involves personal judgment, one cannot have an objective measure of the reliability of the sample results, because the various individuals may have differing and unknown chances of being drawn." When the sampling frame is complete and there is no nonresponse, every population member has a known probability of being in the probability sample. This guarantees that the sample is representative of all subpopulations and thus promotes representation equity. The survey designer can "[d]ecide what information is really needed," define the population of interest, and lay plans "for eliciting clear, intelligible information" (Deming, 1950, p. 5). The survey instrument can measure categories for which disaggregated statistics are desired, promoting feature equity. Even when there is nonresponse, a probability sample has the advantage that the initial sample is selected randomly and is thus not subject to sources of bias that can affect other sample-selection methods.

In the panel's view, probability samples will continue to have an important role for producing equitable and representative data in a new data infrastructure. Alternative data sources, however, can enhance data-equity aspects of survey programs. In some situations, the information currently collected in a survey can be obtained more efficiently, and with more granularity, from another source. Chen (2022, slide 13) highlighted the potential of "[s]caling up the use of objective measurement technologies" for promoting data equity. Administrative data can contribute to statistics about small geographic areas or small demographic groups, and may make it possible to produce statistics for previously unstudied populations by increasing their representation. Some populations, such as persons in nursing homes or prisons, are excluded from many surveys but included in some administrative records. Better representation is also important when data are used to develop algorithms to make decisions about hiring, creditworthiness, criminal justice, or medicine (see Box 3-1). Beyond that, using multiple data sources can help identify areas in which subpopulations are underrepresented or mismeasured in surveys or administrative data sources. Record linkage can add variables needed for producing statistics that are disaggregated by race, ethnicity, or other characteristics measured in a linked data source.

While combining data sources can enhance knowledge about subpopulations, there is also the potential that combining data will increase bias. Records with less information available for linkage are more likely to have linkage errors, and linkage rates vary by participants' age, gender, race, ethnicity, and health and socioeconomic status (see Chapters 2 and 3; Bohensky et al., 2010). Small area and imputation models may also be less accurate for certain population subgroups and geographic areas.

Groves (2022) emphasized that concern about data equity has to be explicitly addressed in day-to-day practice. This includes equity of measurement, applicability of concepts, and coverage across diverse subgroups. In the panel's judgment, data-equity considerations should be a key component of data-collection planning and of regular program reviews. Methodology reports for surveys and other data sources typically include assessment of the quality of the data for producing national statistics. As stated in Chapter 3, several of these quality dimensions map to equity aspects. Addressing equity issues in data documentation will promote transparency and enhance data equity across the federal statistical system (see Conclusion 3-3).

Improving data equity across the federal statistical system will be challenging and will require a broad-based approach that integrates perspectives of federal statistical agencies, other data producers, data users, and community members. Possible short-term activities include:

- Developing standards for equitable data, such as revising U.S. Office of Management and Budget (1997) standards for collecting data on race and ethnicity (currently underway; see Box 3-3) and implementing best practices for measuring sexual orientation and gender identity (NASEM, 2022c);
- Adding standardized items to surveys and administrative data collections to measure characteristics for which disaggregated statistics are desired and to facilitate linkage;
- Increasing subpopulation sample sizes in selected surveys;
- Facilitating increased federal-state-local data sharing; and
- Researching equity impacts of data-collection and record-linkage methods (including investing in training necessary for equity assessment).

In the longer term, new statistical methods may need to be developed to promote data equity when combining data. Specifically, research is needed on methods for producing disaggregated statistics for small population groups while protecting confidentiality, and for ensuring that informed consent to data collection includes all possible uses of the data (see Box 3-5).

Wardell (2022) noted that equitable data is still a new concept and that it encompasses much more than just adding a new variable to a data collection. Executive Order 13895 (2021) charges agencies to understand disparities in the programs they administer and to identify roadblocks for accessing federal services. Data are essential for understanding programs' impact and reach, and can be used to establish "feedback loops" where data inform changes to programs and services at federal, state, and local levels. Wardell (2022) also stressed the importance of building capacity for robust equity assessment—bringing in people with appropriate training and skill sets to work with federal agencies and also building capacity and infrastructure at the local level.

9.2 FUTURE CHALLENGES AND OPPORTUNITIES

This report, building on the framework for a vision of a new data infrastructure in the previous report in this series (NASEM, 2023), concentrates on methods and examples in which using multiple data sources can improve statistics currently collected through major survey programs. Table 9-1 provides a list of all conclusions from this report. There is much work to be done, and future reports in this series will address other aspects of a new data infrastructure: governance and information technology structure, protecting confidentiality, and allowing public use of blended-data products.

For most of the examples in this report, agencies and researchers have already obtained access to the data sources, and the challenges involve how to use them. But of course, one of the primary impediments for using multiple data sources is the difficulty of acquiring or accessing the data. The Uniform Crime Reporting Program exhibits some of the challenges for computing representative statistics when a nonrepresentative part of the population contributes data (see Chapter 7). The previous report in this series (NASEM, 2023) discussed legal issues and

incentives for sharing data to produce official statistics, arguing that organizations holding data will be more likely to share that data if they directly benefit from doing so.

Even when data sources are acquired, however, there is no guarantee that data elements collected now will continue to be collected in the future, or that agencies or private-sector organizations that are willing to contribute data now will keep sharing their data (or, if data are purchased, that the price will remain affordable). If administrative records or private-sector data are used for programs in which it is important to compare statistics over time, continued availability and consistency of information from the data sources are crucial. Federal statistical agencies can play an important role in increasing coordination in this area, both in terms of facilitating access and promoting standard definitions and protocols for measurement (Advisory Committee on Data for Evidence Building, 2022).

Creating useful statistics and data products from combined data sources requires skills in addition to those needed to produce estimates from probability surveys. A new data infrastructure requires investment not only in data sources but also in the people who can work with those data. Section 2.3 lists some of the technical challenges for combining data, and some of the areas of expertise needed to address them. Beyond the technical challenges, there are challenges for promoting data equity and public trust in data, and these areas require additional resources and expertise. Statistical agencies will need investments in personnel, training, and computer infrastructure to take advantage of new data resources.

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.

Groves (2022) noted that the workshop presentations (Appendix A) were the work of pioneers, and that one component of CNSTAT's vision of a redesigned national data infrastructure concerns how to "make this type of work, now done by pioneers, routine rather than cutting-edge."

Today's data world contains amounts of digital information that were inconceivable when the theory of probability sampling was developed in the 1930s. Arora (2022b, p. 24) argued that the ability to use data to answer societal questions is "the real value proposition of a national statistical office. It's not just about putting more data out there. It's trying to make sense of what's happening in society and showing how different parts of it are intricately connected. If we don't do that, someone else will." Groves (2022) concluded the workshop with a vision of the new data infrastructure:

> We are in an unprecedented moment in history—in the history of digital data and information derived from those data…. We seek a vision, in sum, that will protect the privacy of Americans while simultaneously producing for them, and their common good, better statistical information.

[TABLE 9-1 about here]

TABLE 9-1 Report Conclusions

Chapter 2. Types of Data and Methods for Combining Them

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.
CONCLUSION 2-3: Linking survey data with administrative records requires substantial expertise and investment. Decisions need to be made among reasonable alternative methods, and then periodically re-examined as data sources change or new linkage methods are developed. Documentation that assesses the quality of the linkages allows data users to evaluate the possible impact of linkage errors on analyses and to account for uncertainties in the linkage process.
CONCLUSION 2-4: Statistical methods such as small area estimation, imputation, and combining statistics for subpopulations can integrate information from multiple data sources without requiring individual records to be linked.

Chapter 3. Using Multiple Data Sources to Enhance Data Equity

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.

Chapter 4. Creating New Data Resources with Administrative Records

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.
CONCLUSION 4-2: Linking administrative data and sampling frames can enable useful future data linkages for social science research and evidence-based policy analysis. However, combined data sources do not necessarily have either full population coverage for generating national statistics or sufficient sample sizes to investigate differences among population subgroups.
CONCLUSION 4-3: The National Vital Statistics System can serve as a model for assembling state-administered data programs into coordinated, standardized national databases of administrative records that can be linked to other data sources.
CONCLUSION 4-4: Administrative records are a valuable source of information for official statistics and social and economic research. Each administrative records dataset considered for use in creating national statistics needs to be understood in terms of both its original and its proposed uses. This includes assessing the dataset's fitness for use, timeliness, continuing availability, population coverage, measurement of key concepts, and equity aspects.

Chapter 5. Data Linkage to Improve Income Measurement

CONCLUSION 5-1: Comparison of survey data with linked administrative records can provide statistical agencies with valuable information on measurement quality as well as guidance for further investigations and improvements.

Chapter 6. Data Linkage to Supplement Health Surveys

CONCLUSION 6-1: The U.S. National Center for Health Statistics has linked many of its surveys with administrative records datasets, providing valuable resources for investigating long-term health outcomes and promoting evidence-based policy. These linkage procedures and documentation can serve as models for other partnerships between program-oriented and federal statistical agencies.
CONCLUSION 6-2: Longitudinal surveys provide perspectives on individual and household behavior not available in cross-sectional surveys. Data from such longitudinal surveys can be enhanced through data linkages to create new opportunities for social science research.

Chapter 7. Combining Multiple Data Sources to Measure Crime

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.
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.

Chapter 8. Using Multiple Data Sources for County-Level Crop Estimates

CONCLUSION 8-1: The National Agricultural Statistics Service has made substantial progress in the difficult process of developing models to produce crop estimates at different levels of geography. Important advances include producing objective estimates with measures of uncertainty.
CONCLUSION 8-2: Remotely sensed data have great potential for improving agricultural production models. The resolution and quality of the data are important considerations when choosing appropriate geographic units for modeling and analysis. Private-sector data, such as data from precision agriculture, could also be of value if data-sharing mechanisms that protect privacy can be developed. Data sharing could be improved by cross-agency cooperation to develop and use interoperable geographic units, and by development of quality standards for non-survey data.

Chapter 9. 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.

APPENDIX A

Workshop Agenda

Towards a Vision for a New Data Infrastructure for Federal Statistics and Social and Economic Research in the 21st Century

Workshop 2: The Implications of Using Multiple Data Sources for Major Survey Programs

Open Session (Virtual)

Monday, May 16, 2022

11:00 am–11:15 am EDT

Introduction and Welcome

- **Sharon Lohr**, Consensus Panel Chair, *Arizona State University*
- **Cheryl Eavey**, *National Science Foundation*

11:15 am–1:15 pm EDT

Session 1: Opportunities for Using Multiple Data Sources to Enhance Major Survey Programs

Moderator: **Jean-François Beaumont,** *Statistics Canada*

11:20–11:50 Keynote—Census Bureau Modernization: A New Vision for an Enterprise Approach to Statistical Data: **Robert Santos**, *Director, U.S. Census Bureau*

11:50–12:20 Keynote—Combining Multiple Data Sources to Produce Inclusive Official Statistics: **Anil Arora**, *Chief Statistician of Canada*

12:20–12:35 Discussant: **Joseph Salvo**, *University of Virginia*

12:35–12:50 Discussant: **Haoyi Chen**, *United Nations*

12:50–1:15 Panel/Audience Discussion

1:15 pm–1:35 pm EDT *BREAK*

1:35 pm–3:05 pm EDT

Session 2: Measuring Crime in the 21st Century: A Panel Discussion

Moderator: **Elizabeth Stuart,** *Johns Hopkins University*

1:35–2:35 Panel

Janet Lauritsen, *University of Missouri-St. Louis*
Ramiro Martinez, Jr., *Northeastern University*
Erica Smith, *US Bureau of Justice Statistics*
Derek Veitenheimer, *State of Wisconsin*

2:35–3:05 Panel/Audience Discussion

3:05 pm–3:20 pm EDT *BREAK*

3:20 pm–4:50 pm EDT

Session 3: Improving Agriculture Statistics with New Data Sources
Moderator: **Shaowen Wang,** *University of Illinois Urbana-Champaign*

3:25–3:45 The Crops County Estimates Program: Developing Official Statistics Based on Available Data: **Linda Young**, *U.S. National Agricultural Statistics Service*

3:45–4:05 The Evolution of the Use of Satellite and Administrative Data in Estimating Mid-season Field Crop Yields at Statistics Canada: **Herbert Nkwimi-Tchahou**, *Statistics Canada*

4:05–4:25 Layers and Beyond, Modern Use of Connected Spatial and Non Spatial Datasets to Unlock Insights in R&D: **Martin Mendez-Costabel**, *Bayer Crop Science*

4:25–4:35 Discussant: **Michael Goodchild**, *University of California, Santa Barbara*

4:35–4:50 Panel/Audience Discussion

4:50 pm–5:00 pm EDT

Wrap-up Day 1 (**Sharon Lohr**)

5:00 pm EDT *ADJOURNMENT*

Wednesday, May 18, 2022

11:00 am–11:05 am EDT

Introduction and Welcome
Sharon Lohr, Consensus Panel Chair, *Arizona State University*

11:05 am–12:40 pm EDT

Session 4: Data Linkage for Income and Health Statistics
Moderator: **Judith A. Seltzer,** *University of California, Los Angeles*

11:10–11:30 The National Experimental Well-being Statistics Project*:* **Jonathan Rothbaum**, *U.S. Census Bureau*

11:30–11:50 Realizing the Power of Health Data through Linkages:
Lisa Mirel, *U.S. National Center for Health Statistics*

11:50–12:10 Administrative Data Linkages in the Health and Retirement Study: Social Security and Medicare/Medicaid Data: **Jessica Faul and Helen Levy**, *University of Michigan*

12:10–12:20 Discussant: **Hilary Hoynes**, *University of California, Berkeley*

12:20–12:40 Panel/Audience Discussion

12:40 pm–1:00 pm EDT *BREAK*

1:00 pm–3:00 pm EDT

Session 5: Issues in Data Equity
Moderator: **David Mancuso,** *State of Washington*

1:05–1:25 From Data to Decisionmaking: Using Data and Engagement to Advance Racial Equity: **Steven Brown**, *Urban Institute*

1:25–1:45 Administrative Data and Statistics for Small Race Groups and Other Populations: **Randall Akee**, *University of California, Los Angeles*

1:45–2:05 Diversity in Data: How Data Collection Decisions Help and Harm the Outcome: **Frauke Kreuter**, *Ludwig-Maximilians-University of Munich and University of Maryland*

2:05–2:15 Moderator/Discussant Questions and Reflections: **Margaret Levenstein**, *University of Michigan*

2:15–2:45 "Fireside Chat": **Clarence Wardell,** *Chief Data and Equitable Delivery Officer, Executive Office of the President* and **Kimberlyn Leary**, *Harvard University and Urban Institute*

2:45–3:00 Panel/Audience Discussion

3:00 pm–3:30 pm EDT

Discussion and Wrap-up
Moderators: **Sharon Lohr**, *Arizona State University* and **Robert M. Groves**, Chair of Committee on National Statistics, *Georgetown University*

3:30 pm EDT *ADJOURNMENT*

NOTE: Recordings of each session and presentation slides can be found at: https://www.nationalacademies.org/event/05-16-2022/the-implications-of-using-multiple-data-sources-for-major-survey-programs-workshop

APPENDIX B

Biographical Sketches of Panel Members

SHARON LOHR (chair) is a professor emerita at Arizona State University, where she was Dean's Distinguished Professor of Statistics until 2012. Between 2012–2017, as a vice president at Westat, she developed survey designs and statistical analysis methods for use in transportation, public health, crime measurement, and education. Lohr's research interests include sample surveys, design of experiments, hierarchical models, and combining multiple sources of data. She is the author of numerous research articles as well as the books Sampling: Design and Analysis and Measuring Crime: Behind the Statistics. Lohr is an elected fellow of the American Statistical Association, an elected member of the International Statistical Institute, and the inaugural recipient of the Gertrude M. Cox Statistics Award for contributions to the practice of statistics. Her invited presentations include selection as the Morris Hansen Lecturer and the Deming Lecturer. Lohr has served on two previous National Academies committees: The Panel on Improving Federal Statistics for Policy and Social Science Research Using Multiple Data Sources and State-of-the-Art Estimation Methods and the Panel on the Functionality and Usability of Data from the American Community Survey. She earned her B.S. degree in mathematics from Calvin College, and her Ph.D. in statistics from the University of Wisconsin-Madison.

JEAN-FRANÇOIS BEAUMONT is a senior statistical advisor at Statistics Canada. Over his career, he has conducted development and research projects on topics such as statistical data integration, small area estimation, treatment of missing values in surveys, bootstrap variance estimation and other estimation issues for sample surveys. Beaumont is currently the editor of Survey Methodology, an associate editor for Metron, an elected member of the International Statistical Institute, and was the president of the Survey Methods Section of the Statistical Society of Canada. He has delivered many invited presentations and courses, including the opening address of the Colloque Francophone sur les Sondages, and was a Morris Hansen Lecturer. He obtained a B.S. degree in actuarial science and an M.S. in statistics from Laval University.

LAWRENCE D. BOBO is dean of social science, the W. E. B. Du Bois professor of the social sciences and holds the title of Harvard College professor at Harvard University. He has previously served as chair of the Department of African and African American Studies and currently holds appointments in the Department of Sociology as well as the Department of African and African American Studies. His research focuses on the intersection of social psychology, social inequality, politics, and race. Bobo is an elected member of the National Academy of Science and of the American Philosophical Society and a member of the board of directors and board vice-chair of the American Institutes for Research. He is a Guggenheim fellow, an Alphonse M. Fletcher Sr. fellow, a fellow of the Center for Advanced Study in the Behavioral Sciences, the American Academy of Arts and Sciences, as well as the American Association for the Advancement of Science, and is a Russell Sage Foundation visiting scholar. Bobo was elected the W. E. B. Du Bois fellow of the American Academy for Political and Social

Science. He has held tenured appointments in the sociology departments at the University of Wisconsin, Madison, the University of California, Los Angeles, and at Stanford University where he was the Martin Luther King Jr. Centennial professor, chair of the Program in African American Studies and Director of the Center for Comparative Studies in Race and Ethnicity. He is currently working on the Race, Crime, and Public Opinion project as well as book on racial division and American politics. Bobo has an M.A. and Ph.D. in sociology from the University of Michigan.

MICK P. COUPER is a research professor at the Survey Research Center in the Institute for Social Research at the University of Michigan. His current research interests include survey nonresponse, design and implementation of survey data collection, effects of technology on the survey process, and computer-assisted interviewing, including both interviewer-administered (CATI and CAPI) and self-administered (web, audio-CASI, IVR) surveys. Many of Couper's current projects focus on the design of web, smartphone, and mixed-mode surveys. He has served on National Academies studies including the Panel on Redesigning the Bureau of Labor Statistics Consumer Expenditures Surveys, the Panel on the Research on Future Census Methods, and the Oversight Committee for the Workshop on Survey Automation. Couper has an M.Soc.Sc. in sociology from the University of Cape Town, South Africa, an M.A. in applied social research from the University of Michigan, and Ph.D. in sociology from Rhodes University in South Africa.

HILARY HOYNES is professor of public policy and economics and holds the Haas distinguished chair in economic disparities at the University of California Berkeley where she also co-directs the Berkeley Opportunity Lab. She is an economist who works on poverty, inequality, food and nutrition programs, and the impacts of government tax and transfer programs on low-income families. Hoynes' current work examines how access to the social safety net in early life affects children's later life health and human capital outcomes. She is a member of the American Academy of Art and Sciences and a fellow of the Society of Labor Economists. Hoynes has served as co-editor of the American Economic Review and the American Economic Journal: Economic Policy and is on the editorial board of the American Economic Review: Insights. She served on the National Academies committee on Building an Agenda to Reduce the Number of Children in Poverty by Half in 10 Years and the Federal Commission on Evidence-Based Policy Making. Hoynes received her B.A. in economics and mathematics from Colby College, and her Ph.D. in economics from Stanford University.

KIMBERLYN LEARY began her career as a clinical psychologist working as a practitioner to improve access to diverse communities. Her early work on negotiated transactions in psychotherapy expanded to broader research on negotiation, conflict transformation, and change management. Leary is an associate professor of psychology at the Harvard Medical School/McLean Hospital, an associate professor in the Department of Health Policy and Management at the Harvard T.H. Chan School of Public Health, and she was a lecturer in public policy at the Harvard Kennedy School where she continues to direct executive education program for the Bloomberg Center for Cities and the Bloomberg Harvard City Leadership Initiative. She served as an adviser to the Obama White House from, where she worked with the White House Council on Women and Girls to develop the Advancing Equity initiative (which focused on improving life outcomes for women and girls of color). Leary later served on the

Biden-Harris transition as a member of the Agency Review Team for the Office of National Drug Control Policy. She recently completed a detail to the U.S. Office of Management and Budget as a senior equity fellow and the Domestic Policy Council as a senior policy advisor through an Intergovernmental Personnel Act to help implement President Biden's executive order on equity (Executive Order 13985). She is also a senior vice president at the Urban Institute, where she leads program development and research management initiatives across the Institute's policy and research centers.

DAVID MANCUSO is director of the Research and Data Analysis Division of the Washington State Department of Social and Health Services. He leads a team of approximately 100 researchers and IT professionals performing analytical work across the spectrum of publicly funded social and health services in Washington State. Mancuso's division developed and continues to maintain the agency's Integrated Client Databases—a powerful federated data environment linking Medicaid medical, behavioral health, and long-term care data with social service, criminal justice, housing, child welfare, education, employment, and vital statistics data. He has expertise in quasi-experimental program evaluation, performance measurement and the development of predictive modeling technologies to support intervention targeting and care management in Medicaid delivery systems. Mancuso co-developed the Predictive Risk Intelligence System, the predictive modeling tool supporting physical and behavioral health interventions for Medicaid and dual Medicare-Medicaid beneficiaries in Washington State. He received his Ph.D. in economics from Stanford University.

JUDITH A. SELTZER is research professor and professor emerita of sociology at the University of California, Los Angeles. Previously, she directed the California Center for Population Research at UCLA and was on the faculty of the University of Wisconsin-Madison, where she contributed to the development and implementation of the National Survey of Families and Households. She also was president of the Population Association of America, and she previously served on the Board of Overseers for the General Social Survey. Her research interests include kinship patterns, intergenerational obligations, relationships between nonresident fathers and children, and how legal institutions and other policies affect family change. She is especially interested in kinship institutions that are in flux, such as marriage and cohabitation in the contemporary United States or divorced and nonmarital families. She also explores ways to improve the quality of survey data on families, and in 2013 Seltzer and her colleagues added a module with family rosters to the Panel Study of Income Dynamics to provide new data on U.S. family networks. She served on the CNSTAT Standing Committee on Reengineering Census Operations, the Panel on Residence Rules in the Decennial Census, the Panel on the Design of the 2010 Census Program of Evaluations and Experiments, and the Panel to Review the 2010 Census. She has a B.A. in sociology from Princeton University and both an M.A. and Ph.D. in sociology from the University of Michigan.

ELIZABETH A. STUART is Bloomberg professor of American health in the department of mental health at the Johns Hopkins Bloomberg School of Public Health, with joint appointments in the Department of Biostatistics and the Department of Health Policy and Management. She also serves as executive vice dean for Academic Affairs. Stuart has extensive experience in methods for estimating causal effects for program and policy evaluation, particularly as applied to mental health, public policy, and education. Her primary research interests include designs for

estimating causal effects in nonexperimental settings (such as propensity scores), and methods to assess and enhance the generalizability of randomized trials to target populations. She has received research funding from the National Science Foundation, the Patient Centered Outcomes Research Institute, the Institute of Education Sciences, the W.T. Grant Foundation, and the National Institutes of Health and has served on advisory panels for the National Academies of Sciences Engineering and Medicine, the U.S. Department of Education, and the Patient Centered Outcomes Research Institute. She has received the mid-career award from the Health Policy Statistics Section of the American Statistical Association, the Gertrude Cox Award for applied statistics, Harvard University's Myrto Lefkopoulou Award for excellence in Biostatistics, the Rod Little Lectureship from the University of Michigan Department of Biostatistics, and Society for Epidemiologic Research Marshall Joffe Epidemiologic Methods award. She is a fellow of the American Association for the Advancement of Science as well as the American Statistical Association. Stuart received her Ph.D. in statistics in 2004 from Harvard University.

SHAOWEN WANG is professor and head of the department of geography and geographic information science; and an affiliate professor of the department of computer science, department of urban and regional planning, and school of information sciences at the University of Illinois Urbana-Champaign (UIUC). He also has served as founding director of the CyberGIS Center for Advanced Digital and Spatial Studies at UIUC. His research interests include geographic information science and systems, advanced cyberinfrastructure and CyberGIS, complex environmental and geospatial problems, computational and data sciences, geospatial science and technology, high-performance and distributed computing, and spatial analysis and modeling. He received the National Science Foundation CAREER Award, named a Helen Corley Petit Scholar, Centennial Scholar, and Richard and Margaret Romano Professorial Scholar by UIUC's College of Liberal Arts and Sciences. He served as a member of the Committee on Models of the World for the National Geospatial-Intelligence Agency and as a member of the Board on Earth Sciences and Resources of the National Academies of Sciences, Engineering, and Medicine from. He received a B.S. in computer engineering from Tianjin University, an M.S. in geography from Peking University, and a M.S. in computer science and a Ph.D. in geography from the University of Iowa.

References

Abowd, J.M., Stephens, B.E., Vilhuber, L., Andersson, F., McKinney, K.L., Roemer, M., and Woodcock, S. (2009). The LEHD infrastructure files and the creation of the Quarterly Workforce Indicators. In Dunne, T., Jensen, J.B., and Roberts, M.J. (Eds.), *Producer Dynamics: New Evidence from Micro Data* (pp. 149-230). Chicago: University of Chicago Press.

Abowd, J.M. and Vilhuber, L. (2005). The sensitivity of economic statistics to coding errors in personal identifiers. *Journal of Business and Economic Statistics, 23*(2), 133-152.

Abraham, K.G., Haltiwanger, J.C., Hou, C., Sandusky, K., and Spletzer, J.R. (2021). Reconciling survey and administrative measures of self-employment. *Journal of Labor Economics*, *39*(4), 825-860.

Adeyemo, W. and Batchelder, L. (2021). Advancing Equity Analysis in Tax Policy. Available: https://home.treasury.gov/news/featured-stories/advancing-equity-analysis-in-tax-policy

Addington, L.A. (2019). NIBRS as the new normal: What fully incident-based crime data mean for researchers. In *Handbook on Crime and Deviance* (pp. 21-33). Cham, Switzerland: Springer.

Addington, L.A., and Lauritsen, J.L. (2021). Using national data to inform our understanding of family and intimate partner violence victimization: A review of a decade of innovation. *Feminist Criminology*, *16*(3), 304-319.

Advisory Committee on Data for Evidence Building. (2022). *Advisory Committee on Data for Evidence Building: Year 2 Report.* Available: https://www.bea.gov/system/files/2022-10/acdeb-year-2-report.pdf

Ahrens, K.A., Haley, B.A., Rossen, L.M., Lloyd, P.C., and Aoki, Y. (2016). Housing assistance and blood lead levels: Children in the United States, 2005-2012. *American Journal of Public Health*, *106*(11), 2049-2056.

Akee, R. (2022). Administrative data and statistics for small race groups and other populations. Presentation at the National Academy of Sciences, Engineering, and Medicine Workshop on The Implications of Using Multiple Data Sources for Major Survey Programs. May 18, 2022. Available: https://www.nationalacademies.org/event/05-16-2022/the-implications-of-using-multiple-data-sources-for-major-survey-programs-workshop

Akee, R, Jones, M.R., Porter, S.R., and Simeonova, E. (2020). Hispanic and Asian earnings inequality: New workers and immigrants. *AEA Papers and Proceedings, 110*(May), 442-446.

Akee, R., Jones, M.R., and Porter, S.R. (2019). Race matters: Income shares, income inequality, and income mobility for all U.S. races. *Demography, 56*(3), 999-1021.

Akee, R., Mykerezi, E., and Todd, R. (2020). Business Dynamics on American Indian Reservations: Evidence from Longitudinal Datasets. U.S. Census Bureau Center for Economic Studies Working Paper CES 20-38. Available: https://www.census.gov/library/working-papers/2021/adrm/CES-WP-20-38.html

Ali, B. and Dahlhaus, P. (2022). The role of FAIR data towards sustainable agricultural performance: A systematic literature review. *Agriculture*, *12*(2), 309.

Allen, R. (2008). *Agriculture Counts: The Founding and Evolution of the National Agricultural Statistics Service, 1957-2007.* Washington, DC: U.S. Department of Agriculture. Available: https://www.nass.usda.gov/About_NASS/pdf/agriculture_counts.pdf

Angwin, J., Larson, J., Mattu, S., and Kirchner, L. (2016). Machine bias: There's software used across the country to predict future criminals. And it's biased against blacks. Available: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Aram, J., Zhang, C., Golden, C., Zelaya, C. E., Cox, C. S., Ye, Y., and Mirel, L. B. (2021). Assessing linkage eligibility bias in the National Health Interview Survey. *Vital and Health Statistics, 2*(186), 1-28.

Arias, E. (2021). Race crossover in longevity. In Gu, D., and Dupre, M.E. (Eds.), *Encyclopedia of Gerontology and Population Aging.* Cham, Switzerland: Springer, pp. 4119-4128.

Arias, E., Heron, M.P. and Hakes, J.K. (2016). The validity of race and Hispanic origin reporting on death certificates in the United States: An update. *Vital and Health Statistics*, *2*(172), 1-21.

Arias, E., Xu, J., Curtin, S., Bastian, B., and Tejada-Vera, B. (2021). Mortality profile of the non-Hispanic American Indian or Alaska Native Population, 2019. *National Vital Statistics Reports, 70*(12), 1-27.

Arora, A. (2022a). Combining multiple data sources to produce inclusive official statistics. Presentation at the National Academy of Sciences, Engineering, and Medicine Workshop on The Implications of Using Multiple Data Sources for Major Survey Programs. May 16, 2022. Available: https://www.nationalacademies.org/event/05-16-2022/the-implications-of-using-multiple-data-sources-for-major-survey-programs-workshop

Arora, A. (2022b). Modernizing government statistics for the 21st Century. *Amstat News*, *541*(July), 24-26.

Ashby, M.P.J. (2019). Studying crime and place with the Crime Open Database. *Research Data Journal for the Humanities and Social Sciences*, *4*(1), 65-80.

Ashby, M.P.J. (2020). Initial evidence on the relationship between the coronavirus pandemic and crime in the United States. *Crime Science,* 9(1), 1-16.

Asher, J., Resnick, D., Brite, J., Brackbill, R., and Cone, J. (2020). An introduction to probabilistic record linkage with a focus on linkage processing for WTC registries. *International Journal of Environmental Research and Public Health*, *17*(18), 6937.

Banks, D., Planty, M., Couzens, L., Lee, P., Brooks, C., Scott, K.M., and Whyde, A. (2019). Arrest-related deaths program: Pilot study of redesigned survey methodology. *Technical Report NCJ 252675*. Washington, DC: U.S. Bureau of Justice Statistics. Available: https://bjs.ojp.gov/content/pub/pdf/ardppsrsm.pdf

Barnett-Ryan, C. and Berzofsky, M. (2022). Establishing new methods for estimating crime in the U.S.—The transition to incident-based crime reporting through NIBRS [February 17 Webinar]. U.S. Department of Justice, Office of Justice Programs. Available: https://bjs.ojp.gov/media/video/66496

Barnett-Ryan, C. and Swanson, G. (2017). The role of state programs in NIBRS data quality: A case study of two states. *Journal of Contemporary Criminal Justice, 24*(1), 18-31.

Beals, M., Jamieson, A., Rogers, J., and Raiha, N. (2021). *Our Clients Speak: Results from the Social and Health Services Client Survey*. Olympia, WA: Washington State Department of Social and Health Services. Available: https://www.dshs.wa.gov/sites/default/files/rda/reports/research-11-259.pdf

Bearer-Friend, J. (2019). Should the IRS know your race? The challenge of colorblind tax data. *Tax Law Review, 73*(1), 1-68.

Beaumont, J.-F. (2020). Are probability surveys bound to disappear for the production of official statistics? *Survey Methodology, 46*(1), 1-28.

Bee, C.A. and Mitchell, J. (2017). Do Older Americans Have More Income Than We Think? US Census Bureau SESHD Working Paper #2017-39. Available: https://www.census.gov/library/working-papers/2017/demo/SEHSD-WP2017-39.html.

Bee, A. and Rothbaum, J. (2019). The Administrative Income Statistics (AIS) project: Research on the use of administrative records to improve income and resource estimates. U.S. Census Bureau. SEHSD Working Paper 2019-36. Available: https://www.census.gov/library/working-papers/2019/demo/SEHSD-WP2019-36.html

Bee, C.A., Gathright, G.M.R., and Meyer B.D. (2015). Bias from unit non-response in the measurement of income in household surveys. Available: https://harris.uchicago.edu/files/jsm2015_bgm_unit_non-response_in_cps.pdf.

Bell, W.R., Basel, W.W., and Maples, J.J. (2016). An overview of the U.S. Census Bureau's Small Area Income and Poverty Estimates Program. In Pratesi, M. (Ed.), *Analysis of Poverty Data by Small Area Estimation* (pp. 349-378). Hoboken, NJ: Wiley.

Benedetto, G., Haltiwanger, J., Lane, J., and McKinney, K. (2007). Using worker flows to measure firm dynamics. *Journal of Business and Economic Statistics*, *25*(3), 299-313.

Benedetto, G., Stinson, M.H., and Abowd, J.M. (2013). The creation and use of the SIPP Synthetic Beta. Available: https://www.census.gov/content/dam/Census/programs-surveys/sipp/methodology/SSBdescribe_nontechnical.pdf

Benzeval, M., Bollinger, C.R., Burton, J., Couper, M.P., Crossley, T.F., and Jäckle, A. (2020). Integrated Data: Research Potential and Data Quality. Institute for Social and Economic Research, University of Essex: Understanding Society. Working Paper Series No. 2020-02. Available: https://www.understandingsociety.ac.uk/research/publications/525974

Berzofsky, M., Liao, D., Couzens, G.L., Smith, E.L., and Barnett-Ryan, C. (2022). *Estimation Procedures for Crimes in the United States Based on NIBRS Data*. Report NCJ-305108. Washington, DC: Bureau of Justice Statistics and Federal Bureau of Investigation. Available: https://www.ojp.gov/ncjrs/virtual-library/abstracts/estimation-procedures-crimes-united-states-based-nibrs-data

Binette, O. and Steorts, R.C. (2022). (Almost) all of entity resolution. *Science Advances, 8*(12), 1-14. Available: https://www.science.org/doi/abs/10.1126/sciadv.abi8021

Bird, S.M. and King, R. (2018), Multiple systems estimation (or capture-recapture estimation) to inform public policy. *Annual Review of Statistics and its Application, 5*(1), 95-118.

Black, D.A., Hsu, Y.C., Sanders, S.G., Schofield, L.S., and Taylor, L.J. (2017). The Methuselah effect: The pernicious impact of unreported deaths on old-age mortality estimates. *Demography*, *54*(6), 2001-2024. Available: https://www.thieme-connect.com/products/ejournals/html/10.1055/s-0039-1693129

Black, M.C., Basile, K.C., Breiding, M.J., Smith, S.G., Walters, M.L., Merrick, M.T., Chen, J., and Stevens, M.R. (2011). *The National Intimate Partner and Sexual Violence Survey (NISVS): 2010 Summary Report*. Atlanta, GA: National Center for Injury Prevention and Control, Centers for Disease Control and Prevention.

Bohensky, M.A., Jolley, D., Sundararajan, V., Evans, S., Ibrahim, J., and Brand, C. (2011). Development and validation of reporting guidelines for studies involving data linkage. *Australian and New Zealand Journal of Public Health*, *35*(5), 486-489.

Bohensky, M.A., Jolley, D., Sundararajan, V., Evans, S., Pilcher, D.V., Scott, I., and Brand, C.A. (2010). Data linkage: a powerful research tool with potential problems. *BMC Health Services Research*, *10*(1), 1-7.

Bohme, F.G. (1989). *200 Years of Census Taking: Population and Housing Questions, 1790-1990*. Washington DC: U.S. Government Printing Office.

Bollinger, C.R., Hirsch, B.T., Hokayem, C.M., and Ziliak, J.P. (2018). The good, the bad and the ugly: Measurement error, non-response and administrative mismatch in the CPS. Available: https://www.cemmap.ac.uk/wp-content/legacy/uploads/GoodBadUglyFull.pdf

Bollinger, C.R., Hirsch, B.T., Hokayem, C.M., and Ziliak, J.P. (2019). Trouble in the tails? What we know about earnings nonresponse thirty years after Lillard, Smith, and Welch. *Journal of Political Economy*, *127*(5), 2143-2185.

Boman, J.H., and Gallupe, O. (2020). Has COVID-19 changed crime? Crime rates in the United States during the pandemic. *American Journal of Criminal Justice*, *45*(4), 537-545.

Bond, B., Brown, J.D., Luque, A., and O'Hara, A. (2014). The nature of the bias when studying only linkable person records: Evidence from the American Community Survey. CARRA Working Paper No. 2014-08. Washington, DC: Center for Administrative Records Research and Applications. Available: https://www.census.gov/content/dam/Census/library/working-papers/2014/adrm/carra-wp-2014-08.pdf

Boruch, R.F. (2011). *Administrative Record Quality and Integrated Data Systems.* Actionable Intelligence for Social Policy, University of Pennsylvania. Available: https://aisp.upenn.edu/wp-content/uploads/2015/09/0033_12_SP2_Record_Quality_Data_Systems_000.pdf

Boryan, C., Yang, Z., Mueller, R., and Craig, M. (2011). Monitoring US Agriculture: The US Department of Agriculture, National Agricultural Statistics Service, Cropland Data Layer Program. *Geocarto International*, *26*(5), 341-358.

Boryan, C.G. and Yang Z. (2021) Geospatial land use and land cover data for improving agricultural area sampling frames. In Di L. and Üstünda B. (Eds.), *Agro-geoinformatics: Theory and Practice*. Cham, Switzerland: Springer.

Boudreaux, M., Fenelon, A., and Slopen, N. (2018). Misclassification of rental assistance in the National Health Interview Survey: evidence and implications. *Epidemiology*, *29*(5), 716-720.

Boudreaux, M., Noon, J.M., Fried, B., and Pascale, J. (2019). Medicaid expansion and the Medicaid undercount in the American Community Survey. *Health Services Research*, *54*(6), 1263-1272.

Brackstone, G. (1999). Managing data quality in a statistical agency. *Survey Methodology*, *25*(2), 139-149.

Bradley, V.C., Kuriwaki, S., Isakov, M., Sejdinovic, D., Meng, X.-L., and Flaxman, S. (2021). Unrepresentative big surveys significantly overestimated US vaccine uptake. *Nature*, *600*(December), 695-700.

Brayne, S. (2017). Big data surveillance: The case of policing. *American Sociological Review*, *82* (5), 977-1008.

Bregger, J.E. (1984). The Current Population Survey: A historical perspective and BLS' role. *Monthly Labor Review*, *107*(6), 8-14.

Brick, J.M., and Williams, D. (2013). Explaining rising nonresponse rates in cross-sectional surveys. *The ANNALS of the American Academy of Political and Social Science*, *645*(1), 36-59.

Brisbane, J. and Mohl, C. (2014). The potential use of remote sensing to produce field crop statistics at Statistics Canada. Proceedings of the Statistics Canada Symposium 2014. Available: https://www.statcan.gc.ca/eng/conferences/symposium2014/program/14259-eng.pdf

Brown, K.S. (2022). From data to decision-making: Using data and engagement to advance racial equity. Presentation at the National Academy of Sciences, Engineering, and Medicine Workshop on The Implications of Using Multiple Data Sources for Major Survey Programs. May 18, 2022. Available: https://www.nationalacademies.org/event/05-16-2022/the-implications-of-using-multiple-data-sources-for-major-survey-programs-workshop

Brown, K.S., Su, Y., Jagganath, J., Rayfield, J., and Randall, M. (2021). *Ethics and Empathy in Using Imputation to Disaggregate Data for Racial Equity: Landscape Scan Findings*. Washington, DC: Urban Institute. Available: https://www.urban.org/sites/default/files/publication/104678/ethics-and-empathy-in-using-imputation-to-disaggregate-data-for-racial-equity_0.pdf.

Buolamwini, J. and Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, *81*, 77-91. Available: https://proceedings.mlr.press/v81/buolamwini18a.html?mod=article_inline

Cai, Y., Guan, K., Peng, J., Wang, S., Seifert, C., Wardlow, B., and Li, Z. (2018). A high-performance and in-season classification system of field-level crop types using time-series Landsat data and a machine learning approach. *Remote Sensing of Environment*, *210*, 35-47

Calderwood, L. and Lessof, C. (2009). Enhancing longitudinal surveys by linking to administrative data. In Lynn, P. (Ed.), *Methodology of Longitudinal Surveys* (pp. 55-72). Hoboken, NJ: John Wiley & Sons.

Carfagna, E. and Carfagna, A. (2015). Combining list frames with different kinds of area frames. Proceedings of the 60th World Statistics Congress of the International Statistical Institute. Available: https://2015.isiproceedings.org/Files/STS006-P4-S.pdf

Carletto, C., Dillon, A., Zezza, A. (2021). Agricultural data collection to minimize measurement error and maximize coverage. In Barrett, C.B., Just, D.R. (Eds.), *Handbook of Agricultural Economics Volume 5* (pp. 4407-4480). Amsterdam: Elsevier.

Carson, E.A. (2015). Linking administrative BJS data: Better understanding of prisoners' personal histories by linking the National Corrections Reporting Program (NCRP) and CARRA data. Proceedings of the 2015 Federal Committee on Statistical Methodology (FCSM) Research Conference. https://nces.ed.gov/FCSM/pdf/A1_Carson_2015FCSM.pdf

Catalano, S. (2016). Interviewing conditions in the National Crime Victimization Survey,1993-2013. NCJ-249682, Washington, DC: Bureau of Justice Statistics. Available: https://www.bjs.gov/content/pub/pdf/icncvs9313.pdf

Celhay, P.A., Meyer, B.A., and Mittag, N. (2021). Errors in Reporting and Imputation of Government Benefits and their Implications. National Bureau of Economic Research Working Paper 29184. Available: http://www.nber.org/papers/w29184

Chen, H. (2022). Promoting global data equity: The work of the Inter-Secretariat Working Group on Household Surveys. Presentation at the National Academy of Sciences, Engineering, and Medicine Workshop on The Implications of Using Multiple Data Sources for Major Survey Programs. May 16, 2022. Available: https://www.nationalacademies.org/event/05-16-2022/the-implications-of-using-multiple-data-sources-for-major-survey-programs-workshop

Chen, J.T. (2015). Merging survey data with aggregate data from other sources: Opportunities and challenges. In Johnson, T. P. (Ed.), *Health Survey Methods* (pp. 717-754). Hoboken, NJ: John Wiley & Sons.

Chen, L. and Nandram, B. (2022). A Hierarchical Bayesian Beta-Binomial Model for Sub-areas. In Hanagal, D.D., Latpate, R.V., Chandra, G. (Eds.), *Applied Statistical Methods*. ISGES 2020. Springer Proceedings in Mathematics & Statistics, vol 380. Singapore: Springer.

Chen, L., Cruze, N., and Young, L. (2021). Transitioning to model-based official statistics: The case of crops county estimates. Presentation to the Sixth International Conference on Establishment Statistics.

Chen, L., Nandram, B., and Cruze, N.B. (2022). Hierarchical Bayesian model with inequality constraints for US county estimates. *Journal of Official Statistics, 38*(3), 709-732.

Chen, S. and Haziza, D. (2019). Recent developments in dealing with item non-response in surveys: a critical review. *International Statistical Review, 87*(S1), S192-S218.

Chow, M., Fort, T.C., Goetz, C., Goldschlag, N., Lawrence, J., Perlman, E.R., Stinson, M, and White, T.K. (2021). Redesigning the Longitudinal Business Database. U.S. Census Bureau Center for Economic Studies Working Paper 21-08. Available: https://www.census.gov/library/working-papers/2021/adrm/CES-WP-21-08.html.

Christen, P. (2019). Data linkage: The big picture. *Harvard Data Science Review*, *1*(2). Available: https://doi.org/10.1162/99608f92.84deb5c4

Chromy, J., and Wilson, D. (2012). Multiple frame approaches to identify and survey victims of rape and sexual assault. Paper commissioned by the National Research Council Panel on Measuring Rape and Sexual Assault in the Bureau of Justice Statistics Household Surveys, Washington, DC. Available: https://sites.nationalacademies.org/cs/groups/dbassesite/documents/webpage/dbasse_080064.pdf

Citro, C.F. (2014). From multiple modes for surveys to multiple data sources for estimates. *Survey Methodology*, *40*(2), 137-161.

Coble, K.H., Mishra, A.K., Ferrell, S., and Griffin, T. (2018). Big data in agriculture: A challenge for the future. *Applied Economic Perspectives and Policy*, *40*(1), 79-96.

Comenetz, J. (2016). Frequently occurring surnames in the 2010 census. Washington, DC: U.S. Census Bureau. Available: https://www2.census.gov/topics/genealogy/2010surnames/surnames.pdf

Compson, M. (2022). Improving county-level earnings estimates with a new methodology for assessing geographic and demographic information to US workers. *Social Security Bulletin*, *82*(1), 11-28.

Cook, S.L., Gidycz, C.A., Koss, M.P., and Murphy, M. (2011). Emerging issues in the measurement of rape victimization. *Violence Against Women, 17*(2), 201-218.

Cooper, C.R., and Getter, D.E. (2020). *Consumer Credit Reporting, Credit Bureaus, Credit Scoring, and Related Policy Issues*. Washington, DC: Congressional Research Service.

Couper, M.P. (2013). Is the sky falling? New technology, changing media, and the future of surveys. *Survey Research Methods*, *7*(3), 145-156.

Craig, M. (2010). A history of the Cropland Data Layer at NASS. Available: https://pdi.scinet.usda.gov/portal/sharing/rest/content/items/5fb11fe081744dab8d470fdd9b472771/data

Crosby, A.E., Mercy, J.A., and Houry, D. (2016). The National Violent Death Reporting System. *American Journal of Preventive Medicine, 51*(5), S169-S172.

Cruze, N.B., Erciulescu, A.L., Nandram, B., Barboza, W.J., and Young, L.J. (2019). Producing official county-level agricultural estimates in the United States: Needs and challenges. *Statistical Science*, *34*(2), 301-316.

Culhane, D.P., Fantuzzo, J., Rouse, H.L., Tam, V., and Lukens, J. (2010). Connecting the dots: The promise of integrated data systems for policy analysis and systems reform. Intelligence for Social Policy, University of Pennsylvania. Available: https://aisp.upenn.edu/resource-article/connecting-the-dots-the-promise-of-integrated-data-systems-for-policy-analysis-and-systems-reform/

Cummings, J. (1918). Statistical work of the federal government in the United States. In Koren, J. (Ed.), *The History of Statistics* (pp. 573-689). New York: Burt Franklin.

Cunningham, C., Foster, L., Grim, C., Haltiwanger, J., Pabilonia, S.W., and Stewart, J. (2021). Productivity Dispersion, Entry, and Growth in U.S. Manufacturing Industries. U.S. Census Bureau Center for Economic Studies Working Paper 21-21.

Cunningham, J. (2021). Uniting Our Data to Inform Our States. Presented at the *Value of Science: Data, Products, & Use* conference. Available: https://coleridgeinitiative.org/wp-content/uploads/2021/06/Cunningham-KYSTATS.pdf

Cunningham, J, Hui, A., Lane, J., and Putnam, G. (2022). A value-driven approach to building data infrastructures: The example of the MidWest Collaborative. *Harvard Data Science Review, 4*(1), 1-9. Available: https://hdsr.mitpress.mit.edu/pub/mfhpwpxq/release/2.

Czajka, J.L. and Beyler, A. (2016). Declining response rates in federal surveys: Trends and implications. Washington, DC: Mathematica Policy Research. Available: https://mathematica.org/publications/declining-response-rates-in-federal-surveys-trends-and-implications-background-paper.

Czajka, J.L. and Denmead, G. (2008). Income Data for Policy Analysis: A Comparative Assessment of Eight Surveys: Final Report. Mathematica Policy Research, Inc. Report No.: 6302-601. Available: https://mathematica.org/publications/income-data-for-policy-analysis-a-comparative-assessment-of-eight-surveys

Czajka, J.L. and Stange, M. (2018). *Transparency in the Reporting of Quality for Integrated Data: A Review of International Standards and Guidelines*. Washington, DC: Mathematica Policy Research.

Daas P.J., Ossen, S., Vis-Visschers, R., and Arends-Tóth J. (2009). *Checklist for the Quality Evaluation of Administrative Data Sources*. The Hague: Statistics Netherlands.

Dalton, J. (2007). SOI approaches first century in the twenty-first. *Internal Revenue Service Statistics of Income Bulletin,* 27(1), 4-5.

Data Foundation and AGree Initiative. (2022). *Modernizing Data Infrastructure to Improve Economic and Ecological Outcomes*. Washington, DC: Data Foundation. Available: https://www.datafoundation.org/modernizing-agriculture-data-infrastructure-to-improve-economic-and-ecological-outcomes-2022

Davern, M., Call, K.T., Ziegenfuss, J., Davidson, G., Beebe, T.J., and Blewett, L. (2008). Validating health insurance coverage survey estimates: A comparison of self-reported coverage and administrative data records. *Public Opinion Quarterly*, *72*(2), 241-259.

Davern M., Klerman, J.A., Baugh, D.K., Call, K.T., and Greenberg, G.D. (2009). An examination of the Medicaid undercount in the Current Population Survey: Preliminary results from record linking. *Health Services Research, 44*(3), 965–987

Davern, M., Roemer, M. and Thomas, W. (2014). Merging survey data with administrative data for health research purposes. In Johnson, T.P. (Ed.), *Handbook of Health Survey Methods* (pp. 1651-1698). Hoboken, NJ: John Wiley & Sons.

Davies, C. (2009). Area frame design for agricultural surveys. National Agricultural Statistics Service RDD Research Report Number RDD-09-xx. Washington, DC: US Department of

Agriculture. Available: https://www.nass.usda.gov/Publications/Methodology_and_Data_Quality/Advanced_Topics/AREA%20FRAME%20DESIGN.pdf

Davies, P.S. and Fisher, T.L. (2009). Measurement issues associated with using survey data matched with administrative data from the Social Security Administration. *Social Security Bulletin, 69*(2), 1-12.

Davis, S.J., Haltiwanger, J.C., and Schuh, S. (1998). *Job Creation and Destruction*. Cambridge, MA: MIT Press.

Day, H.R. and Parker, J.D. (2013). Self-report of Diabetes and Claims-based Identification of Diabetes Among Medicare Beneficiaries. National Health Statistics Reports #69. Hyattsville, MD: NCHS. Available: https://www.cdc.gov/nchs/data/nhsr/nhsr069.pdf

Decker, S.L., Doshi, J.A., Knaup, A.E., Polsky, D. (2012). Health service use among the previously uninsured: is subsidized health insurance enough? *Health Economics*, *21*(10), 1155-1168.

Deming, W.E. (1950). *Some Theory of Sampling.* New York: John Wiley & Sons.

Doidge, J.C., and Harron, K.L. (2019). Reflections on modern methods: linkage error bias. *International Journal of Epidemiology*, *48*(6), 2050-2060.

Dunn, M., Haugen S.E., and Kang, J.-L. (2018). The Current Population Survey—Tracking unemployment in the United States for over 75 years. *Monthly Labor Review, 141*(1), 1-22.

Dushi, I. and Iams, H.M. (2013). Pension plan participation among married couples. *Social Security Bulletin*, *73*(3), 45-52.

Dushi, I. and Trenkamp, B. (2021). Improving the Measurement of Retirement Income of the Aged Population. Social Security Administration Office of Research, Evaluation, and Statistics Working Paper #116.

Dutwin, D. and Buskirk, T.D. (2021). Telephone sample surveys: Dearly beloved or nearly departed? Trends in survey errors in the era of declining response rates. *Journal of Survey Statistics and Methodology, 9*(3), 353-380.

Egan, K.B., Cornwell, C.R., Courtney, J.G., and Ettinger, A.S. (2021). Blood lead levels in US children ages 1-11 years, 1976-2016. *Environmental Health Perspectives*, *129*(3-037003), 1-11.

Eggleston, J., Klee, M.A., and Munk, R. (2022). Self-Employment Status: Imputations, Implications, and Improvements. U.S. Census Bureau SIPP Working Paper #303 and SEHSD Working Paper #2022-06. Available: https://www.census.gov/content/dam/Census/library/working-papers/2022/demo/sehsd-wp2022-06.pdf

Elliott, M.N., Morrison, P.A., Fremont, A., McCaffrey, D.F., Pantoja, P., and Lurie, N. (2009). Using the Census Bureau's surname list to improve estimates of race/ethnicity and associated disparities. *Health Services and Outcomes Research Methodology*, *9(*2), 69-83.

Elliott, M.R., and Valliant, R. (2017). Inference for nonprobability samples. *Statistical Science*, *32*(2), 249-264.

Elliott, M.R., Raghunathan, T.E., and Schenker, N. (2018). Combining estimates from multiple surveys. *Wiley StatsRef: Statistics Reference Online*, 1-10. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/9781118445112.stat08079

Engelhardt, G.V., and Kumar, A. (2007). Employer matching and 401(k) Saving: Evidence from the Health and Retirement Study. *Journal of Public Economics, 91*(10), 1920-1943.

Equitable Data Working Group. (2022). *A Vision for Equitable Data: Recommendations from the Equitable Data Working Group.* Washington, DC: Office of the President. Available: https://www.whitehouse.gov/wp-content/uploads/2022/04/eo13985-vision-for-equitable-data.pdf

Erciulescu, A.L., Cruze, N.B., and Nandram, B. (2018). Benchmarking a triplet of official estimates. *Environmental and Ecological Statistics*, *25*(November), 523-547.

Erciulescu, A.L., Cruze, N.B., and Nandram, B. (2019). Model-based county level crop estimates incorporating auxiliary sources of information. *Journal of the Royal Statistical Society: Series A*, *182*(1), 283-303.

Erciulescu, A.L., Cruze, N.B., and Nandram, B. (2020). Statistical challenges in combining survey and auxiliary data to produce official statistics. *Journal of Official Statistics*, *36*(1), 63-88.

Eurostat. (2021). *European Statistical System Handbook for Quality and Methods Reports—Re-edition 2021*. Luxembourg: Publications Office of the European Union.

Executive Order 13985 of January 20, 2021. Executive Order on Advancing Racial Equity and Support for Underserved Communities Through the Federal Government. *Federal Register, 86*(14), January 25, 2021, 7009-7013.

Faul, J. and Levy, H. (2022). Overview of HRS Administrative Data Linkages. Presentation at the National Academy of Sciences, Engineering, and Medicine Workshop on The Implications of Using Multiple Data Sources for Major Survey Programs. May 18, 2022. Available: https://www.nationalacademies.org/event/05-16-2022/docs/DBE928F840C122601E5182478F58F71D32FE5478C94E

Fay, R.E. (2021). *Constructing and Disseminating Small-Area Estimates from the National Crime Victimization Survey, 2007-2018*. Report NCJ-300603. Washington, DC: Bureau of Justice Statistics. Available: https://www.ojp.gov/library/publications/constructing-and-disseminating-small-area-estimates-national-crime.

Federal Committee on Statistical Methodology. (2005). *Report on Statistical Disclosure Limitations Methodology*. Statistical Policy Working Paper No. 22. Available: https://www.hhs.gov/sites/default/files/spwp22.pdf

Federal Committee on Statistical Methodology. (2018). *Transparent Quality Reporting in the Integration of Multiple Data Sources: A Progress Report, 2017-2018*. Available: https://nces.ed.gov/FCSM/pdf/Quality_Integrated_Data.pdf

Federal Committee on Statistical Methodology. (2020). *A Framework for Data Quality.* Available: https://nces.ed.gov/FCSM/pdf/FCSM.20.04_A_Framework_for_Data_Quality.pdf

Feldman, J.M., Gruskin, S., Coull, B.A., and Krieger, N. (2017). Quantifying underreporting of law-enforcement-related deaths in United States vital statistics and news-media-based data sources: A capture-recapture analysis. *PLoS Medicine*, *14*(10), e1002399.

Ferguson, A.G. (2017). *The Rise of Big Data Policing*. New York: New York University Press.

Fernandez, L., Ennis, S., Porter, S.R., and Carson, E. (2022). Mortality in a multi-state cohort of former state prisoners, 2010-2015. NCJ-304203. Washington DC: Bureau of Justice Statistics. Available: https://bjs.ojp.gov/library/publications/mortality-multi-state-cohort-former-state-prisoners-2010-2015

Fernandez, L., Shattuck, R., and Noon, J. (2018). The use of administrative records and the American Community Survey to study characteristics of undercounted young children in the 2010 census. Center for Administrative Records Research and Applications Working Paper 2018-05. Washington, DC: U.S. Census Bureau. Available:

https://www.census.gov/content/dam/Census/library/working-papers/2018/adrm/carra-wp-2018-05.pdf

Finlay, K., Mueller-Smith, M., and Papp, J. (2022). The Criminal Justice Administrative Records System: A next-generation research data platform. *Scientific Data 9*(562), 1-11.

Fischer, R.L., Richter, F.G-C., Anthony, E., Lalich, E, and Coulton, C. (2019). Leveraging administrative data to better serve children and families. *Public Administration Review*, *79*(5), 675-683.

Fisher, G.G. and Ryan, L.H. (2018). Overview of the Health and Retirement Study and introduction to the special issue. *Work, Aging and Retirement*, *4*(1), 1-9.

Fitzpatrick, T.B. (1988). The validity and practicality of sun-reactive skin types I through VI. *Archives of Dermatology*, *124*(6), 869-871.

Flegal, K.M, Graubard, B.I., Williamson, D.F., and Gail, M.H. (2005). Excess deaths associated with underweight, overweight, and obesity. *Journal of the American Medical Association*, *293*(15), 1861-1867.

Fouch, D., and Martin, K. (2022, February 15). Shifting the crime reporting paradigm—Lessons learned from the FBI's transition to NIBRS [Webinar]. U.S. Department of Justice, Office of Justice Programs. https://bjs.ojp.gov/media/video/66506

Fox, L. and Burns, K. (2021). The Supplemental Poverty Measure: 2020. U.S. Census Bureau Current Population Report P60-275. Available: https://www.census.gov/content/dam/Census/library/publications/2021/demo/p60-275.pdf

Fox, L., Rothbaum, J., and Shantz, K. (2022). Fixing errors in a SNAP: Addressing SNAP underreporting to evaluate poverty. *American Economic Association Papers and Proceedings*, *112*(May), 330-334. Available: https://doi.org/10.1257/pandp.20221040

Fritz, S., See, L., Bayas, J.K.L., Waldner, F., Jacques, D., Becker-Reshef, I., Whitcraft, A., Baruth, B., Bonifacio, R., Crutchfield, J., Rembold, F., Rojas, O., Schucknecht, A., Van der Velde, M., Verdin, J., Wu, B., Yan, N., You, L., Gilliams, S., Mücher, S., Tetrault, R., Moorthy, I., and McCallum, I. (2019). A comparison of global agricultural monitoring systems and current gaps. *Agricultural Systems*, *168*(January), 258-272.

Gallego J., Carfagna E., and Baruth B. (2010). Accuracy, objectivity and efficiency of remote sensing for agricultural statistics. In Benedetti, R., Bee, M., Espa, G., and Piersimoni, F. (Eds.), *Agricultural Survey Methods* (pp. 193-211). Chichester, UK: Wiley.

Gebbers, R. and Adamchuk, V.I. (2010). Precision agriculture and food security. *Science*, *327*(5967), 828-831.

Genadek, K.R. and Alexander, J.T. (2019). The Decennial Census Digitization and Linkage Project. U.S. Census Bureau ADEP Working Paper 2019-01. Available: https://www.census.gov/library/working-papers/2019/econ/adep-wp-dc-digitization-linkage.html.

Gennetian, L.A., Seshadri, R., Hess, N.D., Winn, A.N., and Goerge, R.M. (2016). Supplemental Nutrition Assistance Program (SNAP) benefit cycles and student disciplinary infractions. *Social Service Review*, *90*(3), 403-433.

Giest, S. and Samuels, A. (2020). 'For good measure': Data gaps in a big data world. *Policy Sciences*, *53*, 559-569.

Gilbert, R., Lafferty, R., Hagger-Johnson, G., Harron, K., Zhang, L.C., Smith, P., Dibben, C., and Goldstein, H. (2018). GUILD: guidance for information about linking data sets. *Journal of Public Health*, *40*(1), 191-198.

Gindi, R. and Cohen, R.A. (2012). Assessing measurement error in Medicare coverage from the National Health Interview Survey. *Medicare Care & Medicaid Research Review*, 2(2), E1-E15.

Global Strategy to improve Agricultural and Rural Statistics (GSARS). (2017). *Handbook on Remote Sensing for Agricultural Statistics*. Rome: GSARS Handbook.

Goerge, R.M. and Lee, B.J. (2002). Matching and cleaning administrative data. Chapter 7 in National Academies of Sciences, Engineering, and Medicine, *Studies of Welfare Populations: Data Collection and Research Issues.* Washington, DC: The National Academies Press.

Goerge, R.M. and Wiegand, E.R. (2019). Understanding vulnerable families in multiple service *systems. RSF: The Russell Sage Foundation Journal of the Social Sciences, 5*(2), 86-104.

Goerge, R.M., Harden, A, and Lee, B.J. (2008). Consequences of teen childbearing for child abuse, neglect, and foster care placement. In Hoffman, S.D. and Maynard, R.A. (Eds.), *Kids Having Kids: Economic Costs and Social Consequences of Teen Pregnancy (2nd edition)* (pp. 257-288). Washington DC: The Urban Institute.

Goerge, R.M., Harris, A., Bilaver, L.M., Franzetta, K., Reidy, M., Schexnayder, D., Schroeder, D., Staveley, J., Kreader, J.L., Obenski, S., Prevost, R.C., Berning, M.E., and Resnick, D.M. (2009). Employment Outcomes for Low-Income Families Receiving Child Care Subsidies in Illinois, Maryland, and Texas. Final Report to U.S. Department of Health and Human Services, Administration for Children and Families, Office of Planning, Research, and Evaluation under Grant Number 90YE0070.

Goerge, R.M., van Voorhis, J., and Lee, B.J. (1994). Illinois's Longitudinal and Relational Child and Family Research Database. *Social Science Computer Review 12*(3), 351-65. Goetz, C. and Stinson, M. (2021). The Business Dynamics Statistics: Describing the Evolution of the U.S. Economy from 1978-2019. U.S. Census Bureau Center for Economic Studies Working Paper 21-33. Available: https://www.census.gov/library/working-papers/2021/adrm/CES-WP-21-33.html

Golden, C. and Mirel, L.B. (2021). Enhancement of health surveys with data linkage. In Chun, A.Y., Larsen, M.D., Durrant, G., and Reiter, J.P. (Eds.), *Administrative Records for Survey Methodology.* Hoboken, NJ: John Wiley & Sons.

Golden, C., Driscoll, A.K., Simon, A.E., Judson, D.H., Miller, E.A., and Parker, J.D. (2015). Linkage of NCHS population health surveys to administrative records from Social Security Administration and Centers for Medicare & Medicaid Services. *Vital and Health Statistics, 1*(58), 1-53.

Goodchild, M. (2022). Discussion: Improving agriculture statistics with new data sources. Presentation at the National Academy of Sciences, Engineering, and Medicine Workshop on The Implications of Using Multiple Data Sources for Major Survey Programs. May 16, 2022. Available: https://www.nationalacademies.org/event/05-16-2022/the-implications-of-using-multiple-data-sources-for-major-survey-programs-workshop

Goodchild, M. and Li, L. (2012). Assuring the quality of volunteered geographic information. *Spatial Statistics* 1: 110-120.

Groves, R.M. (2006). Nonresponse rates and nonresponse bias in household surveys. *Public Opinion Quarterly*, *70*(5), 646-675.

Groves, R.M. (2022). Lessons Learned and Next Steps: Discussion and Wrap-Up Session. Comments at the National Academy of Sciences, Engineering, and Medicine Workshop on The Implications of Using Multiple Data Sources for Major Survey Programs. May 18, 2022.

Available: https://www.nationalacademies.org/event/05-16-2022/the-implications-of-using-multiple-data-sources-for-major-survey-programs-workshop [Video 31].

Groves, R.M. and Lyberg, L. (2010). Total survey error: Past, present, and future. *Public Opinion Quarterly*, *74*(5), 849-879.

Groves, R.M., Fowler, F.J., Couper, M.P., Lepkowski, J.M., Singer, E., and Tourangeau, R. (2011). *Survey Methodology*, 2nd ed. Hoboken, NJ: John Wiley & Sons.

Gustman, A.L., and Steinmeier, T.L. (2005). Imperfect knowledge of Social Security and pensions. *Industrial Relations*, *44*(2), 373-397.

Haas, A., Elliott, M.N., Dembosky, J.W., Adams, J.L., Wilson-Frederick, S.M., Mallett, J.S., Gaillot, S., Haffer, S.C., and Haviland, A.M. (2019). Imputation of race/ethnicity to enable measurement of HEDIS performance by race/ethnicity. *Health Services Research 54*(1), 13-23.

Hamilton, A. (1791). Report on duties arising on tonnage, for the year ending September 30, 1790. In *American State Papers: Commerce and Navigation, Volume VII* (pp. 6-8). Washington, DC: Gales and Seaton.

Hand, D.J. (2020). *Dark Data: Why What You Don't Know Matters. A Practical Guide to Making Good Decisions in a World of Missing Data*. Princeton, NJ: Princeton University Press.

Handley, K., Kamal, F., and Ouyang, W. (2021). A Long View of Employment Growth and Firm Dynamics in the United States: Importers vs. Exporters vs. Non-Traders. U.S. Census Bureau Center for Economic Studies Working Paper 21-38. Available: https://www.census.gov/library/working-papers/2021/adrm/CES-WP-21-38.html

Hansen, M.H., Hurwitz, W.N., and Madow, W.G. (1953). *Sample Survey Methods and Theory: Volume 1.* New York: John Wiley & Sons.

Hanson, E.J. (2021). *The National Incident-Based Reporting System: Benefits and Issues*. Washington, DC: Congressional Research Service.

Harcourt, B.E. (2015). Risk as a proxy for race: The dangers of risk assessment. *Federal Sentencing Reporter*, *27*(4), 237-243.

Harrell, E. (2021). Crimes Against Persons with Disabilities, 2009-2019—Statistical Tables. NCJ 301367. Washington, DC: Bureau of Justice Statistics. Available: https://bjs.ojp.gov/content/pub/pdf/capd0919st.pdf

Harron, K., Goldstein, H., and Dibben, C. (Eds.) (2016). *Methodological Developments in Data Linkage*. Hoboken, NJ: John Wiley & Sons.

Hart, N. and Yohannes, M. (Eds.) (2019). *Evidence Works: Cases Where Evidence Meaningfully Informed Policy*. Washington, D.C.: Bipartisan Policy Center. Available: https://bipartisanpolicy.org/report/evidenceworks/

Haziza, D. (2009). Imputation and inference in the presence of missing data. In Pfeffermann, D. and Rao, C.R. (Eds.), *Sample Surveys: Design, Methods, and Applications. Handbook of Statistics, Volume 29A* (pp. 215-246). Amsterdam: North-Holland.

Hedegaard, H., and Warner, M. (2021). Evaluating the cause-of-death information needed for estimating the burden of injury mortality: United States, 2019. *National Vital Statistics Reports, 70*(13), 1-22.

Helms, V.E., King, B.A., and Ashley, P.J. (2017). Cigarette smoking and adverse health outcomes among adults receiving federal housing assistance. *Preventive Medicine*, *99*: 171-177.

Helms, V.E., Steffen, B.L., Rudd, E., and Sperling, J. (2018). *A Health Picture of HUD-assisted Children, 2006-2012*. Washington, DC: U.S. Department of Housing and Urban Development.

Herz, D.C., Dierkhising, C.B., Raithel, J., Schretzman, M., Guiltinan, S., Goerge, R.M., Cho, Y., Coulton, C., and Abbott, S. (2019). Dual system youth and their pathways: A comparison of incidence, characteristics and system experiences using linked administrative data. *Journal of Youth and Adolescence*, *48*(12), 2432-2450.

Hetzel, A.M. (1997). *History and Organization of the Vital Statistics System*. Hyattsville, MD: National Center for Health Statistics.

Hill, C., Helm, K., Hong, J., and Phan, N. (2022). Census Coverage Estimates for People in the United States by State and Census Operations. 2020 Post-Enumeration Survey Estimation Report, PES20-G-02RV. Washington, DC: U.S. Government Printing Office. Available: https://www2.census.gov/programs-surveys/decennial/coverage-measurement/pes/census-coverage-estimates-for-people-in-the-united-states-by-state-and-census-operations.pdf

Himelein-Wachowiak, M., Giorgi, S., Devoto, A., Rahman, M., Ungar, L., Schwartz, H.A., Epstein, D.H., Leggio, L., and Curtis, B. (2021). Bots and misinformation spread on social media: Implications for COVID-19. *Journal of Medical Internet Research*, *23*(5). Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8139392/

Hokayem, C., Bollinger, C., and Ziliak, J.P. (2015). The role of CPS nonresponse in the measurement of poverty. *Journal of the American Statistical Association*, *110*(511), 935-945.

Hokayem, C., Raghunathan, T., and Rothbaum, J. (2022). Match bias or nonignorable nonresponse? Improved imputation and administrative data in the CPS ASEC. *Journal of Survey Statistics and Methodology*, *10*(1), 81-114.

Honeycutt, A. A., Segel, J. E., Zhuo, X., Hoerger, T. J., Imai, K., and Williams, D. (2013). Medical costs of CKD in the Medicare population. *Journal of the American Society of Nephrology*, *24*(9), 1478-1483.

Horst, M. and Marion, A. (2019). Racial, ethnic and gender inequities in farmland ownership and farming in the US. *Agriculture and Human Values*, 36(1), 1-16.

Hughes, A.G., McCabe, S.D., Hobbs, W.R., Remy, E., Shah, S., and Lazer, D.M. (2021). Using administrative records and survey data to construct samples of Tweeters and Tweets. *Public Opinion Quarterly*, *85*(S1), 323-346.

Humes, K., and Hogan, H. (2009). Measurement of race and ethnicity in a changing, multicultural America. *Race and Social Problems, 1*, 111-131.

Hurd, M.D., Martorell, P., Delavande, A., Mullen, K.J., and Langa, K.M. (2013). Monetary costs of dementia in the United States. *New England Journal of Medicine*, *368*(14), 1326-1334.

Hurst, B. (2016). Testimony in *Big Data and Agriculture: Innovation and Implications*. Hearing before the Committee on Agriculture, House of Representatives, 114th Congress, October 28, 2015, Serial No. 114-32, pp. 6-15. Washington, DC: U.S. Government Publishing Office. Available: https://www.govinfo.gov/content/pkg/CHRG-114hhrg97412/pdf/CHRG-114hhrg97412.pdf

Hyman, M., Sartore, L., and Young, L.J. (2022). Capture-recapture estimation of characteristics of U.S. local food farms using a web-scraped list frame. *Journal of Survey Statistics and Methodology*, *10*(4), 979-1004.

Iachan, R. and Dennis, M.L. (1993). A multiple frame approach to sampling the homeless and transient population. *Journal of Official Statistics, 9*(4), 747-764.

Institute of Medicine. (2009). *Race, Ethnicity, and Language Data: Standardization for Health Care Quality Improvement.* Ulmer, C., McFadden, B., and Nerenz, D.R. (Eds.) Washington, DC: The National Academies Press.

Interagency Technical Working Group on Evaluating Alternative Measures of Poverty. (2021). Final Report of the Interagency Technical Working Group on Evaluating Alternative Measures of Poverty. Available: https://www.bls.gov/cex/itwg-report.pdf

International Association of Chiefs of Police. (1929). *Uniform Crime Reporting: A Complete Manual for Police. Report of the Committee on Uniform Crime Records.* New York, NY: J.J. Little and Ives Company.

International Society of Precision Agriculture (2019). Monthly Newsletter (July). Available: https://ispag.org/site/newsletter/?id=90.

Iwashyna, T.J., Ely, E.W., Smith, D.M., and Langa, K.M. (2010). Long-term cognitive impairment and functional disability among survivors of severe sepsis. *Journal of the American Medical Association*, *304*(16), 1787-1794.

Iwig, W., Berning, M., Marck, P., and Prell, M. (2013). Data quality assessment tool for administrative data. Prepared for the Statistical Uses of Administrative Records subcommittee of the Federal Committee on Statistical Methodology. Available: https://nces.ed.gov/FCSM/pdf/DataQualityAssessmentTool.pdf

Jäckle, A., Beninger, K., Burton, J., and Couper, M. P. (2021a). Understanding data linkage consent in longitudinal surveys. In Lynn, P. (Ed.), *Advances in Longitudinal Survey Methodology* (pp. 122-150). Hoboken, NJ: John Wiley & Sons.

Jäckle, A., Burton, J., Couper, M.P., Crossley, T.F., and Walzenbach, S. (2021b), Understanding and Improving Data Linkage Consent in Surveys. Institute for Social and Economic Research, University of Essex: Understanding Society Working Paper Series No. 2021-01.

Jagadish, H.V., Stoyanovich, J., and Howe, B. (2021a). COVID-19 brings data equity challenges to the fore. *Digital Government: Research and Practice*, *2*(2), 1-7.

Jagadish, H.V., Stoyanovich, J., and Howe, B. (2021b). The many facets of data equity. Central Europe Workshop Proceedings, 2841. Available: http://ceur-ws.org/Vol-2841/PIE+Q_6.pdf

James, N. and Council, L.R. (2008). How crime in the United States is measured. Washington, DC: Congressional Research Service. Available: https://fas.org/sgp/crs/misc/RL34309.pdf

Jarmin, R.S. and Miranda, J. (2002). The Longitudinal Business Database. Center for Economic Studies Working Paper 02-17. Available: https://www2.vrdc.cornell.edu/news/3/20050812-JarminMiranda2002.pdf

Jarvis, J.P. (2015). Examining National Incident-Based Reporting System (NIBRS) data: Perspectives from a quarter century of analysis efforts. *Justice Research and Policy*, *16*(2), 195-210.

Jiang, H., Hu, H., Zhong, R., Xu, J., Xu, J., Huang, J., Wang, S., Ying, Y., and Lin, T. (2019). A deep learning approach to conflating heterogeneous geospatial data for corn yield estimation: A case study of the U.S. corn belt at the county level. *Global Change Biology*, *26*(December), 1754-1766.

Johansson, R., Effland, A., and Coble, K. (2017). Falling response rates to USDA crop surveys: Why it matters. *farmdoc daily*, *7*(January), 9.

Johnson, D.M. (2014). An assessment of pre- and within-season remotely sensed variables for forecasting corn and soybean yields in the United States. *Remote Sensing of Environment 141,* 116-128.

Johnson, D.M. (2016). A Comprehensive Assessment of the Correlations between Field Crop Yields and Commonly Used MODIS Products. *International Journal of Applied Earth Observation and Geoinformation*, *52*(October), 65-81.

Jones, M.R. and Ziliak, J.P. (2022). The antipoverty impact of the EITC: New estimates from survey and administrative tax records. *National Tax Journal, 75*(3), 451-479.

Jones, N., Marks, R., Ramirez, R., and Ríos-Vargas, M. (2021). 2020 census illuminates racial and ethnic composition of the country. Available: https://www.census.gov/library/stories/2021/08/improved-race-ethnicity-measures-reveal-united-states-population-much-more-multiracial.html

Judson, D. (2000). The Statistical Administrative Records System: System design, successes, and challenges. Paper presented at the 2000 NISS/Telcordia Data Quality Conference. Washington, DC: NISS. Available: https://www.niss.org/events/niss-affiliates-workshop-data-quality-challenges-computer-science-and-statistics.

Judson, D. and Popoff, C. (2005). Administrative records research. In Kempf-Leonard, K. (Ed.), *Encyclopedia of Social Measurement* (pp. 17-27). Amsterdam: Elsevier.

Kalton, G. (2020). *Introduction to Survey Sampling (2nd ed).* Thousand Oaks, CA: Sage.

Keisler-Starkey, K., and Bunch, L.N. (2022). *Health Insurance Coverage in the United States: 2021.* Current Population Reports P60-278. Washington, DC: U.S. Census Bureau. Available: https://www.census.gov/content/dam/Census/library/publications/2022/demo/p60-278.pdf.

Keller, S., Prewitt, K., Thompson, J., Jost, S., Barrett, C., Nusser, S., Salvo, J., and Shipp, S. (2022). *A 21st Century Census Curated Data Enterprise: A Bold New Approach to Create Official Statistics*. Charlottesville, VA: University of Virginia Biocomplexity Institute. Available: https://libraopen.lib.virginia.edu/public_view/zw12z549f

Kennedy, C. (2022). Exploring the assumption that online opt-in respondents are answering in good faith. Paper presented at the 2022 Morris Hansen Lecture, March 1. Slides available: https://washstat.org/hansen/2022Kennedy.pdf

Kennedy, C., Hatley, N., Lau, A., Mercer, A., Keeter, S., Ferno, J., and Asare-Marfo, D. (2021). Strategies for detecting insincere respondents in online polling. *Public Opinion Quarterly*, *85*(4), 1050-1075.

Kennel, T. (2021). The Post-Enumeration Survey: Measuring coverage error. Available: https://www.census.gov/newsroom/blogs/random-samplings/2021/12/post-enumeration-measuring-coverage-error.html

Keyes, K.M., Rutherford, C., Popham, F., Martins, S.S., and Gray, L. (2018). How healthy are survey respondents compared with the general population?: Using survey-linked death records to compare mortality outcomes. *Epidemiology, 29*(2), 299-307.

Khubba, S., Heim, K., and Hong, J. (2022). *National Census Coverage Estimates for People in the United States by Demographic Characteristics. 2020 Post-Enumeration Survey Estimation Report PES20-G-01*. Washington, DC: U.S. Government Publishing Office. Available: https://www2.census.gov/programs-surveys/decennial/coverage-measurement/pes/national-census-coverage-estimates-by-demographic-characteristics.pdf

Kilss, B. and Alvey, W. (1984). *Statistical Uses of Administrative Data: Recent Research and Present Projects, Volume 1.* Washington, DC: Department of the Treasury.

Kilss, B. and Scheuren, F.J. (1978). The 1973 CPS-IRS-SSA Exact Match Study. *Social Security Bulletin, 41*(10), 14-22.

Kim, D.Y. and Phillips, S.W. (2021). When COVID-19 and guns meet: A rise in shootings. *Journal of Criminal Justice*, *73*, 101783.

Kinney, S.K., Reiter, J.P., Reznek, A.P., Miranda, J., Jarmin, R.S., and Abowd, J.M. (2011). Towards unrestricted public use business microdata: The Synthetic Longitudinal Business Database. *International Statistical Review, 79*(3), 362-384.

Kitzmiller, E.M. (2013). *IDS Case Study, Chapin Hall: Leveraging Chapin Hall's Mission to Enhance Child Well-Being*. Philadelphia: University of Pennsylvania, Actionable Intelligence for Social Policy. http:// www.aisp.upenn.edu/wp-content/uploads/2015 /08/ChapinHall_CaseStudy.pdf

Kohler, U., Kreuter, F., and Stuart, E.A. (2019). Nonprobability sampling and causal analysis. *Annual Review of Statistics and its Application*, *6*(March), 149-172.

Konny, C.G., Williams, B.K., and Friedman, D.M. (2022). Big data in the US Consumer Price Index: Experiences and plans. In Abraham, K.G., Jarmin, R.S., Moyer, B.C., and Shapiro, M.D. (Eds.), *Big Data for Twenty-First-Century Economic Statistics* (pp. 69-98). Chicago: University of Chicago Press.

Krebs, C. (2014). Measuring sexual victimization: On what fronts is the jury still out and do we need it to come in? *Trauma, Violence, & Abuse, 15*(3), 170-180.

Kreuter, F. (2022). Diversity in data: How data collection decisions help and harm the outcome. Presentation at the National Academy of Sciences, Engineering, and Medicine Workshop on The Implications of Using Multiple Data Sources for Major Survey Programs. May 18, 2022. Available: https://www.nationalacademies.org/event/05-16-2022/the-implications-of-using-multiple-data-sources-for-major-survey-programs-workshop

Kreuter, F., Ghani, R., and Lane, J. (2019). Change through data: A data analytics training program for government employees. *Harvard Data Science Review*, 1(2). Available: https://hdsr.mitpress.mit.edu/pub/0mb0zzlc/release/7

Kuehn, D. (2022a). *Better Data for Better Policy: Lessons Learned from across the Coleridge Initiative's Partnerships.* Washington, DC: Urban Institute. Available: https://www.urban.org/sites/default/files/2022-05/Better%20Data%20for%20Better%20Policy.pdf

Kuehn, D. (2022b). *Better Data for Better Policy: The Coleridge Initiative in Ohio: Steady Progress and Productive Collaboration*. Washington DC: Urban Institute. Available: https://www.urban.org/sites/default/files/2022-05/Better%20Data%20for%20Better%20Policy%20-%20the%20Coleridge%20Initiative%20in%20Ohio.pdf

Lariscy, J.T. (2011). Differential record linkage by Hispanic ethnicity and age in linked mortality studies: implications for the epidemiologic paradox. *Journal of Aging and Health*, *23*(8), 1263-1284.

Lariscy, J.T. (2017). Black-white disparities in adult mortality: implications of differential record linkage for understanding the mortality crossover. *Population Research and Policy Review*, *36*(1), 137-156.

Larrimore, J., Mortenson, J., and Splinter, D. (2021). Household incomes in tax data: Using addresses to move from tax-unit to household income distributions. *Journal of Human Resources, 56*(July), 600-631.

Lauritsen, J.L. (2022a). Crime Measurement in the United States. Presentation at the National Academy of Sciences, Engineering, and Medicine Workshop on The Implications of Using Multiple Data Sources for Major Survey Programs. May 18, 2022. Available:

https://www.nationalacademies.org/event/05-16-2022/the-implications-of-using-multiple-data-sources-for-major-survey-programs-workshop

Lauritsen, J.L. (2022b). Victimization in Different Types of Areas in the United States: Subnational Findings from the National Crime Victimization Survey, 2010-2015. NCJ-252630. Washington, DC: Bureau of Justice Statistics. Available: https://bjs.ojp.gov/library/publications/victimization-different-types-areas-united-states-subnational-findings

Leach, M.A., Van Hook, J., and Bachmeier, J.D. (2018). Using Linked Data to Investigate True Intergenerational Change: Three Generations over Seven Decades. U.S. Census Bureau CARRA Working Paper #2018-09.

Leary, K. (2022). Fireside chat on data equity. Discussion at the National Academy of Sciences, Engineering, and Medicine Workshop on The Implications of Using Multiple Data Sources for Major Survey Programs. May 18, 2022. Available: https://www.nationalacademies.org/event/05-16-2022/the-implications-of-using-multiple-data-sources-for-major-survey-programs-workshop

Lee, S., Nishimura, R., Burton, P., and McCammon, R. (2021). *HRS 2016 Sampling Weights*. Ann Arbor, MI: University of Michigan Institute of Social Research. Available: https://hrs.isr.umich.edu/sites/default/files/biblio/HRS2016SamplingWeights.pdf

Lehnen, R.G. and Skogan, W. (Eds.). (1981). *The National Crime Survey: Working Papers*. Volume I: Current and Historical Perspectives. NCJ-75374. Washington DC: Bureau of Justice Statistics.

Leroux, A., Di, J., Smirnova, E., Mcguffey, E.J., Cao, Q., Bayatmokhtari, E., Tabacu, L., Zipunnikov, V., Urbanek, J. K., and Crainiceanu, C. (2019). Organizing and analyzing the activity data in NHANES. *Statistics in Biosciences*, *11*(2), 262-287.

LeRoy, L., Wasserman M., Rezaee, M., and White, Alan. (2013). *Understanding Disparities in Persons with Multiple Chronic Conditions: Research Approaches and Datasets*. Cambridge, MA: Abt Associates. Available: https://aspe.hhs.gov/reports/understanding-disparities-persons-multiple-chronic-conditions-research-approaches-datasets-0

Lesiv, M, Bayas, J.C.L., See, L., Duerauer, M., Dahlia, D., Durando, N., Hazarika, R., Sahariah, P.K., Vakolyuk, M., Blyshchyk, V., Bilous, A., Perez-Hoyos, A., Gengler, S., Prestele, R., Bilous, S., Akhtar, I.U.H., Singha, K., Choudhury, S.B., Chetri, T., Malek, Z., Bungnamei, K., Saikia, A., Sahariah, D., Narzary, W., Danylo, O., Sturn, T., Karner, M., McCallum, I., Schepaschenko, D., Moltchanova, E., Fraisl, D., Moorthy, I., Fritz, S. (2019). Estimating the global distribution of field size using crowdsourcing. *Global Change Biology*, *25*, 174- 186.

Leslie, D., Katell, M., Aitken, M., Singh, J., Briggs, M., Powell, R., Rincón, C., Perini, A., and Javadeva, S. (2022). *Data Justice in Practice: A Guide for Impacted Communities.* Montréal: The Alan Turing Institute in collaboration with The Global Partnership on AI.

Levenstein, M. (2022). Discussion on 'Issues in data equity'. Presentation at the National Academy of Sciences, Engineering, and Medicine Workshop on The Implications of Using Multiple Data Sources for Major Survey Programs. May 18, 2022. Available: https://www.nationalacademies.org/event/05-16-2022/the-implications-of-using-multiple-data-sources-for-major-survey-programs-workshop

Lewis, K., Ellwood, M., and Czajka, J. (1998). *Counting the Uninsured: A Review of the Literature.* Washington, DC: Urban Institute.

Liao, D., Zimmer, S., and Berzofsky, M. (2021). *Small Area Estimation for the National Crime Victimization Survey: A Guide for Data Processing and Estimation Procedures*. NCJ

300580. Washington, DC: Bureau of Justice Statistics. Available: https://www.ojp.gov/library/publications/small-area-estimation-national-crime-victimization-survey-guide-data

Liebler, C.A., Porter, S.R., Fernandez, L.E., Noon, J.M., and Ennis, S.R. (2017). America's churning races: Race and ethnicity response changes between census 2000 and the 2010 census. *Demography*, *54*(1), 259-284.

Lin P.J., Daly, A.T., Olchanski. N., Cohen J.T., Neumann, P.J., Faul, J.D., Fillit, H,M., and Freund, K.M. (2021). Dementia diagnosis disparities by race and ethnicity. *Medical Care 59*(8), 679-686.

Lloyd, P.C., Helms, V.E., Simon, A.E., Golden, C., Brittain, J., Call, E., Mirel, L.B., Steffen, B.L., Sperling, J., Rudd, E.C., Parker, J.D., and Star, C.S. (2017). Linkage of 1999-2012 National Health Interview Survey and National Health and Nutrition Examination Survey data to U.S. Department of Housing and Urban Development administrative records. *Vital and Health Statistics, 1*(60), 1-31.

Lohr, S.L. (2019). *Measuring Crime: Behind the Statistics.* Boca Raton, FL: CRC Press.

Lohr, S.L. (2021). Multiple-frame surveys for a multiple-data-source world. *Survey Methodology*, *47*(2), 229-263.

Lohr, S.L. (2022). *Sampling: Design and Analysis, third edition.* Boca Raton, FL: CRC Press.

Lohr, S.L. and Raghunathan, T.E. (2017). Combining survey data with other data sources. *Statistical Science*, *32*(2), 293-312.

Lukens, G. and Sharer, B. (2021). *Closing Medicaid Coverage Gap Would Help Diverse Group and Narrow Racial Disparities*. Washington, DC: Center on Budget and Policy Priorities. Available: https://www.cbpp.org/sites/default/files/6-10-21health.pdf.

Lynch, J. (2018). Not even our own facts: Criminology in the era of big data. *Criminology*, *56*(3), 437-454.

Lyu, F., Yang, Z., Xiao, Z., Diao, C., Park, J., and Wang, S. (2022). CyberGIS for scalable remote sensing data fusion. In *Proceedings of Practice and Experience in Advanced Research Computing (PEARC'22)*. Boston, MA: Association for Computing Machinery.

Mahajan, P. (2021). Immigration and Local Business Dynamics: Evidence from U.S. Firms. U.S. Census Bureau Center for Economic Studies Working Paper 21-18.

Mancuso, D. and Huber, A. (2021). *Integrated Client Databases*. Washington State Health and Human Services Research and Data Analysis Division. Available: https://www.dshs.wa.gov/sites/default/files/rda/reports/research-11-205.pdf

Mapes, B.M., Foster, C.S., Kusnoor, S.V., Epelbaum, M.I., AuYoung, M., Jenkins, G., Lopez-Class, M., Richardson-Heron, D., Elmi, A., Surkan, K., Cronin, R. M., Wilkins, C.H., P., Pérez-Stable, E.J., Dishman, E., Denny, J.C., and Rutter, J.L. (2020). Diversity and inclusion for the All of Us research program: A scoping review. *PloS ONE*, *15*(7), e0234962.

Marra, E. and Kennel, T. (2022). Source and Accuracy of the 2020 Post-Enumeration Survey Person Estimates. 2020 Post-Enumeration Survey Methodology Report PES20-J-01. Washington, DC: U.S. Government Publishing Office. Available: https://www2.census.gov/programs-surveys/decennial/coverage-measurement/pes/2020-source-and-accuracy-pes-estimates.pdf

Martin, K. (2021). Sexual Assaults Recorded by Law Enforcement, 2019. NCJ-301236. Washington, DC: Bureau of Justice Statistics. Available: https://bjs.ojp.gov/nibrs/reports/sarble/sarble19

Martínez, R. (2015). *Latino Homicide: Immigration, Violence, and Community (2nd ed.)*. New York: Routledge.

Martínez, R. (2022). Panel Discussion: Measuring Crime in the 21st Century. Presentation at the National Academy of Sciences, Engineering, and Medicine Workshop on The Implications of Using Multiple Data Sources for Major Survey Programs. May 16, 2022. Available: https://www.nationalacademies.org/event/05-16-2022/the-implications-of-using-multiple-data-sources-for-major-survey-programs-workshop

Mathews, K., Phelan, J., Jones, N.A., Konya, S., Marks, R., Pratt, B.M., Coombs, J., Bentley, M. (2017). *2015 National Content Test Race and Ethnicity Analysis Report*. Washington DC: U.S. Census Bureau. Available: https://www2.census.gov/programs-surveys/decennial/2020/program-management/final-analysis-reports/2015nct-race-ethnicity-analysis.pdf

McClure, D., Santos, R., and Kooragayala, S. (2017). Administrative Records in the 2020 US Census: Civil Rights Considerations and Opportunities. Urban Institute Research Report. Available: https://www.urban.org/research/publication/administrative-records-2020-us-census

McGrath-Lone, L., Libuy, N., Etoori, E., Blackburn, R., Gilbert, R., and Harron, K. (2021). Ethnic bias in data linkage. *The Lancet Digital Health, 3*(6), e339.

Medalia, C., Meyer, B., O'Hara, A., and Wu, D. (2019). Linking survey and administrative data to measure income, inequality, and mobility. *International Journal of Population Data Science*, *4*(1), 1-8.

Meitinger, K.M., and Johnson, T.P. (2020). Power, culture and item nonresponse in social surveys. In Brenner, P.S. (Ed.), *Understanding Survey Methodology* (pp. 169-191). Cham, Switzerland: Springer.

Mendez-Costabel, M. (2022). Layers and beyond: Modern use of connected spatial and nonspatial datasets to unlock insights in R&D. Presentation at the National Academy of Sciences, Engineering, and Medicine Workshop on The Implications of Using Multiple Data Sources for Major Survey Programs. May 16, 2022. Available: https://www.nationalacademies.org/event/05-16-2022/the-implications-of-using-multiple-data-sources-for-major-survey-programs-workshop

Meng, X.L. (2018). Statistical paradises and paradoxes in big data (I): Law of large populations, big data paradox, and the 2016 US presidential election. *The Annals of Applied Statistics*, *12*(2), 685-726.

Mercer, A.W., Kreuter, F., Keeter, S., Stuart, E.A. (2017). Theory and practice in nonprobability surveys: Parallels between causal inference and survey inference, *Public Opinion Quarterly*, *81*(S1), 250–271. Available: https://doi.org/10.1093/poq/nfw060.

Mercer, A.W., Lau, A., and Kennedy, C. (2018). *For Weighting Online Opt-In Samples, What Matters Most?* Washington, DC: Pew Research.

Available: https://www.census.gov/library/working-papers/2011/adrm/ces-wp-11-14.html

Meyer, B.D. and Mittag, N. (2019). Using linked survey and administrative data to better measure income: Implications for poverty, program effectiveness, and holes in the safety net. *American Economic Journal: Applied Economics*, *11*(2), 176-204.

Meyer, B.D. and Mittag, N. (2021). Combining administrative and survey data to improve income measurement. In Chun, A.Y., Larsen, M.D., Durrant, G., and Reiter, J.P. (Eds.), *Administrative Records for Survey Methodology* (pp. 297-322). Hoboken, NJ, Wiley.

Meyer, B.D., Mittag, N., and Goerge, R.M. (2022) Errors in survey reporting and imputation and their effects on estimates of food stamp program participation. *Journal of Human Resources*, *57*(5), 1605-1644.

Meyer, B.D., Mok, W.K.C., and Sullivan, J.X. (2015). Household surveys in crisis. *Journal of Economic Perspectives*, *29*(4), 199-226.

Meyer, B.D., Wyse, A., Grunwaldt, A., Medalia, C., and Wu, D. (2021a). Learning about homelessness using linked survey and administrative data. *National Bureau of Economic Research Working Paper 28861.* Available: https://www.nber.org/papers/w28861

Meyer, B.D., Wu, D., Mooers, V., and Medalia, C. (2021b). The use and misuse of income data and extreme poverty in the United States. *Journal of Labor Economics*, *39*(S1), S5-S58.

Miller, E.A., Decker, S.L., and Parker, J.D. (2016). Characteristics of Medicare Advantage and fee-for-service beneficiaries upon enrollment in Medicare at Age 65. *Journal of Ambulatory Care Management*, *39*(3), 231-241.

Miller, E.A., McCarty, F.A., and Parker, J.D. (2017). Racial and ethnic differences in a linkage with the National Death Index. *Ethnicity & Disease*, *27*(2), 77.

Mirel, L.B. (2022). Realizing the Power of Health Data through Linkages. Presentation at the National Academy of Sciences, Engineering, and Medicine Workshop on The Implications of Using Multiple Data Sources for Major Survey Programs. May 18, 2022. Available: https://www.nationalacademies.org/event/05-16-2022/docs/D0A0A5F0C0D78B2553B5DFF653C228B66339A1BE3D76

Mirel, L.B., Arispe, I., Helms, V, and Cox, C. (2019a). Assessing children's health in public and assisted housing. In Hart, N. and Yohannes, M. (Eds.), *Evidence Works: Cases Where Evidence Meaningfully Informed Policy*. Washington, D.C.: Bipartisan Policy Center.

Mirel, L.B., Golden, C., Keralis, J.M., Ye, Y., Lloyd, P.C., and Weeks, J.D. (2019b). Evaluating Survey Report of Social Security Disability Benefit Receipt Using Linked National Health Interview Survey and Social Security Administration Data. National Health Statistics Reports #131. Available: https://www.cdc.gov/nchs/data/nhsr/nhsr131-508.pdf

Mirel, L.B., Simon, A.E., Golden, C., Duran, C.R., and Schoendorf, K.C. (2014). Concordance Between Survey Report of Medicaid Enrollment and Linked Medicaid Administrative Records in Two National Studies. National Health Statistics Reports #72. Available: https://www.cdc.gov/nchs/data/nhsr/nhsr072.pdf

Mirel, L.B., Wheatcroft, G., Parker, J.D., and Makuc, D.M. (2012). *Health Characteristics of Medicare Traditional Fee-for-Service and Medicare Advantage Enrollees: 1999-2004 National Health and Nutrition Examination Survey Linked to 2007 Medicare Data*. National Health Statistics Reports; no 53. Hyattsville, MD: National Center for Health Statistics.

Morgan, R.E. and Thompson, A. (2021). Criminal Victimization, 2020. Report NCJ 301775. Washington, DC: Bureau of Justice Statistics. Available: https://bjs.ojp.gov/library/publications/criminal-victimization-2020

Morgan, R.E. and Thompson, A. (2022). The Nation's Two Crime Measures, 2011-2020. Report NCJ 303385. Washington, DC: Bureau of Justice Statistics. Available: https://bjs.ojp.gov/library/publications/nations-two-crime-measures-2011-2020

Morgan, R.E. and Truman, J.L. (2020). Criminal Victimization, 2019. Report NCJ 255113. Washington, DC: Bureau of Justice Statistics. Available: https://bjs.ojp.gov/content/pub/pdf/cv19.pdf

Moyer, B. (2021). National Center for Health Statistics initiatives to use private sector data. Presentation to the National Academy of Sciences, Engineering, and Medicine Panel on The

Scope, Components, and Key Characteristics of a 21st Century Data Infrastructure. December 16, 2021. Available: https://www.nationalacademies.org/event/12-16-2021/the-scope-components-and-key-characteristics-of-a-21st-century-data-infrastructure-workshop-1b.

Mule, T. (2021). Administrative Records and the 2020 Census. Available: https://www.census.gov/newsroom/blogs/random-samplings/2021/04/administrative_recor.html.

Murphy, S.L., Xu, J., Kochanek, K.D., Curtin, S.C., and Arias, E. (2017). Deaths: Final data for 2015. *National Vital Statistics Reports*, *66*(6), 1-73. Available: https://www.cdc.gov/nchs/data/nvsr/nvsr66/nvsr66_06.pdf

Nandram, B., Cruze, N.B., Erciulescu, A.L., and Chen, L. (2022). Bayesian small area models under inequality constraints with benchmarking and double shrinkage. RDD Research Report Number RDD-22-02. Washington, DC: U.S. Department of Agriculture.

Narayanan, A., Stern, A., and Macdonald, G. (2021). *Spatial Equity Data Tool Technical Appendix*. Washington, DC: Urban Institute.

National Academies of Science, Engineering, and Medicine (NASEM). (2016a). *Modernizing Crime Statistics: Report 1: Defining and Classifying Crime.* Lauritsen, J.L. and Cork, D.L. (Eds.) Washington, DC: The National Academies Press.

NASEM. (2016b). *Reducing Response Burden in the American Community Survey: Proceedings of a Workshop*. Washington, DC: The National Academies Press. https://doi.org/10.17226/23639.

NASEM. (2017a). *Federal Statistics, Multiple Data Sources, and Privacy Protection: Next Steps.* Groves, R.M. and Harris-Kojetin, B.A. (Eds.) Washington, DC: The National Academies Press.

NASEM. (2017b). *Improving Crop Estimates by Integrating Multiple Data Sources*. Bock, M.E. and Kirkendall, N.J. (Eds.) Washington, DC: The National Academies Press.

NASEM. (2017c). *Innovations in Federal Statistics: Combining Data Sources While Protecting Privacy*. Groves, R.M. and Harris-Kojetin, B.A. (Eds.) Washington, DC: The National Academies Press.

NASEM. (2018). *Modernizing Crime Statistics: Report 2: New Systems for Measuring Crime.* Lauritsen, J.L. and Cork, D.L. (Eds.) Washington, DC: The National Academies Press.

NASEM. (2020). *2020 Census Data Products: Data Needs and Privacy Considerations: Proceedings of a Workshop*. Cork, D.L., Citro, C.F., and Kirkendall, N.J., Rapporteurs. Washington, DC: The National Academies Press. https://doi.org/10.17226/25978.

NASEM. (2021a). *A Satellite Account to Measure the Retail Transformation: Organizational, Conceptual, and Data Foundations*. Washington, DC: The National Academies Press.

NASEM. (2021b). *Principles and Practices for a Federal Statistical Agency: Seventh Edition*. Harris-Kojetin, B.A. and Citro, C.F. (Eds.) Washington, DC: The National Academies Press.

NASEM. (2022a). *A Vision and Roadmap for Education Statistics*. Hedges, L., Chiu, M., Stone, C., Chaney, B., and Kirkendall, N. (Eds.) Washington, DC: The National Academies Press.

NASEM. (2022b). *Improving Consent and Response in Longitudinal Studies of Aging: Proceedings of a Workshop*. Harris-Kojetin, B., Rapporteur. Washington, DC: The National Academies Press.

NASEM. (2022c). *Measuring Sex, Gender Identity, and Sexual Orientation*. Bates, N., Chin, M., and Becker, T. (Eds.) Washington, DC: The National Academies Press.

NASEM. (2022d). *Modernizing the Consumer Price Index for the 21st Century*. Washington, DC: The National Academies Press.
NASEM. (2022e). *Transparency in Statistical Information for the National Center for Science and Engineering Statistics and All Federal Statistical Agencies*. Washington, DC: The National Academies Press.
NASEM. (2023). *Toward a 21st Century National Data Infrastructure: Mobilizing Information for the Common Good.* Groves, R.M., Mesenbourg, T., and Siri, M. (Eds.) Washington, DC: National Academies Press.
National Archive of Criminal Justice Data. (2022). *Resource Guide: Uniform Crime Reporting Program*. Ann Arbor, MI: University of Michigan. Available: https://www.icpsr.umich.edu/web/pages/NACJD/guides/ucr.html
National Research Council. (1995). *Modernizing the U.S. Census*. Edmonston, B. and Schultze, C. (Eds.) Washington, DC: The National Academies Press.
National Research Council. (2000). *Small-Area Estimates of School-Age Children in Poverty: Evaluation of Current Methodology*, Citro, C.F. and Kalton, G. (Eds). Washington, DC: The National Academies Press.
National Research Council. (2003). *Planning the 2010 Census: Second Interim Report*. Cork, D.L., Cohen, M.L., and King, B.F. (Eds.) Washington, DC: The National Academies Press.
National Research Council. (2004). *The 2000 Census: Counting Under Adversity*. Citro, C.F., Cork, D.L., and Norwood, J.L. (Eds.) Washington, DC: The National Academies Press.
National Research Council. (2009). *Vital Statistics: Summary of a Workshop.* Siri, M.J. and Cork, D.L., Rapporteurs. Washington, DC: The National Academies Press.
Naudé, W., and Vinuesa, R. (2021). Data deprivations, data gaps and digital divides: Lessons from the COVID-19 pandemic. *Big Data & Society*, *8*(2), 1-12.
Nelson, A.H., Jenkins, D., Zanti, S., Katz, M., Berkowitz, E., Burnett, T.C., and Culhane, D. (2020). *A Toolkit for Centering Racial Equity Throughout Data Integration*. Actionable Intelligence for Social Policy, University of Pennsylvania. Available: https://aisp.upenn.edu/wp-content/uploads/2022/07/AISP-Toolkit_5.27.20.pdf
Newman, C. and Scherpf, E. (2013). *Supplemental Nutrition Assistance Program (SNAP) Access at the State and County Levels: Evidence from Texas SNAP Administrative Records and the American Community Survey*. Economic Research Report No 156. Washington, DC: U.S. Department of Agriculture, Economic Research Service.
Nixon, M., Thomas, S. D., Daffern, M., and Ogloff, J.R. (2017). Estimating the risk of crime and victimisation in people with intellectual disability: A data-linkage study. *Social Psychiatry and Psychiatric Epidemiology*, *52*(5), 617-626.
Nkwimi-Tchahou, H., Mohl, C., Reichert, G., and Bédard, F. (2022). The evolution of the use of satellite and administrative data in estimating mid-season field crop yields at Statistics Canada. Presentation at the National Academy of Sciences, Engineering, and Medicine Workshop on The Implications of Using Multiple Data Sources for Major Survey Programs. May 16, 2022. Available: https://www.nationalacademies.org/event/05-16-2022/the-implications-of-using-multiple-data-sources-for-major-survey-programs-workshop
Noon, J.M., Fernandez, L.E., and Porter, S.R. (2019). Response error and the Medicaid undercount in the Current Population Survey. *Health Services Research, 54*(1), 34-43.
Nwaoha-Brown, F., Eliason, J., Sundukchi, M., Farber, J., and Mattingly, T. (2021). *Source and Accuracy Statement for Calendar Year 2020 Data Collection of the Survey of Income and Program Participation (SIPP), Version 1.0.* Washington, DC: U.S. Census Bureau.

Available: https://www2.census.gov/programs-surveys/sipp/tech-documentation/source-accuracy-statements/2020/sipp-2020-SA-02-DEC21.pdf

O'Hara, A., Bee, A., and Mitchell, J. (2016). Preliminary Research for Replacing or Supplementing the Income Question on the American Community Survey with Administrative Records. US Census Bureau Center for Administrative Records Research and Applications Memorandum Series #16-7.

Office for Victims of Crime. (2013). Vision 21: Transforming victim services final report. NCJ-239957. Washington, DC: Office of Justice Programs, US Department of Justice. Available: https://ovc.ojp.gov/library/publications/vision-21-transforming-victim-services-final-report

Organization for Economic Co-operation and Development. (2012). Quality Framework and Guidelines for OECD Statistical Activities Version 2011/1. Publication STD/QFS(2011)1. Paris: OECD. Available: http://www.oecd.org/officialdocuments/displaydocumentpdf/?cote=std/qfs(2011)1&doclanguage=en

Oronce, C.I.A., Scannell, C.A., Kawachi, I., and Tsugawa, Y. (2020). Association between state-level income inequality and COVID-19 cases and mortality in the USA. *Journal of General Internal Medicine*, *35*(9), 2791-2793.

Orvis, K. (2022). Reviewing and revising standards for maintaining, collecting, and presenting data on race and ethnicity. Available: https://www.whitehouse.gov/omb/briefing-room/2022/06/15/reviewing-and-revising-standards-for-maintaining-collecting-and-presenting-federal-data-on-race-and-ethnicity/

Parker, J.D., Kravets, N., and Vaidyanathan, A. (2018). Particulate matter air pollution exposure and heart disease mortality risks by race and ethnicity in the United States: 1997 to 2009 National Health Interview Survey with mortality follow-up through 2011. *Circulation, 137*(16), 1688-1697.

Parrish, J.W., Shanahan, M.E., Schnitzer, P.G., Lanier, P., Daniels, J.L., and Marshall, S.W. (2017). Quantifying sources of bias in longitudinal data linkage studies of child abuse and neglect: Measuring impact of outcome specification, linkage error, and partial cohort follow-up. *Injury Epidemiology*, *4*(1), 1-13.

Pattavina, A., Hirschel, D., and Scearbo, M. (2013). Reliability in NIBRS reporting of substance use in incidents of intimate partner violence. *Justice Research and Policy*, *15*(2), 21-42.

Pedace, R. and Bates, N. (2000). Using administrative records to assess earnings reporting error in the Survey of Income and Program Participation. *Journal of Economic and Social Measurement*, *26*(3-4), 173-192.

Perlman, J. (1951). The Continuous Work-History Sample: The first 12 years. *Social Security Bulletin*, *14*(4), 3-10.

Peressini, T., McDonald, L., and Hulchanski, J. D. (2010). Towards a strategy for counting the homeless. In Hulchanski, J.D., Campsie, P., Chau, S.B.Y., Hwang, S.H., and Paradis, E. (Eds.). *Finding Home: Policy Options for Addressing Homelessness in Canada*, (pp. 728-751). Toronto: Cities Centre Press. Available: https://www.homelesshub.ca/resource/finding-home-policy-options-addressing-homelessness-canada

Peterson, S. and Will, I. (2021). Source and accuracy statement for the 2020 National Crime Victimization Survey, Version 1.0. In U.S. Bureau of Justice Statistics, *National Crime Victimization Survey, [United States], 2020 (ICPSR 38090)* (pp. 1-24). Ann Arbor, MI: Inter-university Consortium for Political and Social Research. Available: https://www.icpsr.umich.edu/web/ICPSR/studies/38090

Peterson, S., Toribio, N., Farber, J., and Hornick, D. (2021). Nonresponse Bias Report for the 2020 Household Pulse Survey, Version 1.0. Washington, DC: U.S. Census Bureau. Available: https://www2.census.gov/programs-surveys/demo/technical-documentation/hhp/2020_HPS_NR_Bias_Report-final.pdf

Pew Research. (2017). Methods 101: Random Sampling [Video]. Available: https://www.pewresearch.org/methods/2017/05/12/methods-101-video-random-sampling/

Piquero, A.R., Scott, K., Smith, E., and Abraham, E. (2022). *Generating National Estimates of Crime Using NIBRS Data: Understanding the Transition.* Webinar held on October 4, 2022.

Planty, M., Banks, D., and Goree, S. (2018). Web-scraping data for official statistics: Examining the periodicity and quality of indicators of crime from law enforcement web sites. Paper presented at the annual conference of the American Association of Public Opinion Research.

Porter, S.R., Liebler, C.A., and Noon, J.M. (2016). An outside view: What observers say about others' races and Hispanic origins. *American Behavioral Scientist*, *60*(4), 465-497.

President's Council of Advisors on Science and Technology. (2014). *Big Data and Privacy: A Technological Perspective*. Washington, DC: Executive Office of the President. Available: https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/PCAST/pcast_big_data_and_privacy_-_may_2014.pdf

President's Task Force on Environmental Health Risks and Safety Risks to Children. (2016). *Key Federal Programs to Reduce Childhood Lead Exposure and Eliminate Associated Health Impacts*. Available: https://ptfceh.niehs.nih.gov/features/assets/files/key_federal_programs_to_reduce_childhood_lead_exposures_and_eliminate_associated_health_impactspresidents_508.pdf

Prevost, R. and Leggieri, C. (1999). Expansion of administrative records uses at the Census Bureau: A long-range research plan. Paper presented at the November, 1999 meeting of the Federal Committee on Statistical Methodology, Washington, DC.

Pujol, D., and Machanavajjhala, A. (2021). Equity and privacy: More than just a tradeoff. *IEEE Security & Privacy*, *19*(6), 93-97.

Rabin, R. (1989). Warnings unheeded: A history of childhood lead poisoning. *American Journal of Public Health, 79*(12), 1668-1674.

Radin, J.M., Wineinger, N.E., Topol, E.J., and Steinhubl, S.R. (2020). Harnessing wearable device data to improve state-level real-time surveillance of influenza-like illness in the USA: a population-based study. *The Lancet Digital Health*, *2*(2), e85-e93.

Raghunathan T, Ghosh K, Rosen A, Imbriano P, Stewart S, Bondarenko I, Messer K, Berglund P, Shaffer J, Cutler D. (2021). Combining information from multiple data sources to assess population health. *Journal of Survey Statistics and Methodology*, *9*(3), 598-625.

Ramirez, R. and Borman, C. (2021). How we complete the census when demographic and housing characteristics are missing. Available: https://www.census.gov/newsroom/blogs/random-samplings/2021/08/census-when-demographic-and-housing-characteristics-are-missing.html.

Randall, M., Stern, A., and Su, Y. (2021). *Five Ethical Risks to Consider Before Filling Missing Race and Ethnicity Data.* Washington, DC: Urban Institute.

Randall, S., Brown, A., Boyd, J., Schnell, R., Borgs, C., and Ferrante, A. (2018). Sociodemographic differences in linkage error: An examination of four large-scale datasets. *BMC Health Services Research*, *18*(1), 1-9.

Rao, J.N.K. (2021). On making valid inferences by combining data from surveys and other sources. *Sankhyā B: The Indian Journal of Statistics*, *83*(1), 242-272.

Rao, J.N.K., and Molina, I. (2015). *Small Area Estimation*. Hoboken, NJ: John Wiley & Sons.

Rässler, S. and Riphahn, R.T. (2006). Survey item nonresponse and its treatment. *Allgemeines Statistisches Archiv*, *90*(1), 217-232.

Ratcliffe, M. (2021a). *Frames Program Overview.* Presentation to the Census Bureau Scientific Advisory Committee. Washington, DC: U.S. Census Bureau. Available: https://www2.census.gov/about/partners/cac/sac/meetings/2021-03/presentation-frames-project.pdf.

Ratcliffe, M. (2021b). *Frames Program Update.* Presentation to the Census Bureau Scientific Advisory Committee. Washington, DC: U.S. Census Bureau. Available: https://www2.census.gov/about/partners/cac/sac/meetings/2021-09/presentation-frames-program-update.pdf.

Regoeczi, W.C. and Banks, D. (2014). The nation's two measures of homicide. Report NCJ 247060, Bureau of Justice Statistics, Washington, DC. Available: https://www.bjs.gov/content/pub/pdf/ntmh.pdf

Reichert, G., Bédard, F., Mohl, C., Benjamin, W., Jiongo, V.D., Chipanshi, A., and Zhang, Y. (2016). Canada—Crop Yield Modelling using Remote Sensing, Agroclimatic Data, and Statistical Survey Data. Proceedings of the Seventh International Conference on Agricultural Statistics (ICAS-VII). Rome: Food and Agriculture Organization. Available: https://www.istat.it/storage/icas2016/g43-reichert.pdf

Reiter, J.P. (2021). Assessing uncertainty when using linked administrative records. In Chun, A.Y., Larsen, M.D., Durrant, G., and Reiter, J.P. (Eds.), *Administrative Records for Survey Methodology* (pp. 139-153). Hoboken, NJ: John Wiley & Sons.

Roberts, T. and Hernandez, K. (2021). *Open Data for Agriculture and Nutrition: A Literature Review and Proposed Conceptual Framework*. Institute of Development Studies Working Paper 545, Brighton, United Kingdom. Available: https://ieeexplore.ieee.org/document/9592826

Robinson, S. and Willyard, K. A. (2021). *Small Area Health Insurance Estimates: 2019*. Current Population Reports P30-09. Washington, DC: U.S. Census Bureau. https://www.census.gov/content/dam/Census/library/publications/2021/demo/p30-09.pdf

Rogers, R.G., Hummer, R.A., and Everett, B.G. (2013). Educational differentials in U.S. adult mortality: An examination of mediating factors. *Social Science Research, 42*(2), 465-481.

Rose Li and Associates (2016). *Expert Meeting on the Potential Value of Centers for Medicare and Medicaid Services Data as a Resource for National Institute on Aging Studies.* Available: https://www.nia.nih.gov/sites/default/files/d7/2016-05-04_nas_cms_meeting_summary_rla.pdf.

Rose Li and Associates. (2019). *Expert Meeting on the Demography of the Older Residential Care Population: Research Questions and Data Gaps.* Available: https://www.nia.nih.gov/sites/default/files/2019-12/Expert-Mtg-Demography-Older-Res-Care-Final-508.pdf.

Rosenfeld, R. (2022). Was crime up or down in 2021? Why the FBI can't tell us. *The Crime Report* (October 11). Available: https://thecrimereport.org/2022/10/11/was-crime-up-or-down-in-2021-why-the-fbi-cant-tell-us/.

Rosenfeld, R. and Lopez, E. (2022). *Pandemic, Social Unrest, and Crime in U.S. Cities: 2021 Year-end Update*. Washington, DC: Council on Criminal Justice. Available: https://councilоncj.org/crime-trends-yearend-2021-update/

Rothbard, A. (2013). *Quality Issues in the Use of Administrative Data Records.* Actionable Intelligence for Social Policy, University of Pennsylvania. Available: https://aisp.upenn.edu/wp-content/uploads/2015/06/Data-Quality-Paper_Final.pdf

Rothbaum, J. (2019). Using Administrative Records to Improve Income and Resource Estimates. Presentation to the Interagency Technical Working Group on Evaluating Alternative Measures of Poverty. Unpublished manuscript.

Rothbaum, J. (2022). National Experimental Well-being Statistics (NEWS). Presentation at the National Academy of Sciences, Engineering, and Medicine Workshop on The Implications of Using Multiple Data Sources for Major Survey Programs. May 18, 2022. Available: https://www.nationalacademies.org/event/05-16-2022/docs/DBDAFDDC7E26834488AD7373BB7FBE330817DE9349A9

Rothbaum, J. and Bee, A. (2021). Addressing nonresponse bias in household surveys using linked administrative data. Available: https://www.aeaweb.org/conference/2022/preliminary/paper/H3e9AQh9.

Rothbaum, J., Eggleston, J., Bee, A., Klee, M., and Mendez-Smith B. (2021). Addressing nonresponse bias in the American Community Survey during the pandemic using administrative data. U.S. Census Bureau Working Paper SEHSD-#2021-24. Available: https://www.census.gov/library/working-papers/2021/acs/2021_Rothbaum_01.html.

Rothwell, C.J., Freedman, M.A., and Weed, J.A. (2014). The National Vital Statistics System. In Magnuson, J. and Fu, Jr., P. (Eds.), *Public Health Informatics and Information Systems* (pp. 309-327). London: Springer.

Ryan, M. (2019). Ethics of using AI and big data in agriculture: The case of a large agriculture multinational. *The ORBIT Journal*, *2*(2), 1-27.

Sakshaug, J.W. and Antoni, M. (2019). Evaluating the utility of indirectly linked federal administrative records for nonresponse bias adjustment. *Journal of Survey Statistics and Methodology*, *7*(2), 227-249.

Salvo, J. (2022). Better Data for Local Decisions: The Promise of the Curated Data Enterprise. Presentation at the National Academy of Sciences, Engineering, and Medicine Workshop on The Implications of Using Multiple Data Sources for Major Survey Programs. May 16, 2022. Available: https://www.nationalacademies.org/event/05-16-2022/docs/D2E2691D8A27ED5C0DA07F3F952CC970B161027C730F.

Santos, R. (2022). Census Bureau modernization: A new vision for an enterprise approach to statistical data. Keynote address at the National Academy of Sciences, Engineering, and Medicine Workshop on The Implications of Using Multiple Data Sources for Major Survey Programs. May 16, 2022. Available: https://www.nationalacademies.org/event/05-16-2022/docs/DB157CEAAA18655B71315C74DDF97C2F757B072C4FCF.

Saralioglu, E. and Gungor, O. (2020). Crowdsourcing in remote sensing: A review of applications and future directions. *Institute of Electrical and Electronics Engineers Geoscience and Remote Sensing Magazine*, *8*(4), 89-110. Available: doi:10.1109/MGRS.2020.2975132

Schleimer, J.P., Pear, V.A., McCort, C.D., Shev, A.B., De Biasi, A., Tomsich, E., Buggs, S., Laqueur, H.S., and Wintemute, G. J. (2022). Unemployment and crime in US cities during the coronavirus pandemic. *Journal of Urban Health*, *99*(1), 82-91.

Schnepf, R. (2017). *NASS and U.S. Crop Production Forecasts: Methods and Issues*. Congressional Research Service Report 44814. Washington, DC: Congressional Research Service.

Schor, E. L. and Johnson, K. (2021). Child health inequities among state Medicaid programs. *JAMA Pediatrics*, *175*(8), 775-776.

Seeskin, Z.H., Ugarte, G., and Datta, A.R. (2019). Constructing a Toolkit to Evaluate Quality of State and Local Administrative Data. *International Journal of Population Data Science*, *4*(1).

Sehra, S.T., George, M., Wiebe, D.J., Fundin, S., and Baker, J.F. (2020). Cell phone activity in categories of places and associations with growth in cases of COVID-19 in the US. *Journal of the American Medical Association Internal Medicine*, *180*(12), 1614-1620.

Shantz, K. and Fox, L.E. (2018). Precision in measurement: Using State-level Supplemental Nutrition Assistance Program and Temporary Assistance for Needy Families Administrative Records and the Transfer Income Model (TRIM3) to Evaluate Poverty Measurement. U.S. Census Bureau SEHSD Working Paper #2018-30.

Shapiro, M. (2021). Panel Discussion: Lessons learned, opportunities, challenges, and next steps. Presentation to the National Academy of Sciences, Engineering, and Medicine Panel on The Scope, Components, and Key Characteristics of a 21st Century Data Infrastructure. December 9, 2021. Available: https://www.nationalacademies.org/event/12-09-2021/the-scope-components-and-key-characteristics-of-a-21st-century-data-infrastructure-workshop-1a

Shrider, E.A., Kollar, M., Chen, F., and Semega, J. (2021). *Income and Poverty in the United States: 2020*. U.S. Census Bureau Current Population Reports P60-273. Washington, DC: U.S. Government Publishing Office. https://www.census.gov/content/dam/Census/library/publications/2021/demo/p60-273.pdf

Sikstrom, L., Maslej, M.M., Hui, K., Findlay, Z., Buchman, D.Z., and Hill, S.L. (2022). Conceptualising fairness: Three pillars for medical algorithms and health equity. *BMJ Health & Care Informatics*, *29*(100459), 1-11. Available: 10.1136/bmjhci-2021-100459

Simon, A.E., Fenelon, A., Helms, V., Lloyd, P.C., and Rossen, L.M. (2017). HUD housing assistance associated with lower uninsurance rates and unmet medical need. *Health Affairs*, 36(6), 1016-1023.

Singh, L., Traugott, M., Bode, L., Budak, C., Davis-Kean, P.E., Guha, R., Ladd, J., Mneimneh, Z., Nguyen, Q., Pasek, J., Raghunathan, T., Ryan, R., Soroka, S., and Wahedi, L. (2020). *Data Blending: Haven't We Been Doing This for Years?* Washington, DC: Georgetown University Massive Data Institute. Available: https://www.jonathanmladd.com/uploads/5/3/6/6/5366295/mdi_data_blending_white_paper_-_april2020.pdf

Skinner, C., and Wakefield, J. (2017). Introduction to the design and analysis of complex survey data. *Statistical Science*, *32*(2), 165-175.

Smith, C. (1989). The Social Security Administration's Continuous Work History Sample. *Social Security Bulletin, 52*(10), 20-28.

Smith, E. (2017). Estimating Costs for Transitioning to the National Incident-Based Reporting System (NIBRS): Guidance for Local Law-Enforcement Agencies. Available: https://bjs.ojp.gov/sites/g/files/xyckuh236/files/media/document/local_agency_-_estimating_cost_for_transitioning_to_nibrs_01232017.pdf

Smith, E. (2022). Panel Discussion: Measuring Crime in the 21st Century. Presentation at the National Academy of Sciences, Engineering, and Medicine Workshop on The Implications of Using Multiple Data Sources for Major Survey Programs. May 16, 2022. Available: https://www.nationalacademies.org/event/05-16-2022/the-implications-of-using-multiple-data-sources-for-major-survey-programs-workshop

Smith, E., Martin, K., Barrick, K., and Richardson, N. (2018). Research in brief: Leveraging NIBRS to better understand sexual violence. *Police Chief Magazine* [online] Available: https://www.policechiefmagazine.org/rib-leveraging-nibrs-sexual-violence/?ref=7430fa6083fd228a6e33ef329bada5

SNACC. (2007). *Phase I Research Results: Overview of National Medicare and Medicaid Files*. Minneapolis: State Health Access Data Assistance Center. Available: https://www.census.gov/library/working-papers/2007/adrm/snacc-phase-1.html.

SNACC. (2010). *Phase V Research Results: Extending the Phase II Analysis of Discrepancies between the National Medicaid Statistical Information System (MSIS) and the Current Population Survey (CPS) Annual Social and Economic Supplement (ASEC) from Calendar Years 2000-2001 to Calendar Years 2002-2005*. Minneapolis: State Health Access Data Assistance Center. Available: https://www.census.gov/library/working-papers/2010/adrm/snacc-phase-5.html.

Sonnega, A. (2017). *Aging in the 21st Century: Challenges and Opportunities for Americans.* Ann Arbor, MI: University of Michigan. Available: https://hrsonline.isr.umich.edu/sitedocs/databook/inc/pdf/HRS-Aging-in-the-21St-Century.pdf

Sonnega, A., Faul, J. D., Ofstedal, M. B., Langa, K. M., Phillips, J. W., and Weir, D. R. (2014). Cohort profile: The Health and Retirement Study (HRS). *International Journal of Epidemiology*, *43*(2), 576-585.

Sourav, A.I. and Emanuel, A.W.R. (2021). Recent trends of big data in precision agriculture: A review. *IOP Conference Series: Materials Science and Engineering*, *1096*, 012081.

Statistics Canada. (2019). *Statistics Canada Quality Guidelines, Sixth Edition*. Ottawa, ON: Statistics Canada. Available: https://www150.statcan.gc.ca/n1/pub/12-539-x/12-539-x2019001-eng.htm

Statistics Canada. (2020a). *An Integrated Crop Yield Model Using Remote Sensing, Agroclimatic Data and Crop Insurance Data*. Ottawa, ON: Statistics Canada. Available: https://www.statcan.gc.ca/en/statistical-programs/document/3401_D2_V1

Statistics Canada. (2020b). *Integrated Crop Yield Modelling Using Remote Sensing, Agroclimatic Data and Survey Data*. Ottawa, ON: Statistics Canada. Available: https://www.statcan.gc.ca/en/statistical-programs/document/5225_D1_T9_V1

Statistics Canada. (2021). *Field Crop Reporting Series*. Ottawa, ON: Statistics Canada. Available: https://www.statcan.gc.ca/en/dai/btd/fcrs

Statistics Canada. (2022). *Development of a Composite Quality Indicator for Statistical Products Derived from Administrative Sources*. Ottawa, ON: Statistics Canada. Available: https://www150.statcan.gc.ca/n1/pub/46-28-0001/2022001/article/00001-eng.htm

Stock, R. and Gardezi, M. (2022). Arrays and algorithms: Emerging regimes of dispossession at the frontiers of agrarian technological governance. *Earth System Governance*, *12*(April), 100137. Available: https://www.sciencedirect.com/science/article/pii/S2589811622000064#!

Stubbs, M. (2016). *Big Data in U.S. Agriculture*. Congressional Research Service Report R-44331. Available: https://fas.org/sgp/crs/misc/R44331.pdf

Stucky, T.D., Payton, S.B., and Ottensmann, J.R. (2016). Intra-and inter-neighborhood income inequality and crime. *Journal of Crime and Justice*, *39*(3), 345-362.

Studds, S. (2021). Blended data in the Census Bureau's Monthly State Retail Sales Data Product. Presentation to the National Academy of Sciences, Engineering, and Medicine Panel on The Scope, Components, and Key Characteristics of a 21st Century Data Infrastructure.

December 9, 2021. Available: https://www.nationalacademies.org/event/12-09-2021/the-scope-components-and-key-characteristics-of-a-21st-century-data-infrastructure-workshop-1a

Swaine, J. and McCarthy, C. (2016). Killings by US police logged at twice the previous rate under new federal program. *The Guardian* (December 15) Available: https://www.theguardian.com/us-news/2016/dec/15/us-police-killings-department-of-justice-program

Swaine, J. and McCarthy, C. (2017). Young black men again faced highest rate of US police killings in 2016. *The Guardian* (January 8). Available: https://www.theguardian.com/us-news/2017/jan/08/the-counted-police-killings-2016-young-black-men

The Economist. (2022). The pulse of the people. *The Economist (Technology Quarterly), 443*(9295), 11-12. Available: https://www.economist.com/technology-quarterly/2022/05/02/data-from-wearable-devices-are-changing-disease-surveillance-and-medical-research.

Tran, H.N., Gerling, M.W., Mitchell, M., and O'Connor, T.P. (2010). Data tables: Reasons for nonresponse in the 2009 June Area Survey. National Agricultural Research Service RDD Research Report RDD-10-04A. Washington, DC: U.S. Department of Agriculture.

Truman, J.L. and Brotsos, H. (2022). Update on the NCVS instrument redesign. Report NCJ 304055. Washington, DC: Bureau of Justice Statistics. Available: https://bjs.ojp.gov/content/pub/pdf/uncvsir_sum.pdf

Truman, J.L. and Morgan, R.E. (2022). Violent victimization by sexual orientation and gender identity, 2017-2020. Report NCJ 304277. Washington, DC: Bureau of Justice Statistics. Available: https://bjs.ojp.gov/library/publications/violent-victimization-sexual-orientation-and-gender-identity-2017-2020

Turner, A.G. (1983). Research on dual-frame sampling. In Lehnen, R.G. and Skogan, W. (Eds.), *The National Crime Survey: Working Papers. Volume 2: Methodological Studies.* NCJ-90307 (p.120). Washington DC: Bureau of Justice Statistics. Available: https://www.ojp.gov/pdffiles1/nij/90307.pdf

Uhl, S. (2011). Building and maintaining the Master Address File. Available: https://www2.census.gov/geo/pdfs/education/Uhl_CAS_2011.pdf.

United Kingdom Statistics Authority. (2019). Administrative Data Quality Assurance Toolkit. Available: https://osr.statisticsauthority.gov.uk/wp-content/uploads/2019/02/qualityassurancetoolkitupdatedFeb192.pdf

United Nations (2019). *United Nations National Quality Assurance Frameworks Manual for Official Statistics.* New York: United Nations. Available: https://desapublications.un.org/publications/united-nations-quality-assurance-frameworks-manual

United Nations Economic and Social Council. (2019). *In-depth Review of Satellite Imagery/Earth Observation Technology in Official Statistics.* Geneva: United Nations. Available: https://unece.org/DAM/stats/documents/ece/ces/2019/ECE_CES_2019_16-1906490E.pdf

United Nations Inter-Secretariat Working Group on Household Surveys. (2022). *Positioning Household Surveys for the Next Decade*. New York: United Nations Statistical Division. Available: https://unstats.un.org/unsd/statcom/53rd-session/documents/BG-3a-Position-Household-Survey-for-Next-Decade-E.pdf

U.S. Bureau of the Census. (1947a). Estimated population of the United States, by regions, divisions, and states: July 1, 1946. Current Population Reports, Series P-25, No. 2. Washington, DC: U.S. Bureau of the Census.
U.S. Bureau of the Census. (1947b). Suggested procedures for estimating the current population of counties. Population – Special Reports, Series P-47, No. 4. Washington, DC: U.S. Bureau of the Census.
U.S. Bureau of the Census. (1967). Estimates of the population of standard metropolitan statistical areas, July 1, 1965. Current Population Reports, Series P-25, No. 371. Washington, DC: U.S. Bureau of the Census.
U.S. Bureau of the Census. (1968). Estimates of the population of counties, July 1, 1966. Current Population Reports, Series P-25, No. 401. Washington, DC: U.S. Bureau of the Census.
U.S. Bureau of the Census. (1979). The Standard Statistical Establishment List Program. Technical Paper 44. Washington, DC: U.S. Bureau of the Census.
U.S. Bureau of Justice Statistics. (2021a). National Crime Statistics Exchange (NCS-X). Available: https://bjs.ojp.gov/programs/national-crime-statistics-exchange
U.S. Bureau of Justice Statistics. (2021b). National Crime Victimization Survey, [United States], 2020. ICPSR 30890. Ann Arbor, MI: Inter-university Consortium for Political and Social Research
U.S. Bureau of Justice Statistics. (2022a). NCVS Subnational Estimates. Available: https://bjs.ojp.gov/subnational-estimates-program
U.S. Bureau of Justice Statistics. (2022b). NIBRS Estimation Summary. Report NCJ 305107. Available: https://bjs.ojp.gov/library/publications/nibrs-estimation-summary.
U.S. Census Bureau. (2002). *Measuring America: The Decennial Censuses From 1790 to 2000*. Report POL/02-MA. Washington DC: US Department of Commerce. Available: https://www.census.gov/library/publications/2002/dec/pol_02-ma.html
U.S. Census Bureau. (2014). *American Community Survey Design and Methodology*. Washington, DC: US Census Bureau. Available : https://www.census.gov/programs-surveys/acs/methodology/design-and-methodology.html
U.S. Census Bureau. (2019). *Current Population Survey Design and Methodology*. Technical Paper 77. Washington, DC: U.S. Census Bureau. Available: https://www.census.gov/programs-surveys/cps/technical-documentation/complete.html
U.S. Census Bureau (2020a). *Your Guide to the 2020 Census: How to Respond to the 2020 Census Paper Questionnaire*. Washington, DC: U.S. Census Bureau. https://www2.census.gov/programs-surveys/decennial/2020/resources/language-materials/guides/English-Guide.pdf
U.S. Census Bureau (2020b). *2020 Census: Our Mission to Count Everyone.* https://www.test.census.gov/content/dam/Census/library/factsheets/2020/dec/mission-count-everyone/mission-count-everyone.pdf
U.S. Census Bureau. (2021a). *2010-2020 County-Level Estimation Details*. Washington, DC: U.S. Census Bureau. Available: https://www.census.gov/programs-surveys/saipe/technical-documentation/methodology.html
U.S. Census Bureau. (2021b). *2020 Survey of Income and Program Participation Users' Guide.* Available: https://www2.census.gov/programs-surveys/sipp/tech-documentation/methodology/2020_SIPP_Users_Guide_OCT21.pdf
U.S. Census Bureau. (2021c). *Disclosure Avoidance for the 2020 Census: An Introduction. Washington DC: U.S. Census Bureau*. Available:

https://www2.census.gov/library/publications/decennial/2020/2020-census-disclosure-avoidance-handbook.pdf

U.S. Census Bureau. (2021d). *Monthly State Retail Sales Technical Documentation*. Washington DC: US Census Bureau. Available: https://www.census.gov/retail/mrts/www/statedata/msrs_technical_documentation.pdf

U.S. Census Bureau. (2021e). *National Longitudinal Mortality Study (NLMS): Project Overview*. Washington DC: US Census Bureau. Available: https://www.census.gov/topics/research/nlms.html

U.S. Census Bureau. (2021f). *U.S. Decennial Census Measurement of Race and Ethnicity Across the Decades: 1790-2020* [Infographic]. Available: https://www.census.gov/library/visualizations/interactive/decennial-census-measurement-of-race-and-ethnicity-across-the-decades-1790-2020.html.

U.S. Census Bureau. (2022a). March 2021 Annual Social and Economic Supplement (ASEC): Complete Technical Documentation. Available at: https://www2.census.gov/programs-surveys/cps/techdocs/cpsmar21.pdf

U.S. Census Bureau. (2022b). Research to improve data on race and ethnicity. Available: https://www.census.gov/about/our-research/race-ethnicity.html

U.S. Census Bureau. (2022c). Source of the Data and Accuracy of the Estimates for the Household Pulse Survey - Phase 3.5. Washington, DC: U.S. Census Bureau. Available: https://www.census.gov/programs-surveys/household-pulse-survey/technical-documentation.html

U.S. Centers for Disease Control and Prevention (CDC). (2021a). Data Modernization Initiative Strategic Implementation Plan (December 22, 2021 version). Available: https://www.cdc.gov/surveillance/pdfs/FINAL-DMI-Implementation-Strategic-Plan-12-22-21.pdf

U.S. Centers for Disease Control and Prevention. (2021b). U.S. influenza surveillance: Purpose and methods. Available: https://www.cdc.gov/flu/weekly/overview.htm

U.S. Centers for Disease Control and Prevention. (2022). National Violent Death Reporting System (NVDRS). Available: https://www.cdc.gov/violenceprevention/datasources/nvdrs/

U.S. Centers for Medicare and Medicaid Services. (2021). Medicare Current Beneficiary Survey, 2019: Data User's Guide: Survey File Public Use File. Office of Enterprise Data and Analytics. Available: https://data.cms.gov/medicare-current-beneficiary-survey-mcbs/medicare-current-beneficiary-survey-data.

U.S. Commission on Evidence-Based Policymaking. (2017). *The Promise of Evidence-Based Policymaking*. Washington, DC: Commission on Evidence-Based Policymaking.

U.S. Congress (1845). Public Statutes at Large of the United States of America from the Organization of Government in 1789, to March 3, 1845. Available: https://tile.loc.gov/storage-services/service/ll/llsl/llsl-c1/llsl-c1.pdf

U.S. Congress (1992). The Residential Lead-Based Paint Hazard Reduction Act of 1992. Available: https://www.congress.gov/102/statute/STATUTE-106/STATUTE-106-Pg3672.pdf.

U.S. Congress (1994). Census Address List Improvement Act of 1994. Available: https://www.congress.gov/103/statute/STATUTE-108/STATUTE-108-Pg4393.pdf.

U.S. Congress (2016). Evidence-Based Policymaking Commission Act of 2016. Available: https://www.congress.gov/114/statute/STATUTE-130/STATUTE-130-Pg317.pdf

U.S. Congress (2019). Foundations for Evidence-Based Policy Making Act of 2018. P.L. 115-435. Available: https://www.congress.gov/115/plaws/publ435/PLAW-115publ435.pdf.

U.S. Department of Agriculture. (1969). The story of U.S. agricultural estimates. Statistical Reporting Service Miscellaneous Publication No. 1088. Washington, DC. Available: https://www.nass.usda.gov/About_NASS/pdf/The%20Story%20of%20U.S.%20Agricultural%20Estimates.pdf

U.S. Department of Agriculture. (2019). 2017 Census of Agriculture: Appendix A. Washington, DC: US Department of Agriculture. Available: https://www.nass.usda.gov/Publications/AgCensus/2017/Full_Report/Volume_1,_Chapter_1_US/usappxa.pdf

U.S. Department of Housing and Urban Development (HUD). (2021). *American Healthy Homes Survey II Lead Findings*. Washington, DC: HUD. Available: https://www.hud.gov/sites/dfiles/HH/documents/AHHS_II_Lead_Findings_Report_Final_29oct21.pdf

U.S. Environmental Protection Agency (EPA). (2022). Biomonitoring - Lead, in America's Children and the Environment. Washington, DC, US EPA. Available: https://www.epa.gov/americaschildrenenvironment/biomonitoring-lead

U.S. Federal Bureau of Investigation (FBI). (2013). *Summary Reporting System (SRS) User Manual,Version 1.0*. Washington, DC: U.S. Department of Justice. Available: https://le.fbi.gov/file-repository/summary-reporting-system-user-manual.pdf/view

U.S. Federal Bureau of Investigation. (2021). *2021.1 National Incident-Based Reporting System User Manual*. April 15, 2021 version. Washington DC: FBI. Available: https://www.fbi.gov/file-repository/ucr/ucr-2019-1-nibrs-user-manua-093020.pdf

U.S. Federal Bureau of Investigation. (2022a). *NIBRS Estimation FAQs.* Washington, DC: FBI. Available: https://cde.ucr.cjis.gov.

U.S. Federal Bureau of Investigation. (2022b). *The Transition to the National Incident-Based Reporting System (NIBRS): A Comparison of 2020 and 2021 NIBRS Estimates.* Washington, DC: FBI.

U.S. Federal Trade Commission. (2022). *Consumer Sentinel Network Data Book 2021*. Washington, DC: Federal Trade Commission. https://www.ftc.gov/reports/consumer-sentinel-network-data-book-2021

U.S. Government Accountability Office. (2016). *Sexual Violence Data: Actions Needed to Improve Clarity and Address Differences Across Federal Data Collection Efforts*. GAO-16-546. Washington, DC: U.S. Government Accountability Office. Available: https://www.gao.gov/assets/680/678511.pdf

U.S. Government Accountability Office. (2020). *2020 Census: The Bureau Concluded Field Work but Uncertainty about Data Quality, Accuracy, and Protection Remains*. GAO Report 21-206R. Washington, DC: Government Accountability Office. Available: https://www.gao.gov/assets/gao-21-206r.pdf

U.S. Internal Revenue Service (IRS). (2019). *Federal Tax Compliance Research: Tax Gap Estimates for Tax Years 2011-2013*. Publication 1415 (Rev. 9-2019). Washington, DC: Internal Revenue Service. Available: https://www.irs.gov/pub/irs-pdf/p1415.pdf.

U.S. National Agricultural Statistics Service (NASS). (2012). The Yield Forecasting Program of NASS. Statistical Methods Branch Staff Report Number SMB 12-01. Available: https://www.nass.usda.gov/Publications/Methodology_and_Data_Quality/Advanced_Topics/Yield%20Forecasting%20Program%20of%20NASS.pdf

U.S. NASS. (2022). Annual Crop Production Methodology and Quality Measures. Available: https://www.nass.usda.gov/Publications/Methodology_and_Data_Quality/Crop_Production/01_2022/cranqm22.pdf

U.S. National Center for Health Statistics (NCHS. (2009). *The National Health Interview Survey (1986-2004) Linked Mortality Files, Mortality Follow-up through 2006: Matching Methodology*. Hyattsville, MD: National Center for Health Statistics. Available: https://www.cdc.gov/nchs/data/datalinkage/matching_methodology_nhis_final.pdf.

U.S. National Center for Health Statistics. (2013). National Death Index User's Guide. Hyattsville, MD: National Center for Health Statistics. Available: https://www.cdc.gov/nchs/ndi/resources.htm

U.S. National Center for Health Statistics. (2016). *The Linkage of National Center for Health Statistics Surveys to Medicare Enrollment and Claims Data (1999-2013) - Methodology and Analytic Considerations*. Hyattsville, MD: National Center for Health Statistics. Available: https://www.cdc.gov/nchs/data-linkage/cms/nchs_medicare_linkage_methodology_and_analytic_considerations.pdf

U.S. National Center for Health Statistics. (2019). The Linkage of National Center for Health Statistics Survey Data to Medicaid Enrollment and Claims Data: Methodology and Analytic Considerations. Hyattsville, MD: National Center for Health Statistics. Available: https://www.cdc.gov/nchs/data/datalinkage/nchs-medicaid-linkage-methodology-and-analytic-considerations.pdf

U.S. National Center for Health Statistics. (2020a). National Death Index. NCHS Fact Sheet. Available: https://www.cdc.gov/nchs/data/factsheets/factsheet_ndi.pdf

U.S. National Center for Health Statistics. (2020b). National Health and Nutrition Examination Survey. NCHS Fact Sheet. Available: https://www.cdc.gov/nchs/data/factsheets/factsheet_nhanes.pdf

U.S. National Center for Health Statistics. (2020c). National Health Interview Survey. NCHS Fact Sheet. Available: https://www.cdc.gov/nchs/data/factsheets/factsheet_nhis.pdf

U.S. National Center for Health Statistics. (2021a). National Vital Statistics System. NCHS Fact Sheet. Available: https://www.cdc.gov/nchs/data/factsheets/factsheet_NVSS.pdf

U.S National Center for Health Statistics. (2021b). National Vital Statistics System Improvements. Hyattsville, MD: National Center for Health Statistics. Available: https://www.cdc.gov/nchs/data/factsheets/2020-NVSS-improvement-factsheet-508.pdf

U.S. National Center for Health Statistics. (2021c). *The Linkage of National Center for Health Statistics Survey Data to Medicare Enrollment, Claims/Encounters and Assessment data (2014-2018): Linkage Methodology and Analytic Considerations*. Hyattsville, MD: National Center for Health Statistics. Available: https://www.cdc.gov/nchs/data-linkage/cms/nchs_medicare14_18_linkage_methodology_and_analytic_considerations.pdf

U.S. National Center for Health Statistics. (2022a). About the National Death Index. Hyattsville, MD: National Center for Health Statistics Available: https://www.cdc.gov/nchs/ndi/about.htm

U.S. National Center for Health Statistics. (2022b). *National Health Interview Survey, 2021: Survey Description*. Hyattsville, MD: National Center for Health Statistics. Available: https://ftp.cdc.gov/pub/Health_Statistics/NCHS/Dataset_Documentation/NHIS/2021/srvydesc-508.pdf

U.S. National Center for Health Statistics. (2022c). *The Linkage of the National Center for Health Statistics (NCHS) Survey Data to U.S. Department of Housing and Urban*

Development (HUD) Administrative Data: Linkage Methodology and Analytic Considerations. Hyattsville, MD: National Center for Health Statistics. Available: https://www.cdc.gov/nchs/data/datalinkage/NCHS-HUD-Linked-Data-Methodology-and-Analytic-Considerations.pdf.

U.S. National Center for Health Statistics. (2022d). *The Linkage of National Center for Health Statistics Survey Data to the National Death Index — 2019 Linked Mortality File (LMF): Linkage Methodology and Analytic Considerations.* Hyattsville, Maryland: National Center for Health Statistics. Available: https://www.cdc.gov/nchs/data/datalinkage/2019ndi-linkage-methods-and-analytic-considerations-508.pdf.

U.S. Office of Management and Budget (OMB). (1997). Revisions to the standards for the classification of federal data on race and ethnicity. *Federal Register, 62*(210), 58782-58790. Available: https://www.govinfo.gov/content/pkg/FR-1997-10-30/pdf/97-28653.pdf

U.S. Office of Management and Budget. (2002). Guidelines for ensuring and maximizing the quality, objectivity, utility, and integrity of information disseminated by federal agencies. *Federal Register, 67*(36), 8452-8460. Available: https://www.govinfo.gov/content/pkg/FR-2002-02-22/pdf/R2-59.pdf

U.S. Office of Management and Budget. (2006). *Standards and Guidelines for Statistical Surveys*. Washington, DC: Office of Management and Budget. Available: https://obamawhitehouse.archives.gov/sites/default/files/omb/inforeg/statpolicy/standards_stat_surveys.pdf

U.S. Office of Management and Budget. (2014). Statistical Policy Directive No. 1: Fundamental Responsibilities of Federal Statistical Agencies and Recognized Statistical Units. *Federal Register, 79*(231), 71610-71616. Available: https://www.govinfo.gov/content/pkg/FR-2014-12-02/pdf/2014-28326.pdf

U.S. Office of Management and Budget. (2016). Standards for maintaining, collecting, and presenting federal data on race and ethnicity. *Federal Register, 81*(190), 67398-67401. Available: https://www.govinfo.gov/content/pkg/FR-2016-09-30/pdf/2016-23672.pdf

U.S. Office of Management and Budget. (2019a). Federal Data Strategy—A Framework for Consistency. OMB Memorandum M-19-18. Available: https://www.whitehouse.gov/wp-content/uploads/2019/06/M-19-18.pdf

U.S. Office of Management and Budget. (2019b). Improving Implementation of the Information Quality Act. OMB Memorandum M-19-15. Available: https://www.cdo.gov/assets/documents/OMB-Improving-Implementation-of-Info-Quality-Act-M-19-15.pdf

U.S. Office of Management and Budget. (2021). Evidence-Based Policymaking: Learning Agendas and Annual Evaluation Plans. OMB Memorandum M-21-27. Available: https://www.whitehouse.gov/wp-content/uploads/2021/06/M-21-27.pdf

Urban Institute. (2021a). Introducing the Spatial Equity Data Tool Version 2. Available: https://urban-institute.medium.com/introducing-the-spatial-equity-data-tool-version-2-f2a8e900f84

Urban Institute. (2021b). Spatial Equity Data Tool. Available: https://apps.urban.org/features/equity-data-tool/

Van Buuren, S. (2018). *Flexible Imputation of Missing Data*. Boca Raton, FL: CRC Press.

Veitenheimer, D. (2022). Panel Discussion: Measuring Crime in the 21st Century. Presentation at the National Academy of Sciences, Engineering, and Medicine Workshop on The Implications of Using Multiple Data Sources for Major Survey Programs. May 16, 2022.

Available: https://www.nationalacademies.org/event/05-16-2022/the-implications-of-using-multiple-data-sources-for-major-survey-programs-workshop

Wadsworth, T., and Roberts, J.M. (2008). When missing data are not missing: A new approach to evaluating Supplemental Homicide Report imputation strategies. *Criminology*, *46*(4), 841-870.

Wagner, D. and Layne, M. (2014). The Person Identification Validation System (PVS): Applying the Center for Administrative Records Research and Applications' (CARRA) Record Linkage Software. U.S. Census Bureau CARRA Working Paper #2014-01. Available: https://www.census.gov/content/dam/Census/library/working-papers/2014/adrm/carra-wp-2014-01.pdf

Wang, S. (2016). CyberGIS and spatial data science. *GeoJournal*, 81(6), 965-968.

Wang, S., and Goodchild, M.F. (2019). *CyberGIS for Geospatial Innovation and Discovery.* Dordrecht, Netherlands: Springer.

Wardell, C. (2022). Fireside chat on data equity. National Academies of Sciences, Engineering, and Medicine Workshop on The Implications of Using Multiple Data Sources for Major Survey Programs. May 18, 2022. Available: https://www.nationalacademies.org/event/05-16-2022/the-implications-of-using-multiple-data-sources-for-major-survey-programs-workshop

Weir, D., Faul, J., and Langa, K. (2011). Proxy interviews and bias in the distribution of cognitive abilities due to non-response in longitudinal studies: A comparison of HRS and ELSA. *Longitudinal and Life Course Studies*, *2*(2), 170-184.

Williams, B.A., Brooks, C.F. and Shmargad, Y. (2018). How algorithms discriminate based on data they lack: Challenges, solutions, and policy implications. *Journal of Information Policy, 8*(March), 78-115.

Williams, D. and Brick, J.M. (2018). Trends in U.S. face-to-face household surveys and level of effort. *Journal of Survey Statistics and Methodology*, *6*(2), 186-211.

Wu, C. (2022). Statistical inference with non-probability survey samples. *Survey Methodology, 48*(2), 283-311.

Xu, J., Murphy, S.L., Kochanek, K., and Arias, E. (2021). *Deaths: Final Data for 2019*. Hyattsville, MD: National Center for Health Statistics. Available: https://stacks.cdc.gov/view/cdc/106058.

Young, L.J. (2019). Agricultural crop forecasting for large geographical areas. *Annual Review of Statistics and Its Application*, *6*(March), 173-196.

Young, L.J. (2022). The crops county estimates program: Developing official statistics based on available data. Presentation at the National Academy of Sciences, Engineering, and Medicine Workshop on The Implications of Using Multiple Data Sources for Major Survey Programs. May 16, 2022. Available: https://www.nationalacademies.org/event/05-16-2022/the-implications-of-using-multiple-data-sources-for-major-survey-programs-workshop

Young, L.J. and Chen, L. (2022). Using small area estimation to produce official statistics. *Stats 5*,881-897.

Young L.J., Hyman, M., and Rater, B.R. (2018). Exploring a big data approach to building a list frame for urban agriculture: A pilot study in the City of Baltimore. *Journal of Official Statistics, 34*(2), 323-340.

Young, L.J. and Jacobsen, M. (2022). Sample design and estimation when using a web-scraped list frame and capture-recapture methods. *Journal of Agricultural, Biological and Environmental Statistics*, *27*, 261-279.

Zablotsky, B. and Black, L.I. (2019). Concordance between survey reported childhood asthma and linked Medicaid administrative records. *Journal of Asthma*, *56*(3), 285-295.

Zhang, C. and Lu, Y. (2021). Study on artificial intelligence: The state of the art and future prospects. *Journal of Industrial Information Integration*, *23*(100224), 1-11.

Ziliak, J.P., Hokayem, C., and Bollinger, C.R. (2022). Trends in Earnings Volatility using Linked Administrative and Survey Data, *Journal of Business & Economic Statistics*, 1-11. Available: https://doi.org/10.1080/07350015.2022.2102023

Zook, M., Barocas, S., Boyd, D., Crawford, K., Keller, E., Gangadharan, S.P., Goodman, A., Hollander, R., Koenig, B.A., Metcalf, J., Narayanan, A., Nelson, A., and Pasquale, F. (2017). Ten simple rules for responsible big data research. *PLoS Computational Biology, 13*(3), e1005399.

COMMITTEE ON NATIONAL STATISTICS

The Committee on National Statistics was established in 1972 at the National Academies of Sciences, Engineering, and Medicine to improve the statistical methods and information on which public policy decisions are based. The committee carries out studies, workshops, and other activities to foster better measures and fuller understanding of the economy, the environment, public health, crime, education, immigration, poverty, welfare, and other public policy issues. It also evaluates ongoing statistical programs and tracks the statistical policy and coordinating activities of the federal government, serving a unique role at the intersection of statistics and public policy. The committee's work is supported by a consortium of federal agencies through a National Science Foundation grant, a National Agricultural Statistics Service cooperative agreement, and several individual contracts.