USING ADMINISTRATIVE RECORDS IN STATISTICS AND SOCIAL SCIENCE RESEARCH

April 8, 2022

Teja Pristavec
Research Scientist,
Coleridge Initiative
https://orcid.org/0000-0003-2021-4682
teja.pristavec@coleridgeinitiative.org

Stephanie Shipp
Deputy Director and Professor,
Social and Decision Analytics Division
https://orcid.org/0000-0002-2142-2136
sss5sc@virginia.edu

Sallie Keller
Distinguished Professor in Biocomplexity,
Division Director
Social and Decision Analytics Division
https://orcid.org/0000-0001-7303-7267
sak9tr@virginia.edu

UVA | BIOCOMPLEXITY INSTITUTE

Acknowledgments: We would like to thank the Alfred P. Sloan Foundation, Grants G-2019-11316 and G-2020-14002 for their support of this research. We would also like to thank Ken Prewitt, Columbia University, Steve Jost, Subject Matter, John Thompson, Biocomplexity Distinguished Institute Fellow, Joseph Salvo, Biocomplexity Institute Fellow, and Sarah Nusser, Biocomplexity Institute Research Professor for their review of this work.

Citation: Pristavec, T., Shipp, S., and Keller, S. (2022). Technical Report: Using Administrative Records in Official Statistics and Social Science Research. Proceedings of the Biocomplexity Institute, TR 2022-025. https://doi.org/10.18130/b7b9-x755

Table of Contents

Abstract

National statistical offices typically rely on census population counts and surveys to estimate social, demographic, and economic changes, as well as program use and other indicators crucial for funding allocation and effective decision-making. These same data are foundational to social science research. However, declining response rates, increasing costs, and societal change outpacing field data collection challenges survey quality. This timely topic lays the groundwork for a new way of thinking about information about our country –people, places, and the economy. To address these issues, statistical offices, state and local governments, and social scientists have been exploring the use of government and private sector-generated administrative data to provide estimates that are timelier, more geographically granular, and less costly.

Technical Report: Using Administrative Records in Official Statistics and Social Science Research

Introduction

National statistical offices typically rely on census population counts and surveys to estimate social, demographic, and economic changes, as well as program use, and other indicators crucial for funding allocation and effective decision-making. These same data are foundational to social science research. However, declining response rates, increasing costs, and societal change outpacing field data collection pose challenges to survey quality (Miller, 2017). This is a timely topic that lays the groundwork for a new way of thinking about information about our country –people, places, and the economy.

To address these issues, statistical offices, state and local governments, and social scientists have been exploring the use of government and private sector-generated administrative data to provide estimates that are timelier, more geographically granular, and less costly (Allard et al., 2018; Groves & Schoeffel, 2018; NASEM, 2017a; NASEM 2017B; Playford et al., 2016). In the last two decades, the potential for administrative data to supplement or replace designed surveys has continued to grow (Auerbach et al., 2019). For example, administrative records capture increasingly larger population shares (Bauder & Judson, 2003; Rastogi & O'Hara, 2012), and technical advances in record linkage and entity resolution have improved our ability to follow individuals, businesses, and processes across time (Harron et al., 2017). The Evidence-Based Policymaking Commission spurred further interest in the potential value of integrating federal survey and administrative data and making those data available to researchers for program evaluation (US House of Representatives 2019). The foundation for implementing the Commission's recommendations has been passed in recent legislation (US House of Representatives, 2019).

In light of these advancements, work is underway to reconceptualize how survey, administrative and private sector data may be integrated to provide more timely, accurate and granular information to policy makers and the public. Toward that end, this report reviews the proliferation, challenges, and future directions of using administrative data sources in national statistical offices and social science research. We first document the growing success and accelerated progress of using and linking administrative records to survey and private sector data to provide estimates and official statistics at local, state, and federal levels. We then summarize both the current technical and societal challenges that will need to be overcome for administrative data to become a common source of such estimates. Finally, we conclude with best practices, recommendations, and future directions for expanding administrative data applications in official statistics and social science research.

The Proliferation of Administrative Data Use as a Response to Survey Challenges, Societal Transformation, and Policy Environment Changes

Surveys have long been the primary data source for official statistics and research studies in social sciences. However, rapid societal and technological changes present opportunities for monitoring and measuring demographic, economic, and other aspects of social life beyond traditional data collection methods like surveys. Statistical agencies that rely on such data must increasingly address the need for more complex and timely data in producing their estimates while facing falling response rates and increased fielding costs coupled with restricted budgets (Bostic et al., 2016). In addition, statistical agencies and other actors exist in a changing policy environment that now emphasizes efficient data use. These three trends—societal change giving rise to new data opportunities, challenges to survey data quality, and new data-related policies—provided grounds for increasing the use of administrative sources in official statistics and social science research.

First, the quick pace of societal change in the past two decades gave rise to new technologies, increased digitization, and new data sources and types (World Economic Forum, 2011). Techniques like rapid scanning, text recognition, user-friendly uploads, and new devices, sensors, and systems can now record and transcribe data in real time. Using these techniques, governments and corporations now routinely and instantaneously collect and store data on behaviors and states as varied as purchases and transactions, climate or road conditions, health care plan utilization, or land use and zoning. Extensive digitization and recording, better system connectedness and interactivity, and increased human-computer interaction also result in faster data accumulation (Brady, 2019). These processes facilitate the creation of digitized administrative data sources that can provide valuable and timely insights increasingly hard to capture using surveys. These changes and opportunities creates the need to pay active attention to data security that is discussed later in this report. There is also a cadre of people who advocate for individual data rights, e.g., patient's rights in deciding how data should be shared.

Second and coupled with such social change, survey response rates in general and federal survey response rates in particular are in decline. Individuals are challenging to reach by phone due to new technologies like voice mail and caller identification, an increase in the number of households with mobile phones and smart devices only, and new legislation like the Telephone Consumer Protection Act that bans using automatic phone dialing systems without consent (Couper, 2017). Data collection partly adapted to these new trends by shifting from in-person and landline phone to mixed-mode, cell phone call and text, and web-based collection (Couper, 2017; Miller, 2017), but the declining trend holds. For example, the National Health Interview Survey response rate dropped from 91% in 1997 to 73% in 2014; the Medicare Current Beneficiary Survey response rate dropped from 87% in 1991 to 72% in 2013; and Medical Expenditure Panel Survey Household Component response declined from 93% in 1996 to 76% in 2014 (Czajka & Beyler, 2016). These declining response rates can lead to increased survey fielding costs, requiring more resources to increase participation and avoid insufficient and unrepresentative samples (Miller, 2017).

Third, the policy environment responded to technological, social, and survey changes by encouraging efficient use of existing data, reuse, sharing, and furthering open data principles. In the last decade, documents like the 2009 Open Government Directive in the US mandated that federal agencies release at least three non-sensitive data sources for public use online, identify other data sources that could eventually be released, and devise a timeline for doing so (US Office of Management and Budget [OMB], 2009). The bilateral Open Government Partnership, founded in 2011 by the US, UK, Brazil, Indonesia, Mexico, Norway, Philippines, and South Africa, prompted the development of Open Government National Action Plans that included facilitating public data access, releasing FOIA request data online, improving record management, and constructing online data tools. The 2013 US Open Data Policy mandated that federal government agencies strive for open data as their default (OMB, n.d.; OMB, 2013). Finally, the OMB released memoranda on data sharing while protecting privacy, open data, and providing and using administrative data for statistical purposes (OMB, 2000; OMB, 2014). The National Academies of Sciences issued two reports outlining current and future federal statistical uses of administrative data and propose the creation of a new entity to facilitate secure access to data for statistical purposes to enhance the quality of federal statistics (NASEM 2017a, NASEM 2017b, Commission on Evidence-Based Policymaking 2017).

More recently, two further developments reshaped attitudes towards data sharing at the U.S. federal government level. In 2017, the Commission on Evidence-Based Policymaking released a report recommending the creation of a data infrastructure that would become a routine part of government operations and facilitate efficient, evidence-based policy- and decision-making while encouraging survey and program data sharing between statistical and regulatory agencies (Commission on Evidence-Based Policymaking, 2017). The Foundations for Evidence-Based Policymaking Act (H.R.4174), signed into law in 2019, addressed and incorporated ten of the Commission's recommendations. The Act facilitates public access to government data—including data based on administrative records—and promotes its use by requiring agencies to develop data inventories and respond to data access requests. The Act also addresses disclosure concerns, instituting a committee on privacy protection, mandating confidentiality risk assessments, and reauthorizing the Confidential Information Protection and Statistical Efficiency Act (CIPSEA). Taken together, these two Acts provided grounds and impetus for the discovery, use, and sharing of new data sets within official statistics and social science communities.

In response to these changes and rather than dedicating a greater part of their budgets to maintaining sample representativeness, statistical offices and researchers began exploring alternative data sources to supplement or replace survey data (Couper, 2017; Miller, 2017; NASEM, 2017a; Wallgren & Wallgren, 2014). Although surveys remain a valuable source of contextual and attitudinal information, administrative records emerge as an inexpensive and well-structured but under-used data source that often captures large parts of the population and are already housed within government and private sector data infrastructures. Administrative data can be used to improve the quality of household survey data. These include evaluating survey data quality, improving sampling frames and weights, linking data sources, imputing missing data, providing covariates for small

area estimation, and replacing survey questions directly (Citro (2014, p. 152). The following section provides an overview of current administrative data uses in U.S. official statistics at the federal and state levels.

Current Uses of Administrative Data

The Census Bureau defines administrative records and third-party data as: "micro data records contained in files collected and maintained by administrative (i.e., program) agencies and commercial entities. Government and commercial entities maintain these files for the purpose of administering programs and providing services. Administrative records are distinct from systems of information collected exclusively for statistical purposes, such as those the U.S. Census Bureau produces under the authority of Title 13 of the United States Code (U.S.C.). The Census Bureau uses, and seeks to use, administrative records developed by federal agencies, tribal, state, and local governments as well as data from commercial entities." (US Census Bureau, no date) Frequently Asked Questions, What are administrative records and third-party data?

Multiple administrative records comprise administrative data, which are data used to administer an organization, program, or service process. Examples include data accumulated as government agencies administer social programs, as companies track customer orders, or as universities monitor student enrolment (Keller et al., 2016; Keller et al., 2017). Institutions and companies frequently produce administrative data to record, monitor, and evaluate their programs and services. Statistical agencies can make effective use of these data by cleaning, processing, and repurpose these administrative data to develop statistical registers and produce estimates about the populations that administrative records capture (Wallgren & Wallgren, 2010; Wallgren & Wallgren, 2014).

Using administrative data in research and statistical systems is not new. Federal agencies and researchers already employ these data, either as standalone or linked to other sources—other administrative files, surveys, and third-party data—to provide official statistics, examine data quality, construct sampling frames, develop weights, and conduct population studies. Below, we provide illustrations of how administrative data are currently used as standalone information or in combination with other administrative sources, surveys, and third-party data.

Using Administrative Data in US Official Statistics

Administrative data can serve standalone as a basis for official statistics and reports. For example, the Social Security Administration uses its four large administrative systems:

- social security number application data are in the Numident file
- individuals' lifetime wages and earnings collected through the Internal Revenue Service's Form W-2 are in the Master Earnings File;
- benefits programs administration are in the Master Beneficiary Record; and

- individuals' disability and Supplemental Security Income program information are in the Supplemental Security Income record. They are collected to report on the beneficiaries and use of agency programs (Maxfield, 2008).

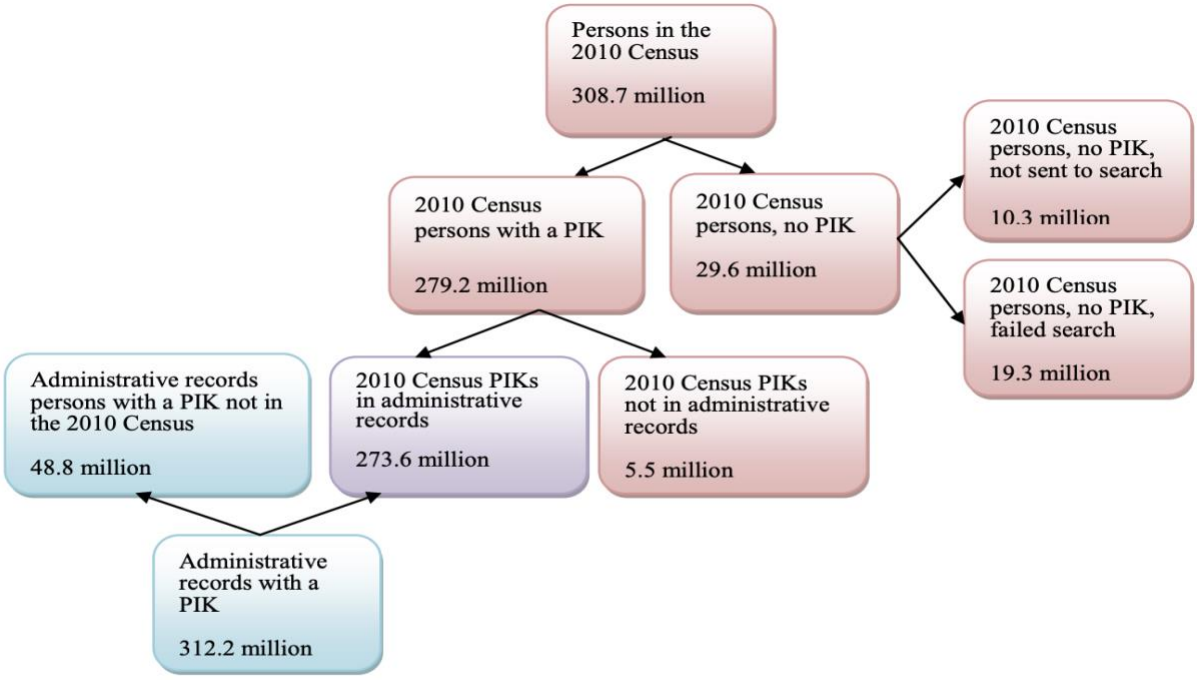
Similarly, the Internal Revenue Service's Statistics of Income Division employs its administrative data, based on individual Form 1040 and corporate Form 1120 tax returns, to fulfill its mandate to regularly produce statistics on income, annual reports for the public, and articles in its quarterly bulletin (Greenia, 2008). The Division also researches performance and compliance using these data and produces special tabulations by request for defined research uses (*ibid*).

Federal Statistical Agencies

Federal statistical agencies frequently link their administrative data assets to support research efforts (NASEM, 2017a). In 1999, the US Census Bureau conducted one such major linkage, developing the Statistical Administrative Records System (StARS) as part of its Administrative Records Experiment, which simulated an administrative records-based 2000 decennial census. StARS linked administrative data from the Internal Revenue Service, Department of Housing and Urban Development, Centers for Medicare and Medicaid Services, the Indian Health Service Patient Registration System, and the Selective Service System Registration System (Bauder & Judson, 2003; Berning, 2003). This experiment only relied on the six aforementioned federal data files. It did not use any state, local, or third-party data, yet found an 81% match between administrative records and census housing units counts and good census tract-level agreement between administrative counts and the 2000 census enumeration (*ibid.*). While the project also identified coverage and accuracy issues in capturing children, multi-unit households, and racial and ethnic minorities, it concluded that large administrative records data processing operations are technically feasible and worth investing time in working with them.

The Bureau's 2010 follow-up Census Match Study expanded StARS, linking it with additional federal administrative records, state data, and third-party files to compare demographic characteristics of the population captured in this system with the population covered in the 2010 decennial census (Rastogi & O'Hara, 2012). This study used 22 data files from eight federal agencies and five commercial vendors. A more extensive set of administrative records available, combined with the Bureau's improved linkage capabilities, resulted in the finding that administrative data are reliable for decennial census address and count confirmation. Approximately 89% of administrative records matched 2010 census records—the percentage was higher at 95% when only considering records with a Protected Identification Key (PIK) (Wagner & Layne 2014). A similarly high percentage of census records, 93%, had an administrative record address match with the 2010 census. The study concluded that high administrative records coverage could assist in improving decennial census address and person coverage and inform the determination of housing unit occupancy status and household population counts (*ibid.*). See Figure 1 for a review of count and match comparison of persons enumerated in the 2010 decennial census with administrative data sources.

Figure 1. Count and Match of 2010 Census and Administrative Records Persons (figure from Rastogi & O'Hara, 2012).

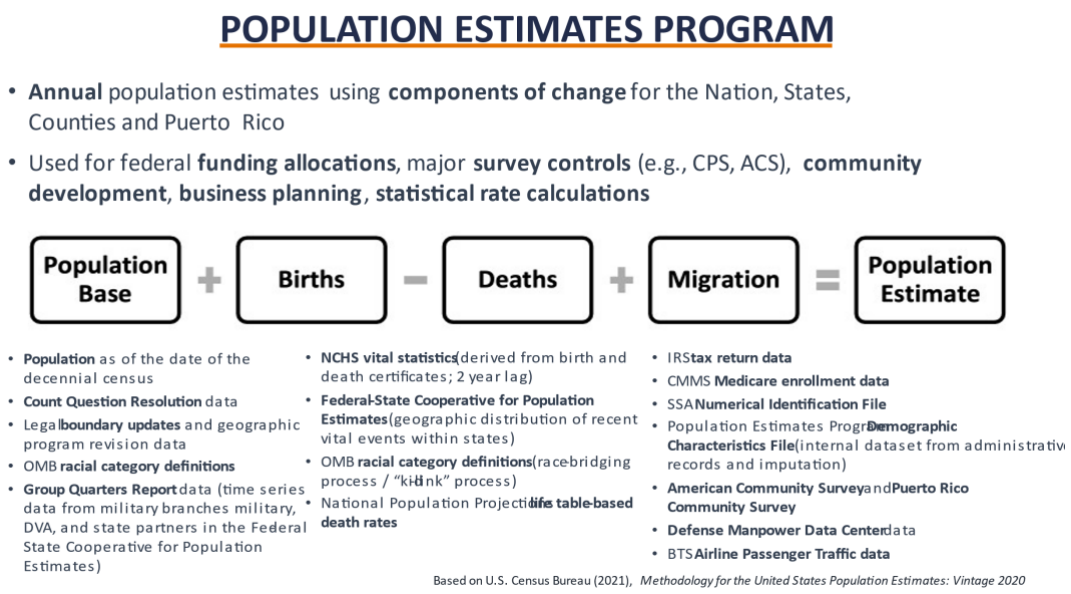


Sources: 2010 Census and 2010 Census Match Study Administrative Records Data.

Another major Census Bureau program that regularly conducts data linkage is its population estimates effort. It provides annual population estimates at multiple geographies used for funding allocation, developing survey controls, business planning, and statistical rate calculations. Although this program does not track individuals over time, it invests considerable effort into integrating data from federal and state agencies, as well as third-party providers. It links files from the decennial census, Count Question Resolution data,[1] Group Quarters Reports from military branches, Department of Veterans Affairs, and state partners, vital statistics, Internal Revenue Service tax returns, airline passenger traffic data, and others. On average, its estimates are highly accurate; for example, the average absolute difference between the 2010 final total resident population estimates and the 2010 decennial census counts was 3.1% across all US counties (US Census Bureau, 2021b). Figure 2 provides an overview of the population estimates program and the mix of survey and administrative data sources used.

[1] The Count Question Resolution (CQR) operation provides an opportunity for tribal, state, and local governmental units to request that the Census Bureau review their boundaries and/or housing counts by block to correct geographical errors. https://www.census.gov/about/policies/quality/corrections/cqr.html#:~:text=The%20Count%20Question%20Resolution%20(CQR,block%20to%20correct%20geographical%20errors.

Figure 2. A Summary Description of the US Census Bureau's Population Estimates Program



Other statistical agencies conduct similar linkages. For example, the Internal Revenue Service Statistics of Income Division links administrative records from Forms 990, 990-EZ, and 990-PF to support its research and provide public use data on tax-exempt organizations and charitable activities, called the Exempt Organization Financial Extract. The Division also linked 2.5 billion tax returns from 1996 to 2015 to build an individual-level Data Bank (Johnson et al., 2018). Further, the Department of Education Rehabilitation Services Administration links Case Service Report (RSA-911) data with 1980-1988 Social Security Administration earnings files, matching individuals' earning histories with disability and vocational rehabilitation information.

Administrative data can also form the foundation for developing public use data products and estimates on key social and economic indicators. In this vein, the Bureau of Labor Statistics (BLS) maintains the Business Employment Dynamics (BED) program and Quarterly Workforce Indicators (QWI), both based on administrative data from government programs, primarily unemployment insurance programs. The BED dataset links establishment and firm data from unemployment insurance quarterly contributions reports, and QWI data is created from records on unemployment insurance wage use. The BLS Quarterly Census of Employment and Wages (QCEW) is also based on state unemployment insurance tax system records, covering approximately 97% of all civilian, non-farm wage and salary jobs (Robertson, 2017). These data are the foundation for the BLS internal research and official estimates and an information source for social science researchers. See Table 1 for an overview of BLS examples of using survey and administrative data to improve their data products.

Table 1. Bureau of Labor Statistics examples of combining survey and administrative data in employment and establishment data sources

Unemployment Insurance (UI) System

Quarterly Census of Employment and Wages (QCEW)

+ Annual Refiling Survey (ARS): industry, geography, and respondent information.
+ Multiple Worksite Report (MWR): employment and wage data for individual establishments under multi-unit businesses.

UI records (~**9.6 million establishment records quarterly**) serve as:

- **sampling frame** for BLS establishment surveys;
- **source of information** about national, state, and metropolitan economies.

UI program operational in 50 States, DC, Puerto Rico, Virgin Islands.

- Ensuring **record uniformity**, software interoperability.

Example: **Business Employment Dynamics** (BDM)

- **Record linkage across quarters** using unique UI identification numbers, predecessor and successor information, probability-based matching, and analyst review.

State Agencies

Administrative record use and linkage are common in state agencies due to their program and benefit eligibility establishment, monitoring, and reporting responsibilities. State human services agencies frequently use Temporary Assistance for Needy Families, Supplemental Nutrition Assistance Program, Medicaid Insurance, and Unemployment Insurance administrative data. They also report linking these sources to create integrated data systems and linking them to internal files like a client site visit and interview data, or external sources like the decennial census and American Community Survey public use files (Allard et al., 2018). Examples of successful state-level administrative data linkage include:

- Rhode Island's *RI360*, an anonymized administrative records database that links data products across all state agencies, contains over 2.7 billion records on over 4 million residents and facilitates examining poverty and economic opportunity, education, social programs, and health care in the state (Hastings et al., 2019).
- The University of Denver's Colorado Evaluation and Action Lab's Linked Information Network of Colorado integrates data from multiple state and local agencies for evaluation and research (Leboeuf, 2020).
- Multiple states are also using Department of Health and Human Services grants to expand administrative data integration to meet operational, policy, and research goals.
- Recently, Connecticut built on its existing integrated Unemployment Insurance quarterly wage, Quarterly Census of Employment and Wages, and Department of Motor Vehicles database, matching records to Jobs First Employment Services and Temporary Family Assistance program data.
- Wisconsin expanded an existing Client Assistance for Re-Employment and Economic Support database with Supplemental Security Income, Temporary

Assistance for Needy Families, Unemployment Insurance quarterly wage, childcare, child support, food stamps, Medicaid, and Children's Health Insurance Program records. They created a longitudinally linked infrastructure with records of participants across all programs dating back to 1990.

- Indiana matched Temporary Assistance for Needy Families (TANF) administrative records to the National Directory of New Hires, the State Directory of New Hires, and the Unemployment Insurance quarterly wage data. Similarly, South Carolina linked longitudinal TANF, food stamp, and employment service use and work activity records with Unemployment Insurance quarterly wage data, child welfare, Medicaid, and Children's Health Insurance Program data (Wheaton et al., 2012).
- For more than a decade, New York City has created an alternate poverty measure that better accounts for family resources including in-kind transfers and tax credits, and expenditures such as housing and medical costs using a variety of survey and administrative data (City of New York, 2020).

States and local governments collect administrative data that would be useful to the federal statistical system to supplement or enhance existing surveys or to create new data products. Federal government incentives can often spur state and local governments to provide access to their administrative data. The National Center for Educational Statistics in the U.S. Department of Education funds states through grants to create a Statewide Longitudinal Data System (SLDS), which is an integrated system of administrative data on student achievement and performance at the state level. These grants provide funding to hire staff and acquire hardware, software, and technical assistance for states to collect information on students over time. The Department of Education requires states share these data at all levels of government and to make them available for research (SLDS 2021, NASEM 2017a).

Another example of state data is about the U.S. prison population. State departments of correction collect data on these populations and share the data with the Bureau of Justice Statistics. BJS uses the data to provide national statistics by state. These data are used to design federal programs that inform prison construction as well as programs for the reintegration into civilian life for prisoners who have served their sentences. These data include descriptive data on the people admitted to and released from the institutions. Specific elements of the data are

- conviction offense,
- date of admission; date of release,
- the person’s age, race, and ethnicity,
- the offender’s criminal justice status at admission; and
- the county of conviction.

There are many examples of matching local and state administrative data to current federal data sources and programs. In addition, the federal government can create grants and other programs to spur states to create useful data products out of data already collected for administration of schools, prisons, and other state and local services. Most

recently, administrative data on skilled nursing facilities, university dorms and a variety of other records from licensing and accreditation agencies were utilized by the Census Bureau, as well as state and local governments, all in an effort to support the enumeration and evaluation of the population in group quarters facilities. One example is the New York State Nursing Home Weekly Bed Census, which provides information on occupancy of nursing homes since 2009 (New York State Department of Health, 2021).

Local Data

There are other data sources and types of data collected at the local level. Examples include data from sensors such as weather conditions and water quality data, and videos from traffic cameras. These other data are collected to administer programs or to track local conditions and would be valuable for federal statistics or new statistical measures. Some data sources are well structured and can be more easily used in statistical analysis, such as weather data and taxi meter data. Unstructured data are more challenging to use such as traffic videos. Combining traffic video data and taxi meter data could be used to better understand traffic flows by time of day (NASEM 2017a). Another surprising source of data that is useful for evaluating traffic flows is fire-EMS 911 data by time of day and day of week (Keller et al., 2016; Keller et al., 2017).

State and local administrative records provide new sources of data for federal statistics. Challenges of using these data include establishing uniform national standards for reporting, protecting the confidentiality of the data, and developing standards for fitness of use (NASEM 2017a). One good example is the provision of address information to the U.S. Census Bureau through the Local Update of Census Addresses (LUCA) program, where state and local governments use their lists of addresses from property, taxation, and other files that support governmental functions, in an effort to improve the address list used for the decennial census and, ultimately, for the census and the plethora of surveys that use the Master Address File of the Census Bureau to select samples, as with the American Community Survey (U.S. Census Bureau, 2020).

Third Party Data

Third party data from the private sector can provide access to data that would be very difficult for statistical agencies to collect, such as actual transactions in lieu of aggregate data currently reported on federal surveys. There are risks as well. Business data contain proprietary information and implicit firm strategies and sometimes personally identifiable data or data that may be more easily identified when multiple sources are linked to create new data products or analysis. (NASEM 2017a).

To categorize and assess the challenges of using third party data, NASEM (2017a) groups private data into five types:

1. Structured data from censuses and probability surveys, e.g., customer satisfaction surveys, marketing research surveys, and media use surveys.
2. Structured data from administrative records, e.g., data produced by businesses, such as commercial transactions, banking and stock records. credit card records, insurance claims, and data from machines rented or leased to local governments, farmers, and nonprofit organizations to collect administrative data

3. Other structured data that are organized and can easily be placed in a database, although they may still require profiling to assess quality and restructuring to be usable, e.g., commerce transactions, mobile phone location sensors, Global Positioning System sensors, utility company sensors, and weather and pollution sensors.
4. Semistructured data that cannot be placed in a database (e.g., relational, spreadsheet, or other type) and for which profiling and wrangling is usually more difficult than for structured data because the fields are not always fixed or are in a form that requires translation. Examples include Extensible Markup Language (XML) files, data from computer systems, logs, eb logs, mobile phone content and text messages, E-mail, data from Internet of things, and sport activity sensors from watches, etc.
5. Unstructured data, such as in text, images, and videos, that require that information must first be extracted and then placed in a structured table for further processing and analysis.

As the data are increasingly unstructured, they present new challenges in preparing the data for statistical uses. Each of these types of data sources must be evaluated for their relevance, accuracy, completeness, trustworthiness, and other characteristics that we describe in more detail in the sections below.

To use third party data, statistical agencies negotiate how they will receive the data (NASEM 2017a). For example, approaches may include

- the company prepares and provides the data in an aggregated form. The statistical agency can use these data to benchmark or supplement other data sources or use directly in statistical products;
- the company transfers the data to the agency for the agency to compute the statistics, e.g., BLS is negotiating with large companies to provide payroll and other internal company data from which BLS will extract relevant information, rather than asking the company to complete its surveys; or
- data collection, processing, and analysis are outsourced to the private firm.

Another approach is to benchmark federal statistics with published third-party data. Some private-sector companies use their administrative records to produce and publish statistics such as the National Employment Report from Automatic Data Processing, Inc. (ADP), which precedes the Bureau of Labor Statistics (BLS) release of the employment situation each month (ADP 2021[2], NASEM 2017a). ADP attempts to match BLS definitions for the reporting periods to provide national level employment growth statistics by industry. Taking into account differences in the population coverage, and other characteristics, BLS can use these estimates as one source to benchmark their results.

[2] ADP. 2021. ADP Employment Report Methodology. Automatic Data Processing, Inc. (ADP), https://adpemploymentreport.com/common/docs/ADP-NER-Methodology-Full-Detail.pdf

Acquiring and using third-party data sources require new skills from data discovery, negotiating data acquisition, data profiling to evaluate data quality, benchmarking, and a putting the data to use in statistical outputs and analysis.

Using Administrative Data with Surveys

Administrative data used alongside survey data can support survey-related operations, that is, improve survey development, lessen response burden, address attrition, and reduce fielding costs. In addition, administrative data can be used to identify and alleviate survey data quality or coverage issues or present an alternative to survey data collection altogether (Zanutto & Zaslavsky, 2002). The USDA National Agricultural Statistics Service, for example, is exploring the use of geospatial data, weather data, and other environmental data as spatially-linked auxiliary data in models with survey data collected from area sampling units (Cruze, 2015, NASEM 2017a). Other examples of these combined data use cases are presented below.

Using Administrative Data to Support Surveys

Administrative records also have the potential to enhance surveys from two perspectives: by improving survey design or addressing its limitations, and adding new relevant data elements that increase the potential of the survey data to address policy and research questions. Administrative data differ from survey data in that they are not collected with a research aim. In contrast, surveys define the population, sample, and instruments before data collection. Data in administrative systems are recorded before defining a population and sample for research purposes. Accordingly, the two data sources have different strengths and can act as complements (Couper, 2017; Miller, 2017; Wallgren & Wallgren, 2014).

Administrative data can support and improve survey development. A few examples follow:

- The Centers for Medicare and Medicaid Services' Medicare Current Beneficiary Survey's sampling frame is based on administrative records.
- The Census Bureau uses administrative data to enhance the sampling frame for the Survey of Minority-Owned Business Enterprises (Williams & Moore, 1998).
- The Agency for Healthcare Research and Quality's Medical Expenditure Panel Survey uses administrative records for nonresponse weight calculation.

Administrative sources can also increase coverage of a target population in surveys. They frequently better capture persons who are typically excluded from surveys, like institutionalized individuals, and they can provide near-complete coverage of special interest populations like program participants.

Further, administrative data can enhance surveys by providing program and contextual information about survey respondents. Conversely, supplementing administrative records with survey data can add socioeconomic and attitudinal variables

to administrative records. Many federal statistical agencies link their administrative records with other agencies. For example, the US National Center for Health Statistics partnered with the Department of Housing and Urban Development (HUD) to link National Health and Nutrition Examination Survey (NHANES) data with HUD administrative records (National Center for Health Statistics, 2021). This linkage of two data sources allows researchers to examine how housing conditions shape health outcomes (e.g., Ahrens et al., 2016). The Center also linked Centers for Medicare and Medicaid Services administrative files with National Long-Term Care Providers Survey data, thus facilitating research on program utilization across a broader range of long-term health facilities (see Elliott et al., 2012).

The U.S. Census Bureau has created the Supplemental Poverty Measure (SPM), which 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. For example, the SPM incorporates survey data on tax credits and income transfers in alleviating poverty, using data from the Current Population Survey Annual Social and Economic Supplement (Fox, 2020). Further, administrative records have been used to evaluate the level of coverage of SNAP and TANF participation and the effectiveness of models aimed at addressing under coverage of program participation in household surveys (Shantz and Fox, 2018).

Social science research projects also are linking to data from official statistics. The University of Michigan's Health and Retirement Study, with support from the NIH National Institute on Aging and the Social Security Administration, has been linked to administrative data from Centers for Medicare and Medicaid Services' Medicaid Analytic Extracts and Medicare Claims and Summary Data, to Social Security Administrations' earnings and benefits data, and to Veterans Affairs' health care data (Health and Retirement Study, 2021). Researchers can now integrate social program context when studying middle age and older adult life course. In a further example, the Census Bureau linked its Survey of Income and Program Participation to Social Security Administration records.

Using Administrative Records to Understand Data Quality and Coverage

Administrative data used in tandem with survey data can inform quality and coverage assessments for both data sources. Particularly in federal statistical agency work, issues like misreporting and under- or overcounts have implications for program funding, needs assessments, and official statistics. Studies that combine administrative records and survey sources for quality assessment purposes frequently focus on discrepancies in information that can be unreliable when self-reported and on the coverage of populations captured.

We provide examples using income and racial and ethnic origin to illustrate how researchers employ administrative records in studying the quality of survey data information that can be unreliable. Multiple studies investigate the consistency of income reporting. For example, Pedace & Bates (2000) linked longitudinal Survey of Income and

Program Participation data with Social Security Administration's Social Security Summary of Earnings administrative records to assess the magnitude of income misreporting. The authors find that information on earnings receipt was congruent between survey data and administrative records in approximately 90% of cases. The survey data underestimated earnings by an average of $450, with high earners most likely to underreport. The authors further identified sociodemographic characteristics associated with earnings under- and overreporting. Older adults and individuals with low education were likely to underreport earnings, suggesting administrative records may be a better source of data on these subgroups.

Cristia & Schwabish (2009) similarly matched longitudinal Survey of Income and Program Participation data with Social Security Administration's Detailed Earning Records administrative data, finding that survey respondents underreported earnings by approximately $2,800, with low earners more likely to overreport and high earners again more likely to underreport. This approach of creating a synthetic or modeled micro dataset from partially observed data has not been tried in the federal statistical system except in two instances, both undertaken to protect confidentiality. As noted below, the Survey of Income and Program Participation Synthetic Beta, Synthetic Longitudinal Business Database, and Survey of Consumer Finance also use synthetic data methods in developing publicly released files (NASEM 2017b).

Brummet et al. (2018), linking two waves of Consumer Expenditure Survey data to Internal Revenue Service Forms 1040, W-2, and 1099 administrative records, found an underreport of payroll and retirement income in the survey, but similar levels of income reports for self-employment income in the survey and administrative data.

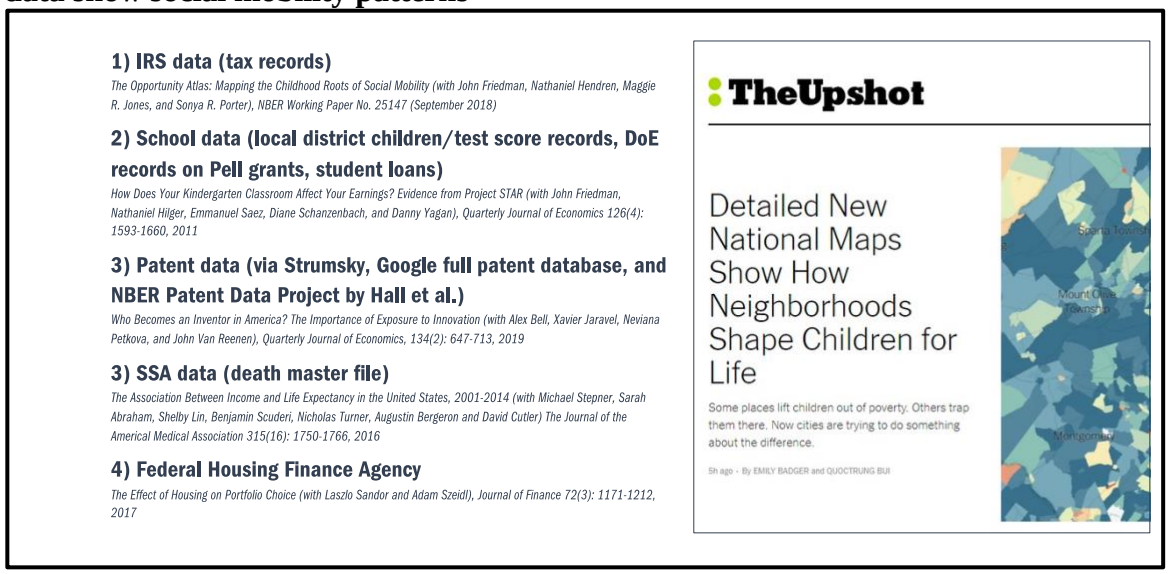
A second large group of discrepancy studies explores issues in reported racial and ethnic origin across sources. Ennis et al. (2018) link administrative (e.g., Housing and Urban Development Public and Indian Housing Information Center, Temporary Assistance for Needy Families, and Medicaid Statistical Information System data), survey (e.g., American Community Survey), and decennial census data to examine racial and ethnic origin information congruence. They find that 79% of linked records had no discrepancies in Hispanic origin responses. Among the group showing discrepancies, the authors identify renters and single-parent households as those more likely to have incongruent racial and ethnic identification information recorded across sources (ibid.).

In evaluating administrative record quality, understanding the coverage of populations and their characteristics is critical to adopting administrative data for social science research. The share of the population captured in administrative data has increased over time. For example, while the 2000 decennial census administrative records experiment resulted in an 81% match rate between the records and census enumeration (Bauder & Judson, 2003), in 2010, the match rate was higher at 89% for the entire census population and 98% for the population of individuals with PIK identifiers (Rastogi et al., 2012). Related research further shows that 96% to 98% of the US population can be found in tax records (Chetty et al., 2014; Mortenson et al., 2009). See Figure 3 for a case

study using tax records and other administrative data to trace how the neighborhoods that children grow up in influence future earnings.

However, while overall population coverage in administrative data can be high, the data can under-capture particular sociodemographic groups. Bhaskar et al. (2014) pooled census records, five commercial datasets, and administrative records from nine federal programs (e.g., Centers for Medicare and Medicaid Services Medicare Enrollment Database, Selective Service System Registration File, Housing and Urban Development Tenant Rental Assistance Certification System, and others). They assessed whether and to what extent these records capture demographic information recorded in 2010 American Community Survey (ACS) responses. Administrative records captured 93% of ACS respondents' Hispanic ethnicity, 81% of respondents' race, and 93% respondents' gender and age. However, Hispanic ethnicity was more often missing in administrative data for Hispanics than non-Hispanics and race was more often missing for minorities. In contrast, age and gender coverage was high and comparable across groups.

Figure 3. Using IRS data combined with local education. Patent, and housing data show social mobility patterns



Researchers also link administrative records with a combination of other administrative, survey, and commercial data to assess population coverage between sources. For example, Fernandez et al. (2018) use the 2010 Census Match Study administrative records file, which links multiple federal agency surveys and administrative records with third-party data and supplement it with Medicaid administrative records to examine the coverage of children in the 2010 decennial census. The authors find that ethnic and racial minority children and those living in multi-unit buildings, multigenerational, or low-income households are more likely to be found in administrative and survey records but not found in decennial census data.

Using Administrative Data to Explore Survey Alternatives

Statistical agencies increasingly employ administrative records to examine their value as a complement or substitute for the surveys they typically conduct. The Bureau of Labor Statistics (BLS) currently produces Import and Export Indexes based on an international price survey. It is exploring how to incorporate over two million monthly administrative records on exports and unit values into the series to improve its accuracy, with initial tests suggesting that a blended dataset could provide the agency with more robust indexes (Fast & Fleck, 2019). Further, BLS linked Internal Revenue Service Forms 1040, W-2, and 1099 data to two waves of its Consumer Expenditure Survey to assess improvements in income survey estimates (Brummet et al., 2018).

BLS also supplements Quarterly Census of Employment and Wages (QCEW) data, based on unemployment insurance administrative records, with County Business Patterns, Annual Survey of Public Employment and Payroll, Railroad Retirement Board data to construct estimates for jobs not captured in QCEW (Robertson, 2017).

In a similar vein, the Bureau of Economic Analysis began producing an experimental Health Care Satellite Account using administrative claims data to address the limitations of its Medical Expenditure Panel Survey and better understand health care spending (National Research Council, 2009). The US Energy Information Administration uses energy supplier records to obtain information on energy consumed to supplement its Residential Energy Consumption Survey (US Energy Information Administration, 2021).

Other agencies are exploring ways to replace surveys with administrative data; in 2017, the National Center for Science and Engineering Statistics, the National Science Foundation's statistical agency, redesigned its Survey of Graduate Students and Postdoctorates in Science and Engineering to reduce reporting burden. It transitioned from collecting survey responses from institutions to collecting direct administrative data uploads (Gordon et al., 2018). Instead of institution staff completing questionnaires for which they had to pool and tally information across university departments, respondents now submit a records file using a standardized set of codes that facilitate the process.

Social science researchers are also examining how administrative data can enhance or supplement surveys. One study comparing local government administrative real estate assessment data with American Community Survey (ACS) housing variables finds that property year built, value, and taxes paid measures are in agreement between the two sources and could be used interchangeably (Molfino et al., 2017). In fact, research on the value of benefits as reported in the early test data for the American Community Survey (ACS) were shown to be deficient when compared with actual data for SNAP recipients in Maryland, which was one study that led to the decision to exclude a question on the value of benefits (Taeuber et al., 2004).

Further, the assessment data are available at the parcel level, a lower level of geography than ACS housing data. Thus, replacing ACS survey data with the administrative source could provide advantages for officials making real estate and

planning decisions for county and city neighborhoods (*ibid*.) as well as reduce respondent burden. The Census Bureau explored the use of Internal Revenue Service tax information to replace the income questions in the American Community Survey (O'Hara, 2016, Houser & Sanders 2016). In other cases, agencies can replace survey by using administrative data and data from other sources. For example, in 2012, the National Center for Health Statistics stopped collecting the National Nursing Home Survey and the National Home and Hospice Survey and provided the same information using administrative data from the Centers for Medicare & Medicaid Services.

Administrative data can be a useful source for similar assessments, helping survey researchers determine what items to include in survey questionnaires and what items can be replaced with administrative data, thus reducing survey respondent burden. Other approaches are to link administrative data with third-party data, described next.

Using Administrative Data for Linkage with Third-Party Data

In addition to linking administrative data to surveys, records can also be linked to third-party sources. Third-party data refers to "micro-level data on persons, families, or households [that are maintained] by commercial, private-sector entities" (National Academies of Sciences, Engineering, and Medicine, 2019; 7). Although third-party data typically match administrative records at a lower rate than survey data (Brummet, 2014; Kingkade, 2013; Rastogi & O'Hara, 2012), it can similarly be used to expand administrative record coverage, add contextual information to records, or facilitate data quality assessments.

Federal statistical agencies employ third-party data for all the purposes mentioned above. Within the US Census Bureau, CoreLogic and Black Knight property and tax data, various multiple listing services data files, and Veteran Service Group of Illinois consumer and household data are part of designing and quality-checking decennial census counts and operations (*ibid*) and benchmarking American Community Survey data (Kingkade, 2013). To reduce respondent burden, provide more timely information, and improve quality, the Census Bureau also integrates its survey and administrative data with third-party NPD Group's point-of-sale information from over 300,000 retailers into its Monthly Retail Trade Survey. This experimental product models retail sales (US Census Bureau, 2021a). The third-party file covers both physical and online retailers across multiple industries. It adds to the Monthly Retail Trade Survey information on units sold, product sales, average prices, and total sales (Hutchinson, 2017). See Table 2 for types of data the Census Bureau uses to improve statistical data products.

Table 2. Examples of Third-Party Data Sources examined by the U.S. Census Bureau to reduce the burden on respondents, to validate Census' mapping databases, and to lower the cost of the Census.	
Tax Assessment	Stand Alone Mortgage
Mortgage Assessment	Eviction
Automated Valuation	Property Tax and Deeds
Deeds and Loans	Multiple Listing Services
Master Address	Foreclosures
Notice of Delinquency	Household Member Data
Parcel Boundary	Telephone Data
Release of Mortgage	Education data

There are already several creative approaches to integrating administrative data with federal statistics. BEA links its data with third-party insurance claims files to provide improved payment estimates and treatment price indices by disease condition in experimental Health Satellite (or "Blended") Account files (Bureau of Economic Analysis, 2021). Integrating the private claims data allows BEA to address potential quality issues arising from the Medical Expenditure Panel Survey's small sample (National Research Council, Committee on National Statistics, 2008).

The U.S. Department of Agriculture (USDA) links administrative data on SNAP participation and purchases with survey responses from the probability sample of 5,000 households in the National Household Food Acquisition and Purchase Survey (FoodAPS). The linked information provides information about SNAP eligibility in the 30 days prior to the survey, helped to address data discrepancies, and provide information on use of the electronic benefits card usage (Ver Ploeg et al., 2015, NASEM 2017b).

The U.S. Bureau of Justice Statistics (BJS) links state correctional facilities admissions and releases with other administrative record data to examine the reasons for prisoners' return to prison or success in not returning (Carson, 2015, NASEM 2017b). By linking these data to other sources, BJS can tell if former inmates have a job using Social Security data or whether they are married using ACS data.

However, linking data is not always easy or accurate as the match rate varies depending on the study or population and can lead to biased conclusions (Harron et al. 2014). There are studies that propose statistical methods to account for linkage bias (Lahiri and Larsen, 2005; Hof and Zwinderman, 2012; Judson et al., 2013), but these still do not always remove the bias in selected variables used (NASEM 2017a).

Technical Challenges in Administrative Record Use

Using administrative records for official statistics and social science research faces four key challenges (Amaya et al., 2020; Auerbach et al., 2019; Groves & Schoeffel, 2018; Wallgren & Wallgren, 2010):

- Data source coverage: Identifying relevant sources that maximize the coverage of the population of interest.
- Data linkage: Linking often incompatible units, raising issues of unique identifiers and continuity over time.
- Variable compatibility: Making appropriate decisions about similar variable mismatches, as these can be due to either inconsistencies and errors in recording or disparate variable definitions.
- Security, access, and sharing: Fostering transparency and accountability of data and uses.

We review these challenges in turn.

Data Source Coverage

Administrative data typically capture a set of individuals, objects, or other entities relevant to the program collecting the data, and only collect information while these entities are part of the program. Such data sources alone do not usually contain all the entries for general research populations of interest. They are limited in their use for statistical purposes like facilitating analytic approaches that address causal questions (Groves & Schoeffel, 2018). Researchers should be aware of definitions that limit the entities for which administrative records are collected. Examples follow:

> The Database on Ideology, Money in Politics, and Elections administrative records does not capture campaign contribution donations less than $200 (Hill & Huber, 2017).
> Corruption cases are not captured in administrative data if they are not pursued in the federal court (Cordis & Milyo, 2016).
> Unauthorized housing units, potentially non-permanent dwellings like boats and timeshares, and non-taxed housing units like housing built on college campuses, military bases, or tribal lands, may be recorded as housing units in Census Bureau records but not in tax assessor data (Jarosz & Hofmockel, 2013).

In using administrative records, researchers must evaluate whether the representativeness of available records is suitable for their research purpose. Data sample and coverage determinations can become more complex in cases of administrative data linked to other administrative or survey files. Although linkage match rates can be high overall, they may be disproportionate for particular subgroups, aggravating under- or over-coverage that might be present in standalone data sources. For example, Cristia & Schwabish (2009) successfully match 85% of cases in their survey and administrative data. However, the match was lower for ethnic minorities, resulting in coverage limitations for the population of interest in the final dataset.

Similarly, reliance on administrative records that are biased towards particular subgroups for the program or other reasons may result in a linked data source that is not useful for studying the general population. For example, administrative data that

primarily capture households with young children should be linked to appropriate other data sources that cover childless households to ensure representative coverage of the population (Zanutto & Zaslavsky, 2002). Following linkage, researchers should assess the resulting data coverage and potential biases by benchmarking to other data sources. For example, the Census Bureau conducts an exhaustive coverage evaluation using post-enumeration surveys and demographic analysis estimates developed from administrative records (National Research Council, 2009).

As it concerns studies affecting civil rights that make use of administrative records, McClure et al (2017) have discussed how the underrepresentation of hard-to-reach subpopulations in administrative records may put them at a significant disadvantage in the deployment of procedures that are heavily based on such records. The bias introduced by differences in the representation of vulnerable and non-vulnerable populations can lead to large over- and undercounts across populations, affecting the allocation of resources, political representation, community investments, and the ability of government to evaluate disparities among racial-ethnic minorities, American Indians on tribal lands, and immigrants, among others.

Data Linkage

Record linkage refers to matching two or more records, typically stored in different data sources, that refer to the same object, person, organization, or other entity (GAO, 2001). Two or more matched records can be correctly linked or true positives, incorrectly linked or false positives, or matches that remain unlinked or false negatives. In linking administrative data to other administrative records (government and private sources) and surveys, researchers must consider the availability of matching variables, the representativeness and population coverage of each data source, and the comparability and congruence of their variables (Min et al., 2019).

Unique Identifiers

Matching variables facilitate linkage with other sources employing the same identifiers. When implemented across multiple data files, matching variables can be used for integrating separate sources into a system of (statistical) registers to obtain a complete listing of all persons, organizations, objects, or other entities that belong to a defined set (Wallgren & Wallgren, 2010). Depending on the type of matching variables available, researchers can employ deterministic or probabilistic record linkage to join data sources (Fellegi & Sunter, 1969; Winkler, 1995; Winkler, 2006; Yancey, 2002).

Deterministic or exact record linkage refers to matching records for individuals, organizations, or other entities based on exact agreement on all matching variables in a given dataset. Typically, deterministic record linkage is based on unique identifiers or a combination of personal or organizational information (e.g., name, birth date, and social security number or business name and address). For example, Pedace & Bates (2000) linked longitudinal Survey of Income and Program Participation data with Social

Security Administration's Social Security Summary of Earnings records using social security number information present in both files.

In the absence of unique identifiers or a failed deterministic match, researchers can employ probabilistic linkage, matching records on a set of variables where some can be required to be exact matches, and others are assigned importance weights for a probabilistic match. In this approach, cases are assigned agreement weights given their variable values and importance weights. The matches are evaluated against a match threshold value given willingness to accept false positives versus false negatives.

Similarly, hierarchical linkage evaluates records based on a list of variables with a priority order, attempting to maximize the number of variables that match; the more variables that match between records, the more likely the cases refer to the same entity. In this way, Graham et al. (2018) linked US Patent and Trademark Office Patent Data Extract, Census Bureau Business Register, Longitudinal Business Database, and Longitudinal Employer-Household Dynamics Employment History Files using Census Bureau's unique identifier information in combination with inventor names and locations, assignee information, business names, and cleaned information from patent documents. Similarly, the Bureau of Labor Statistics links Business Employment Dynamics administrative records on establishments over time using a multi-step process. They first attempt to match on unique identification numbers, followed by using predecessor and successor information relation, a probabilistic match based on business name and contact information, and finally manual review (Parker, 2006).

Research outside the US federal statistical agencies also employs creative linkage strategies in the absence of unique identifiers. Lunn et al. (2020) linked over one million records on Irish farms from multiple administrative data systems spanning ten years. This was possible because each farm's uniqueness could be established through its association with herd identification numbers mandated through a European Union directive. YouGov, a private organization, successfully matched Cooperative Congressional Election Survey respondent data to state election records and campaign contribution data even though these sources were not designed for linkage and did not contain unique identifiers; instead, the company used respondents' names and addresses to conduct matching (Hill & Huber, 2017).

Given considerable challenges with data that lacks unique identifiers, many federal statistical agencies have developed in-house systems that facilitate identifier assignment. The U.S. Census Bureau uses a Person Identification Validation System to assign unique identifiers, or Protected Identification Keys (PIKs), to administrative and survey files and facilitate linkage across data files for Bureau research and other data users. The identification keys are based on name, birth date, and address information. The Bureau uses probabilistic matching to assign the keys, first attempting to match by social security number, followed by name, address, and birth date information, and finally using an age range (Wagner & Layne, 2014).

At the housing level, the Census Bureau assigns every known housing unit a Master Address File unique identification (MAFID) number; the assignment procedure first cleans and standardizes housing unit addresses, and then uses a probabilistic algorithm to assign MAFIDs (Rastogi & O'Hara, 2012; Wagner & Layne, 2014). Once cases in each data file have been assigned a unique identifier like PIK or MAFID, researchers can conduct linkage across files. For example, building on the Bureau's work, Brummet et al. (2018) matched two waves of person-level Consumer Expenditure Survey data and Internal Revenue Service Forms 1040, W-2, and 1099 administrative records. Bhaskar et al. (2019) linked the Current Population Survey Annual Social and Economic Supplement with Medicare Enrollment Database administrative data. In addition to the Census Bureau, the US Department of Agriculture Economic Research Service, Census Bureau, and Food and Nutrition Service jointly maintain a Next-Generation Data Platform that links administrative records on food assistance programs with surveys and other data for research purposes (US Department of Agriculture, 2021). The Centers for Medicare and Medicaid Services house the Integrated Data Repository, which supports linking data resources across the agency (Centers for Medicare and Medicaid Services, 2021).

Longitudinal Linkage and Continuity Over Time

Data products for official statistics, including linked administrative files, should be sustainable over time. Achieving sustainability is difficult to maintain if data collection methods, instruments, or definitions change in response to regulatory, commercial, or other reasons (Groves & Schoeffel, 2018; Struijs et al., 2014; Vichi & Hand, 2019). When linking administrative records with other data longitudinally, researchers must evaluate whether the context of administrative data collection changes, how it may affect variable values, consistency, and the overall quality of records linked over time. For example, the Internal Revenue Service Statistics of Income Division must consider the changing tax code and legislative definitions in maintaining their longitudinal administrative data infrastructure (Johnson et al., 2018).

Entity-level changes that can affect unique identifiers are a further challenge in longitudinal linkage. The same units can change over time and become difficult to match across administrative records. The Bureau of Labor Statistics faces this issue in linking business establishment data over time. The agency conducts multi-step linkage to ensure proper establishment matching for the Business Employment Dynamics and Quarterly Workforce Indicators series. It must keep current on ownership changes, openings and closings, mergers and acquisitions, and other processes that can change the assignment of unique identifiers for a given business. (Parker, 2006). Similarly, persons can change their given or last name at the individual level or appear in records as individuals or part of a unit. The Internal Revenue Service Statistics of Income Division must create derived statistical units to address this problem when matching, since filers can submit individual or joint tax returns at different times, making the observation unit inconsistent across raw records (Johnson et al., 2018).

Variable Comparability

Once data are linked, researchers must consider the consistency and comparability of recording and variable coding across sources and decide which data source to draw on when relevant information is recorded in more than one source.

For example, the administrative (e.g., Housing and Urban Development Public and Indian Housing Information Center, Temporary Assistance for Needy Families, and Medicaid Statistical Information System data), survey and census (e.g., American Community Survey, decennial census), and third- party data commonly linked for the US Census Bureau's internal research capture racial and ethnic identification in different ways. Race and ethnicity were captured with disparate numbers of categories and were not always recorded as separate variables (Bhaskar et al., 2014; Ennis et al., 2018).

Social science researchers face the same challenge. Stansbury et al. (2004) linked the Department of Veterans Affairs' stroke rehabilitation patient treatment files with its Integrated Stroke Outcomes Database administrative data to examine the congruence of racial and ethnic identification information in the two sources. While the patient treatment files recorded self-reported ethnic and racial identification across five categories, administrative records coded by clinicians and therapists captured six categories. Further, the authors found that racial and ethnic designation congruence between the two sources was weaker for nonwhite patients. Despite multiple recoding strategies, the authors observed different effects on outcomes depending on whether they drew on patient or administrative records as their source of race and ethnicity data. In these cases, researchers must assess the quality of responses from each data source given their use case and decide on which source to designate as the primary source for a variable in question in analyses (Ludlum, 2004). One example of such data quality assessment is the BLS Consumer Expenditure Survey, benchmarked against the Residential Energy Consumption Survey, National Health Expenditure Accounts, Medical Expenditure Panel Survey, Current Population Survey, and others (US Bureau of Labor Statistics, 2021).

Security, Access, and Sharing

Data sharing across government agencies, between federal actors and academic researchers, and the public promotes transparency and accountability, supports evaluation and decision-making, and encourages new insights, growth, and innovation. However, providing secure access to administrative data containing individuals' records can pose significant challenges that require data providers to balance information sharing with disclosure controls that limit the amount of data that other parties can access. The lack of effective security controls can lead to unauthorized use, data leaks, and breaches; these can result in identity theft or financial fraud at the individual level. Agencies typically use one of five main sharing models to allow other parties to access and analyze their data in a secure way.

First, researchers can get direct access to sensitive data through licensing or research data centers. Licenses allow individuals to apply for access, sign a data use agreement, retrieve an entire data file or one limited to information relevant for a given project after

the screening, and process information locally. Research data centers require researchers to either visit a physical location to process their queries or to access one virtually (Altman et al., 2016, Karr, 2016; Groves & Schoeffel, 2018). For example, the National Center for Education Statistics provides access licenses (NASEM, 2021), and the Census Bureau maintains the Federal Statistical Research Data Centers for its sensitive and restricted data.

Second, researchers may work with public use data files that have been edited to protect privacy and prevent reidentification using disclosure avoidance procedures. Preparing a data file for public release can involve simple reduction techniques like suppression, redaction, censoring, rounding, or aggregation; perturbative techniques like swapping or injecting noise; or creating a synthetic dataset, which is imputed and replaces original variable values but retains the statistical properties of the underlying data, as an alternative (Karr, 2016). For example, the Internal Revenue Service Statistics of Income Division currently releases a sample-based, time-lagged, and aggregated public use tax data file with a restricted number of variables. In partnership with the Brookings Institution, the Division also recently proposed to build a fully synthetic tax database that would broaden and make more timely access to all data while protecting taxpayer privacy (Burman et al., 2018).

The Survey of Income and Program Participation Synthetic Beta, Synthetic Longitudinal Business Database, and Survey of Consumer Finance also use synthetic data methods in developing publicly released files (NASEM 2017b). However, better computing power and data linkage pose considerable disclosure threats to these strategies. Government agencies employing simple approaches like removing sensitive information will need to adopt more robust approaches to avoid data privacy and security risks (Altman et al., 2016).

Verification servers are a third approach to offering data access while limiting disclosure risk. Researchers can use publicly released data files to conduct their work and subsequently check the consistency of their results against the unedited data via such a server. Verification servers run researchers' analyses and output a measure of how closely the results from the unedited and edited data align (Karr, 2016). The Internal Revenue Service Statistics of Income Division and Brookings Institution proposed to make available the option of running and checking data programs piloted on public use data against IRS' confidential data sources (Burman et al., 2018).

Data access can also be provided through data analysis systems hosted on remote servers (Keller-McNulty & Unger, 1993; 1998). Data analysis systems have a user interface that allows individuals to run simple descriptive analyses on publicly released data. Similarly, restricted data analysis systems give such functionality for restricted data use and typically only permit basic data operations (Karr, 2016). The Census Bureau's *data.census.gov* portal uses the former approach, offering users the ability to interact with multiple data sources that the agency maintains; for example, to view summaries and descriptive, obtain tables at different levels of aggregation, or download files. A similar approach is a system whereby the data required for a given analysis is copied to a central

location and kept there only for the statistical computations. This approach is called data minimization. Security risks are moderated by constraining the scope and duration for which the data are held (NASEM 2017b). Another example when computation on multiple data sources can be decentralized: the data never leave their individual original locations, and the multiple locations engage in a cryptographic protocol to cooperatively compute the outcome of the statistical computations (NASEM 2017b).

Finally, differential privacy is an approach that introduces a controlled level of noise to the data to limit disclosure risk, balanced against the desired degree of accuracy (Dwork, 2006; JASON, 2020). A privacy budget defines the level of acceptable disclosure risk. Because statistics published with differential privacy controls represent an approximation of the underlying data, this approach can pose issues for small populations. Differential privacy can also apply privacy-protecting transformations to an analyst's query rather than to the underlying data (NASEM Committee on National Statistics, 2019). The Census Bureau's OnTheMap tool employs this approach along with distributed record linkage to visualize workforce data for users (Machanavajjhala et al., 2008).

To test the integrity of their security systems and disclosure limitation controls, institutions frequently employ simulated reconstruction and reidentification attacks, which use aggregate data from multiple tables or sources to arrive at underlying individual-level records. For example, the Census Bureau simulated an attack that used publicly available 2010 census tabulations to correctly reconstruct individual-level demographic characteristics and location for 46%, or almost half, of the US population (Abowd, 2018; JASON, 2020). The Bureau then linked these records to third-party data, successfully reidentifying a total of 17% of the population by first and last name (ibid.).

Current strategies for assessing whether and how data can be released or shared include data disclosure board reviews, using disclosure risk checklists, building disclosure models and scenarios that could pose a threat for reidentification. The strategies also include establishing disclosure expert panels that provide regular input and enhancing agency coordination in data releases that would limit the availability of data files that could jointly lead to reidentification (Hawes, 2020; NASEM Committee on National Statistics, 2019).

Societal Challenges in Administrative Record Use

In addition to technical challenges that must be addressed when using administrative data for official statistics and social science research, employing administrative records and linking them with new data sources also raises societal questions about privacy and consent, and trust in government.

Privacy and Confidentiality

Privacy violations in the context of statistical data analysis using multiple sources of data are of two types. The first are threats to the security of the raw data and the second are threats using statistical findings to identify an individual or organization (NASEM 2017b). Several privacy laws apply to data handling, both survey and administrative, within federal statistical agencies and are well documented in many papers and reports (Keller et al., 2016; NASEM, 2017b; NASEM, 2021). The two laws most relevant to administrative data use are discussed here.

The 2002 Confidential Information Protection and Statistical Efficiency Act (CIPSEA) (44 U.S. Code § 101) provides a uniform approach across federal statistical agencies for record confidentiality protection and preventing reidentification. The act mandates that individuals' data can only be used for statistical purposes unless respondents provide informed consent for other uses. The act permits limited data sharing of business data between the Census Bureau, Bureau of Labor Statistics, and Bureau of Economic Analysis; however, they have not been able to implement this provision because of lack of corresponding authorization in the federal tax code (NASEM, 2017b).

Title 26 (26 U.S. Code § 6103) provides protection for sensitive information contained in Internal Revenue Service tax records. The Title specifies conditions for disclosing record contents to other federal agencies and permits record sharing with the Census Bureau for lawfully authorized statistical purposes like assisting in decennial census enumeration. Under restricted conditions and only for regulated and authorized purposes, other statistical agencies have also been able to access tax data; for example:

- the Congressional Budget Office can access tax data to analyze Social Security and Medicare programs,
- the Bureau of Economic Analysis can receive business data to build its survey sampling frames, and
- the National Agricultural Statistical Service can receive a subset of data to validate its Census of Agriculture frame (Greenia, 2008).

A Disclosure Review Board must oversee any record sharing and statistical products derived from these records (Department of Health and Human Services Office for Human Research Protections, 2018).

The Internal Revenue Service working with tax records must follow regulations of the US Tax Code, which governs privacy restrictions for tax record use, stating these can only be used for tax administration, research and analysis, and can only be conducted by a specified set of individuals. However, most rules governing record sharing, use, and linkage were put in place before the proliferation of devices and platforms like social media that generate large amounts of new data, which can then be integrated with administrative records and surveys.

Despite efforts to apply uniform approaches to privacy and confidentiality (e.g., CIPSEA), other steps must still be implemented agency by agency (e.g., negotiating data sharing agreements with each of the 50 states). One proposal to streamline the efforts by each statistical agency and advance the use of administrative data in federal statistics is the creation of a "secure environment" (NASEM 2017a). This secure environment would centralize administrative data quality access and analysis, link these data to surveys and other administrative data, evaluate it to ensure privacy, and allow a common approach for researcher access. This secure environment would need to be accessible by all statistical agencies and is necessary to scale expertise to achieve the vision for using administrative and private sector data to supplement and enhance existing surveys and create new data products (NASEM 2017b).

Consent

In addition to posing privacy issues, administrative data for research purposes, and linked administrative data, also complicate consent. In cases where administrative data are linked to survey data, survey respondents may have provided consent for the use of their survey responses but were not informed about or not asked to consent to uses involving linked administrative records. Even if individuals grant consent for administrative record linkage, linked data use, or other secondary uses involving additional data, the longevity of consent or the necessity to re-consent participants may be unclear.

Research also suggests there is evidence for consent bias in survey and administrative record linkage. Examples of linking administrative with survey and third-party data show that the propensity to give consent for record linkage varies by consent domain, respondent sociodemographic characteristics, attitudes towards privacy, interviewer characteristics, and survey characteristics (Mostafa, 2016; Sala et al., 2012; Sakshaug et al., 2012; Yang et al., 2019). Low consent and consent bias rates are a concern for the routine use of linked data files for official statistics. In addition to interviewer training and consent request framing, strategies like consent request placement across survey waves may increase the likelihood that individuals consent to record linkage (Eisnecker & Kroh, 2017). Current practical guidance on obtaining linkage consent is inconsistent, with opinions ranging from informed consent being necessary in all cases to the possibility of obtaining majority consent, not requiring consent in the presence of safeguards, or not requiring consent for linkage at all (GAO, 2001).

Trust and Transparency

Public trust in government is essential for official statistics and data products to be viewed as legitimate (NASEM, 2021). However, trust in government has decreased in the past decade (Hogan, 2020; Pew, 2020). A recent study of citizens' attitudes towards the 2020 decennial census found that between 47% and 59% of respondents reported they do not trust their local, state, or federal government. An earlier study indicated over 60% of the respondents expected the decennial census enumeration to be used to "help the police and FBI keep track of people who break the law," and over half assumed it would be used

to identify undocumented immigrants (Vines, 2018). Individuals often have concerns about the potential for surveillance and commercial exploitation of their data. They are wary of entrusting their records to an abstract and complex network of government actors. They are skeptical of sharing data when gains from doing so, like improved social infrastructure, are not immediately apparent or realized.

Public support for data sharing and linkage depends on trust. Citizens expect governments as data custodians to be trustworthy, accountable, transparent, and to provide avenues for public engagement where data policies can be assessed and shaped collectively (Sexton et al., 2017). Such support may also depend on knowledge and issue framing. When asked about data sharing and linkage for informed policy- and decision-making, Gallup survey respondents' attitudes were less favorable than when citing specific reasons like government accountability or efficient use of taxpayer funds (Fobia et al., 2019). Respondents cite benefits like increased funding for public services, particularly health, safety, education, and other public infrastructure, and contributing to a better future for their communities, as important motivators in their plans to provide their information in government surveys and the decennial enumeration (Vines, 2018).

Vichi and Hand (2019) and European Statistical System (ESS, 2019) define the role of trust in using "smart statistics." The term "smart statistics" is defined as a system that has autonomous processing that is infrastructural to the data. Adding 'trusted' to the term means, that the smart statistics with the adjective "trusted" implying that the decisions are based on sound data and information extraction; that is, the data must be properly representative of the system being described." Vichi and Hand 2019, page 606). Smart statistics have the following characteristics)

- Autonomous and automated data collection from sensors
- Continuous data collection autonomously data- driven
- Adaptively responding, by data-driven processing, to environmental changes
- Statistics extractable in real-time or in as close to real-time as makes sense
- A rational decision-making process when statistics are used for decisions

This definition of smart statistics introduces the possibility of using data beyond administrative data. ESS has established working groups to tackle the use of data from specific sectors, e.g., online job vacancies; enterprise characteristics; measuring electricity consumption, identifying energy consumption patterns; maritime and inland waterways statistics, and environmental statistics. We introduce the idea of smart statistics here because of the focus on identifying characteristics of trusted smart statistics has broad applicability for all types of non-survey data. A related concept is opportunity data that are derived from Internet-based information, such as websites and social media and captured through application programming interfaces (APIs) and Web scraping (Keller et al. 2020).

To meet the statistical criteria of trustworthiness, data quality, and value (statistics that support society's needs for information), smart statistics must be verifiable, have known representation of groups in the data, and an awareness of what is missing (Vichi

and Hand 2019). Specific challenges include the use of social media or web scraped data that are likely to not be representative of the overall population but can be useful for analyzing specific subsets of the population. Other challenges include verifying the ac curacy and validity and quality of raw data. These challenges are described below.

Addressing Technical and Societal Challenges in Administrative Data Use

Addressing Technical Challenges

As we have shown, administrative records, (and looking to the future, smart statistics, and opportunity data) and their linkage provide unique opportunities for reducing survey response burden, addressing the challenges of missing data and improving population coverage in the production of official statistics, adding contextual information to records from survey and other data files, assessing data quality, and other uses. At the same time, these data pose unique challenges compared to survey files that statistical agencies and social science research have typically employed. Administrative data can vary in completeness, accuracy, relevance, timeliness, coherence, and interpretability relative to a given use case (Seeskin, Datta, & Ugarte 2019).

Third-party data face similar challenges. In addition, many of these data sources are high-dimensional in that they often have massive numbers of features. This presents challenges for statistical uses (NASEM 2017a). For example, in terms of population coverage, there are often concerns about sample bias with these data sources, because data may exist only for those that can afford to buy a product or service. Or in the case of social media, data are available only for those who use the application (Couper, 2013). Retail scanner data may seem to be straightforward but can also have measurement issues depending on what is recorded. For statistical uses in a prices index, scanner data would provide the price, quantity, and description of the product. However, companies may be more interested in market share measures, customer response to promotions and price changes, or reactions to advertising and their scanner data will record those measures (NASEM 2017a). Another challenge is that as the size and assortment of data grows, the ability to identify individuals also grows. Creating anonymized data sources is more challenging as the techniques that are being developed for many legitimate applications of these data sources have also been shown to be used to identify individuals in the data (PCAST 2014, NASEM 2017a).

Several data quality assessment frameworks developed in both US and European statistical agencies, some prepared specifically for administrative data, can provide systematic guidance for evaluating administrative data sources, and detecting and correcting potential errors (Iwig et al., 2013; Lavigne & Nadeau, 2014; NASEM, 2017a; UN Economic Commission for Europe, 2014; (Seeskin, Datta, & Ugarte 2019). For example, the Data Quality Assessment Tool for Administrative Data (Iwig et al., 2013) covers the discovery, initial acquisition, and repeated acquisition phases of working with administrative records, and proposes over 40 questions pertaining to data relevance, accessibility, interpretability, coherence, accuracy, and institutional environment, that

data users should answer to better understand and anticipate possible data issues. The tool helps users gather information on data file characteristics including temporal and spatial coverage, available metadata and documentation, item consistency, missingness, record and processing changes over time, and on the credibility of the establishment that produced the data. Other frameworks cover similar stages and data dimensions, with adaptations for intended use (e.g., Lavigne & Nadeau, 2014). Seeskin, Datta, & Ugarte (2019) describe the process step-by-step with the focus on state and local staff applying the criteria to the use of their administrative data.

> "Some common features of state and local data include that they: require care as they often lack clear metadata; are prepared and stored in computing systems not designed for traditional statistical datasets; may have varying quality for different variables based on their importance for program administration; represent special populations without ready official statistics available; and are subject to changes in eligibility rules over time with groups differentially affected by policy changes." (Seeskin, Datta, & Ugarte 2019, page 2).

In cases where agencies and institutions obtain administrative data that are large-scale or comes from third-party sources, they may consider enhancing the Data Quality Assessment Tool for Administrative Data with insights from assessment frameworks developed for big data. Since third-party administrative data can be less stable over time in terms of contents and definitions, big data quality assessment frameworks—tailored to often unstructured, fleeting, and quickly changing datasets—help highlight potential issues not addressed in survey-based or traditional administrative record quality tools. One such big data quality framework (UN Economic Commission for Europe, 2014) is based on the principles of:

- fitness for use (i.e., data should be relevant and appropriate to answer project questions),
- generic and flexible nature (i.e., the framework should be adaptable to the variety of big data sources, including administrative ones), and
- effort-gain tradeoff (i.e., resources needed to work with the proposed new data source should be justified given potential benefits).

This framework considers data characteristics across the dimensions of institutional environment, privacy and security, complexity, completeness, usability, time factors, accuracy, coherence, and validity. It structures data assessment according to stages of working with the data, like the Data Quality Assessment Tool for Administrative Data, but focuses more on processing characteristics, i.e., the input or acquisition stage quality, throughput or transformation stage quality, and output quality. Seeskin, Datta, & Ugarte (2019) facilitate implementation of a data quality framework by proving code in the form of R markdowns that focuses on data accuracy, completeness of the data, and comparability for longitudinal data, but that can also be used for cross-sectional or time series data.

Similar guidance exists for data linkage frameworks. For example, the GUILD guidelines specify the types of information about each step of the linkage process that

should be provided alongside any linked data files to facilitate reproducibility and result validity (Gilbert et al., 2018). They recommend that at the data provision stage, researchers should include details about population and geographic coverage, data generation and quality control processes. At the linkage stage, researchers should describe linkage identifiers, missingness, data transformations, linkage algorithms, and successful matches, final sample representativeness, and disclosure limitation procedures. When conducting data analyses with linked files, users should report how they addressed linkage errors along with sensitivity analyses. Finally, this linkage information should also be made available in study reports (*ibid*).

In addition to employing quality assessment and linkage frameworks when working with administrative data internally, statistical agencies and social science institutions producing public use administrative data products may also consider providing data quality profile documents for external users and stakeholders. These data quality profiles could be structured in a way similar to documents commonly available for surveys like the Consumer Expenditure Survey (Bureau of Labor Statistics, 2021a), Schools and Staffing Surveys (National Center for Education Statistics, 2021), or the Residential Energy Consumption Survey (Energy Information Administration, 1996). While some information in current data quality profiles is survey-specific—for example, it includes information on frame development and sampling—other sections like data collection procedures, potential errors, data processing, and estimate evaluation are also applicable to and could be adapted for administrative data.

Taken together, employing data quality assessment frameworks, linkage frameworks, and communicating data quality clearly through data quality profiles can help minimize error and biases in working with administrative sources. Good practices advocated in these frameworks, like checking the veracity and accuracy of raw data, triangulating estimates, and thoroughly documenting data, measures, and transformations (Bostic et al., 2016; NASEM, 2017a; Vichi & Hand, 2019), can begin to address common sources of administrative data errors, e.g.,

- coverage error due to under-or over-coverage, sampling error due to large sample sizes,
- specification error due to conceptual mismatch between research questions and available data,
- missing data error due to item and unit missingness, measurement error due to entry mistakes, processing error due to coding and linkage issues, estimation error due to statistical adjustments, and
- analytic errors due to inappropriate modeling all threaten the integrity of administrative record-based estimates and data (Amaya et al., 2020; Groves & Schoeffel, 2018).
-

Addressing Societal Challenges

Building Trust and Transparency

Developing formal guidance on protecting individuals' privacy, informed consent requirements, and disclosure avoidance processes can raise trust in administrative records use (NASEM, 2021). Data stewards are integral to these processes and are a key role that statistical agencies and social science research establishments can institute to facilitate the socialization of administrative data usage.

Data stewards oversee linkage efforts with particular attention to data privacy and security issues, ensure compliance with relevant regulations, maintain accountability, and foster a privacy- and security-conscious organizational culture (GAO, 2001). Stewards must evaluate linkage privacy and ethical risks against potential study or project outcomes and benefits; such decision-making also requires appropriate technical, methodological, and substantive expertise. Independent assessments of proposed record linkage, soliciting input from users, subjects, privacy experts, and topic experts can inform linkage decision-making that results in fewer security risks (ibid). Researchers can also follow privacy analysis frameworks to systematically evaluate data release uses and benefits, identify privacy threats and vulnerabilities across the stages of data use, and select and implement security controls at each stage (Altman et al., 2016).

In addition to practicing good data stewardship, sharing results and clear communication with data users can also facilitate transparency and build trust in statistical and research institutions, and their administrative data usage. Timely dissemination of statistical products that meet the evolving needs of users at all levels of technical skill is a priority for federal statistical agencies (NASEM, 2021). To better understand their user base, agencies can gather user input and feedback on the administrative data they use, how they use it, and what relevant products—microdata, tabulations, reports, graphics, and others—they can provide to the community. Facilitating user access to documentation about data sources and statistical processing, openness about administrative data use, data limitations, sharing restrictions, and disclosure of potential biases and errors can further alleviate user and stakeholder concerns about issues like surveillance, misrepresentation, and commercial exploitation (*ibid*).

Building Capacity and Stakeholder Relationships

As more communities adopt a culture of data-driven decision-making, capacity building can help facilitate better use of administrative records in such processes. State agencies vary in whether they employ research, evaluation or data management staff in their budget to support data work, and agency staff has varying levels of analytic capacity (Allard et al., 2018). Without the necessary expertise and resources to support training or skill development, agencies may not make full use of their administrative data resources.

The complex tasks involved in working with administrative records also frequently expand beyond a single organization and sharing and integrating administrative records from disparate systems require multiple stakeholders' coordination (NASEM, 2021). Researchers who successfully completed challenging data linkages note that close and cooperative relationships between researchers, policymakers, and stakeholders are crucial during the process (Lunn et al., 2020). When technical experts, leaders, administrators, and other stakeholders are all engaged with developing data practices and sharing

procedures, the technical and societal administrative record use challenges may be more easily resolved.

Good relationships between data partners are essential at all stages of data-related processes, beginning with establishing data-sharing agreements (NASEM, 2021). State agencies note that they are more successful in securing such agreements with parties they already had existing relationships with, particularly in cases involving sensitive data related to health or child welfare (Allard et al., 2018). Support and expertise from data partners can also facilitate technical work like resolving variable inconsistencies. When similar variables from different data sources have incompatible values, individuals familiar with the data and collection process can assist in determining whether differences are due to variable definitions or coding errors and in deciding which set of variables to use (Wallgren & Wallgren, 2010).

Trust and a culture of data-driven decision-making can be deciding factors in whether large data linkage projects succeed or fail, like in the case of the Linked Information Network of Colorado, in which stakeholders explicitly attributed positive outcomes to close collaboration (Leboeuf, 2020). Conducting qualitative interviews, holding focus groups, clearly communicating, and piloting changes with stakeholders was also crucial in successfully transitioning from collecting survey responses to collecting administrative records for the Survey of Graduate Students and Post-doctorates in Science and Engineering (Gordon et al., 2018).

In developing good working relationships, all stakeholders can benefit. For example, statistical and academic institutions can enhance their survey and administrative sources with third-party data or develop new methods by addressing its technical challenges. Corporations can gain new insights, validate products, and learn from statistical and academic know-how (Struijs et al., 2014). With data linkage becoming standard practice in official statistics and social science research, federal agencies, academic institutions, and corporations should enhance collaboration for mutual benefit (GAO, 2001; Struijs et al., 2014).

One proposal mentioned briefly above is the creation of a new secure environment for analysis of data from multiple sources, coordinated acquisition and use of data, and the conduct of research about statistical agencies challenges. The entity should follow the principles and practices for federal statistical agencies and permit data access only for statistical purposes to researchers and statistical agencies. A major objective for this secure data enterprise is to identify data sources that can inform and improve national statistics, help develop techniques to use those data to compute national statistics while respecting privacy and other protection obligations on the data, and nurture the expertise required for these activities (NASEM 2017b, Hart & Potok 2020).

Conclusion

In this report, we demonstrate the increasing relevance of administrative data for producing official statistics and informing social science research. Current uses of

administrative records are described. These include standalone and linked surveys with other administrative data, survey data, and third-party or unstructured sources for creating sampling frames, as a substitute for survey data collection or missing values, for adding contextual variables to existing records, and for assessing data quality, among others. We then describe a summary review of the technical and societal challenges of administrative data source use that will have to be overcome for them to become a routine part of official and social science research. Although administrative records already inform many statistical agencies and institutions' research portfolios, further research is needed for developing sound linkage methods, evaluating data quality, and better socializing administrative data with its characteristics and limitations to facilitate the integration of these sources into statistical products.

References

Abowd, J. M. (2018). The U.S. Census Bureau adopts differential privacy. *Proceedings of the 24th Association for Computing Machinery International Conference on Knowledge Discovery and Data Mining*, 2867. doi:10.1145/3219819.3226070

ADP. 2021. ADP Employment Report Methodology. Automatic Data Processing, Inc. (ADP), https://adpemploymentreport.com/common/docs/ADP-NER-Methodology-Full-Detail.pdf

Ahrens, K. A., Haley, B. A., Rossen, L. M., Lloyd, P. C., & 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. doi:10.2105/AJPH.2016.303432

Allard, S. W., Wiegand, E. R., Schlecht, C., Datta, A. R., Goerge, R. M., & Weigensberg, E. (2018). State agencies' use of administrative data for improved practice: Needs, challenges, and opportunities. *Public Administration Review*, *78*(2), 240–250. doi:10.1111/puar.12883

Altman, M., Wood, A., O'Brien, D., Vadhan, S., & Gasser, U. (2016). Towards a modern approach to privacy-aware government data releases. *Berkeley Technology Law Journal*, *30*(2), 1068–2072. doi:10.15779/Z38FG17

Amaya, A., Biemer, P. P., & Kinyon, D. (2020). Total error in a big data world: Adapting the TSE framework to big data. *Journal of Survey Statistics and Methodology*, *8*(1), 89–119. doi:10.1093/jssam/smz056

Auerbach, J., Brummet, Q., Czajka, J., Hough, G. C., Hunsinger, E., & Salvo, J. (2019). Will administrative data save government surveys? *Significance*, *16* (5), 35–39. doi:10.1111/j.1740-9713.2019.01319.x

Bauder, M. & Judson, D. H. (2003). Administrative Records Experiment in 2000 (AREX 2000): Household level analysis. https://www.census.gov/pred/www/rpts/AREX2000_Household%20Analysis.pdf (retrieved 2/18/2021).

Berning, M. A. (2003). Administrative Records Experiment in 2000 (AREX 2000): Request for physical address evaluation. https://www.census.gov/pred/www/rpts/AREX2000_Physical_Address.pdf (retrieved 2/18/2021).

Bhaskar, R., Luque, A., Rastogi, S., & Noon, J. (2014). *Coverage and Agreement of Administrative Records and 2010 American Community Survey Demographic Data - Census Bureau* (#2014-14; CARRA Working Paper Series).

Bhaskar, R., Noon, J., & O'Hara, B. J. (2019). The errors in reporting medicare coverage: A comparison of survey data and administrative records. *Journal of Aging and Health*, *31*(10), 1806–1829. doi:10.1177/0898264318797548

Bostic, W. G., Jarmin, R. S., & Moyer, B. (2016). Modernizing federal economic statistics. *American Economic Review*, *106*(5), 161–164. doi:10.1257/aer.p20161061

Brady, H. E. (2019). The challenge of big data and data science. *Annual Review of Political Science*, *22*(1), 297–323. doi:10.1146/annurev-polisci-090216-023229

Brummet, Q. (2014). Comparison of survey, federal, and commercial address data quality (Working Paper 2014-06). Washington, DC: U.S. Census Bureau Center for Administrative Records Research and Applications.

Brummet, Q., Flanagan-Doyle, D., Mitchell, J., Voorheis, J., Erhard, L., & Mcbride, B. (2018). *Investigating the Use of Administrative Records in the Consumer Expenditure Survey - Census Bureau & BLS* (2018-01). https://www.bls.gov/opub/hom/cex/pdf/cex.pdf

Bureau of Economic Analysis (2022). Health Care website. https://www.bea.gov/data/special-topics/health-care

Bureau of Labor Statistics (2021). About the Health Care Satellite Account. Available online at https://www.bea.gov/data/special-topics/health-care (accessed 4/2/2021).

Bureau of Labor Statistics (2021a). Data Quality in the Consumer Expenditure Surveys. Available at https://www.bls.gov/cex/cecomparison.htm (accessed 5/10/2021).

Burman, L. E., Engler, A., Khitatrakun, S., Nunns, J. R., Armstrong, S., Iselin, J., Macdonald, G., & Stallworth, P. (2018). *Safely expanding research access to administrative tax data: creating a synthetic public use file and a validation server.* Washington, DC*:* Urban Institute & Brookings Institution. https://www.urban.org/sites/default/files/publication/99247/safely_expanding_research_access_to_administrative_tax_data_creating_a_synthetic_public_use_file_and_a_validation_server_2.pdf (retrieved 2/18/2021).

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. (retrieved 6/22/2021)

Centers for Medicare and Medicaid Services. (2021). Integrated Data Repository (IDR). Available online at https://www.cms.gov/Research-Statistics-Data-and-Systems/Computer-Data-and-Systems/IDR (retrieved 2/18/2021).

Chetty, R., Hendren, N., Kline, P., & Saez, E. (2014). Where is the land of opportunity? The geography of intergenerational mobility in the United States. The Quarterly Journal of Economics, 129(4), 1553–1623. doi:10.1093/qje/qju022

City of New York (2020). Office of Economic Opportunity. NYC Government Poverty Measure, 2018. Available at https://www1.nyc.gov/site/opportunity/poverty-in-nyc/poverty-measure.page (accessed 8/9/2021)

Citro, C. (2014). From multiple modes for surveys to multiple data sources for estimates. Survey Methodology, 40 (2), 137-161.

Commission on Evidence-Based Policymaking. (2017). The promise of evidence-based policymaking: Report of the Commission on Evidence-Based Policymaking. Available online at https://www.cep.gov/report/cep-final-report.pdf (retrieved 2/18/2021).

Cordis, A. S., & Milyo, J. (2016). Measuring public corruption in the United States: Evidence from administrative records of federal prosecutions. *Public Integrity*, *18*(2), 127–148. doi:10.1080/10999922.2015.1111748

Couper, M. (2013). Is the sky falling? New technology, changing media, and the future of surveys. Survey Research Methods, 7(3), 145-156.

Couper, M. P. (2017). New developments in survey data collection. *Annual Review of Sociology*, *43*(1), 121–145. doi:10.1146/annurev-soc-060116-053613

Cristia, J., & Schwabish, J. A. (2009). Measurement error in the SIPP: Evidence from administrative matched records. *Journal of Economic and Social Measurement*, *34*(1), 1–17. doi:10.3233/JEM-2009-0311

Cruze, N.B. (2015). Integrating survey data with auxiliary sources of information to estimate crop yields. In Proceedings of the Survey Research Methods Section (pp. 565–578).Washington, DC: American Statistical Association.

Czajka, J. L., & Beyler, A. (2016). Declining response rates in federal surveys: Trends and implications. *Mathematica Policy Research*, *I*(202), 1–54. https://www.mathematica-mpr.com/our-publications-and-findings/publications/declining-response-rates-in-federal-surveys-trends-and-implications-background-paper

Department of Health and Human Services Office for Human Research Protections. (2018). Institutional Review Board written procedures: Guidance for institutions and IRBs. Available at https://www.hhs.gov/ohrp/regulations-and-policy/guidance/institutional-issues/institutional-review-board-written-procedures/index.html (accessed 3/4/2021).

Dwork C. (2006). Differential privacy. In: Bugliesi M., Preneel B., Sassone V., Wegener I. (eds) *Automata, Languages and Programming*. ICALP 2006. Lecture Notes in Computer Science, vol 4052. Springer: Berlin, Heidelberg. doi:10.1007/11787006_1

Eisnecker, P. S., & Kroh, M. (2017). The informed consent to record linkage in panel studies: Optimal starting wave, consent refusals, and subsequent panel attrition. *Public Opinion Quarterly*, *81*(1), 131–143. doi:10.1093/poq/nfw052

Elliott, A. F., Davidson, A., Lum, F., Chiang, M. F., Saaddine, J. B., Zhang, X., Crews, J. E., & Chou, C.-F. (2012). Use of electronic health records and administrative data for public health surveillance of eye health and vision-related conditions in the United States. *American Journal of Ophthalmology*, *154*(6), S63–S70. doi:10.1016/j.ajo.2011.10.002

Energy Information Administration. (1996). Residential Energy Consumption Survey Quality Profile. Energy Consumption Series, DOE/EIA-0555(96)/1. Washington, DC: US Department of Energy. Available at https://www.eia.gov/consumption/residential/data/1993/pdf/555961a.pdf (accessed 5/10/2021).

Ennis, S. R., Porter, S. R., Noon, J. M., & Zapata, E. (2018). When race and Hispanic origin reporting are discrepant across administrative records and third party sources:

Exploring methods to assign responses. *Statistical Journal of the IAOS*, *34*(2), 179–189. doi:10.3233/SJI-170374

European Statistical System (ESS). (2019). Quality Assurance Framework of the European Statistical System. https://ec.europa.eu/eurostat/documents/64157/4392716/ESS-QAF-V1-2final.pdf/bbf5970c-1adf-46c8-afc3-58ce177a0646

Fast, D., & Fleck, S. (2019). Unit Values for Import and Export Price Indexes – A proof of concept. In *NBER Working Paper No. 26373*. doi:10.3386/w26373

Fellegi, I. P., & Sunter, A. B. (1969). A theory for record linkage. *Journal of the American Statistical Association*, 64, 1183-1210. doi:10.1080/01621459.1969.10501049

Fernandez, L., Shattuck, R., & Noon, J. (2018). *The Use of Administrative Records and the American Community Survey to Study the Characteristics of Undercounted Young Children in the 2010 Census - Census Bureau* (#2018-05; CARRA Working Paper Series). https://www.census.gov/content/dam/Census/library/working-papers/2018/adrm/carra-wp-2018-05.pdf

Fobia, A. C., Holzberg, J., Eggleston, C., Childs, J. H., Marlar, J., & Morales, G. (2019). Attitudes towards data linkage for evidence-based policymaking. *Public Opinion Quarterly*, *83*(Special), 264–279. doi:10.1093/poq/nfz008

Fox, Liana (2020). The Supplemental Poverty Measure: 2019. U.S. Census Bureau, Current Population Reports P60-272 September.

Gilbert, R., Lafferty, R., Hagger-Johnson, G., Harron, K., Zhang, L. C., Smith, P., Dibben, C., & Goldstein, H. (2018). GUILD: GUidance for Information about Linking Data sets. *Journal of Public Health*, *40*(1), 191–198. doi:10.1093/pubmed/fdx037

Gordon, J., Eckman, S., Einaudi, P., Sanders, H., & Yamaner, M. (2018). Using administrative records to increase quality and reduce burden in the Survey of Graduate Students and Postdoctorates in Science and Engineering. *Statistical Journal of the IAOS*, *34*(4), 529–537. doi:10.3233/SJI-180450

Government Accountability Office. (2001). *Record Linkage and Privacy: Issues in Creating New Federal Research and Statistical Information (GAO-01-126SP)* (Issue April 2001). http://www.gao.gov/new.items/d04845sp.pdf

Graham, S. J. H., Grim, C., Islam, T., Marco, A. C., & Miranda, J. (2018). Business dynamics of innovating firms: Linking U.S. patents with administrative data on workers and firms. *Journal of Economics & Management Strategy*, *27*(3), 372–402. doi:10.1111/jems.12260

Greenia, N. H. (2008). Statistical use of US federal tax data. Paper presented at the International Seminar on the Use of Administrative Data for Economic Statistics and Register-based Population Census, May 19-20, Daejeon, South Korea. Available at www.oecd.org/sdd/41143235.pdf (accessed 4/2/2021).

Groves, R. M., & Schoeffel, G. J. (2018). Use of administrative records in evidence-based policymaking. *Annals of the American Academy of Political and Social Science*, *678*(1), 71–80. doi:10.1177/0002716218766508

Harron, K., Dibben, C., Boyd, J., Hjern, A., Azimaee, M., Barreto, M. L., & Goldstein, H. (2017). Challenges in administrative data linkage for research. *Big Data and Society*, *4*(2), 1–12. doi:10.1177/2053951717745678

Hart, N., & Potok, N. (2020). *Modernizing US Data Infrastructure: Design Considerations for Implementing a National Secure Data Service to Improve Statistics and Evidence Building*. Washington, DC: Data Foundation. Available online at https://www.datafoundation.org/cover-page-modernizing-us-data-infrastructure-design-considerations-for-implementing-a-national-secure-data-service-2020 (accessed 3/19/2021).

Hastings, J. S., Howison, M., Lawless, T., Ucles, J., & White, P. (2019). Unlocking data to improve public policy. *Communications of the ACM*, *62*(10), 48–53. doi:10.1145/3335150

Hawes, M. B. (2020). Implementing differential privacy: Seven lessons from the 2020 United States Census. *Harvard Data Science Review*, *2*(2). doi:10.1162/99608f92.353c6f99

Health and Retirement Study. (2021). Available restricted data products. Available online at https://hrs.isr.umich.edu/data-products/restricted-data/available-products (retrieved 2/18/2021).

Hill, S. J., & Huber, G. A. (2017). Representativeness and motivations of the contemporary donorate: Results from merged survey and administrative records. *Political Behavior*, *39*(1), 3–29. doi:10.1007/s11109-016-9343-y

Hof, M. H., & Zwinderman, A. H. (2012). Methods for analyzing data from probabilistic linkage strategies based on partially identifying variables. Statistics in medicine, 31(30), 4231-4242. Hogan, H. (2020). Distrust in the governments brings risk to the Census. *Harvard Data Science Review*, *2*(1). doi:10.1162/99608f92.11f3e977

Houser, K. A., & Sanders, D. (2017). The use of big data analytics by the IRS: Efficient solutions or the end of privacy as we know it? *Vanderbilt Journal of Entertainment & Technology Law*, *19*(4), 817–872.

Hutchinson, R. J. (2017). Reducing survey burden through third-party data sources. Paper presented at the Annual Conference of European Statisticians, October 10-12. Available online at https://unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.44/2017/mtg3/DC2017_1-4_Hutchinson_USA_AD.pdf (accessed 4/2/2021).

Iwig, W., Berning, M., Marck, P., and Prell, M. (2013). Data quality assessment tool for administrative data. Available at https://nces.ed.gov/fcsm/pdf/DataQualityAssessmentTool.pdf (accessed 5/10/2021).

Jarosz, B., & Hofmockel, J. (2013). Research note: What counts as a house? Comparing 2010 Census counts and administrative records. *Population Research and Policy Review*, *32*(5), 753–765. doi:10.1007/s11113-013-9290-9

JASON. (2020). *Formal Privacy Methods for the 2020 Census (JSR-19-2F)*. Available at https://www2.census.gov/programs-surveys/decennial/2020/program-management/planning-docs/privacy-methods-2020-census.pdf (accessed 3/19/2021).

Johnson, B. W., Ludlum, M., & Rib, T. (2018). Using administrative records to improve official statistics produced by the Statistics of Income Division, IRS. *Statistical Journal of the IAOS, 34*(1), 25–32. doi:10.3233/SJI-170415

Judson, T. J., Bennett, A. V., Rogak, L. J., Sit, L., Barz, A., Kris, M. G., ... & Basch, E. (2013). Feasibility of long-term patient self-reporting of toxicities from home via the Internet during routine chemotherapy. Journal of Clinical Oncology, 31(20), 2580.

Karr, A. F. (2016). Data sharing and access. *Annual Review of Statistics and Its Application, 3*(1), 113–132. doi:10.1146/annurev-statistics-041715-033438

Keller, S. A., Shipp, S. S., Schroeder, A. D., & Korkmaz, G. (2020). Doing Data Science: A Framework and Case Study. Harvard Data Science Review, 2(1). https://doi.org/10.1162/99608f92.2d83f7f5

Keller, S. A., Shipp, S., & Schroeder, A. (2016). Does big data change the privacy landscape? A review of the issues. *Annual Review of Statistics and Its Application, 3,* 161–180. doi:10.1146/annurev-statistics-041715-033453

Keller, S., Lancaster, V., & Shipp, S. (2017). Building capacity for data driven governance: Creating a new foundation for democracy. *Statistics and Public Policy, 4*, 1–11. doi:10.1080/2330443X.2017.1374897

Keller-McNulty, S., & Unger, E. (1993). Database systems: inferential security. *Journal of Official Statistics, 9*, 475-499.

Keller-McNulty, S., & Unger, E. (1998). A database system prototype for remote access to information based on confidential data. *Journal of Official Statistics, 14*, 347-360.

Kingkade, W. (2013). Self-assessed housing values in the American Community Survey: An exploratory evaluation using linked real estate records. Paper presented at the 2013 Joint Statistical Meetings, Montréal, Québec, Canada, August 3-8.

Lahiri, P., & Larsen, M. D. (2005). Regression analysis with linked data. Journal of the American statistical association, 100(469), 222-230.

Lavigne, M. & Nadeau, C. (2014). A framework for the evaluation of administrative data. Proceedings of Statistics Canada Symposium Beyond Traditional Survey Taking: Adapting to a Changing World. Available at https://www.statcan.gc.ca/eng/conferences/symposium2014/program/14284-eng.pdf (5/11/2021).

Leboeuf, W. (2020). How the linked information network of Colorado (LINC) is bringing cross-system data to prevention conversations. *Policy & Practice*, *25*, 25–27.

Ludlum, M. (2004). Data interpretation across sources: A study of Form 990-PF information collected from multiple databases. Paper presented at the Joint Statistical Meetings, Toronto, Ontario, Canada, August 11. Available at https://www.irs.gov/pub/irs-soi/2004preprintar06.pdf (accessed 4/2/2021).

Lunn, P. D., Lyons, S., & Murphy, M. (2020). Predicting farms' noncompliance with regulations on nitrate pollution. *Journal of Environmental Planning and Management, 63*(13), 2313–2333. doi:10.1080/09640568.2020.1719050

Machanavajjhala, A., Kifer, D., Abowd, J., Gehrke, J., & Vilhuber, L. (2008). Privacy: Theory meets practice on the map. *Proceedings of the IEEE 24th International Conference on Data Engineering,* 277-286. doi: 10.1109/ICDE.2008.4497436.

Maxfield, L. D. (2008). The uses of administrative data at the US Social Security Administration. Paper presented at the International Seminar on the Use of Administrative Data for Economic Statistics and the Register-Based Population and Housing Census, Daejeon, South Korea, May 19-20. Available at www.oecd.org/sdd/41143137.pdf (accessed 4/2/2021).

McClure, Dave, Santos, R., Kooragayala, S. (2017) Administrative Records in the 2020 US Census: Civil Rights Considerations and Opportunities, Washington, DC: Urban Institute. https://www.urban.org/research/publication/administrative-records-2020-us-census (Accessed 8/11/2021).

Miller, P. V. (2017). Is There a Future for Surveys? *Public Opinion Quarterly*, *81*(S1), 205–212. doi:10.1093/poq/nfx008

Min, J., Gurka, K. K., Kalesan, B., Bian, J., & Prosperi, M. (2019). Injury burden in the United States: Accurate, reliable, and timely surveillance using electronic health care data. *American Journal of Public Health*, *109*(12), 1702–1706. doi:10.2105/AJPH.2019.305306

Molfino, E., Korkmaz, G., Keller, S. A., Schroeder, A., Shipp, S., & Weinberg, D. H. (2017). Can administrative housing data replace survey data?. *Cityscape*, *19*(1), 265–292.

Mortenson, J. A., Cilke, J., Udell, M., & Zytnick, J. (2009). Attaching the left tail: A new profile of income for persons who do not appear on federal income tax returns. Paper presented at the 102nd Annual Conference on Taxation, November 12-14, Denver, CO. Available at https://ntanet.org/wp-content/uploads/proceedings/2009/011-mortenson-attaching-left-tail-2009-nta-proceedings.pdf (accessed 4/6/2021).

Mostafa, T. (2016). Variation within households in consent to link survey data to administrative records: evidence from the UK Millennium Cohort Study. *International Journal of Social Research Methodology*, *19*(3), 355–375. doi:10.1080/13645579.2015.1019264

National Academies of Sciences, Engineering, and Medicine [NASEM]. (2021). *Principles and Practices for a Federal Statistical Agency: Seventh Edition*. Washington, DC: The National Academies Press. doi:10.17226/25885.

National Academies of Sciences, Engineering, and Medicine [NASEM]. (2019*). Improving the American Community Survey: Proceedings of a Workshop*. Washington, DC: The National Academies Press. doi:10.17226/25387.

National Academies of Sciences, Engineering, and Medicine [NASEM]. (2017a). *Federal Statistics, Multiple Data Sources, and Privacy Protection: Next Steps*. Washington, DC: The National Academies Press. doi:10.17226/24893.

National Academies of Sciences, Engineering, and Medicine [NASEM]. (2017b). Innovations in Federal Statistics: Combining Data Sources While Protecting Privacy. Washington, DC: The National Academies Press. doi: 10.17226/24652.

National Academies of Sciences, Engineering, and Medicine [NASEM]. Committee on National Statistics (2019). Workshop on 2020 Census Data Products: Data Needs and Privacy Considerations. Available online at https://sites.nationalacademies.org/DBASSE/CNSTAT/DBASSE_196518 (accessed 3/19/2021).

National Center for Education Statistics (2021). A quality profile for SASS: Aspects of the quality of data in the Schools and Staffing Surveys. Available at https://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=94340 (accessed 5/10/2021).

National Center for Health Statistics. (2021). NCHS Data Linked to HUD Housing Assistance Program Files. Available online at https://www.cdc.gov/nchs/data-linkage/hud.htm (retrieved 2/18/2021).

National Research Council. (2009). Coverage Measurement in the 2010 Census. Washington, DC: The National Academies Press. doi:10.17226/12524.

National Research Council (2009). Strategies for a BEA Satellite Health Care Account: Summary of a workshop. Washington, DC: The National Academies Press. doi:10.17226/12494.

National Research Council Committee on National Statistics. (2008). Strategies for a BEA Satellite Health Care Account: Summary of a Workshop. Washington, DC: National Academies Press. Available at https://www.ncbi.nlm.nih.gov/books/NBK214859/ (accessed 4/2/2021).

New York State Department of Health (2021). Nursing Home Weekly Bed Census: Beginning 2009. Available at https://health.data.ny.gov/Health/Nursing-Home-Weekly-Bed-Census-Beginning-2009/uhyy-xp9s (Accessed 8/9/2021)O'Hara, A. (2016). Preliminary Research for Replacing or Supplementing the Income Question on the American Community Survey with Administrative Records. Available: https://www.census.gov/content/dam/Census/library/working-papers/2016/acs/2016_Ohara_01.pdf (accessed 4/2/2021).

Parker, R. P. (2006). Employment dynamics: BLS and Census Bureau use administrative records to provide new data. *Business Economics*, *41*(2), 55–61.

PCAST. (2014). Big Data and Privacy: A Technological Perspective. President's Council of Advisors on Science and Technology https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/PCAST/pcast_big_data_and_privacy_-_may_2014.pdf (accessed 4/2/2021).

Pedace, R., & Bates, N. (2000). Using administrative records to assess earnings reporting error in the survey of income and program participation. In *Journal of Economic and Social Measurement* (Vol. 26). IOS Press.

Pew Research Center. (2020). Americans' views of government: Low trust, but some positive performance ratings. Available online at

https://www.pewresearch.org/politics/2020/09/14/americans-views-of-government-low-trust-but-some-positive-performance-ratings/ (retrieved 2/18/2021).

Playford, C. J., Gayle, V., Connelly, R., & Gray, A. J. G. (2016). Administrative social science data: The challenge of reproducible research. *Big Data and Society*, *3*(2), 1–13. doi:10.1177/2053951716684143

Rastogi, S., & O'Hara, A. (2012). *2010 Census Match Study*. 2010 Census Memoranda Series, Report 247. Washington, DC: US Census Bureau Center for Administrative Records Research and Applications.

Robertson, K. (2017). Benchmarking the Current Employment Statistics survey: perspectives on current research. *Monthly Labor Review*, *11*, 1–21. doi:10.21916/mlr.2017.27

Sakshaug, J. W., Couper, M. P., Ofstedal, M. B., & Weir, D. R. (2012). Linking survey and administrative records: Mechanisms of consent. *Sociological Methods and Research*, *41*(4), 535–569. doi:10.1177/0049124112460381

Sala, E., Burton, J., & Knies, G. (2012). Correlates of obtaining informed consent to data linkage: Respondent, interview, and interviewer characteristics. *Sociological Methods and Research*, *41*(3), 414–439. doi:10.1177/0049124112457330

Schantz, Kathryn, 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. Available at https://www.census.gov/content/dam/Census/library/working-papers/2018/demo/SEHSD-WP2018-30.pdf (Accessed on 8/9/2021).

Seeskin, Z. H., Ugarte, G., & Datta, A. R. (2019). Constructing a toolkit to evaluate quality of state and local administrative data. International Journal of Population Data Science, 4(1). https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7479930/

Sexton, A., Shepherd, E., Duke-Williams, O., & Eveleigh, A. (2017). A balance of trust in the use of government administrative data. *Archival Science*, *17*(4), 305–330. doi:10.1007/s10502-017-9281-4

SLDS. (2021). Statewide Longitudinal Data Systems Grant Program. Institute for Education Sciences (IES). National Center for Education Statistics (NCSES). https://nces.ed.gov/programs/slds/about_SLDS.asp

Stansbury, J. P., Reid, K. J., Reker, D. M., Duncan, P. W., Marshall, C. R., & Rittman, M. (2004). Why ethnic designation matters for stroke rehabilitation: Comparing VA administrative data and clinical records. *The Journal of Rehabilitation Research and Development*, *41*(3A), 269. doi:10.1682/JRRD.2004.04.0046

Struijs, P., Braaksma, B., & Daas, P. J. H. (2014). Official statistics and big data. In *Big Data and Society* (Vol. 1, Issue 1, p. 205395171453841). SAGE Publications Ltd. doi:10.1177/2053951714538417

Taeuber, Cynthia, Resnick, D.M. Love, S.P., Staveley, J., Wilde, P., Larson R. (2004). Differences in Estimates of Food Stamp Program Participation Between Surveys and Administrative Records, A Joint Project of the U.S. Census Bureau, University of Baltimore, U.S. Department of Agriculture, and the Economic Research Service, and Maryland Department of Human Resources, Family Investment Administration. Available at https://www.ubalt.edu/jfi/jfi/reports/fstampfinrept273004.pdf .

UN Economic Commission for Europe. (2014). A Suggested Framework for the Quality of Big Data. Available at https://statswiki.unece.org/download/attachments/108102944/Big%20Data%20Quality%20Framework%20-%20final-%20Jan08-2015.pdf?version=1&modificationDate=1420725063663&api=v2 (accessed 5/10/2021).

US Bureau of Labor Statistics (2021). Data quality in the Consumer Expenditure Surveys. Available at https://www.bls.gov/cex/cecomparison.htm (accessed 5/4/2021).

US Census Bureau (2020). Local Update of Census Addresses Operation (LUCA). Available at https://www.census.gov/programs-surveys/decennial-census/about/luca.html (accessed 8/9/2021).

US Census Bureau (2021a). Monthly State Retail Sales Technical Documentation. Available online at https://www.census.gov/retail/mrts/www/statedata/msrs_technical_documentation.pdf (accessed 4/2/2021).

US Census Bureau (2021b). Methodology for the US population estimates: Vintage 2020. Available at www2.census.gov/programs-surveys/popest/technical-documentation/methodology/2010-2020/methods-statement-v2020-final.pdf (accessed 4/2/2021).

US Census Bureau, (no date) Frequently Asked Questions, What are administrative records and third-party data? https://www.census.gov/content/dam/Census/about/about-the-bureau/adrm/data-linkage/Data%20Acquisitions%20Frequently%20Asked%20Questions.pdf (accessed on 2/12/2022)

US Department of Agriculture. (2021). Census-FNS-ERS Joint Project. Available online at https://www.ers.usda.gov/topics/food-nutrition-assistance/food-assistance-data-collaborative-research-programs/census-fns-ers-joint-project/ (retrieved 2/18/2021).

US Energy Information Administration (2021). Residential Energy Consumption Survey (RECS). Available at https://www.eia.gov/consumption/residential/ (accessed 5/4/2021).

US House of Representatives (2019), Foundation for Evidence-Based Policymaking Act of 2018. https://www.congress.gov/bill/115th-congress/house-bill/4174

US Office of Management and Budget. (2000). Guidance for providing and using administrative data for statistical purposes (M-14-06). Available online at

https://obamawhitehouse.archives.gov/sites/default/files/omb/memoranda/2014/m-14-06.pdf (retrieved 2/18/2021).

US Office of Management and Budget. (2000). Guidance on inter-agency sharing of personal data - protecting personal privacy (M-01-05). Available online at https://obamawhitehouse.archives.gov/omb/memoranda_m01-05/ (retrieved 2/18/2021).

US Office of Management and Budget. (2009). Open Government Directive (M-10-06). In *Obama's Open Government Initiative: Transformation through Transparency*. https://obamawhitehouse.archives.gov/sites/default/files/omb/assets/memoranda_2010/m10-06.pdf

US Office of Management and Budget. (2013). Open data policy – managing information as an asset (M-13-13). Available online at https://www.whitehouse.gov/sites/whitehouse.gov/files/omb/memoranda/2013/m-13-13.pdf (retrieved 2/18/2021).

US Office of Management and Budget. (n.d.) Open government national action plans. Available online at https://obamawhitehouse.archives.gov/open/partnership/national-action-plans (retrieved 2/18/2021).

US Office of Management and Budget. (2014.) Guidance for Providing and Using Administrative Data for Statistical Purposes. Available online at https://obamawhitehouse.archives.gov/sites/default/files/omb/memoranda/2014/m-14-06.pdf (retrieved 2/18/2021).

Ver Ploeg, M., Mancino, L., Todd, J.E., Clay, D.M., and Scharadin, B. (2015). Where Do Americans Usually Shop for Food and How Do They Travel to Get There? Initial Findings from the National Household Food Acquisition and Purchase Survey. Economic Information Bulletin Vol. 138. Washington, DC: U.S. Department of Agriculture.

Vichi, M., & Hand, D. J. (2019). Trusted smart statistics: The challenge of extracting usable aggregate information from new data sources. *Statistical Journal of the IAOS*, *35*(4), 605–613. doi:10.3233/SJI-190526

Vines, M. (2018). 2020 Census Barriers, Attitudes, and Motivators Study (CBAMS) Survey and focus groups key findings. Available at https://becountedmi2020.com/wp-content/uploads/CBAMS_Presentation_MNA_121218.pdf (accessed 4/6/2021).

Wagner, D. & Layne, M. (2014). The Person Identification Validation System (PVS): Applying the Center for Administrative Records Research and Applications' (CARRA) record linkage software (CARRA Working Paper Series, #2014-01). Available online at https://www.census.gov/content/dam/Census/library/working-papers/2014/adrm/carra-wp-2014-01.pdf (retrieved 2/18/2021).

Wallgren, A., & Wallgren, B. (2010). Agricultural survey methods. In R. Benedetti, M. Bee, G. Espa, & F. Piersimoni (Eds.), *Agricultural Survey Methods*. John Wiley & Sons, Ltd. doi:10.1002/9780470665480

Wallgren, A., & Wallgren, B. (2014). The nature of administrative data. In A. Wallgren & B. Wallgren (Eds.), *Register-Based Statistics* (Second, pp. 25–36). John Wiley & Sons, Ltd. doi:10.1002/9781118855959.ch2

Wheaton, L., Durham, C., & Loprest, P. (2012). *TANF and Related Administrative Data Project: Final Evaluation Report to Administration for Children and Families (ACF)*. http://www.urban.org/sites/default/files/publication/25511/412590-TANF-and-Related-Administrative-Data-Project-Final-Evaluation-Report.PDF

Williams, A. C., & Moore, R. A. Jr. (1998). Using administrative data to enhance the sampling frame for the 1997 Survey of Minority-Owned Business Enterprises. Proceedings of the American Statistical Association Survey Research Methods Section. Available at www.asasrms.org/Proceedings/papers/1998_084.pdf (accessed 4/2/2021).

Winkler, W. E. (1995). Matching and record linkage. *Business Survey Methods, 1*, 355–384.

Winkler, W. E. (2006). Overview of record linkage and current research directions (Research Report Series, Statistics #2006-2). Bureau of the Census. Washington, DC: US Census Bureau.

World Economic Forum (2011). Personal data: The emergence of a new asset class. Available at www3.weforum.org/docs/WEF_ITTC_PersonalDataNewAsset_Report_2011.pdf (accessed 5/4/2021).

Yancey, W. E. (2002). BigMatch: A Program for Extracting Probable Matches from a Large File for Record Linkage (Research Report Series, Computing #2002-1). Washington, DC: US Census Bureau.

Yang, D., Fricker, S., & Eltinge, J. (2019). Methods for exploratory assessment of consent-to-link in a household survey. *Journal of Survey Statistics and Methodology*, *7*(1), 118–155. doi:10.1093/jssam/smx031

Zanutto, E., & Zaslavsky, A. (2002). Using Administrative Records to Improve Small Area Estimation: An Example from the U.S. Decennial Census. *Journal of Official Statistics*, *18*(4), 559–576. http://www.scb.se/contentassets/ca21efb41fee47d293bbee5bf7be7fb3/using-administrative-records-to-improve-small-area-estimation-an-example-from-the-u.s.-decennial-census.pdf (retrieved 2/18/2021).