FINAL PROJECT REPORT

To Big Data or Not: Determining the Use of Big Data

CONTRACT NO. HHSP233201500048I

NOVEMBER 30, 2018

PRESENTED TO:

Jeongsoo Kim
Office of The Assistant Secretary for Planning and Evaluation

Department of Health and Human Services
200 Independence Avenue, SW,
Room 440G.7
Washington, DC 20201

PRESENTED BY:

Felicia LeClere
Zachary Seeskin
Jaehoon Ahn
Samantha Rosner
NORC at the University of Chicago

55 East Monroe Street
30th Floor
Chicago, IL 60603
(312) 759-4000 Main

NORC at the UNIVERSITY of CHICAGO

Acknowledgments

We thank our partners at The Office of The Assistant Secretary for Planning and Evaluation, Department of Health and Human Services for their support and guidance throughout this project. We particularly thank Jeongsoo Kim, Amanda Cash, Rashida Dorsey, Jim Sorace, and Joshua Williams. All views expressed are solely those of the authors and are not necessarily those of NORC at the University of Chicago or the U.S. Department of Health and Human Services, Office of The Assistant Secretary for Planning and Evaluation.

We are grateful to the experts who generously provided their time and insight to participate in interviews for this project: Chaintanya Baru, Paul Biemer, Christine Borgman, Philip Bourne, Charlie Catlett, Jeffrey Chen, Michael Chernew, John Eltinge, Gordon Gao, Robert Gibbons, Juergen Klenk, David Lazer, Girish Mundada, Amy O'Hara, Nathaniel Osgood, Alejandro Reti, Lucy Savitz, and Jean-Ezra Yeung.

In addition, we thank the following individuals for their comments and assistance which have greatly improved this report: Martin Barron, Lynette Bertsche, Mike Cohen, Prashila Dullabh, Sherry Emery, Don Jang, Daniel Lawrence, Yoonsang Kim, Katie O'Doherty, Megha Ravanam, Ben Skalland, and Ernie Tani.

Table of Contents

List of Exhibits

Executive Summary

NORC at the University of Chicago presents this final report of activities associated with the "Big Data or Not: Determining the Use of Big Data" to the Office of the Assistant Secretary for Planning and Evaluation (ASPE), Department of Health and Human Services (HHS), Contract # HHSP233201500048I. The purpose of this contract was to provide the Office of Science and Data Policy at ASPE with some informed observations concerning the use of new data sources and data management strategies in policy research, evaluation, and decision-making at the federal level. A secondary goal was to identify successful training models in data science for the federal workforce. To that end, we undertook two research activities aimed at gathering information about the use of new data sources, called both "big data" and "alternative data" in the literature. First, we performed a literature review that focused on recent work in the area with a concentration on the opportunities and challenges of "big data" sources for policy research. We then conducted a series of focused interviews with experts in the field of data science and data management, many of whom have a special concentration in health and health care research. Details about the interview process are included in the body of the report.

The literature review and interviews provided interesting insights regarding the uses of alternative data sources to support federal policy research and decision-making. Our observations fall into four categories. First, we identify areas where new data sources are likely to provide the most powerful and useful replacement or supplement to existing data and measurement. Second, we describe the importance of data quality assessment as well as particular issues affecting non-traditional data sources. Many of the areas we highlight already have active and ongoing investment by HHS agencies, and we would like to reinforce their importance. Third, issues of data governance, which include access to privately held data and privacy protection, have always been central to the federal government's approach to data policy and are of even greater importance with these new data sources. Finally, we identified training and infrastructure development models to help the federal government take advantage of new technologies and data.

Our cumulative evidence suggests that the most fruitful use of new data sources lies in four main areas: (1) improved measurement, (2) more granularity by time, geography, and outcomes, (3) better and more robust prediction models, and (4) more accurate data structures. It then makes sense to invest in data systems that are used in related areas, such as disease and impact surveillance, measurement of physical activity and diet, measurement of income and health care utilization, economic and social forecasting, and studies of social network impact. We note the current very successful integration of administrative records into data systems that measure health care utilization and costs, the use of user-reported adverse drug events, and the real time monitoring of traffic flows as examples of these activities.

Our second set of observations suggest that an investment in data quality standards that are specific to new data sources are an important next step in the maturation process for these data types. The federal data infrastructure has traditionally set very clear and systematic guidelines on data quality requirements for many of its data systems. Our experts were quite clear that for many of these new data types, it will be important to invest in data quality standard setting before full-scale integration of new data sources can

occur. The Common Data Model currently being pursued both at National Institutes of Health and the Office of the National Coordinator for Health Information Technology is an example of one such exercise.

Third, it is clear that the data governance issues associated with new data sources are complicated in ways that require careful and deliberate thought and guidance. One of the critical issues regarding data access will be to identify the next steps in developing public-private partnerships. Our experts provided us with a variety of hopeful examples of effective and productive partnerships. One area of particular promise lay in computing infrastructure. Large technology companies with cloud computing infrastructure have shown a substantial and ongoing willingness to share this infrastructure to facilitate the high dimensional computing necessary to analyze many of the new data sources. Additionally, our experts provided resource sharing and data sharing examples where there was mutual benefit either because of shared incentives or because the analysis of the shared data yielded substantial benefit to private companies. A systematic review of HHS public-private data partnerships would also be a useful effort to identify the successful partnerships and how they arose. This may provide a roadmap to forging those partnerships in the future.

Data privacy and individuals' rights to privacy are still seen as a substantial issue at the core of data access and data integration. There are legal, ethical, and technical issues that will remain part of the dialogue for many years. Our experts agree that data privacy requires new solutions and offered some interesting suggestions including honest brokers, individual data archives, and blockchain technology. At the center of their solutions is the notion of giving data rights back to the individuals who produce the data.

Finally, we reviewed workforce training models and asked our experts to identify what types of skills the next generation of federal employees and analysts will need to take advantage of new data types. One of the most interesting and productive models is the in-house training institutes. The computational and computing challenges that initially required highly trained computer scientists and data scientists are rapidly disappearing. The challenge is to train current staff to understand the nature of the data and their uses. Further, it is critical to put together effective teams with substantive, data, and statistical expertise. As data scientists move through the educational pipelines that provide public policy analysts and decision-makers, these in-house training efforts for federal employees can expedite the transition.

This project has been an opportunity to address selected issues in the integration of new types of data into the data infrastructure that supports federal policy evaluation and decision-making. The breadth of opportunities these new data sources provide for analysts to generate timelier, more accurate, and robust evidence to support the process of federal decision-making should not be understated. It is, in fact, already occurring. The challenges moving forward are to make selective investments where these types of data can have the most value and to begin the careful and painstaking processes of establishing clear data quality standards for data used in support of public decision-making. Additionally, the development of data governance models for gaining access to important privately held data requires great care as well.

Introduction

Overview of Project

This project consisted of two main tasks: 1) a literature review and 2) interviews with experts in the fields of data science and data use. These tasks were completed in consecutive order, with the literature review providing the foundation to inform the strategy for the expert interviews. In this section, we provide an overview of both tasks.

Literature Review

For the first task of this project, NORC conducted a literature review of uses of big data relevant to ASPE and HHS policy areas to evaluate its use in future research and to identify opportunities for methods of integrating new data sources. The literature review first discussed different uses of big data sources as they relate to health policy analysis and then focused on three data types: data maintained by the public sector, data that come from one or more private sector organizations (a category in which we include combined public and private data), and user-generated data. The definitions of "big data" are varied and constituted an important part of this literature review. We used the term "health policy" to refer to both what is more commonly known as "public health policy" (such as population-level issues like disease burden, prevalence, and risk factors) and "health policy" (such as health care financing and delivery issues). The review then focused on insights from the literature on the technological skills and capabilities needed to manage big data. One important aspect of this review is that many of the data sources reviewed are currently in use by ASPE and HHS, but this review highlighted a few relevant and important use cases to illustrate the promises and challenges of big data sources and to inform future health policy analysis. The report demonstrated that understanding the quality of a big data sources is critical to the successful application of the data to support statistical inferences and to provide insight for policy.

Subject Matter Expert Interviews

For the second task of this project, NORC conducted interviews with data science experts in order to supplement the literature review findings. Together NORC and ASPE identified a diverse pool of big data experts, with backgrounds across academia, government, and industry, who were then recruited by NORC to participate in phone interviews. These interviews focused on a variety of topics agreed upon by ASPE and NORC, focusing on topics from the experts' professional backgrounds to their perspectives of big data in key areas of interest to this project. An interview protocol was developed that was approved by the Office of Management and Budget (OMB). At the completion of each interview, NORC summarized the results and provided an interview summary to both ASPE and the expert. The expert interviews provided the project team with a current view of the complexities of the rapid change in the types and uses of data that are occurring in all areas of research, surveillance, and policy analysis. These interviews illustrated the rapid transformation of the data landscape.

Define Big Data

In our literature review, we found that "big data" is a term that is used with an array of definitions, which are often inconsistent and unfocused. This is a theme our experts also underscored as they had a diversity of opinions about the definitions of big data that often differed sharply from one another. To simplify and reflect the array of definitions of interest to ASPE, our literature review defined big data broadly by using Groves's (2011) definition of "organic data," *data that are not collected for statistical purposes*, but are nevertheless used for description, estimation, or analysis for populations or groups. We chose this definition as it recognizes the challenges when data are not collected for statistical purposes. This is in contrast to "designed" data sources, such as surveys, censuses, or randomized experiments whose sole purpose is for statistical inference. Additionally, this definition of big data focuses on the benefits and challenges that arise from how the data were created. Repurposing data, which is the essence of the application of big data to analytic problems, requires understanding the origins of the data and the intrinsic consequences. Upon discussions with ASPE on the breadth of data sources of interest for the project, in the scope of big data, we included administrative data from managing government programs, in addition to any data that are not from designed data sources. This is consistent with Groves's (2011) description, as administrative data are designed for operational purposes and must be reconfigured in many cases to be of use for analysis. Administrative data of all types, including that collected for monitoring federal programs, fall under the definitional rubric of organic data and are defined by many as big data because the data are being used in research and for policy analysis even though they are not collected for that purpose.

Our expert interviews brought further insight and variety to our definition of big data. Each expert had their own perceptions of big data, and most refined the definition by sharing their thoughts on what data sources count as big data. Each definition was nuanced in part by the expert's own work and intellectual history, however, their definitions did share some similarities that are inclusive of the "found and organic" data, descriptors, technology, and data integration.

Characteristics of Big Data

From our literature review, big data are described as having characteristics encompassed by the four V's (Beyer 2011, Japec et al. 2015, NIST 2017): volume, velocity, variety, and veracity. The four V's were chosen in our review, as they help organize both the opportunities and challenges of using big data for analytic purposes. The volume, velocity, and variety of some big data sources suggest how data sources may mitigate weaknesses of and/or add value to surveys and censuses. Veracity, which can be characterized as the quality of measurement, is also an important descriptor and may either be a benefit or a drawback of a big data source. On the one hand, big data sources may not take the same level of care in measurement that "designed" data sources do, which can lead to substantial measurement error. Alternatively, big data may more closely approach the "ground truth" of a phenomena, thus avoiding many of the reporting errors common to surveys.

Our second theme in the literature is that the origin of big data is critical to our evaluation of data quality. Our literature review discussed two distinct pathways by which big data are produced. These include

through "digital life" versus "digital trace," which have very different implications for how the data source can be used and evaluated (Lazer and Radford 2017). The definitions are below:

- **Digital trace**: Data that constitute recordkeeping or chronicling of actions at one or more organizations. Both the public and private sectors produce digital traces. These represent records of actions, but not the actions themselves.
- **Digital life**: Data reflecting a direct action by a user. This often reflects the use of online platforms, including social media. Data from health trackers, like Fitbit, would be another example.

In our review, we discuss that big data sources can usually be classified as either digital trace or digital life and that the difference between the two is critical for assessing the quality of the data and determining how data can be used for decision-making.

Importance of Data for HHS

The motivation for investigating the use of big data is to improve existing data products and expand evidence-building at HHS. Based on guidance from The President's Management Agenda, OMB, and the HHS Strategic Plan (FY 2018-2022), using existing data for analysis and evidence-building are critical to the achievement of the federal government's strategic goals. Furthermore, using data sources for statistics and evidence-building has proven to be of interest across the federal government. One recent example is the Foundations for Evidence-Based Policymaking Act passed by the U.S. House of Representatives in 2017 to support increased use of alternative data sources, including administrative records for policy analysis. Additionally, the results of two important panels at the National Academy of Sciences detailing the future of data collection and analysis in the federal statistical system were published in the fall of 2017.

Through our expert interviews we found that there is also substantial potential for big data to add value in many areas where HHS conducts evaluations and engages in decision-making, as the data can provide both variables and data structures that improve research, including through measurement, the predictive accuracy of statistical models, the timeliness of measures, and more granular inference resulting from data volume. There are still some notable challenges that remain, however, due to the volume of data produced and challenges with representativeness and data quality. The value of big data is quite evident both in current analysis and research conducted by the government and HHS, but further opportunities are swiftly presenting themselves. The striking challenges of data quality, computational requirements, and statistical integrity also remain and are the subject of this report.

Methods

For the two main tasks of this project, the literature review and the expert interviews, NORC and ASPE agreed upon methods for each task to ensure both were focused on achieving the main goals of this project. The following sections discuss the methods used for each task.

Literature Review

The literature review (see Appendix A) was conducted in four main steps: 1) define goals and key search parameters, 2) conduct the search, 3) evaluate relevancy and selected articles, and 4) review articles and lessons learned.

The goals of this review were discussed to identify areas of focus, including but not limited to: 1) technical issues with big data, 2) how to access big data and policy issues of interest to HHS, 3) workforce needs to support the use of big data, and 4) methods of evaluating the quality of big data sources. With these areas of focus in mind, NORC submitted a literature review outline to ASPE before conducting the search.

After identifying the goals of this review, NORC used its expertise to conduct a search for articles in both the published and grey literature; identifying and reviewing a set of papers that provided either an overview or research application with big data for statistical purposes and decision-making. We then reviewed the results of the search and selected articles related to the key issues. To facilitate the review of the selected articles, NORC prepared a database inclusive of the research topics and data types used in the article with fields to record the citation information. Each article was then reviewed and the requisite information entered into the database from the article for any of the topics or questions it addressed. Recording information in this database allowed NORC to easily summarize what was learned.

Finally, our review highlighted important use cases to represent the promises and challenges of big data sources. Use cases were defined as successful examples of the big data applications used in areas of interest to public policy. While the literature review could not showcase all use cases, or even all successful ones, those selected represent both successes and promising developments for which some questions remain. NORC also chose use cases that illustrated different data types and different kinds of uses and challenges emerging form the literature.

Expert Interviews

After the completion of the literature review, NORC shifted focus to the expert interviews. The process for conducting these interviews was developed into four main steps: 1) develop a list of experts, 2) prepare interview materials, 3) conduct the interviews, and 4) summarize interview data.

Develop Expert List

The first step was to develop a formal list of experts through an iterative process. We began this step by creating a list that consisted of academic, government, and industry experts, with the goal of creating a comprehensive and diverse list to yield an adequate number of interviews in each organizational area. The initial list was developed based on internal recommendations at NORC. This initial list was presented to ASPE and was further refined based on recommendations from the Contracting Officer's Representative (COR).

Once a final, comprehensive list (see Appendix B) was agreed upon, NORC and ASPE prioritized experts for interviewing by creating primary and secondary lists. Throughout the interview process, further iterations of this list were created. The primary list consisted of experts preferred for interviewing, while the secondary list was used to provide backups should a primary expert decline or not respond to the interview invitation. As the interviews began, new experts were added based on recommendations from our experts during their interviews.

Preparation of Interview Materials

As the list of experts was compiled, interview materials were prepared that consisted of recruitment materials, interview protocol, and data collection rubrics. The recruitment materials included an initial invitation email, follow-up invitation email (both a second and third follow-up email), interview time poll, and an interview reminder email.

NORC then developed an interview protocol (see Appendix C) that was used as a guideline for each interview. The questions in this protocol were developed from lessons learned from the literature review, which assisted us in identifying key issues related to big data that would benefit from input from the experts. Next, NORC reviewed the project's research questions and goals and solicited input from the COR about key issues. The final interview protocol focused on the topics of the expert's professional background, their current organization, their interests in big data, the importance of big data in research, the evolution of big data, big data used in their organization, experience with data quality, and their definitions of big data.

While developing the interview protocol, NORC also developed data collection rubrics. These rubrics were used to supplement the interview protocol to ensure that information was collected systematically and efficiently. The information in these rubrics was reviewed in order to identify patterns of response and any anomalous responses.

Finally, the project team submitted packages for both Internal Review Board (IRB) review and OMB clearance. NORC submitted an IRB package for review by NORC's IRB team and received a "Determination of Not Human Subjects Research" in December 2017. An information collection request was also submitted and approved by the OMB in April 2018 (OMB Control Number 0990-0421).

Conduct Expert Interviews

NORC developed a schedule for conducting the expert interviews in order to achieve the 21 to 25 interviews in a four-month window. Interviews took place across four batches, beginning in May 2018 through August 2018, with the goal of conducting approximately six interviews in each batch. Each batch consisted of three weeks of interviews, followed by one week of documentation catch-up, during which NORC finalized the interview documentation materials and reviewed the findings.

The interviews were one hour in length with one interviewer, one interviewee, three note takers, and ASPE team members attending when available. Ultimately, 18 one-hour expert interviews were conducted in the four batches (see Appendix D). There were three experts from government, nine experts from academia, and six experts from industry. Seventeen of 18 interviews were conducted via conference call, while one interview was conducted in person at the NORC downtown Chicago office. All interviews were recorded for note-taking purposes, and those recordings were made available for ASPE staff to review. While the project team fell slightly short of the interview goal, NORC and ASPE agreed the interviews were comprehensive and provided substantial insight into the topics of interest. The NORC project team and ASPE project representatives agreed that additional interviews were likely to elicit similar views and that additional interviews were unnecessary.

Summaries of Interviews

After conducting each interview, and after the conclusion of each interview batch, NORC summarized the information collected. Note takers reviewed the interview recording and updated information they recorded in the data collection rubric. This information was then used to write a case summary for each interview. These case summaries included a detailed description of the interview findings and issues that emerged, in addition to basic information about the interview.

NORC used the case summaries to develop key takeaways and conclusions from each interview. NORC further compared and contrasted interviews in the current batch and compared conclusions to those from previous batches. At the end of the documentation week, NORC finalized the interview case summaries and sent the summaries to both ASPE and the expert for review. NORC provided both ASPE and the expert a response rubric to fill out for each summary. This was an opportunity for both ASPE and the expert to provide feedback and to ensure our records of the expert's views were accurate. Additionally, this check allowed interviewees time to reflect on the information they shared and add to their original comments, improving the accuracy and completeness of the information collected. If an expert or ASPE requested an update, NORC updated the summary accordingly (see Appendix E for all individual summaries).

Summary of Literature Review

Literature Review Strategy

For the literature review, NORC first reviewed a set of published articles and some grey literature providing overviews regarding big data for statistical uses and decision-making. NORC aimed to identify papers published or released since 2015 that use big data sources for policy analysis or population studies. Further details of the literature review are available in the full report in Appendix A. NORC used its expertise to identify papers that demonstrated the most promising, successful uses of big data as well as a set of uses that could demonstrate the breadth of data types, benefits, and challenges of using big data sources. Table 1 presents a list of many of the search terms used in the literature review.

Table 1. Partial List of Search Terms Used for Literature Review

Health Policy Topics	Alternative Data and Data Science	Data Types
Population health	Big data	Health administrative data
Public health	Large data	Medicare enrollment
Health care quality, access, evaluation	Data science	Medicaid enrollment
Preventive health services	Data quality	Insurance claims
Demography	Data collection	Immunization registry
Health planning	Methods	Electronic health records
Health expenditures	Analytics	Electronic medical records
Health services	Surveillance	E-pharmacy
Health status indicators	Early warning	Surescripts
Social determinants of health		Consumer purchase data
Population characteristics		Environmental monitor
Social environment		Health monitor
Health services accessibility		Mobile phone
Health disparities		GPS
Urban health		Patient-generated health data
Rural health		Sensors
		Wearable technology

Background – Issues with Alternative Data Sources

One of the important aspects of assessing any new data source for uses beyond their original intent is to assess whether the data are of adequate quality for the proposed research or evaluation activity. New data sources as described under the rubric "big data" are largely not designed for use in many of the policy-related research activities of interest to HHS. One of the critical steps in understanding the suitability of

data for any research activity is to assess the potential sources of inferential error that may either be intrinsic to the data type or an outcome of using data not designed for the proposed activity. Data quality was one of the major foci of this literature review as a consequence. An important obstacle to the introduction of new data sources to existing research programs is uncertainties about data quality.

Data Quality

Data quality is multidimensional, with elements reflecting different aspects needed to support valid statistical inferences. Data quality assessment is also heavily dependent on the context in which the data are used. For example, predictive models have fundamentally different data quality requirements than statistical estimates. In this section, we present a framework to understand the quality of a big data source and its strengths and weaknesses. Table 2 describes different aspects of data quality needed of either traditional or big data sources to support policy analysis, grouped into five categories.[1] This data quality framework has been proposed by agencies in the federal statistical system, as noted in the National Academy of Sciences reports, as a standard against which data for use in federal policy evaluation and decision-making should be evaluated.

Table 2. Data Quality Framework for Assessing Data Sources' Fitness for Policy Research

Data Quality Aspect	Description
Accuracy	Data values reflect their true values (low measurement error). Data are processed correctly (low processing error). Concept measured is concept of interest (construct validity). Data are representative of population (external validity).
Relevance	Data meet requirements of users to study topic of interest.
Timeliness	Data available when expected and in time for policy analysis purposes.
Accessibility	Data can be readily obtained and analyzed by users.
Clarity	Data and statistics presented in clear, understandable format.
Transparency	Methodologies for data preparation and statistical processes available.
Coherence	Enough metadata is available to understand data structure and allow for combination with other statistical information.
Comparability	Data over time and from different records or sets of records reflect the same concept.

In addition to data quality, users should also thoroughly examine big data sources for data accuracy (Biemer 2016), statistical validity (Shadish, Cook, Campbell 2002), and fitness for use. Big data hubris is especially prevalent when working with large data. Lazer et al. (2014) addresses big data hubris as a mistaken belief that the volume of data compensates for any deficiencies in data quality.

Government agencies developing official statistics and formulating and evaluating policy depend on their reputation as fair, impartial, objective, and neutral (Kitchin 2015). If an agency cannot provide data or analysis with adequate data quality and documentation, then the public can lose confidence in the

[1] See Hansen et al. (2010), Japec et al. (2015), National Academies (2017a).

agency's ability to provide evidence for decision-making. Thus agencies must be risk averse in adopting new methods or working with new datasets as the understanding of the issues with certain big data sources is still evolving.

Cleaning/Business/Technological Issues

In addition to the inferential challenges of analyzing big data sources and the extensive data cleaning that can be required, there are important issues involved with analyzing big data sources in a world with rapidly changing technology.

First, it is critical for organization(s) maintaining big data sources to provide users with transparent and detailed information on data maintenance needed to understand data quality. Otherwise, the comparability of the data becomes compromised when organizations adjust their data maintenance processes over time to fit business needs. Standardization can also be critical for making big data sources usable for health policy research by improving data quality. When data are combined from multiple organizations, standards help improve the comparability and coherence of the ultimate data product. Standardization can also guide what checks are needed to verify the processes and quality of a source. These principles of data management are consistent across types of big data and include administrative data, even though there can be substantial variation in data standards and documentation.

Additional challenges with using big data sources can arise when the evolution of users' inputs to the data production system are not understood. For example, many social media sites rapidly change their algorithms and the way users interact with social media platforms. Because these processes are often undocumented and opaque, understanding the data quality of such sources is particularly challenging. Such algorithm dynamics (Japec et al. 2015, Biemer 2016) affect several kinds of big data sources (e.g., search queries) and greatly harm statistical validity.

Violation of "ideal user assumption" (Lazer and Radford 2017) can also lead to problems. In traditional data sources, such as survey data, records mostly reflect single, unique people who express themselves honestly in the data. However, in more novel data sources, such as social media sites, data may not accurately reflect the actions of real users. For example, users can easily misrepresent their identity or bots may generate online traffic that is not an accurate representation of individual activities or views.

Findings and Use Cases by Data Type

Different kinds of alternative data sources have different strengths and face different challenges. They also have different readiness levels for statistical uses and decision-making. Thus, we have grouped our review into three categories, as identified in Table 3: 1) data maintained by the public sector, including administrative records, 2) data coming from one or more private sector organizations, as well as combined public and private sector data, and 3) data that are user-generated. While HHS is already engaging with various types of data across these three categories, we review current such use cases to inform further future investigations.

Table 3. Description of Three Types of Alternative Data Sources

	Administrative Records	Private Sector Data	User-Generated Data
Examples	Medicare and Medicaid enrollment Insurance claims State registries	Electronic health/medical records Insurance claims E-prescription data Consumer purchase data	Social media Environmental and health sensors Mobile phone data/GPS
Veracity	Higher	←	Lower
Digital Trace/Life	Digital trace	Digital trace	Digital life
Maturity of Data Standards	Proven history of successful use Data quality framework from many statistical agencies	Some successful use cases enabled by HL7 Standards and Common Data Model	Proof-of-concept studies New measures of data quality emerging
Characteristics	Primary data source used in conjunction with censuses and surveys Large government efforts to: ■ improve quality ■ harmonize ■ link ■ disseminate	Often in data siloes (e.g., hospitals) Vary in structure and complexity Public and private partnerships are underway to: ■ standardize ■ share technology ■ integrate data platforms	Nonrepresentative Lack of metadata Technological challenges: ■ algorithm dynamics ■ violation of ideal user assumption
Common Uses	Direct estimation Design and calibration of surveys Imputation Record linkage Second survey frame	Some direct estimation Monitoring Surveillance	Monitoring Surveillance Communication

Public Sector

Public sector sources of big data have long been used alone or in conjunction with sample surveys to support policy analysis and evaluation. Administrative data are one of the most critical and promising sources of public sector big data for policy analysis. This is because there has already been a great deal of research and collaboration in developing administrative data as a tool for decision-making. Further, administrative data have also become more robust and timely in recent years due to increased automation, improved data quality checks, and harmonization efforts. While administrative data are originally collected for administrative, regulatory, and law enforcement functions, they can be and are subsequently used for numerous statistical purposes.

Administrative data come from various sources, such as transactions, registries, and federal programs. Transactional data, such as Medicare enrollment and claims, have long been analyzed to understand health care among the elderly. Other transactional data include Social Security earnings and benefits to assess work history and well-being, and voluntary reporting from police departments to understand crime and victimization. Another form of administrative data is registries, which may arise from mandatory

reporting requirements for notifiable diseases such as HIV (e.g., Enhanced HIV/AIDS Reporting System) or voluntary state reports of childhood immunization (e.g., Immunization Information Systems). Federal programs are also a source of administrative data. Examples of such programs include Temporary Assistance for Needy Families and Medicaid.

Federal statistical and regulatory agencies can use administrative data as full replacements for survey data when administrative records fully capture population characteristics. However, the most successful uses of administrative data have been cases when administrative data were used to complement survey data. Such complementary uses of administrative data include data validation, sampling frames, supplemental information on missing and nonresponse data, data linkage, and production of blended statistics.

Two current data systems used for monitoring public health and health care utilization provide an interesting contrast in data linkage and its consequences for inferential quality. Both the National Health Interview Survey (NHIS) linked files and the Medicare Current Beneficiary Survey (MCBS) provide information about health status, health care use, and barriers to care through a combination of in-person survey questionnaires and linked Medicare enrollment and claims data.

The NHIS is a post-hoc linkage of a household survey to administrative records, while the MCBS is a survey designed with data linkage in mind, where the sampling frame is the enrollment data from the Medicare program. The NHIS-Medicare linkage supplements an individual respondent record with information from claims and enrollment files. Additional measures are added to a selection of records that are capable of being linked. Conversely, MCBS adds individual items from claims and health insurance plan enrollment to improve and correct items collected from an individual. The administrative records serve as verification and correction of data collected from respondents.

Private Sector

Private sector data, with some exceptions, traditionally have not been used in research and evaluation in the public sector. This can be attributed to the variety and complexity of privately held data that prevent easy summary or assessment of their overall usefulness for decision-making. Privately held data are more often utilized outside the United States, as other countries have relatively greater access to private sources. Statistics Netherlands, for instance, has organized and captured traffic sensor data, which has become ubiquitous enough to produce national estimates of traffic flow (Puts et al. 2016). Efforts are also underway to generate Consumer Price Indices (CPI) to assess inflation in 22 countries, using daily web-scraped prices for 5 million items to track price shifts. These statistics are being considered as a source of national CPI by many statistical agencies worldwide.

Classifications of private data are helpful to understand some of the quality challenges that may limit their use for policy analysis. Depending on the degree to which the data are structured, standardized, and uniform in nature, data sources can be classified as structured, semi-structured, or unstructured.[2] Structured data share common fields with defined lengths and known characteristics. Such data include

[2] National Academy of Sciences: *Innovations in Federal Statistics: Combining Data Sources while Protecting Privacy* (2017), *Federal Statistics, Multiple Data Sources, and Privacy Protection Next Steps* (2017).

sales data from retail transactions, which are structured by the Universal Product Code, and residential real estate information available from sites such as Zillow, which are structured by the Multiple Listing Service legal requirements for real estate transactions.

Semi-structured data lack the implicit shared organizational structure but coexist with metadata or business rules that can be used to process the data. Twitter data, for instance, is semi-structured in that there are metadata fields such as time, date, and hashtags that can be used to provide a method for structuring contents and fields. Finally, unstructured data—such as videos, pictures, or unstructured text on social media—do not share a common set of characteristics. Structuring the data, then, becomes both an exercise in regularizing data for analysis, but also in identifying shared structures and building data standards. Not surprisingly, most efforts to integrate big data into ongoing data systems in the federal government focus primarily on structured and semi-structured data with agreed upon standards.

In the United States, statistical agencies such as the Bureau of Justice Statistics (BJS) and the Bureau of Labor Statistics (BLS) are experimenting with new sources of private data to augment existing statistics. BJS began to use web-scraped news article in the redesign of the 2015-2016 Census of Arrest Related Deaths (Banks, Ruddle, and Kennedy 2016). BLS uses data from retail scanners, web-based price scraping, and JD Power car prices to adjust and calculate the CPI (Horrigan 2013).

One prominent example of a successful use of private data in health policy analysis is the Health Care Cost and Utilization Project (HCUP). HCUP is a long-term successful collaboration between the Agency for Health Care Research and Quality (AHRQ), states, private organizations, and hospitals. AHRQ provided the conduit and resources for the collaboration between private and public health care institutions and partnerships. While the final data system is provided for use by a federal agency, the original data sources in many cases were private health care systems. HCUP provides individual-level encounter data from hospitals in 48 states and the District of Columbia based on hospital discharge records. The current Nationwide Inpatient Sample (NIS), which is an annual representative sample, includes over 7 million hospital discharges with data on diagnoses and discharges. The data can be used to make national, regional, and local estimates of hospital costs and services. HCUP has spurred large numbers of policy and policy-related research, with over 500 research articles published on the NIS sample alone by 2016. Further, AHRQ issues annual data summaries from the datafiles that describe national trends in hospitalization rates, treatment costs, and readmissions.

User-Generated Data

User-generated data can be defined as data reflecting direct user interactions with a website, platform, product, device, or service. A diverse set of data types falls in this category: social media, data produced by mobile phones, reports on online message boards, data collected by web scraping, data from environmental and health sensors, data produced by the Internet of Things, and many others.

Much of this category of data types encompasses data resulting from online interactions. In general, the data can have both high volume and velocity. The volume of data may allow for monitoring trends in different geographic areas more easily than surveys or censuses. The velocity of data may allow for more timely and affordable collection of user-generated data.

However, due to some substantial challenges, there are fewer mature uses of user-generated data for policy analysis. The veracity of user-generated data can be questionable and difficult to ascertain. This is because users of a service or website are often not representative of a population of interest, and because datasets may have coverage error. In addition, user-generated data may be the most affected by technological challenges, including algorithm dynamics, lack of metadata, and violation of ideal user assumption. Further, the standards for using and analyzing user-generated data are not as mature as for the other two data types.

Many of the most promising use cases for user-generated data are for surveillance and monitoring. Two examples demonstrate emerging uses of user-generated data: the use of mobile phone, GPS, and crowdsourced data for syndromic surveillance; and the use of social media data for adverse drug event monitoring. First, Boston Children's Hospital's Computational Epidemiology Group developed HealthMap (Brownstein et al. 2008) to support applications for monitoring and surveillance of disease outbreaks and emerging public health threats. HealthMap's applications primarily use algorithms to accumulate web-accessible information: news aggregators, eyewitness reports, expert-curated discussions, and validated official reports. The algorithms pull data from these sources through an automated process, constantly updating the system. HealthMap's apps are used by public health departments and government agencies, including the Centers for Disease Control and Prevention, Department of Defense, and World Health Organization.

Second, text mining of social media data can be a promising tool for monitoring adverse drug reactions (ADR) and related events, but, like other uses of user-generated data, is also subject to some substantial challenges. Researchers have recognized the potential for big data sources to enhance pharmacovigilance, including the development of the Federal Drug Administration's (FDA) Sentinel Initiative. The recognition of social media platforms as a place where people may share possible ADRs led to investigating the possibility of text mining social media data for digital pharmacovigilance. Freifeld et al. (2014) studied 6.9 million tweets from Twitter and, using a combination of manual and semi-automated techniques, found 4,401 possible ADRs. Although assessing the validity of the findings was difficult, the researchers compared their findings to those from the FDA Adverse Event Reporting System and found similarities in patterns between the two data sources.

Summary of Interview Data

Introduction

In this section of the report, we provide a thematic summary of our interviews with 18 experts drawn from academia, industry, and government. By and large, these experts were selected both because of the diversity of their technical and analytic backgrounds and because of their long careers in research, data analytics, or data policy. One exception is our early career expert, whose path through the public sector, technology startup, hospital systems, and insurance providers provided us with an important view of the analytic and data issues from the vantage point of a young analyst and entrepreneur. The interviews were focused on a series of interrelated questions designed to elicit both the experts' career and work with alternative data sources and also their perceptions of the opportunities and challenges associated with the substantial shift in the data paradigm that has occurred over the last 20 years. Several of our experts have been at the forefront of federal data policy to promote the use of new data and techniques in the research and policy arena while others are actively engaged in the integration of new data sources in the private sector. Several of our experts have also been instrumental in guiding federal initiatives on data collection and estimation.

The summaries follow a format that identifies the area of interest for the project, a thematic summary of the views expressed, and a short description of the overall conclusion. The areas of summary were chosen to reflect major topic areas that emerged from the literature review and were consistent with our discussion guide. This allows readers to understand our intent in focusing on this topic area and to identify quickly the areas of consensus and departure. Overall, all of our experts agree that there is no longer a question of whether alternative data sources should be used in public decision-making and policy evaluation. This is already occurring, and the question for public institutions now is how to define, shape, and assess the use of data sources that may depart from the data quality and structures that are familiar to policymakers and analysts. The academic and industry experts in the group were particularly aware that the next step is to bring knowledge and processes already in place in their wider communities to the public sector. Our experts from the federal sector or who support federal decision-making provided us with a variety of frameworks for moving forward, particularly in the area of data quality assessment and workforce training.

This summary by definition cannot capture both the complexity and heterogeneity of the views of these experts. We have attached the individual summaries and delivered audio transcripts of the interviews for use by ASPE. Both resources provide insights well beyond the summary below. We have organized the summary around themes identified in the literature review in order to provide a systematic integration of the themes of our investigation. These topics include our experts' opinions about the uses and definitions of big data; the important challenges of integrating alternative data sources into federal decision-making, including issues of standardization and data quality; successful examples of the use of other sources of data in decision-making in the public sector; the experts' perceptions of private-public partnerships; data privacy; and training models for federal workers. We attempted to represent both areas where there was agreement among our experts and those where other perspectives were voiced. Because our experts

included a broad range of data types in their discussion of big data, we have broadened the big data rubric to "alternative data sources." Federal administrative data sources are included in this definition as many of the experts suggested that, while these data have been used in federal decision-making in the past, new methods and uses of the data have changed how these data sources are evaluated.

Brief Summary of Subject Matter Expert Backgrounds

The diversity of perspectives, research experience, and backgrounds of these experts provided us with important diversity in perspectives. In addition to their current employment areas and research interests, their training and experiences shaped their views on the substantial shift in data sources and their use. Nine of our experts are currently in academic institutions, although both O'Hara and Bourne were very recently in the Census Bureau and the National Institutes of Health (NIH), respectively. Three of the experts are currently in the employ of the federal government, although Baru retains his appointment at the San Diego Supercomputing Center, and Chen has worked for the city of New York, among other employers. Six of our experts are currently employed in the private sector, with three in health care analytics, health care systems, and insurers (Reti, Savitz, and Yeung), while the others work for organizations that provide contractual support for the federal government (Biemer, Klenk, and Mundada). They are similar to other experts in that they have moved between academia, the private sector, and an affiliation with the federal government. For example, Klenk is now working on projects to support the Data Commons Pilot at NIH, and Biemer is one of the founding directors of the Joint Program of Survey Methodology (JPSM) that has provided technical and statistical training for the federal workforce for nearly 30 years.

The diversity in their training and areas of expertise also provides an interesting method for understanding their perspectives on the definition of alternative data sources and their value. Gao, Borgman, Catlett, Osgood, Chen, Baru, Klenk, and Mundada are trained as computer scientists or computer engineers with particular foci. Their career paths have been diverse, but their focus has primarily been on the computational and technical development of data. Gao and Osgood expanded their original concentration in computer engineering to a focus on bioinformatics and health data modeling. Catlett, Mundada, and Baru have been thinking and working on data systems and integration across disciplines and data sources for many years and offered interesting perspectives on systems integration and the development of large-scale computing systems. Some of our experts are trained as social scientists, including Lazer, Chernew, and O'Hara, and focus on the analytic challenges and opportunities of using alternative data to identify new methods of measurement and new research questions. Reti, Savitz, and Yeung are trained in health services research and have a clear focus on the use of data to improve patient care, population health, and health systems. Finally, the statisticians in the group, Gibbons, Biemer, and Eltinge, offered thoughtful and full perspectives on the challenges for statistical inference of alternative data. Biemer and Eltinge provided careful guidance on methods of assessing data quality.

Two of our experts, Borgman and Yeung, are unique in their perspectives and experience. Borgman views the development and curation of data as part of scholarly communication. Her perspective on data provenance and data management is unique in that she views it as inseparable from the development of a research problem. Yeung is unique in that he is quite early in his career but has been through the full arc

of private industry from a not-for-profit to a technology startup to a hospital system to an insurance company. He has a very in-depth view of the constraints of the current data environment in each of these settings. Since he remains embedded in the mechanics of engineering the data transitions our other experts discussed, Yeung gave us a clear view of the organizational and operational feasibility of many of the recommendations made by the other experts.

Reasons for Using Alternative Data Sources

Interest for Project

One of the primary goals of this project is to understand why new data sources provide so much additional utility for decision-making. Thus, we asked our experts about how they view these new data sources and what they see as the greatest areas of promise relative to traditional data sources such as surveys, censuses, and experimental data. Understanding the particular value of new data sources can guide which investments might be most promising. We use the terms "big data" and "alternative data sources" interchangeably because some of our experts had very strict definitions of "big" data, which is meant to be data large in volume only, while others extended big data into a metaphor for all data from sources not designed for statistical purposes. The definitional ambiguity intrinsic to this report is a consequence of a lack of consensus in the research and policy community.

Five Reasons to Use Alternative Data Sources

Our experts proposed a broad set of reasons for using alternative data sources, including to: 1) provide novel variables and data structures for research, 2) improve measurement, 3) improve the predictive accuracy of statistical models, 4) provide more timely measures, and 5) enable more granular inferences due to data volume. One expert, Borgman, is a skeptic of alternative data sources and advises caution in using data outside of the context in which they were created.

Areas of Promise for Alternative Data Sources

- Novel variables and data structures
- Measurement
- Improving prediction accuracy
- Timeliness of measures
- Supporting granularity inferences

A commonly held view is that alternative data sources provide new variables and data structures for research that are not possible to find in traditional data sources. Chernew, who researches health insurance design and payment models using electronic health records (EHR) and public and private insurance claims, emphasizes that above all else he looks to new data sources to provide the right variables for the research question. The integration of structured electronic medical records into analyses of claims data has extended the analytic and inferential framework available to health care researchers such as Chernew. Additionally, Gao described using data sources to examine patient experiences, through online provider ratings. Eltinge comes from the perspective of a government statistician seeking to use the best data available to agencies to provide statistical information for the public and stakeholders. He in particular discussed the government's long-standing use of data directly from firms and establishments to produce business statistics.

One theme among several of the experts was that the structure of alternative data sources can enable richer analysis of social networks. Both Gao and Lazer discussed that the emergence of social media data changed how social network analysis research is conducted, as these data inherently incorporate and represent network structure. The use of data integration to bring diverse data sources together to enhance research was a theme among multiple experts. Savitz described the integration of data sources at Kaiser Permanente's Center for Health Research as important for solving substantial questions in health research and for improving patient care. Bourne finds great value from data integration as well as reuse of data for new purposes in order to bring the right evidence to answer research questions.

A second theme is the promise of alternative data sources for improving measurement. This was discussed both in the context of integrating administrative records into the development of official statistics and in the use of personal tracking devices and mobile phone data in health research. In her prior work at the Census Bureau, O'Hara found that measures from administrative records, such as income measures from tax data, can provide more accurate information than surveys, which are subject to substantial reporting errors. Biemer argued that the level of error in surveys is high and, therefore, using non-survey data sources, such as administrative records, can reduce overall error despite the data quality challenges. For systems modeling to study population health and health policy, Osgood finds that data from health trackers and similar sources can provide more accurate information for research than surveys and, in addition, provide measures that are closer to the intended concept of interest. For example, the number of steps one takes in a week as measured by a tracker may be a more relevant measure of energy expenditure than a survey report of the number of times exercising per week. In Klenk's experience working in biomedical research, he shares Osgood's view that alternative data sources, particularly passive measures of activity and medical records, provide more accurate and granular measures of human behavior.

Multiple experts find value in alternative data to improve the predictive ability of statistical and machine learning models. Machine learning is an application of artificial intelligence (AI) that uses statistical modeling to improve data classification iteratively through use of data classification improvements in each round. Specifically, some suggested that because alternative data sources provide many more variables that are correlated with outcomes of interest, adding such variables to a statistical model substantially reduces the error of the model in predicting the outcome. Osgood uses the size and variety of big data to provide rich information for modeling dynamic systems. In particular, he discussed this in reference to an example of infectious disease outbreak. He finds that even for low-quality data, big data has enough resolution that can provide substantial signal, or correlation with the outcome, to inform his research and provide parameters for predictive outcomes and interventions. For conducting predictive modeling for economic statistics at the Bureau of Economic Analysis, Chen similarly finds that big data sources from third-party providers provide some signal that is correlated with the outcome of interest and improves the predictive ability of models. Reti's team at Optum Analytics is exploring integrating new data sources, such as third-party consumer purchases with EHR and claims data to improve predictive modeling. In all cases, the volume and variety inherent in the data, even given the poorer quality of the data, adds predictive capacity to the statistical models. While all noted the inherent difficulties in sorting out the statistical noise from the true signal, the experts agreed that the computational tradeoffs were worthwhile.

Two other advantages of alternative data sources noted by our experts were the ability of big data sources to provide a timely indication of an issue and the advantages of data granularity to support inferences for rare events. Klenk, in discussing the example of studying the opioid epidemic, finds that the volume and velocity of "big" sources can help provide a timely indication of where the epidemic might be getting worse to allow more rapid response. Chen described the advantages of the frequency of data from some big data sources to obtain a richer picture of changes over time. Gibbons sees some advantages from the volume of big data sources that provide enough data to study rare events to estimate variation in treatment response.

A Different Perspective

While 17 out of 18 experts see areas of promise for uses of alternative data sources, Borgman differed from others because of her strong skepticism of alternative data sources. Given her expertise in Information Studies, she argued that it is challenging to analyze data outside of the context in which they are created. Borgman's perspective is that without knowledge of the many details involved in collecting and creating a dataset, it is easy for a data user to have a false sense of the precision or accuracy of their inferences. For her, detailed knowledge of the origins of the data are not separable from the research process. Therefore, Borgman is skeptical of data reuse and of open data policies, a contrasting viewpoint from other experts, particularly Bourne who is a strong advocate of data reuse. This perspective is not inconsistent with the position that one of the primary difficulties with using alternative data sources is the lack of transparency in documentation.

Findings

There is alignment between the findings of the literature review and the interviews regarding benefits of using alternative data sources, although there is different emphasis on which benefits are more important. Most see some areas of benefit for alternative data sources, although the interview with Borgman demonstrates that some perceive great risk with analyzing data outside of the context in which they were created.

The most commonly cited benefit among the experts regarding alternative data sources was the value of novel information in the data for research. Experts emphasized not only the variables available in the data, but the research value in the structure of certain datasets, such as the social network information embedded in social media data. Both the literature review and the expert interviews found that there is great value from combing different types of data sources. In addition, both the literature review and the interviews found that alternative data sources, in some cases, can be used to improve or validate measurement. This is true both for administrative data and for sensor and tracker data for providing "ground truth" measures that are not subject to reporting errors.

The expert interviews placed more emphasis on the role of alternative data sources as covariates to improve predictive and statistical models. Also, similar to the literature review use, some experts found that the timeliness and granularity of these data were of primary value. For example, Gibbons offered the perspective that the benefit of granularity goes beyond small geographies and also pertains to inferences

regarding rare events and variation in treatment response. Chen additionally drew attention to the value of data frequency.

Expert Perspectives on Big Data

Interest for Project

To understand the trends the experts see emerging in new data sources, we asked each expert to provide a definition of big data. The responses are valuable both for guiding definitions of different categories of data sources and for understanding how data science has changed as new data has emerged. Many of our experts quite rightly noted that the term "big data," because of its widely varied use in both popular and academic literature, has lost much of its original meaning and has become a hackneyed and outdated phrase. We attempt nevertheless to capture their perceptions of how the data ecosystem for policy research has changed. Their responses include some of the original definitions of big data, which focus on the characteristics of the data, but extend to the uses of combined data sources or data sources not originally designed for statistical purposes.

On Views of the Term "Big Data"

When asked to provide a definition of big data, many of the experts noted that they avoid using the term due to its overuse. Nonetheless, all experts described how the concept is defined to them, particularly in light of how the use of data sources are evolving.

As we described, many have specific and clear definitions of which data sources count as big data and which do not. While all interviewed have expertise in alternative data sources, some suggested that they do not consider themselves as working with big data, given the prevailing descriptions of big data.

Four Categories of Definitions of Big Data

Four categories of definitions of big data were provided by the experts, including about 1) the origins of the data source, 2) descriptors applying to a data source such as the V's, 3) the technology needed to interact with the data, and 4) data integration and solving research problems.

Categories of Definitions of Big Data

- Data origin: "Organic" or "Designed"
- The descriptors: The V's
- Technology needed
- Integrating data to solve research problems

A few of the experts view the defining feature of big data to be their origin, in that the data collection was not designed to support statistical inferences but rather that the data are taken from a source and reused for statistical or research purposes. These are sometimes referred to as "found" or "organic" data sources. This aligns with the definition NORC used in conducting the literature review and provides a broad encompassing definition of big data. Eltinge, who has spent his career in the Federal Statistical System, also defines big data as found or organic data and cited Bob Groves and Mick Couper as the source of his definition. Baru described big data as found data but specified that big data are "digital exhaust" or by-products of online

activity. Gibbons described the defining feature of big data as observational rather than experimental, noting that this observational, non-experimental nature leads to traditional threats to causal inference.

Many of the experts defined big data in terms of descriptors of the data, including using the V's described by the literature review. Among the three V's, volume and variety were more commonly mentioned than velocity. Lazer suggested that big data are defined by their scale and the complexity of their formats. He used an example of medical image data that have both substantial volume and variety. Mundada and Gao applied fairly strict definitions and suggested that big data can only be defined if all three V's—volume, velocity, and variety—apply. Chernew was among those who do not view themselves as primarily working with big data, but defines big data as encompassing non-traditional formats and structures, such as text, image, and video data. Osgood and Klenk added veracity as a descriptor in that they believe that the non-traditional structures of big data provide better measures of human behavior to inform research and modeling. Biemer, with his interest in administrative records, views administrative data as non-traditional big data as he sees aspects of the V's applying. For example, administrative records, while not as large as traditional big data sources, are not small datasets.

Another common theme was that big data required substantial technology to interact with the data. This definition may relate to the descriptors of big data. For example, big data may be so large in volume that advanced computation is needed to manage the data. Still, experts differed on whether the descriptors and V's are the defining feature of big data or whether the technology itself makes the dataset non-traditional. Borgman suggested that when computation, rather than inspection, is needed to understand the data and turn it into evidence, then the data source can be classified as a new form of datafile.

Catlett, who manages the computational resources at Argonne National Laboratory and is now running the City of Chicago's Array of Things, describes big data as something that is defined by the computational resources and technology needed to manage the data. More specifically, he said that if one can buy the computing resources to manage a data source at an Apple Store, then the data source is not big data. Gao also extended his definition of big data beyond the descriptors to include the technology needed for the data. He cites the origin of big data as the introduction of Hadoop technology to manage large-scale databases. In the context of prior work at the Census Bureau, O'Hara also described big data in terms of the technology needed to manage the data. Yeung defined big data in terms of the technical infrastructure needed, in addition to the methods needed to manage the data. He specified Hadoop and Spark as examples of technology needed to manage big data. Chen defined big data both as whether the computer can handle the size of the data in memory and also in terms of the computational power needed for modeling. He specified that a "big data problem" may refer to the computation needed for modeling even when working with a dataset that may be large in volume.

A few of the experts view big data as primarily defined by data integration and their use to address research problems. Bourne views big data as being about discovering insights from the integration of data sources. To answer important research questions, there is a need to reuse other data available. Savitz also sees big data as problem-driven and as about the need to integrate data sources to find a solution. In his work at Optum Analytics, Reti views the big data movement as the use of new, non-traditional data sources to solve problems. For example, for research in health, one can go beyond EHR and claims data

and examine contacts between patients and providers, unstructured data in health records, and marketing data as future frontiers. All three recast the definitional issue as an opportunity in research design rather than as a characteristic of particular types of data.

Findings

The literature review uses an operational definition of big data as data not collected for statistical purposes. This definition was shared by three of the 18 experts. This definition is useful for providing a broad outline that can be applied to the array of data sources of interest to ASPE for this project. The literature review also provided the V's as descriptors of features of big data that make it unique. The most common definitions of big data among the experts related to either these data characteristics or to the technology needed to use the data. As noted, these definitions have some commonalities, including that advanced technology is needed to manage the data because of either high volume, high frequency, or in non-traditional data structures. Different experts had different views on which features or combination of features make a data source "big," but volume and variety were more commonly mentioned than velocity. A couple of experts added the veracity of big data in their definitions. Some also discussed that big data technology not only applies to the size of the dataset, but also to the methods applied. Finally, a few experts view the big data movement in terms of the need to use new data sources to solve research problems rather than being about the origins or descriptors of the data.

Many of the experts noted that they avoid using the term "big data" because it is now overused and often misconstrued. They all see changes in how new data sources present new opportunities for research as well as new challenges, but they apply different terminology to describe these changes. Most of our experts, in fact, no longer see the new or big data movement as distinct from existing research and data paradigms. They suggest, and we agree, that the big data movement has fully arrived and that they view businesses as now having a mature understanding of how to incorporate such data in their operations. They view other frontiers, such as AI, as the current challenge facing organizations. Nonetheless, as discussed in this report, the pace of adoption of alternative data sources varies among different organization types. Particular reasons for caution in government agencies and for researchers in adopting alternative data sources include ethical considerations to ensure that the analysis guides fair treatment, the issue of representativeness, and the requirements for transparency and access to the data used to make important decisions about government programs and policies. The challenges now have less to do with *whether* to adopt alternative data sources and AI for use in decision-making, but rather *how* to adopt these resources.

Statistical Modeling for "Big" Data

Interest for Project

To help guide data policy analysts on the use of new data sources, it is critical to understand the statistical modeling considerations that are needed to properly use non-traditional data to draw valid statistical conclusions. In discussing modeling with experts, NORC gauged what modeling methods are emerging for new data sources and what the best modeling practices are for large diverse datasets.

In this section, our experts were focused on data that is truly defined as "big," in that the models depend on data sources that are large in volume and velocity. This is a subgenre of data identified by our experts that truly adheres to the "big" data definition.

Expert Uses of Statistical Modeling

Experts shared their perspectives on what statistical and machine learning models they find useful for working with "big" data. There were interesting similarities and differences among the experts. Several of our experts see AI, which is broadly defined as the integration of learning, rational inference, reasoning, and perception into data analysis, as an emerging area that has the chance to change operational practice in business and possibly in government. Experts also expressed some cautionary viewpoints regarding the challenges for modeling with "big" data sources.

Our experts have diverse experiences with applying statistical and machine learning models to "big" data sources. One theme was the importance of dimensionality reduction when working with high-dimensional data with many variables. This was mentioned by Chen, Gibbons, and Yeung. Chen and Gibbons both advocate regularization techniques such as the LASSO and Ridge regression, while Yeung applies unsupervised learning and clustering approaches to reduce dimensionality before conducting analysis. Chen conducts extensive predictive modeling at the Bureau of Economic Analysis and makes substantial use of ensemble methods based on tree-based models, such as random forests, for prediction. Gibbons applies generalized linear mixed-effect models for health applications, including for longitudinal analysis. Yeung conducts text mining for unstructured data in health records and uses deep learning for such applications. He added that depending on the context, domain knowledge may or not be important for modeling—there are some situations where the models will provide valuable insight without the need for domain knowledge and others where domain knowledge is critical to the construction of the model. Osgood primarily focuses on simulating large-scale populations at the individual level using systems modeling and uses "big" data sources to calibrate his models.

A few experts discussed AI as an area of growth for acting on insights from "big" data. Mundada, in his leadership at HP HAVEn at Hewlett Packard, used AI for various clients. He views AI as an area that will grow in the future, particularly for the business sector. In Reti's work at Optum Analytics, he finds AI useful for automating some processes that are otherwise labor intensive, such as identifying text strings to conduct natural language processing for unstructured data. Catlett also viewed AI as intrinsic to the ongoing and useful applications of the Array of Things used to monitor urban activity and provide meaningful real-time guidance to city governments on policing, traffic management, and other activities that require continuous monitoring.

Points of Caution for Modeling with "Big" Data

Experts did express caution when conducting modeling from "big" data sources. First, while many experts see value in using "big" data sources to improve the predictive ability of models, Reti has found that sometimes even very large incremental amounts of data yield only small incremental predictive power. In addition, Reti sees risk in relying on black box models, which can miss data quality issues. He believes descriptive statistics will continue to be needed to provide background on the data before

investing in advanced statistical models. Finally, Lazer is concerned about the role of algorithms and automated models in providing defaults that lead researchers to avoid exercising control over the research process. Thus, while the addition of "big" to predictive models has proven to be valuable to the accuracy of the predictions, there is some skepticism that predictive accuracy is enough.

Findings

The literature review and expert interviews both drew attention to the importance of dimensionality reduction for conducting modeling on high-dimensional data when the number of variables is very large. The literature review found that without dimensionality reduction, noise accumulation can harm prediction accuracy. Experts recommended regularization approaches and unsupervised learning as approaches to address this challenge. Experts apply a range of models to Big Data sources, including ensemble tree-based methods, deep learning, generalized linear mixed-effects models, and systems modeling. Multiple experts see AI as an area for growth in the future. Experts shared a few warnings about conducting modeling with Big Data, including that incorporating Big Data in predictive models sometimes yields only minimal improvement in predictive ability, that black box models can miss data quality issues, and that relying on algorithms to provide defaults for research can harm the research process.

On Challenges Using Alternative Data Sources

Interest for Project

While it is clear that alternative data sources do have great value for decision-making and research, understanding the particular challenges of using such data sources relative to traditional, "designed" data sources is important. Based on the findings of the literature review, NORC gauged our experts' perspectives on such issues as data quality, cleanliness, and the impact of business and technological issues on the usefulness of alternative data sources. The findings can inform our understanding of when a data source is fit for conducting policy analysis and best practices for the evaluation of new data.

While experts see great promise in the use of alternative data sources, they also recognize many challenges. Similar to the literature review, experts discussed a few categories of challenges with alternative data sources related to data quality, transparency, data cleanliness, and business and technological issues.

Data Quality Challenges for Alternative Data Sources

Experts identified multiple challenges for data quality of alternative data sources that warrant careful consideration of their use. A common theme was that the context of data collection has a strong impact on the ultimate data quality and that challenges emerge when the data are collected for some purpose other than to inform policy research. Borgman discussed concerns for comparability over time, particularly when breaks in data series occur because of transitions from one data source to another. This happens when using the new data source is deemed more important than the threat to continuity. O'Hara is also concerned about breaks-in-series, particularly because administrative records and other alternative data

sources may be sensitive to policy changes of which agencies and researchers must be made aware. Administrative data, when used as a sole or secondary source, are influenced by programmatic and legal changes that are often not transparent to those using the data for decision-making or surveillance. Gao is also concerned with comparability in the context of EHR and software systems, such as Epic, Cerner, or McKesson. The lack of comparability between technology outputs is critical. Eltinge drew attention to the challenges of different levels of aggregations of units in the data and rules of aggregation. He discussed this in the context of business data, where it is difficult to combine data when different businesses provide either firm-level or establishment-level data.

Some experts, including Baru, Gibbons, and Reti, mentioned the representativeness of alternative data sources and the bias that can result. Baru discussed that it is easy to lose sight of those not represented by a dataset and the resulting concerns for fair and ethical treatment. Gibbons sees alternative data sources as observational data, which present all the accompanying threats to inference, including inferential bias. He recommends repeated replication of findings on other datasets as a way to overcome the limitations of a single dataset.

Challenges Due to Data Transparency, Cleanliness, and Technological Issues

A second set of concerns about alternative data sources relates to business and technological issues with these data sources. Multiple experts expressed the importance of the transparency of the data source, particularly for third-party data, and the necessity for clear documentation. Without proper documentation, experts found challenges for determining whether the data will ever be fit for decision-making. Baru noted that the transformation process of data sources with non-traditional structures needs to be clearly documented and fully transparent to the researcher or data user. Chen and Osgood also both emphasized the importance of transparency and good documentation for a data source from its data collection process to the production of the final version of the data. Klenk emphasized that messiness is the defining feature of non-traditional data sources and stressed the importance of parallel standardization efforts in metadata.

Some experts specified concerns with business and algorithmic rule changes that affect the consistency and reliability of the data source for policy analysis. Reti was concerned about the cleanliness of third-party data sources and sees many areas for potential problems due to the role of black box algorithms in analyzing and managing large-scale datasets. Lazer emphasized the need for transparency of private sector and social media data sources. The use of algorithms for such data sources, including the role of bots, search optimization, and false profiles, presents a threat to conclusions from all analyses of alternative data sources. Yeung is also concerned that actors in hospitals and medical provider settings have very poor incentives to provide the quality reporting needed for EHR and related sources. For example, doctors and nurses may have errors in their reports. Further, providers have financial incentives to report certain medical codes more frequently and may "gamify" certain data elements and processes to achieve goals other than the creation of data useful for decision-making.

Alternative Perspectives

While the majority of experts had concerns for the quality of alternative data sources that merited careful evaluation, four of our experts had interesting perspectives on why data quality should be less of a concern than it currently is. One perspective was shared by Chen, Klenk, and Osgood. They are less concerned about noise, or "messiness," in alternative data sources and use modeling to decompose the signal from the noise. All three are interested in applications where alternative data sources improve the predictive ability of models, and they view noise in alternative data sources due to error as something that can be addressed in modeling

A different perspective was shared by Chernew. As a researcher primarily interested in having the right measures for the research problem, Chernew does not view data quality as being nearly as important. Thus, he is less concerned with the quality of alternative data sources because he views data quality issues as secondary to other features of data to support research.

Findings

There is largely alignment between the findings of the literature review and the expert interviews regarding challenges for alternative data sources, with challenges falling into two categories: 1) those related to data quality and 2) those related to business and technological issues resulting in concerns for data transparency and cleanliness. An overarching theme is that because alternative data sources are collected for a number of reasons other than for conducting public sector decision-making, concerns naturally emerge for the data's fitness for use. Among data quality concerns, experts emphasized threats to comparability, the lack of representativeness, and potential bias introduced by alternative data sources.

Several experts are concerned about the transparency and cleanliness of alternative data sources. Because the methods of data collection may be evolving, the collection and curation of the data may not be well established or documented. Lazer drew attention to how technological and algorithmic changes lead to particular concerns for social media data sources. Yeung provided clear examples of how incentives in health care settings can lead to errors in EHR data.

The view that data quality is important to assess is not universal, particularly as demonstrated by Chernew. Also, some experts who are largely interested in predictive modeling have the perspective that modeling can address some data quality issues by accounting for some sources of error and statistical noise in modeling. Because of the data volume, the noise in the data is less threatening to analytic work as each observation and variable has less value than do traditional data sources. The volume also brings with it the ability to model away the noise in the data effectively.

Data Quality Assessment

Interest for Project

After understanding the concerns for using alternative data sources for decision-making, a framework to assess data quality is needed to evaluate the fitness for use. We asked the experts about their practices for

assessing data quality and their advice for best practice for such an assessment. The findings can help inform the development of guidelines and standards for data quality assessment in the future.

Expert Perspectives on Assessing Data Quality

Among the experts, Biemer proposed one of the more comprehensive frameworks for assessing the quality of alternative data sources. Biemer has a very long and active research program in data error structures and has recently developed a research program in error sources in alternative data sources. His perspective derives from the total survey error literature and decomposes sources of error affecting variables or rows in the dataset from errors affecting columns or units. Beyond that, there are errors affecting specific cells. By this approach, Biemer motivates the understanding that different sources of error may affect alternative data sources, such as measurement error when values are incorrect; coverage error when, for example, units that should be included are not; and validity error when measures do not match the research concepts of interest. While discussing that the level of error in surveys can be high, Biemer discussed that the non-designed or organic nature of alternative data sources can lead these data to have risks for such errors. He has worked on developing risk profiles to assess the potential for such errors in administrative data in particular.

However, other experts suggested different practices for data quality assessments based on the context of their experiences. Eltinge argued that data quality analysis can be challenging, with measures being indirect. He proposed that agencies and researchers who conduct data quality assessments should focus on data quality components that are most uncertain or potentially large. Bourne discussed that using the community to review the quality of a dataset can be informative for other researchers. He compared this model to Yelp as well as to the peer review process for publication in academic journals. Bourne proposed that a system for citations could be developed for trusted data sources similar to citations for academic articles.

In some situations, particularly when alternative data sources are used for covariates in modeling, experts did not emphasize the importance of data quality assessment. Chen and Osgood, in particular, argued that analyzing data quality was not a focus for the predictive applications they are interested in.

Findings

Biemer provided the most mature and detailed perspective on the potential sources of error for alternative data sources. His perspective matches what we found in the literature review. However, data quality assessment is varied in practice and may depend on the data source. Predictive models have fundamentally different data quality requirements than do official statistical estimates from government agencies. Bourne's perspective on using data citations may reflect a practical approach that can more rapidly increase the use of new datasets but that relies on the community to assess data sources rather than through the establishment of best practices.

Data Standardization in Practice

Interest for Project

The use of standardization for alternative data sources has the potential to increase data sharing and support usability by establishing common coding and measurement. For standardization, we are specifically interested in how data processing workflows can be established to develop common data formats that can be readily shared and adopted in a given research community. Learning from past experiences can guide the thoughtful development of standardization rules. We asked experts about their perspectives on data standardization to learn how it has occurred in practice and what positive and negative effects these standardization processes have had on research.

Expert Experiences with Data Standardization

One theme from our discussions with the experts is that standardization is not always well planned and is affected by how research and fields evolve. Savitz, who works primarily with EHR data at Kaiser Permanente's Center for Health Research, views standardization as needed to harmonize measures coming from different sources and systems, as the lack of agreement on harmonized measures harms cumulative medical science. From her perspective, however, what ultimately drives data quality is not only the standards developed but also the behaviors resulting from the standards and what is actually implemented by practitioners. For example, there is variation in whether and how documentation is produced once the standards are determined. Baru discussed how the maturity of data practices in different fields affects standardization. Specifically, fields with mature, well-defined theoretical models at the core of the discipline develop standardization for data and measures more quickly. For example, physics and astronomy have well-defined conceptual models and have mature data standards relative to fields that remain in the discovery phase, such as geology and archaeology. Borgman mentioned that data standardization practice often reflects whose scientific practice prevails within a field. Standardization and requirements for metadata are, then, developed to enforce one academic perspective on an entire field. This is an interesting perspective on data standardization, which is often viewed as unimpeachable evidence of progress in the regularization of data sources.

Findings

The literature review found that the emergence of the Common Data Model and the HL7 standards enabled successful integration of health data sources to support such use cases as the FDA's Sentinel Initiative for monitoring adverse drug reactions and the Centers for Disease Control and Prevention's National Syndromic Surveillance Program. However, because new data sources may be constantly evolving, the path toward standardization may not be clear or may quickly become outdated. The experts in general did not find that data standardization always leads toward scientific progress. In particular, Savitz is concerned that standards must also consider the resulting behaviors of data practitioners impacted by the standards. Borgman argues that data standardization is a way to enforce a particular scientific perspective on an entire field, and Baru suggests that the maturity of the conceptual and theoretical models in the scientific field will substantially influence the pace and quality of data standardization.

Criteria for Adopting Alternative Data Sources for Official Statistics

Interest for Project

A particularly important area of application for alternative data sources is their use for official estimates commonly coming from federal statistical agencies. These estimates are critical for policy research and decision-making for a variety of reasons. Among other things, they provide an indication of the need for certain services or disparities or health and well-being outcomes. We asked the experts their views on the path to adoption of new data sources for official statistics to inform recommended practices in the future.

Expert Perspectives on Adopting Alternative Data Sources for Official Statistics

Biemer proposed that agencies use three factors to decide whether and how to adopt a new data source for federal statistics: 1) the cost and affordability, 2) the impact on the series based on how the data source will be used, and 3) the importance of the data series. Regarding the second criterion, agencies should consider whether the use of the data source will cause a break-in-series and what impact that will have on decision-making. If a data series has important policy uses that will be adversely affected by a break-in-series, extra care must be taken by the agency. Eltinge discussed that to successfully transition to a new data source, both a "backbone" (or a compelling reason why the data series improves measurement) and a "bridge" (or a process for calibrating the new data source) are needed. The extant data series using the traditional data source will be the backbone, and the new data will be used to build the bridge to the next part of the series. O'Hara pointed to examples across the federal statistical system where there is prospect of the integration of new data sources into the statistical system, including at the Internal Revenue Service, Census Bureau, Bureau of Economic Analysis, and BLS. She argued that agencies need to support research to fully evaluate and test the impacts of adopting the new data sources.

Findings

All experts agreed that a detailed research process is needed to guide adoption of alternative data sources for official statistics. Careful approaches are needed to ensure that any changes will not adversely and unfairly impact certain groups and due to aforementioned challenges with new data sources as they are not collected and curated for the purpose of policy analysis. Particularly the risk of breaks-in-series must be thoughtfully considered by statistical agencies. Note that these criteria are fundamentally different than those discussed by experts such as Chen and Osgood who are primarily interested in statistical modeling. This reflects that the criteria for adopting an alternative data source depend heavily on the use and the dataset's fitness for that purpose.

On Values and Challenges of Public-Private Partnerships

Interest for Project

Data have been the cornerstone of decision-making and public policy evaluation, and their origins have always been at the core of statistical policy. Access to data, then, is an important first step in all research. One of the most important structural problems with new data sources is that a large volume of new data

sources, with the exception of program administrative records, is in the private sector. Retail sales data, electronic medical records, user-generated social media data, and transaction data are all data assets held by private entities. For many of these private companies, the data are in fact a business asset. Developing partnerships between the public and private sectors that are sustainable and yield critical resources for research and decision-making will be the challenge for federal policymakers. We asked our experts to identify the challenges of having vital data resources in private hands and to provide successful examples of public-private partnerships.

Consensus

With some exceptions, our experts agreed that access to privately held data was viewed as a critical element in moving public policy forward. Klenk had substantial optimism about the open science model, particularly for industries where there is substantial revenue to be gained through expanded data universes, which can improve the research or testing of products. His example comes from the pharmaceutical industry where sharing data on drug development and testing has proven of value to the companies involved, and they have begun to engage substantially in data sharing. The remaining experts all agree that this is a challenge that needs to be addressed proactively as privately held data hold enormous promise for assessing public well-being and aiding in decision-making.

There are several very hopeful examples of private-public partnerships noted by our experts that are worth mentioning. First, Lazar described a collaboration between the social media companies and researchers brokered by the Social Science Research Council called Social Science One, which is designed to give researchers the ability to understand issues related to democracy using social media data. The second effective collaboration has been brokered by the National Science Foundation (NSF) to provide researchers with cloud computing capacity. Private sector vendors have donated computing infrastructure and provided shared infrastructure through the NIH Data Commons Pilot and the Kaiser Virtual Data Warehouse. These collaborations have developed as a result of shared goals and benefits. Finally, Mundada also suggested that the federal government may want to begin to work with an "honest broker" model to use a third party to receive and reformulate data by aggregation or others methods from the private sector for use in decision-making.

There are two cautionary notes sounded by experts that are important. First, Chernew noted that data are an asset to most private entities and that for health care insurers and providers in particular the movement to take this data into the public domain represents an important challenge. The second caveat comes from Borgman, whose generalized critique of open data and data reuse, suggests that the privately held data sources will lose their provenance when exported from their environment and may lose the important metadata that makes them intelligible.

Findings

All of our experts agreed that private sector data are a critical and necessary next step in the development of the data universe important for public decision-making. Some offered cautionary notes concerning the ownership and provenance of private data, but all agreed that this is a critical issue for the federal government. None offered concrete steps forward, with the exception of Mundada, who viewed the

honest broker model as a way to provide private entities with the security and protection necessary to make privately held data available. Some hopeful signs were noted by our experts and were clustered around shared computing resources and specific research problems. All the experts agreed that there is likely no systematic solution possible given the complexities of different industries and different data types.

Training the Workforce

Interest for Project

In order to keep pace with the rapidly changing technological landscape, training the workforce has become an important topic for the federal government. We asked the experts about the training needed to keep federal employees up to date with evolving data sources as well as the technical skills to analyze such data.

Structures for Training

Currently, there are different structures in place for training the workforce on skills related to alternative data sources across academia, government, and industry. In academia, formal programs on such data skills are being created at the vocational, undergraduate, and graduate level. Baru emphasized that skills related to alternative data sources should be offered to a broad audience without being limited to those in traditional STEM fields or those with advanced mathematical training. Baru and his team at NSF believe that there should be data science jobs at all levels, from those with vocational degrees to those with advanced graduate training. NSF has also commissioned a study from the National Academy of Sciences,[3] which seeks to organize the undergraduate data education in a more systemized manner. Biemer recommends the Joint Program in Survey Methodology at the University of Maryland, which has a data science track, as a model academic program. He notes the program's emphasis on the total error framework in understanding data quality. He believes the framework provides a grounding in survey analyses that will remain important in the future.

In government, various efforts are taking place simultaneously to train the workforce. When Chen was the Chief Data Scientist at the Department of Commerce, he established an academy where interested staff could learn to use open-source technology. In addition to such academies, Biemer suggests government agencies actively engage in academic conferences and develop short courses. In industry, Reti mentions Optum Analytics's training institute that is available to company staff. Several hundred Optum employees are enrolled in the institute, which trains staff with latest skills for data analyses. Optum specifically targets staff with mathematical and statistical backgrounds for such training.

[3] National Academies of Sciences, Engineering, and Medicine: *Envisioning the Data Sciences Discipline: the Undergraduate Experience*.

Computational and Analytical Skills

Experts from academia, government, and industry shared the computational and analytical skills they believed were important in working with alternative data sources. Catlett described four essential roles that are necessary for working on large-scale alternative data projects such as the Array of Things at the Urban Center for Computation and Data. These roles include 1) computer hardware engineers to build the machinery, 2) database and technical infrastructure managers, 3) computer scientists to design databases that capture complex data from a variety of sources and the algorithms that provide results from the data, and 4) statisticians, mathematicians, computer scientists, and substantive experts who understand the algorithms and the substance in order to provide thorough guidance for programming and analyses.

Further, Chen identified four essential skills for working on alternative data projects. First is the ability to determine whether the project is a data project. Often, people in non-quantitative roles will confuse social and behavioral problems with technical problems. Only the latter can be solved using a pure data-driven approach. Second is the necessity of good programming skills. Since 70 to 80 percent of data projects involve data manipulation, proficiency in data manipulation frees the analyst to spend more time on the truly challenging technical problems of any project. Third is the ability to question doctrine and dogma. Previously, Chen used economic theories to construct models that would predict outcomes. In recent research, however, he found that theories often do not withstand empirical scrutiny. Chen said that a data scientist's job is to figure out how to best represent the process from a data perspective. Finally, good communication skills are needed to present the analytic results of the data.

Findings

Academia, government, and industry are currently implementing various different and novel structures to train the workforce on skills related to using alternative data sources. Some of the efforts include reorganizing data training in post-secondary educational institutions and establishing training academies within federal departments and private companies. Further, methods, access, and use of alternative data are being rapidly democratized—new technologies and widespread training efforts are beginning to bring this type of analytic activity to larger audiences. Both expert interviews and NORC's literature review suggest that managing alternative data requires a unique blend of domain knowledge, technical skills, behavioral attributes, and personality traits from an organization's staff. This requires building data science teams with diverse capabilities that can complement each other, such as domain expertise, research methods, computer programming, and system administration.

Privacy Methods for the Future

Interest for Project

One of the biggest challenges with using alternative data is protecting individual privacy. Current uses of granular, linked individual data for research and policy analysis pose substantial threats to individual privacy. We asked the experts for potential solutions to harnessing the power of alternative data while protecting the privacy of individuals.

Different Ideas for Individual Privacy Protection

Different experts envision different privacy methods going forward. Osgood states that individuals should be given a tiered opt in or opt out privacy option similar to that of Apple Inc.'s privacy agreements. Such a choice would be offered in a broad effort to escrow data for use in select urgent and important analytic and surveillance activities. This idea of a tiered voluntary data escrow suggests a more nuanced approach to data privacy than is currently applied in research. He discusses blockchain and homomorphic encryption technology as tools to facilitate such an approach.

Similar to Osgood, Yeung suggests that blockchain technology could greatly empower patients to have ownership of their own data. Blockchain technology maintains a history of transaction records across several computers that are linked in a peer-to-peer network. Such technology could decentralize health care data from payers and providers into the hands of patients. Thus, each patient would be able to have full control over whom they share their data with. Mundada, on the other hand, suggests an honest broker model, where a trusted and neutral third party will help assimilate various different data sources by attaining data at the lowest possible cost and protecting it at the highest level.

Findings

While all experts addressed the importance of individual privacy protection, there was not yet a consensus on best practices. Many experts suggest novel approaches to ensuring privacy, such as using blockchain technology and the honest broker model. Experts agree that private-public partnerships are crucial to advancements in individual privacy protection and that the field will see faster improvements as the government and the private sector negotiate terms of engagement on these issues.

Opportunities

The results of our literature review and expert interviews provide some interesting guidance and direction for the role of new data in policy research and decision-making at HHS. As with all data policy, it is difficult to provide broad and comprehensive guidance on how to approach new and emerging data sources, especially given the heterogeneity of data use and applications in public policy development and evaluation. This is particularly true given the breadth of the HHS research and policy portfolio. Nevertheless, there are some generalizable and useful lessons to be drawn from our focused examination of the definitions, use and data quality, and data access issues associated with integrating new data sources into policy evaluation and research as well as decision-making. Substantial advances have already been made in the introduction of new data sources into the federal data ecosystem as evidenced by the National Academy of Science reports of 2017. As many of our experts suggested, it is no longer the case that we should ask the question of whether alternative or big data sources should be used in evaluation and decision-making in the public sector. We should be asking the question of how and where these data sources can supplement, replace, or expand our ability to understand the health and well-being of the American population and the impact of public programs in the future.

To that end, our observations and recommendations fall into four categories. First, we will address where new data sources are likely to provide the most powerful and useful replacement or supplement to existing data and measurement. These observations come both from our literature review and the voices of our experts. Second, we address areas of importance in data quality assessment and the next steps in the maturation of data sources. Many of the areas we highlight already have active and ongoing investment by HHS agencies, and we would like to reinforce their importance. Third, the issues of data governance, which include access to privately held data and privacy protection, have always been central to the federal government's approach to data policy. With new data sources, these issues have become even more critical. Finally, we will address our summary of training models and infrastructure development to help the federal government take advantage of new technologies and data.

Effective Use of Alternative Data

Our cumulative evidence suggests that the most fruitful use of new data sources lies in four main areas: 1) improved measurement, 2) more granularity by time, geography, and outcomes, 3) better and more robust prediction models, and 4) more accurate data structures. Our experts provided particular insight into how these four attributes of new data sources can improve many of our analytic tasks and the power of our analyses. It makes sense, then, to invest in first slowly transitioning many of the data systems that are currently used to do disease and impact surveillance, measure behaviors such as physical activity and diet or measures of income and health care utilization, economic and social forecasting, and social network impact. These are areas of investment where the return to introducing new data will be the highest and where there are already successful pilot programs. We noted the very successful integration of administrative records into data systems that measure health care utilization and costs, the use of user reported adverse drug events, and the real-time monitoring of traffic flows as examples of these activities.

Investment in Data Quality Assessment and Standards

Our second set of observations suggest that an investment in data quality standards that are specific to new data sources are an important next step in the maturation process for these data types. The federal data infrastructure has traditionally set very clear and systematic guidelines on data quality requirements for many of its data systems. Our experts were quite clear that for many of these data types, it will be important to invest in data quality standard setting before full scale integration of new data sources can occur. Biemer and Eltinge, in particular, were clear in their guidance on this score. The challenges of transparency, population representativeness, and algorithmic and technological stability were of particular concern. Two important areas, in which HHS is already active are data standard setting and dimensionality reduction. The Common Data Model currently being pursued both at NIH and the Office of the National Coordinator for Health IT is an example of one such exercise. Dimensionality reduction often occurs through machine learning models and, increasingly, both research and policy experts are pursuing the use of such techniques to make data of this type more useful.

Investment in Data Governance

Third, it is clear that the data governance issues associated with new data sources are complicated in ways that require careful and deliberate thought and guidance. Data governance broadly defined includes issues of data access, data sharing, privacy protection, and technological protection. In our summary, we are not prepared to discuss the full range of these issues but do have some examples and suggestions related to data access and privacy protection that surfaced from our experts. One of the critical issues in data access will be to identify the next steps in developing public-private partnerships. Our experts provided us with a variety of hopeful examples of effective and productive partnerships. One area of particular promise lay in computing infrastructure. Large technology companies with cloud computing infrastructure have shown a substantial and ongoing willingness to share this infrastructure to facilitate the high dimensional computing necessary to analyze many of the new data sources. Additionally, our experts provided resource sharing and data sharing examples where there were mutual benefits, either because of shared incentives or because the analysis of the shared data yielded substantial benefit to private companies. Our expert Klenk described it most effectively when he said that open source software used to be seen as a losing proposition for most companies and now is central to much of the innovation that occurs in the technology world. He envisions the same transformation of the open data movement in the not-too-distant future. A systematic review of HHS public-private data partnerships would also be a useful effort to identify the successful partnerships and how they arose. This may provide a roadmap to forging those partnerships in the future.

Data privacy and individuals' rights to privacy are still seen as a substantial issue at the core of data access and data integration. There are legal, ethical, and technical issues that will remain part of the dialogue for many years. Our experts agree that data privacy requires new solutions and offered some interesting suggestions including honest brokers, individual data archives, and block chain technology. At the center of their solutions is the notion of giving data rights back to the individuals who produce the data. This idea, while it has a host of complications, sits in the middle of debates about data privacy. Like the other observations presented here, there is likely no generalized data privacy solution. Every data type will need a solution tailored to how the data originates as the degree to which individuals acknowledge,

are made aware of, and relinquish their rights to the final data varies substantially. The value of the data, however, must always be weighed against the rights of the individuals.

Investment in Workforce Training

Finally, we reviewed workforce training models and asked our experts to identify what types of skills the next generation of federal employees and analysts will need to take advantage of new data types. One of the most interesting and productive models is the in-house training institutes such as the ones described by Chen and Reti. The computational and computing challenges that initially required highly trained computer scientists and data scientists are being democratized rapidly according to many of our experts. The challenge is to train current staff to understand the nature of the data and its uses and to put together effective teams with substantive, data, and statistical expertise. As data scientists move through the educational pipelines that provide public policy analysts and decision-makers, these in-house training efforts or institutes that focus on federal employees modeled on JPSM may be an interim solution to providing the federal workforce with the skills to use new data types and formats.

Conclusion

This project has been an opportunity to address selected issues in the integration of new types of data into the data infrastructure that supports federal policy evaluation and decision-making. The breadth of opportunities these new data sources provide for analysts to generate timelier, accurate, and robust evidence to support the process of federal decision-making should not be understated. It is, in fact, already occurring. The challenges moving forward are to make selective investments where these types of data can have the most value and to begin the careful and painstaking processes of establishing clear data quality standards for data used in support of public decision-making. These activities are also already occurring in many agencies across the federal government, and they are likely to gain substantial momentum over the next decade.

References

Banks, D., Ruddle, P., & Kennedy, E. (2016). *Arrest-Related Deaths Program Redesign Study, 2015-16: preliminary findings*. Washington, DC: Bureau of Labor Statistics. Retrieved January 2018, from https://www.bjs.gov/index.cfm?ty=pbdetail&iid=5864

Beyer, M. (2011). Gartner says solving big data challenge involves more than just managing volumes of data. *Gartner*. Retrieved January 12, 2018, from http://www.gartner.com/newsroom/id/1731916

Biemer, P. P. (2016). Errors and inference. In Foster, I., Ghani, R., Jarmin, R. S., Kreuter, F., & Lane, J. (Eds.). *Big data and social science: A practical guide to methods and tools* (pp. 265-297). Boca Raton, FL: CRC Press.

Brownstein, J. S., Freifeld, C. C., Reis, B. Y., & Mandl, K. D. (2008). Surveillance sans frontieres: internet-based emerging infectious disease intelligence and the HealthMap project. *PLoS Medicine*, *5*(7), e151.

Department of Health and Human Services (HHS). (2018, February 28). Strategic Plan FY 2018 - 2022. Washington, DC: HHS. Retrieved from https://www.hhs.gov/about/strategic-plan/index.html

Freifeld, C. C., Brownstein, J. S., Menone, C. M., Bao, W., Filice, R., Kass-Hout, T., & Dasgupta, N. (2014). Digital drug safety surveillance: monitoring pharmaceutical products in twitter. *Drug Safety*, *37*(5), 343-350.

Groves, R. M. (2011). Three eras of survey research. *Public Opinion Quarterly, 75*(5), 861–871.

Hansen, S. E., Benson, G., Bowers, A., Pennell, B. E., Lin, Y., & Duffey, B. (2010). Survey quality. University of Michigan Institute for Social Resarch Cross-Cultrual Survey Guidelines. Retrieved January 12, 2018, from http://projects.isr.umich.edu/csdi/quality.cfm

Horrigan, M. (2013). *Big Data and Official Statistics*. Retrieved January 12, 2018 https://www.bls.gov/osmr/symp2013_horrigan.pdf

Japec, L., Kreuter, F., Berg, M., Biemer, P., Decker, P., Lampe, C., Lane, J., O'Neil, C., & Usher, A. (2015). Big data in survey research: AAPOR task force report. *Public Opinion Quarterly*, 79(4), 839-880.

Kitchin, R. (2015). The opportunities, challenges and risks of Big Data for official statistics. *Statistical Journal of the IAOS*, *31*(3), 471-481.

Lazer, D., Kennedy, R., King, G., & Vespignani, A. (2014). The parable of Google flu: traps in big data analysis. *Science*, *343*(6176), 1203-1205.

Lazer, D., & Radford, J. (2017). Data ex machina: introduction to big data. *Annual Review of Sociology*, *43*, 19-39.

National Academies of Sciences, Engineering, and Medicine. (2018). *The Undergraduate Perspective: Interim Report*. Washington, DC: National Academies Press.

National Academies of Sciences, Engineering, and Medicine. (2017). *Federal Statistics, Multiple Data Sources, and Privacy Protection: Next Steps*. Washington, DC: National Academies Press.

National Academies of Sciences, Engineering, and Medicine. (2017). *Innovations in Federal Statistics: Combining Data Sources while Protecting Privacy*. Washington, DC: National Academies Press.

NIST Big Data Public Working Group. (2017). *Draft NIST Big Data Interoperability Framework: Volume 1, Definitions*. NIST Special Publications 1500-1, Version 2, Draft 2. Retrieved January 12, 2018, from https://bigdatawg.nist.gov/_uploadfiles/M0613_v1_3911475184.docx

Puts, M. J. H., Tennekes, M., Daas, P. J. H., & de Blois, C. (2016). Using Huge Amounts of Road Sensor Data for Official Statistics. In *Proceedings of the European Conference on Quality in Official Statistics (Q2016)*, Madrid, Spain. Retrieved November 2016, from http://www.pietdaas.nl/beta/pubs/pubs/q2016Final00177.pdf

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston, MA: Houghton Mifflin.

APPENDIX A
Literature Review
To Big Data or Not: Determining the Utility of Big Data for DHHS Policy Research

LITERATURE REVIEW

To Big Data or Not: Determining the Utility of Big Data for DHHS Policy Research

CONTRACT NO. HHSP233201500048I

FEBRUARY 7, 2018

PRESENTED TO:

James Sorace
Jeongsoo Kim
Office of The Assistant Secretary for Planning and Evaluation
Department of Health and Human Services
200 Independence Avenue, SW,
Room 415F
Washington, DC 20201

PRESENTED BY:

Zachary H. Seeskin, Felicia LeClere, and Jaehoon Ahn
NORC at the University of Chicago
55 East Monroe Street
30th Floor
Chicago, IL 60603
(312) 759-4000 Main
(312) 759-4004 Fax

NORC at the UNIVERSITY of CHICAGO

Table of Contents

Executive Summary

Big Data holds great promise for expanded use at the Department of Health and Human Services to support evidence-building for health policy research. Different Big Data sources can provide novel information for policy and may have advantages relative to surveys and censuses due to the speed of data collection, the potential for geographic granularity, and having less measurement error. However, Big Data sources must be evaluated carefully for policy research so that the data meet the quality standards needed. When data do not come from a government agency, the accuracy, reliability, and consistency of what is measured from a Big Data source can be difficult to determine. This literature review examines the benefits and challenges of using Big Data for evidence-building and reviews different uses of Big Data. While administrative records have been used successfully for some government data products, the understanding of how to evaluate the quality of other Big Data sources, from electronic health records to data from environmental and health sensors to social media, is still maturing. However, the carefully designed standards used for electronic health records have allowed for many successful policy uses. These standards can serve as a model for other uses of Big Data sources. The strengths and weaknesses of different kinds of data should inform how they are applied. For example, data that are used directly for construction of estimates have different quality requirements than data that are used as auxiliary information in conjunction with surveys and censuses. The speed of data collection for some Big Data sources makes the data promising for real-time surveillance and monitoring and for generating hypotheses to investigate with higher-quality data sources. In addition, we discuss the analytical challenges of working with Big Data, the algorithms that have been used in Big Data applications, and the skills needed among teams using Big Data statistically.

Introduction

There is a substantial opportunity for government agencies and policymakers to use new data sources to inform policymaking and evidence-building, including in the area of public health. With computational power rapidly increasing and large amounts of data being produced, the term “Big Data” has emerged to describe the new kinds of data sources emerging. While the potential for Big Data to help provide new insight has been recognized, the science for evaluating and analyzing Big Data is still evolving and maturing.

Statistical use of data, that is, the description, estimation, or analysis for populations or groups, guides policymaking and evidence-building. For any agency or policymaker seeking to use Big Data statistically, careful guidance is needed to understand the strengths and limitations of the data sources. In this report, NORC at the University of Chicago (NORC) presents a literature review of the uses of Big Data for health policymaking to guide the Department of Health and Human Services’ (DHHS) Office for the Assistant Secretary of Planning and Evaluation (ASPE). This literature review provides an informational resource to help facilitate DHHS decision-making about the integration of Big Data into public policy formation and evaluation.

Defining Big Data

"Big Data" is a term that has been used with an array of definitions. Of interest to DHHS is a variety of alternative data sources that could provide additional value to data sources traditionally used for evidence-building, specifically surveys, censuses, and randomized experiments. For this literature review, we define Big Data broadly, using Groves' (2011) definition of "organic data," *data that are not originally collected for statistical purposes*. This definition recognizes that large databases are being produced with great frequency and often rapidly and inexpensively. However, these data sources differ from surveys, censuses, and randomized experiments, or "designed" data, in that their data collection is not curated for the purposes of statistical use or evidence-building. The definition we use for Big Data focuses on benefits and challenges arising from how data were created.

This definition of Big Data encompasses a wide range of datasets that can be used for health policymaking, all facing the challenge that the data collection was not designed to support statistical uses. Among the Big Data sources reviewed here are private and public health insurance claims, Medicare and Medicaid enrollment records, state health registries, electronic health records, consumer purchase data from businesses, electronic prescription data, social media, data generated by mobile phone use, and data produced by electronic and health sensors.

Describing Big Data

Big Data is often described as having characteristics encompassed by three V's (Beyer 2011, Japec et al. 2015, NIST 2017):

Volume: The sheer amount of data available.

Velocity: The speed at which data collection events occur.

Variety: The complexity of the formats in which data sources exist.

Across different sources, other descriptors of Big Data sources have been suggested. We draw attention to a fourth V (Japec et al. 2015, IBM 2017):

Veracity: The ability to trust that the data can support accurate statistical inferences.

The four V's help describe both the promise and the challenge of Big Data. The volume, velocity, and variety of some Big Data suggests how these data sources may mitigate weaknesses of and/or add value to surveys and censuses. These attributes, however, also make Big Data hard to manage computationally. Specific expertise may be required to manage large datasets, and processing time may be longer. Because of the complexity of formats for some Big Data sources, extensive data cleaning may be required to make the data usable.

In addition, this review provides guidance toward determining when the veracity of a Big Data source can be trusted and what uses of a data source are still promising when veracity is questionable. A major focus of this review is understanding and maintaining the data quality needed to support statistical inferences for policymaking.

One common issue with the veracity of Big Data sources is the lack of representativeness of the population of interest. For example, consider a dataset that includes *all* Twitter users. Because Twitter users are different from the rest of the population, it is challenging even with statistical adjustment to extrapolate a finding on Twitter users to the population of interest for a policy question. Also, because data collection is not originally designed for statistical use, the information needed to use a Big Data source for policymaking may not be available. For example, metadata on the variables and data structure may not be available. Data on demographic characteristics needed to adjust the estimates may not be collected or available for enough records.

The origin of Big Data is critical to our evaluation of data quality. There are two distinct kinds of Big Data that have different implications for the use of such data sources (Lazer and Radford 2017).

Digital trace: Data that constitute recordkeeping or chronicling of actions at one or more organizations. Both the public and private sectors produce digital traces. These represent records of actions, but not the actions themselves.

Digital life: Data reflecting a direct action by a user. This often reflects the use of online platforms, including social media. Data from health trackers, like Fitbit, would be another example.

In our review, we have found that Big Data sources can usually be classified as either reflecting digital traces or digital life. Digital traces usually involve one or more organizations that curate and maintain the data sources for a purpose other than statistical inference or policy analysis. The data are likely to be highly structured and systematic as they are used to manage processes or define eligibility for use of services. In contrast, digital life data tend to be more complex to use and analyze as they are less structured and curated because they are captured as part of online activity. These differences are critical for assessing the issues with using the data and determining how the data can be used for policymaking.

The Importance of Data for DHHS Policymaking

In order to assist the Department in identifying policy priorities and coordinating policy planning across the range of issues and populations served by DHHS, ASPE conducts research and evaluation that provides targeted analyses on rising legislative, budgetary, and regulatory issues related to health; disabilities, aging and long-term care; human services; and science and data policy.

Current health and human services policy focuses on systemic change to provide high-value services to DHHS program participants, promote widespread diffusion and adoption of innovation and knowledge, and ensure better transparency and accountability of publicly funded programs. Current policy issues include supporting physician payment reform under the Medicare Access and CHIP Reauthorization Act of 2015, addressing the opioid epidemic, understanding the impact of social determinants of health, and improving the quality of life for people with Alzheimer's, among many others.

To this end, ASPE requires information for a range of policy questions. Using new data sources for statistics and evidence-building has been a recent focus of both policymakers and researchers. The U.S. House of Representatives passed the Foundations for Evidence-Based Policymaking Act of 2017 in November to support the increased use of alternative data sources, including administrative records for

policymaking.4 These data sources can provide novel information for policymaking that cannot be collected by censuses or surveys. The sound statistical analysis of data is critical for estimating needs for services in the U.S. population and for evaluating the effectiveness of policies. Thus, understanding how Big Data can add value to the evidence provided by traditional data sources like surveys and censuses is of tremendous importance. Further, with response rates tending to decline and survey costs increasing over time, there is a need to investigate alternative data sources that may be less costly (Groves 2011). This review will help ASPE provide guidance to DHHS policymakers on the technical, workflow, and business issues to consider and instruct their use and production of Big Data sources for policy research.

For guiding policymaking, it is important to DHHS that statistical uses of data ensure fairness so that different groups are treated equitably. Policymakers must also understand the limitations of the data sources they use in order to avoid doing harm to those affected by their policies. Government agencies rely greatly on public trust in the information they provide, so understanding when and how Big Data sources can be used for policymaking based on their level of veracity is of the utmost importance.

Literature Review Strategy

For this literature review, NORC first identified and reviewed a set of papers providing overviews regarding Big Data for statistical uses and policymaking. Then, NORC identified papers published or released since 2015 that use Big Data sources for policymaking or population studies focusing on applications both in and outside of government. This literature review does not constitute a systematic review, which would have been a considerable undertaking due to the vast number of papers on this topic. Instead, NORC used its expertise to identify papers that demonstrated the most promising, successful uses of Big Data, as well as a set of uses that could demonstrate the breadth of data types, benefits, and challenges of using Big Data sources. The review includes both published articles and grey literature, and we explore both social science and medical literature. Table 1 presents a list of many of the search terms used in the literature review.

Table 1. Partial List of Search Terms Used for Literature Review

1. Health Policy Topics	2. Big Data and Data Science	3. Data Types
4. Population health	5. Big data	6. Health administrative data
7. Public health	8. Large data	9. Medicare enrollment
10. Health care quality, access, evaluation	11. Data science	12. Medicaid enrollment
13. Preventative health services	14. Data quality	15. Insurance claims
16. Demography	17. Data collection	18. Immunization registry
19. Health planning	20. Methods	21. Electronic health records
22. Health expenditures	23. Analytics	24. Electronic medical records
25. Health services	26. Surveillance	27. E-pharmacy
28. Health status indicators	29. Early warning	30. Surescripts
31. Social determinants of health	32.	33. Consumer purchase data
34. Population characteristics	35.	36. Environmental monitor

[4] This bill was passed based on the recommendations of the Report of the Commission on Evidence-Based Policymaking from September 2017.

1. Health Policy Topics	2. Big Data and Data Science	3. Data Types
37. Social environment	38.	39. Health monitor
40. Health services accessibility	41.	42. Mobile phone
43. Health disparities	44.	45. GPS
46. Urban health	47.	48. Patient-generated health data
49. Rural health	50.	51. Sensors
52.	53.	54. Wearable technology

Appendix Table 1 includes a list of some of the notable use cases examined in the literature review.

Overview

This literature review discusses different uses of Big Data sources as they relate to health policymaking. In our review, a number of themes emerged that can help guide ASPE and DHHS to thoughtful use of Big Data for evidence-building:

Different kinds of Big Data sources have different strengths and face different challenges. Therefore, guidance for one kind of Big Data source may not apply to other kinds of data. Thus, we have grouped our review into three categories: 1) data maintained by the public sector, including administrative records, 2) data coming from one or more private sector organizations, including hybrid systems combining public and private sector data, and 3) data that is user-generated, e.g., from online platforms. NORC has found that, within these categories, the issues with using these data sources are more similar. As will be discussed, the first two typically represent digital traces whereas user-generated data typically represent digital life.

Different data types have different readiness levels for statistical uses and policymaking. The maturation of standards and requirements for processing and documentation of Big Data sources can be critical to assure strong data quality and to guide the successful use of Big Data sources for policymaking. A summary of NORC's observations on the different data types is provided in Table 2.

Table 2. Description of Three Types of Big Data Sources

55.	56. Administrative Records	57. Private Sector Data	58. User-Generated Data
59. Examples	60. Medicare and Medicaid enrollment 61. Insurance claims 62. State registries 63.	64. Electronic health/medical records 65. Insurance claims 66. E-prescription data 67. Consumer purchase data	68. Social media 69. Environmental and health sensors 70. Mobile phone data/GPS
71. Veracity	72. Higher	73. ←	74. Lower
75. Digital Trace/Life	76. Digital trace	77. Digital trace	78. Digital life
79. Maturity of Data Standards	80. Proven history of successful use 81. Data quality framework from many statistical agencies	82. Some successful use cases 83. Emergence of HL7 Standards/Common Data Model	84. Proof-of-concept studies 85. New measures of data quality emerging

55.	56. Administrative Records	57. Private Sector Data	58. User-Generated Data
86. Characteristics	87. Primary data source used in conjunction with censuses and surveys 88. Large government efforts to: 89. - improve quality 90. - harmonize 91. - link 92. - disseminate	93. Often in data siloes (e.g., hospitals) 94. Varies in structure and complexity 95. Public and private partnerships are underway to: 96. - standardize 97. - share technology 98. - integrate data platforms	99. Nonrepresentative 100. Lack of metadata 101. Technological challenges: 102. - algorithm dynamics 103. - violation of ideal user assumption
104. Common Uses	105. Direct estimation 106. Design and calibration of surveys 107. Imputation 108. Record linkage 109. Second survey frame	110. Some direct estimation 111. Monitoring 112. Surveillance	113. Monitoring 114. Surveillance

In some scenarios, data from Big Data sources may have less measurement error than surveys that are affected by respondent recall. Record linkage of Big Data sources with surveys and other data sources can sometimes improve data accuracy while reducing respondent burden. Big Data sources often have the advantage of being timely and affordable. Thus, some Big Data sources can help provide timely policy guidance that surveys, censuses, and randomized experiments cannot due to the time and expense of collecting and processing the data. Thus, they are particularly promising for monitoring and surveillance to enable rapid response to an emerging health issue. In scenarios where timeliness is more important than accuracy, as is sometimes the case for surveillance and monitoring, Big Data can play an important role.

While the volume of a data source alone may not improve inferences when data veracity is a concern, the volume of a Big Data source sometimes enables producing estimates with geographic granularity. While traditional data sources rarely collect enough data in small geographic areas to be informative, the thoughtful use of Big Data sources can sometimes support the tracking of emerging issues in small geographic areas.

In this review, NORC highlights important use cases chosen to represent the promise and challenges of Big Data sources. Because of the many kinds of applications being explored, NORC cannot possibly highlight all use cases or even all prominent successful uses. However, the use cases chosen represent both successes and promising developments for which questions remain. The use cases are also chosen to represent the different data types and different kinds of uses and challenges emerging from the literature.

The remainder of this literature review proceeds as follows. Section II reviews the general themes emerging from the literature on how to evaluate Big Data sources for statistical use and what issues can arise. This section provides the framework for describing the benefits and challenges of the use cases discussed in the subsequent sections. Sections III to V investigate use cases for the three data types described earlier: Section III discussing public sector data, Section IV discussing private sector data and hybrid systems, and Section V discussing user-generated data. Section VI describes the skills and capabilities needed among teams working with Big Data sources. Section VII then concludes and

summarizes NORC's observations for ASPE to inform guidelines for uses of Big Data sources for health policymaking.

Evaluating Big Data Sources for Use in Health Policy Research

This section presents considerations important for evaluating uses of Big Data sources for health policy research, including for data quality, analytics, and technology. Here we establish the themes and the framework that will be discussed in the context of specific data types and use cases in the following sections of this literature review.

Statistical Validity and Data Quality

Survey and census data collections are designed to both minimize sources of data error and to achieve the data quality needed for research and policymaking. In contrast, Big Data sources are not typically collected for evidence-building and policymaking purposes and may not even support ready measurement of data quality. Even when the level of data quality is difficult to determine, thinking through a framework to understand the quality, a particular data source can help suggest what that dataset's strengths and weaknesses are. In this section, we present a data quality framework that can be applied to assess alternative data sources' fitness for statistical uses.

Data are used to support statistical inferences, which can then be used for conclusions and policymaking. Shadish, Cook, and Campbell (2002) discuss four main aspects of the validity of statistical inferences:

- **Statistical conclusion validity**: Approximate truth about correlations between variables.
- **Internal validity**: Approximate truth about cause-effect relationships.
- **Construct validity**: Whether the concept being measured is the same as the concept targeted for measurement.
- **External validity**: Approximate truth about whether an inference holds for a population.

Often for population studies, construct validity and external validity are of greatest interest. On the other hand, internal validity is of primary importance for assessing the effects of a social program or intervention, but is less important for observational or descriptive studies.

Data quality is multidimensional, with elements reflecting different aspects needed to support valid statistical inferences. Table 3 describes different aspects of data quality needed of either traditional or Big Data sources to support policymaking, grouped into five categories.[5]

[5] See Hansen et al. (2010), Japec et al. (2015), NAS (2017a).

Table 3. Data Quality Framework for Assessing Data Sources' Fitness for Policy Research

115. Data Quality Aspect	116. Description
117. Accuracy	118. Data values reflect their true values (low measurement error). 119. Data are processed correctly (low processing error). 120. Concept measured is concept of interest (construct validity). 121. Data are representative of population (external validity).
122. Relevance	123. Data meet requirements of users to study topic of interest.
124. Timeliness	125. Data are available when expected and in time for policy purposes.
126. Accessibility **127. Clarity** **128. Transparency**	129. Data can be readily obtained and analyzed by users. 130. Data and statistics are presented in clear, understandable format. 131. Methodologies for data preparation and statistical processes are available.
132. Coherence **133. Comparability**	134. Enough metadata is available to understand data structure and allow for combination with other statistical information. 135. Data over time and from different records or sets of records reflect the same concept.

Data accuracy is often the focus of statistical assessments of data sources, sometimes referred to as total survey error (TSE) in the context of surveys (Biemer 2016). TSE considers different errors that can affect inferences, including:

Sampling error: Error due to having a subset of cases from a larger population. This error can either be variable, due to having a random sample, or systematic, when certain cases are more likely to be included in the sample and estimates are not appropriately adjusted.

Measurement error: Errors arising from recorded data values, differing from their true values.

Processing error: Errors introduced by the processing of data, including any transformation, editing, and coding.

Coverage error: Errors when units in the population have a chance of either not being in the sample or of being in the sample multiple times.

Nonresponse error: Error in inferences due to missing entire or partial responses for records. When nonresponse is related to the missing data values, inferences can have systematic errors.

Data accuracy is directly related to veracity in the context of Big Data and is often a concern for a Big Data source. Since the data are convenience samples and not collected for statistical purposes, systematic sampling error and coverage error can be major concerns, leading to inferences that are not generalizable to the population of inference. Further, the variables available from a Big Data source may not directly correspond to those of interest for policymaking. On the other hand, Big Data sources sometimes have lower measurement error than surveys, as the data come directly from a record or transaction and are not dependent on the recall of a survey respondent.

Lazer et al. (2014) addresses the mistaken belief that the volume of data compensates for any deficiencies in data quality, referred to as "Big Data hubris." In fact, measurement, construct validity, representativity, and other elements contributing to data accuracy remain just as important. To demonstrate this, Meng (2014) compares the mean square error (MSE) of a simple random sample (SRS) of size 100,

producing statistically unbiased estimates, to a Big Data source with millions of records. The Big Data source has certain cases that are more likely to be included than others, leading to systematic error and statistical bias. Specifically, let there be a weak 0.1 correlation in the population between a variable of interest and whether a response is present for that variable in the Big Data source. Meng shows that the Big Data source can have MSE hundreds of times as large as the SRS of 100. Statistical bias is the main contributor to the MSE in this instance. Unless the Big Data source includes records for 50 percent or more of the population, the Big Data source's MSE will be less than the SRS's. Further, with the same parameters, the Big Data source would need to contain 96 percent of the population to produce an estimate with lower MSE than a SRS of 2,400.

A number of other challenges emerge for the quality of a Big Data source because the data are not collected for statistical purposes. Even when the data are highly relevant for the policy topic, the methodologies and processes by which the data are produced may not be clear or transparent. Changes in how a dataset is curated mean that data may not be comparable over time or across different kinds of records. There also may not be enough metadata available to use the data with other statistical information available. This lack of transparency and continuity will bring into question any of the statistics used to inform policy and evaluate change over time.

Big Data, however, often have the advantage of timeliness or velocity. Traditional data collection can take time to collect, process, and review before estimates are available to policymakers. Some Big Data sources can provide data at enough speed to allow for rapid estimation. When the validity of these inferences can be ensured, Big Data can offer new opportunities for more timely policy action and intervention. Further, the information available from a Big Data source is sometimes more relevant for policy purposes than what can be collected in a census or survey.

Fitness for Use

There are a variety of uses of Big Data sources, and different uses require different strengths from the data source. Therefore, any evaluation of a data source for health policy research must depend on the context. Common statistical uses of alternative data sources include: for direct estimation, for record linkage, to assist with design and calibration of surveys, for imputation, as a second survey frame, and for small area estimation. Some reviews of statistical uses of alternative data sources are provided in Johnson, Massey, and O'Hara (2015) and Lohr and Raghunathan (2017).

This review views the potential of Big Data sources to be used for health policymaking from the perspective of "fitness for use"—that is, whether the data have the strengths needed for a specific use. For example, to estimate the prevalence of a disease in a population, data accuracy and veracity are critical, and timeliness may be less important. For surveillance, by contrast, timeliness is relatively more important compared with accuracy. In order to respond to a need where an epidemic may be emerging, it is relatively more important to have a timely indication than a statistically accurate inference.

Analytical Challenges with Large Datasets

Large datasets yield unique analytical challenges, particularly for high dimensional datasets with many variables. The sheer volume of data can lead to patterns emerging in the data source that are not

meaningful. Fan et al. (2014) reviews three challenges that commonly emerge in statistical analysis of large datasets:

- **Spurious correlation**
- **Noise accumulation**
- **Incidental endogeneity**

Spurious correlation refers to when a high dimensional dataset contains a set of variables that are uncorrelated in the population, but, due to the large number of variables, some have high correlations in the sample dataset. The larger the number of variables, the more frequently spurious correlations tend to occur. This means that relationships found in the dataset may not have statistical conclusion validity. Thus, when analyzing a high dimensional dataset, it can be critical to adjust statistical conclusions for multiple hypothesis tests.

Noise accumulation refers to the scenario where a predictive model with Big Data has high classification error due to the inclusion of a large set of independent variables that are unrelated to the outcome of interest. This is because the spurious correlations in the data used to fit the model are not meaningful for classification or prediction. When a high dimensional dataset contains some variables with meaningful correlations with the outcome and other variables with spurious correlations, the inclusion of variables with spurious correlations overfits the model and decreases the classification accuracy.

Incidental endogeneity also occurs in high dimensional settings and can lead to incorrect scientific discoveries. Endogeneity refers to when one or more independent variables in a statistical model are correlated with the residual error term. This can result from an uncontrolled confounding variable causing both the dependent variable and an independent variable, or reverse causality of the dependent and independent variable. This is a critical assumption of many statistical models and can invalidate scientific interpretations of model coefficients. In a high dimensional dataset, incidental endogeneity can occur by chance because there are so many independent variables in the model. Thus, the interpretation of model coefficients must be conducted with care for high dimensional datasets due to incidental endogeneity.

Two sets of methods that can help address spurious correlations and noise accumulation in high dimensional data sources include dimension reduction and variable selection. Dimension reduction involves reducing the number of random variables under consideration for analysis by obtaining a set of principal variables that are combinations of variables. Some commonly used techniques include principal components analysis and linear discriminant analysis. Variable selection involves fitting a statistical model with a penalization for including more variables in the model, a process referred to as "regularization" (NAS 2017c). Some common examples are Ridge regression and the Least Absolute Shrinkage and Selection Operator.[6][7]

[6] For further discussion of these methods, see Friedman, Hastie, and Tibshirani (2001) or Fan, Han, and Liu (2014).

[7] Challenges to statistical inference due to spurious correlations, noise accumulation, and incidental endogeneity can be compounded when a Big Data source lacks requirements needed for data accuracy or veracity. See Biemer (2016) for details.

Big Data and a Fast-Changing Technological World

In addition to the computational challenges of analyzing Big Data sources and the extensive data cleaning that can be required, there are important issues involved with analyzing Big Data sources in a world with rapidly changing technology. Further, lack of transparency in how a Big Data source is produced and maintained can present challenges for analyses (Lazer and Radford 2017).

Business Processes and Big Data

It is critical for the organization(s) maintaining Big Data sources to provide researchers and government agencies with detailed information regarding data maintenance needed to understand data quality. First, there can be barriers for researchers or government agencies to work with and obtain the full detailed documentation and metadata needed to assess the usefulness of a Big Data source for policymaking. Businesses typically do not have expertise in policy research and can be unaware of what is needed from a data source to be used for policymaking. Further, a business or other organization may have a financial interest in keeping their methodologies secret, compromising the clarity or coherence of the Big Data source. These organizations may adjust their data curation and maintenance processes over time to fit business needs, thus compromising the comparability needed to use the data for policymaking purposes.

One theme among the successful use cases reviewed in this study is the use of standardization to improve the quality of Big Data sources. When data are used from different organizations that are not collecting the data for policy purposes, having requirements for how the data are maintained and for providing clear documentation can help improve several dimensions of data quality. When data are combined from multiple organizations, standards help improve the comparability and coherence of the ultimate data product. Standardization can also guide what quality control checks are needed to verify the processes and quality of a Big Data source. Standardization can be critical for making Big Data sources usable for health policy research.

Algorithm Dynamics

Additional challenges can arise with using Big Data sources when the evolution of users' inputs to the data production system is not understood. Taking the example of social media, the way users interact with social media platforms and the characteristics of users can change rapidly and frequently. Companies maintaining social media websites may alter their algorithms to adapt to such changes. Because these processes are often undocumented and opaque, understanding the data quality of such data sources is particularly challenging.

Thus, algorithm dynamics (Japec et al. 2015, Biemer 2016) can be a concern for several kinds of Big Data sources. One particular example comes from Google Flu Trends, which used Google searches to track flu prevalence in different areas of the United States. Lazer and Radford (2017) believe it was likely that Google changed its search algorithm at a certain point to make it easier for users to find health-related information. This is believed to have led users to change their patterns of how they searched for the flu, which harmed the comparability of Google's estimates of flu prevalence over time, a possible cause of Google's overestimate of flu prevalence in 2013. Thus, algorithm dynamics can greatly harm statistical validity.

Ideal User Assumption

Additional challenges can emerge when the "ideal user assumption" is violated (Lazer and Radford 2017). In typical data sources, records reflect single, unique people who express themselves honestly in the data. However, users can easily misrepresent their identity and/or have multiple accounts. Further, much traffic on sites is generated by bots, which may or may not be malicious actors, rather than human users. There is even a possibility for users with some understanding of how a Big Data source is being used for policy purposes to intentionally corrupt the data. As will be discussed, ability to verify the ideal user assumption is particularly pronounced in user-generated data sources. By contrast, surveys and censuses, while subject to some false reporting, exercise great control over data collection and data source inputs.

Implications for Policymaking

Government agencies developing official statistics and formulating policy depend on their reputation as fair, impartial, objective, and neutral (Kitchin 2015). If an agency cannot provide data or analysis with adequate data quality and documentation, then the public can lose confidence in the agency's ability to provide evidence for policymaking. When Big Data sources do not have adequate transparency and data analysis may be affected by algorithm dynamics or violations of the ideal user assumption, agencies will not have sufficient understanding of the data to guide fair, objective policy development. Agencies must be risk-averse in adopting new methods or working with new datasets, as the understanding of the issues with certain Big Data sources is still evolving.

Public Sector Data

Considerations for Public Sector Data

Public sector sources of Big Data have long been used alone or in conjunction with sample surveys to support policymaking and evaluation. It is, however, only recently that they have been "labeled" as such, and it has become more straightforward to use them as a primary or ancillary source of data because of improvement in timeliness and data quality. Administrative data, the primary form of public Big Data, is one of the most critical and promising sources of public sector data as there has already been a great deal of research and collaborative development of administrative data as a tool for policy analysis. Several recent developments have pushed administrative data into the forefront of efforts by policy analysts to harness Big Data. Two National Academy Reports (2017a, 2017b) have emphasized the role of the coordination and integration of the government's resources, including administrative data sources to improve the inferential quality and coverage of extant data. These reports recommended centralized access to administrative data from all affiliated agencies.

The U.S. Census Bureau, moreover, which has long used administrative records to improve and expand the federal data systems, received federal funding in 2016 to build upon their well-developed Federal Statistical Research Data Center (FSRDC) to help provide broad access to administrative records from all state, local, and federal agencies willing to participate in record access (Jarmin and O'Hara 2016, Lane 2016). One of the primary recommendations made by the Commission of Evidence-Based Policy Making is to establish a National Secure Data Service (NSDS). The NSDS would be similar to the

FSRDC concept, working as a service to facilitate secure access and data linkage across the government to facilitate systematic access and usage of administrative records for statistical purposes.[8]

Administrative data, by and large, have as their origin information collected for administrative, regulatory, or law enforcement purposes. They are records of transactions that are required either by law or to provide services. As "organic" data, administrative data are not originally designed to be used for statistical purposes, but rather are a byproduct or digital trace of other activities. Examples of administrative data that have subsequently been used for statistical purposes are numerous. Transactional data such as Medicare enrollment and claims data have long been analyzed to understand health care among the elderly. Social Security earnings and benefits have been used to assess work history. Uniform crime statistics, based on voluntary reporting from police departments, are analyzed to understand crime and victimization. Registries, another form of administrative data, often have their origins in either mandatory reporting for notifiable diseases such as HIV (such as the Enhanced HIV/AIDS Reporting System) or voluntary state-level reporting of childhood immunization (such as the Immunization Information Systems). States and local municipalities also collect and use data from federal programs that are regulated and funded by states and municipalities such as the Supplemental Nutrition Assistance Program and Medicaid. Both programs have recently undergone changes that have improved federal access and data quality that will provide the federal government with a more comprehensive picture of these federally mandated programs.

All major statistical and regulatory agencies of the federal government look to administrative sources to characterize the populations of interest in aggregate tabulations or as full replacement for survey data where the populations can be fully characterized by the administrative record. In recent years, because of increased automation, improved data quality checks, and harmonization efforts, administrative data have become more robust and timely for policy analyses. Current uses of administrative data for statistical purposes within the federal system, however, have been most successful when used in operational conjunction with or as a complement to survey data. Since administrative data are likely a full census of all participants or transactions for a federal program, they are often used as sources of data validation or supplemental data for extant surveys. Program enrollment and eligibility rosters are often used as sampling frames, either as the source of complete information or ancillary information for identifying the inferential population for sample surveys.

Similarly, administrative data can be used to improve statistical estimates in post-processing through imputation models or nonresponse adjustment, both of which are substantially improved by the availability of ancillary information on missing data items and survey nonresponders. This is particularly true if the sampling frame used for the survey is from the same source. As importantly, aggregate tabulations from administrative data provide important sources of data validation for surveys where survey estimates of enrollment characteristics can be systematic compared to aggregate tables generated from the original sources.

One of the most important statistical uses of administrative data is through data linkage and the production of blended statistics in which the administrative data provides either ancillary or alternative

[8] One notable difference of the NSDS from the FSRDC model would be the facilitation of temporary data linkages for research purposes as opposed to operating as a data warehouse storing linked data long term.

measures to extend survey data. Registries and administrative records can provide passive follow up in cohort studies for disease incidence and mortality. Transaction data can provide administrative detail for self-reported outcomes such as medical events and costs. Citro (2014) and Lohr and Raghunathan (2017) describe three specific ways in which administrative data and survey data can be analytically linked. First, and perhaps most widely used, is individual record linkage in which the data from administrative records are thought to be a direct match of the individual from whom data were collected in a survey. There are many examples of individual record linkage, the mostly widely used of which is to link survey records to the National Death Index. Deterministic and probabilistic methods are used and have been substantially refined to minimize record matching error, which can occur because of errors and omission in the origin data sources.

The second blended use of administrative and survey records is to add or correct single data items in the survey or administrative record by combining individual fields. The source of error in this case is item or unit non-comparability between the data sources, which can impede effective statistical harmonization. The third use of administrative data is to extend the inferential use of survey data to smaller levels of geography—providing policy analysts with information that is applicable to the local area. These estimates use administrative data at various levels of geography as covariates in sophisticated models to correct for local area population compositions. Error is introduced by model assumptions and incomplete measurement.

Administrative data are characterized by substantial sources of error and, in part, because of their organic nature. Using the data quality framework described in Table 3, administrative data suffer primarily from limitations in accuracy, timeliness, accessibility, and comparability. Many administrative data systems are not well designed or standardized, and many lack quality control and attention to missing items or records. This is particularly true of administrative systems that were born in pre-digital eras and have only recently been transformed for digital transmissions and inputs. The lack of comparability between records or lack of standardization will slow down the production of analytically sound data and introduce long delays from the relevant data year and the availability of data for analysis or linkage. Additionally, when state and local data are being combined at the national level, the lack of a standardized data structure, measurement, and method may substantially hinder harmonization and slow down data availability. Finally, administrative data are often collected under circumstances where later statistical use of the data has not been envisioned. The data may be substantially protected by law (CIPSEA, HIPAA, or Title 13), and agencies may interpret this as restrictive. Similarly, data linkage may be substantially hindered by concerns about confidentiality when personal identifying information (PII) is necessary for linkage.

Uses of Administrative Data for Health Policy

There are numerous examples of successful efforts to integrate administrative records into ongoing survey data collection in the Federal Statistical System. The long history of using administrative data in support of survey systems and as a source of linked records is well documented. Nevertheless, two current data systems, which are used for monitoring the public's health, health care utilization, and sources of health insurance coverage and payment, provide an interesting and effective contrast on how administrative data can be linked together and the consequences for inferential quality. In part, the contrast between the two examples arises from the design of the underlying data collection, but also from differences in access and use of the supplemental administrative data. The National Health Interview Survey (NHIS) linked files

and the Medicare Current Beneficiary Survey (MCBS) both provide information about health, health care use, and barriers to care through a combination of in-person survey questionnaires and linked Medicare enrollment and claims data. The NHIS is a post-hoc linkage of a household survey to administrative records, while the MCBS is a survey designed with data linkage in mind, as the sampling frame is the enrollment data from the Medicare program. The NHIS–Medicare linkage is an example of the first use of linked blended data sources—it supplements an individual respondent record with information from claims and enrollment files. Additional measures are added to a selection of records that are capable of being linked. The MCBS conversely is an example of the second type of use made of administrative records, which is to add individual items from claims and health insurance plan enrollment to improve and correct items collected from an individual. The administrative records serve as verification and correction of data collected from respondents.

The National Health Interview Survey

The NHIS, an in-person annual cross-sectional household survey conducted by the National Center for Health Statistics (NCHS), has long been used as a benchmarking survey for measures of population health and well-being. The NHIS is used throughout DHHS to monitor Healthy People goals, which are 10-year objectives for improving Americans' health and monitoring trends in health and disability. It has been in continuous data collection since 1957 and is also widely used by analysts to understand the epidemiology and etiology of many acute and chronic diseases. Along with other surveys at NCHS—including the Longitudinal Study of Aging, National Health and Nutrition Examination Survey, National Home and Hospice Care Survey, and National Nursing Home Survey—the NHIS is linked to Medicare enrollment and claims under an interagency agreement with the Centers for Medicare & Medicaid Services (CMS) (NCHS 2017). This is the third such collaboration. Previous linkage for the NHIS was facilitated by ASPE and the Social Security Administration.

The most recent linkage for the NHIS provided individual record matches between the 1994–2013 NHIS survey respondents to a variety of Medicare eligibility and claims files, including the Master Beneficiary Summary File, which is an annual file that contains demographic and enrollment information for beneficiaries enrolled in Medicare in the calendar year (including segments associated with enrollment in Part A/B and Part D, and Cost and Utilization and Chronic Conditions segments, which summarize utilization and Medicare payment and the presence of chronic health conditions, respectively). Additionally, Medicare utilization files are linked and include summaries of inpatient stays: Medicare Provider Analysis and Review, Part D Prescription Drug Events, Outpatient files, Home Health Agency, Carrier (summaries of physician claims), and Durable Medical Equipment. The linkage was done in the CMS Virtual Research Data Center for eligible NHIS survey participants. Deterministic methods of record linkage are used to make the linkage with variations in the methods of linkage depending on the completeness of the PII provided. For those persons found to be eligible in a previous round of linkage, approximately 98 percent of records were matched deterministically (Zhang, Parker, and Schenker 2016).

To be considered eligible, an NHIS respondent must have provided consent as well as PII, such as a full or partial Social Security number (SSN) or Medicare Health Insurance Claim (HIC). During an earlier round of linkage activities, NCHS considered a refusal to provide an SSN as a refusal to consent to linkage. The combination of a decline in response rates to the NHIS and an increase in the proportion of respondents who refused to provide SSNs led NCHS researchers to investigate the value of a partial SSN

match, as well as separable consent for those who refused to supply a SSN or HIC number (Dahlhamer and Cox 2007). This revision in 2007 has improved the number of respondents eligible for linkage, as well as the proportion who were matched. For the 2006 NHIS, 22.7 percent of the total sample age 65 and over where linked to the Medicare administrative data, which was down from 43.6 percent in 2005. Those figures rose to 44.3 percent in 2007 and 51 percent in 2008, in part as a function of the change in methods of gathering PII and informed consent.

Like other record linkage attempts, the NHIS Medicare administrative data linkage creates datafiles that are enriched by the linkage, but also subject to error due to a variety of factors that include (as noted above): 1) records not linked due to missing PII, 2) item missingness due to incomplete coverage of administrative records, and 3) missingness created by changes in program eligibility and program characteristics that lead to inconsistent data sources. Zhang, Parker, and Schenker (2016) used an earlier version of the NHIS–Medicare claims link to understand and compensate for these sources of error by multiple statistical imputation. They use as an example an estimate of the annual prevalence of mammography for women over 65 for the 2004–2005 NHIS respondents. The NHIS asks ever prevalence of mammography and relies on self-reported data, whereas the Medicare claims data provides information about annual claims for procedures conducted. Thus, claims data provide a better measure of the true annual incidence of mammography, but the linkage is incomplete for a variety of reasons. Less than half of the women age 65 and over in the NHIS are eligible for linkage due to consent or PII issues. Additionally, women enrolled in managed care (MA) plans for Medicare (approximately 20 percent in 2006) do not have detailed claims and, therefore, no record of mammography from claims is available. Finally, eligibility gaps or death may limit the records available to identify the appropriate claims. While the paper successfully imputes annual rates of mammography for Medicare beneficiaries, it required substantial attention to the sources of error and the potential inferential limits of linked datafiles for statistically sound estimates.

Medicare Current Beneficiary Survey

In contrast to the NHIS-Medicare linkage, the MCBS begins with the Master Beneficiary Enrollment File as the sampling frame, and its respondents are completely matched to claims files by design. The original design was premised on a full partnership between the survey data collection and the administrative records. The MCBS is a continuous, in-person, multi-purpose longitudinal survey covering a representative national sample of the Medicare population, including the population of beneficiaries age 65 and over and beneficiaries age 64 and below with disabilities, residing in the United States and its territories. The MCBS is designed to aid CMS in administering, monitoring, and evaluating the Medicare program. A leading source of information on Medicare and its impact on beneficiaries, the MCBS provides important information on beneficiaries that is not available in CMS administrative data and plays an essential role in monitoring and evaluating beneficiary health status and health care policy. Respondents for the MCBS are sampled from the Medicare administrative enrollment data. The sample is designed to be representative of the Medicare population as a whole and by the following age groups: under 45, 45 to 64, 65 to 69, 70 to 74, 75 to 79, 80 to 84, and 85 and over.

As part of data collection, respondents are asked detailed questions that focus on use of medical services and the resulting costs, and are asked essentially the same way every time a section is administered. The respondent is asked about new health events and to complete any partial information that was collected in

the last interview. For example, the respondent may mention a doctor visit during the health care utilization part of the interview. In the cost section, an interviewer will ask if there are any receipts or statements from the visit. The interview also includes sections about health insurance. During each interview, the respondent is asked to verify ongoing health insurance coverage and to report any new health insurance plans. During three rounds of data collection every year, respondents are asked to provide a full accounting of all health care visits, medical encounters, and expenses and, then, to detail the amount each activity costs and who provided payment—be it Medicare, other private, or public insurance plans—or if the cost was paid out-of-pocket.

Designed in 1991, the goal of the MCBS was to extend the government's understanding of how Medicare beneficiaries received and paid for care. As health care became more expensive, it was critical for policy purposes to understand all costs and sources of payment. The expansion of supplemental insurance and the rise of out-of-pocket costs means that a claims-only approach to characterizing health care costs among Medicare enrollees would not sufficiently characterize the entirety of their costs. Additionally, both respondent recall of events and costs are notably subject to bias. Beginning in 1992, the MCBS began linking the survey data directly to enrollment and claims datafiles through a direct matching process and subsequent reconciliation of the costs of care with adjutant imputation. Data are collected for both Medicare and non-Medicare covered services in the interview and later matched and reconciled with a direct match using a unique Medicare beneficiary ID.

Unlike the NHIS, the MCBS does not suffer from linkage error because of the direct match made possible by identified records on both sides of the match (Eppig and Chulis 1997). The MCBS rather suffers from matching error associated with missing or incorrect data on the survey or claims side. The matching process uses the survey data, which is reorganized to resemble claims files, with dated events that are used to link to Medicare claims records. Records that include a Medicare claim number are matched directly on the claim number, while the remaining records are matched based on an iterative method that aligns service date, event type, and provider. The resulting file contains data for medical event types and services and contains fields for survey only, claims only, and survey and claims combined. The final payment amounts and source are generated from a combination of the available data.

The sources of error in the estimates arise, in part, from the same source of error in the linked NHIS–Medicare claims files. Medicare Advantage participants (now approximately 30 percent of Medicare beneficiaries) do not have claims files. Thus, there are only survey file reports for the cost of care for persons who are enrolled in MA plans. MA enrollees are, in fact, likely different from those enrolled in traditional fee-for-service (FFS) plans. State-level variation in enrollment in MA plans in 2015 was quite large, with Medicare beneficiaries in states such as California, Hawaii, Minnesota, and Oregon at or just below 40 percent of beneficiaries and states such as Nebraska, Illinois, and Maine under 20 percent of enrollees (Kaiser Family Foundation 2017). Match error, where the claims record or the survey report is incomplete, adds additional room for error and is likely not independent of the health of the individual whose recall of events and dates, as well as the likely payer, may be problematic.

Substantial research using the MCBS, including Park et al. (2017), which examines the potential strategies health care providers who offer Medicare Advantage use to shift high-cost enrollees off their plans, relies on the accuracy of the matches and the quality of the enrollment data. Park et al. (2017) use information about plan switching and the health of MA and FFS beneficiaries to assess whether MA plan

providers are "pushing" respondents to traditional plans when their health declines. This analysis, which has substantial policy relevance, depends on the match quality and the data quality in order to draw this inference. It is critical in all analyses of matched data of this type to understand the sources of error.

Private Sector Data

Considerations for Private Sector Data

Private sector data, with some exceptions, has not been traditionally used in policy research and evaluation because they lack important qualities that make them fit for use. In the two volumes from the National Academy of Sciences review (NAS 2017a, 2017b), the committee lays out both criteria for classifying private sector data and a quality framework for understanding their use. The variety and complexity of privately held data prevent easy summary or assessment of their overall usefulness for policy. As with the publically held administrative data described above, privately held data are generated for diverse purposes that often do not meet the basic standards for data used for statistical purposes by the federal government.

Current uses of privately held data are more widespread internationally as many countries have more substantial access to data from private sources than does the United States. Statistics Netherlands, for instance, has organized and captured traffic sensor data, which has become ubiquitous enough to provide nearly complete coverage of national roadways, to characterize traffic and road conditions nationally in real time. The sensor data are processed and concatenated to produce national estimates of traffic flow (Puts et al. 2016). Another example are recent efforts to generate Consumer Price Indices (CPI) to assess inflation in 22 countries, using web-scraped prices for 5 million items daily to track price shifts. These statistics are being considered as a source of national CPI by many statistical agencies worldwide. Validation exercises and ongoing assessment of data quality and cleaning are currently being used to assess fitness of use (Cavallo 2017; NAS 2017a, 2017b).

In the United States, statistical agencies such as the Bureau of Justice Statistics (BJS) and the Bureau of Labor Statistics (BLS) are experimenting with new sources of data to augment existing statistics. BJS, in the redesign of the Census of Arrest Related Deaths, conducted in the 2015-2016 data year, began to use web-scraped news articles from a variety of sources to develop a broader canvas of information about deaths to persons arrested or in custody (Banks, Ruddle, and Kennedy 2016). BLS uses data from retail scanners, web-based price scrapping, and JD Power car prices to adjust and calculate the CPI (Horrigan 2013). Their use of retail scanner data to augment traditional price gathering mechanisms began in the late 1990s, but has expanded as access to retail scanner data has been routinized by several private market research firms.

Classifications of private data are helpful to understand some of the quality challenges that may limit their use for policy analysis. The NAS volumes classify data sources by the degree to which the data are structured, standardized, and uniform in nature. Structured data in the private sector, that is data that share common fields with defined lengths and known and agreed upon characteristics, include sales data from retail transactions, which are largely structured by the Universal Product Code, or residential real estate information available from sites such as Zillow, which is structured by the Multiple Listing Service and

legal requirements for real estate transactions. Structured data have the benefit of available metadata and more limited requirements for data processing and cleaning, thus making them easier to aggregate, disseminate, or use as input to other estimates. As discussed in the next section, mobile phone data and GPS tracking data are also highly structured and share common metadata and thus are an easy source of complimentary data for policy research.

Semi-structured data lack the implicit shared organizational structure, but they coexist with metadata or business rules that can be used to process the data. Twitter data, for instance, is semi-structured in that there are metadata fields such as time, date, and hashtags that can be used to provide a method for structuring some content and providing methods for summarizing or searching certain fields. Finally, unstructured data such as videos, pictures, or unstructured text on social media may not share a common set of characteristics partly because of the way in which the digital object is created and partly because there are no agreed upon standards by which data can be regularized. Structuring the data, then, becomes both an exercise in regularizing data for analyses as with the semi-structured data, but also identifying the shared structure empirically and building the data standards. Not surprisingly, most efforts to integrate Big Data into ongoing data systems in the federal government focus primarily on structured and semi-structured data with agreed upon standards.

Aside from issues of data standardization, privately held data present additional challenges to their use for policy. First and foremost, access to private data may be quite limited as data are often viewed as a business asset. Second, a lack of transparency and documentation often render privately held data unfit for use for statistical purposes as the information necessary to provide the public with an adequate explanation of the sources and limitations of the data is not possible. Third, private entities tend not to share similar technologies or data elements so that aggregation across vendors or users is quite difficult. This limits the generalizability of the data beyond a single vendor or user base if they cannot be systematically integrated. Finally, data quality is always a challenge as private data are collected, optimized, and used for purposes other than statistical inference. As such, there may be little incentive to impose quality control standards for items such as demographics or place of residence if neither represents an important determinant of the success of the product or process.

One of the most challenging and dynamic private-public partnerships has been the rapid adoption of electronic health records systems. The Health Information Technology for Economic and Clinical Health Act of 2009 and the introduction of the Medicare Electronic Health Record (EHR) Incentive Program, which provides incentives and penalties to eligible clinicians and hospitals to adopt EHR, changed the fundamental data landscape of private and public health care. This combined, in the early years of the Affordable Care Act, with various stages of Meaningful Use certification, pushed most health care providers such as hospitals, practice-based physicians, and clinics into the use of EHR. Based on a supplement to the American Hospital Association Survey, in 2009, 12.2 percent of acute care non-federal hospitals had functioning basic EHR systems, but by 2015, 96 percent reported having certified technology (Swain et al. 2015). Similar rates of rapid adoption have occurred among physicians, with 76 percent reporting use of a certified EHR systems on the 2015 National Electronic Health Records Survey (ONC 2016).

This swift and universal adoption of electronic means of providing health care have led many to speculate about the future of integrated medical care and research, as well as call for the substantial integration of

electronic medical records into policy research (Mooney, Westreich, and El-Sayed 2015, Binder and Blettner 2015). Examples of recent research include the standardization of digital breast imaging data through text mining (Margolies et al. 2016) and predictive modeling with machine learning of hospital readmission rates for heart failure (Shameer et al. 2017). In the following sections, we discuss important case studies in which data standardization, shared technologies, and careful integration have begun to broaden the scope of collaboration between private health care providers, researchers, and policy analysts. Some additional collaborations, which rely heavily on the maturation of this private-public partnership, have yet to come to fruition, including the National Institutes of Health's (NIH) *Cancer Moonshot* and All of Us Precision Medicine Initiative, but will be heavily dependent on the ongoing efforts at standardization of data items and exchange.

Uses of Private Sector Data

In this section, we describe a use case of private data for policy purposes. The Health Care Cost and Utilization Project, is a long-term successful collaboration between the Agency for Health Care Research and Quality (AHRQ), states, and private organizations and hospitals to provide individual-level encounter data from hospitals in almost every state in the nation. It evolved prior to the advent of the widespread adoption of the electronic medical records but has become substantially successful by ameliorating many of the challenges public-private partnerships present for the development of effective data collaborations.

The Health Care Cost Utilization Project (HCUP)

HCUP was developed beginning in 1988 using an administrative data source, the hospital discharge record, whose main purpose is for billing, but also contains basic demographic information, services provided, disease status, and the cost of services and payers. The data source is an all-payer data, as it contains a complete census of patients discharged from the hospitals. States, private organizations, and municipalities receive this information either as a voluntary donation or because of state mandates (Steiner, Elixhauser, and Schnaier 2002). The partnership has expanded dramatically since the initial collaborations. A total of 49 states now participate in at least one of the data products, with some states' participation occurring through the state health departments (such as Florida and Illinois), state hospital associations (such as Indiana and Iowa), and others through affiliated organizations (Connecticut and Hawaii). The HCUP partners act as a liaison to the hospital to gather the discharge records and communicate with the AHRQ to provide standardized data.

The full roster of data products now include a substantial expansion of the original Nationwide Inpatient Sample (NIS) (1988), which is an annual representative sample of records for the hospitals in the annual file. The sample includes over 7 million hospital discharges with data-related diagnoses and discharges, as well as the severity of the discharge disease, and can be used to make local, regional, and national estimates of hospital costs and services. The size of the datafile allows researchers and policymakers to identify rare diseases and costly hospitalizations, as well estimates for states and local areas for those states that provide additional geographic information. This core NIS annual datafile uses data items that all states share in common. The State Inpatient Database (added in 1995) is the universe of all data items and all discharges for the participating state. These files contain data items that are standardized at the state level, but may be missing from the national level. They are useful for state health policy and regulation. Additional cuts on the datafiles have been created since the addition of this data product,

including The Kids' Inpatient Database (KID), which is a national sample similar to the NIS for children discharged from the hospital, as well as nationwide samples for emergency departments and readmissions (added in 2013). State-level data are available for ambulatory surgery and outpatient services as well as emergency departments. The degree to which states participate in all of these data products is variable, but all have data in the KID and NIS files (AHRQ 2018).

Policy and policy-related research have been successfully based on HCUP, both by government researchers and academic researchers who have access to the data. In 2016, approximately 500 research articles were published from the NIS sample alone (Khera and Krumholz 2017). ARHQ issues annual data summaries from the datafiles that describe the national trends in hospitalization rates, treatment, costs, and readmissions. The success of the collaboration has a great deal to do with how it is operationalized. First and foremost, the items shared from all the partners are based on a standard digital format with agreed upon parameters. These shared fields are then edited, cleaned, and realigned by AHRQ. The HCUP partners receive back from the collaboration both edited files, as well as technical assistance in developing and maintaining their data. The hospitals that participate, many of whom are in the private sector, receive in return free data cleaning support and technical assistance. Moreover, AHRQ provides comparative assessments of data quality, measurement, and effort. As part of the partnership, they provided a software tool to transition from ICD-9 to ICD-10 coding in records.

The success of the collaboration is also a direct consequence of simultaneous national harmonization and standardization and state and local data and confidentiality and data sharing requirements. The State Inpatient Databases files return to the states and organizations all of the customized data sources necessary for state-level policymaking, while the NIS puts trends in national perspective. Moreover, the states and organizations work with the hospitals to determine the level of detail they are willing to submit and which datafiles they will contribute to. This flexibility makes the partnership easy to maintain as the partners feel that the hospitals they represent in these data transactions are protected.

User-Generated Data

Considerations for User-Generated Data

User-generated data, which we define as data reflecting direct user interactions with a website, platform, product, or service, and reflecting digital life, present some different challenges for informing public health policymaking than the previously discussed two sets of data types. We include a diverse set of data types in this category: social media, data produced by mobile phones—sometimes with GPS data, reports on online message boards, data collected by web scraping, data from environmental and health sensors, data produced by the Internet of Things, and many others.

Much of this category of data types encompasses data resulting from online interactions. In general, the data can have both very high volume and velocity. The volume of data may allow for monitoring trends in different geographic areas more easily than surveys or censuses. Collecting user-generated data can be affordable and rapid.

However, due to some substantial challenges, there are fewer mature uses of user-generated data for policymaking. The veracity of user-generated data can be questionable and difficult to ascertain. Users of a service or website are often not representative of a population of interest. For example, it is well known that younger generations tend to use the internet more than older generations. Further, datasets may have coverage error. According to estimates from the 2015 American Community Survey, 13 percent of U.S. households do not have a computer and 23 percent do not have any internet subscription (Ryan and Lewis 2017). Thus, a large set of U.S. households are not covered by data sources relying on internet use.

In addition, user-generated data may be the most affected by technological challenges. Algorithm dynamics may be a concern when a platform changes its algorithm, causing more searches or uses of a keyword of a certain type. Many of these platforms do not make available metadata needed for transparency about all processes affecting the data source. Inferences relying on the ideal user assumption will mislead when users have multiple accounts or bots account for a large share of traffic.

The standards for using and analyzing user-generated data are not as mature as for the other two data types. Kim, Huang, and Emery (2016) discuss standards for analysis of social media data. Social media analysis often uses queries and filters to collect relevant data for topics of interest. The effectiveness of these filters can be affected by privacy settings, involve complex application programming interfaces (API), and depend on computationally intensive machine learning algorithms. Kim et al. (2016) propose reporting standards for tracking retrieval precision (how much of the retrieved data is relevant) and retrieval recall (how much of the relevant data is retrieved). These standards can help assure that an analysis neither has undercoverage of relevant content nor is so broad as to contain irrelevant information. Kim et al. (2016) also emphasize the importance of transparency of all processes, including describing the data sources and how the data were accessed or collected. In general, these early developments in standardization reflect that the understanding of best practices for collection and analysis of user-generated data is very much still maturing.

In general, there are far fewer successful uses of user-generated data sources for policymaking and in use by government agencies. Lack of representativeness, algorithm dynamics, and violations of the ideal user assumption are among a few of the challenges that are particularly pronounced for these data sources. Uses of data types described in Sections III and IV, in general have much more developed standards for assuring high data quality.

Nonetheless, user-generated data sources have particular strengths due to volume and velocity. The speed with which data become available can allow for real-time insight and rapid reaction to an emerging issue. Thus, the use of user-generated data for early warning systems, surveillance, and monitoring is promising. User-generated data may also be promising for generating hypotheses that can be tested with higher-quality data sources.

The use of user-generated data is an exciting development for public health policymaking and research. The examples we highlight are just a subset of the potential uses of user-generated data for public health policymaking, but were chosen by NORC to reflect the range data types, applications, benefits, and challenges of these data sources.

Uses of User-Generated Data for Health Policy

In this section, we focus on two examples demonstrating emerging uses of user-generated data: the use of mobile phone, GPS, and crowdsourced data for syndromic surveillance, and the use of social media data for adverse drug event monitoring. The choice to highlight these cases reflects NORC's conclusion from the literature review that many of the promising use cases for user-generated data are for surveillance and monitoring. In general, user-generated data sources come with many questions about data veracity that are exacerbated by the challenges of limited transparency, algorithm dynamics, and violations of the ideal user assumption. However, the volume and velocity of the user-generated datasets make them valuable for real-time recognition of and reaction to an emerging issue.

Both of these examples are substantially less developed than hybrid systems of non-user generated public and private sector data that accomplish similar aims—the Centers for Disease Control and Prevention's (CDC) National Syndromic Surveillance Program (NSSP),[9] and the U.S. Food and Drug Administration's (FDA) Sentinel Initiative.[10]

CDC NSSP (Simonsen et al. 2016) integrates electronic health information from emergency departments, urgent care, ambulatory care, inpatient care, pharmacy data, and lab data, with standardized analytic tools to support detection of and rapid response to hazardous events and disease outbreaks. The sheer volume of data helps support surveillance with high spatial and temporal resolution. The BioSense platform allows for cloud-based sharing of health information with tools to capture, store, and analyze data. Standards are set for data cleaning, and quality control checks are included (English 2017).

FDA's Sentinel Initiative (Robb et al. 2012) combines electronic health records, insurance claims data, and registries for adverse event monitoring to ensure safety of drugs and other regulated medical products. A distributed data infrastructure allows for rapid analysis across the database of more than 193 million patients (Popovic 2017). The use of the Common Data Model helps ensure standardization and maintain data quality. Methods are refined to get more precise estimates when an issue is detected, a process called "signal refinement."

The Big Data sources used for CDC's NSSP and FDA's Sentinel Initiative present some challenges, but both are hybrid systems that have mature data standards to support surveillance and monitoring. This section discusses two sets of related examples using user-generated data, demonstrating their promise while reflecting that data standards are not as developed.

The Use of Web, Mobile Phone, GPS, and Crowdsourced Data for Syndromic Surveillance

A number of tools have emerged for using crowdsourcing for syndromic surveillance. Boston Children's Hospital's Computational Epidemiology Group developed HealthMap[11] (Brownstein et al. 2008) to support applications for monitoring and surveillance of disease outbreaks and emerging public health

[9] https://www.cdc.gov/nssp/index.html

[10] https://www.fda.gov/safety/fdas-sentinel-initiative

[11] http://www.healthmap.org/

threats. HealthMap's applications primarily use algorithms to accumulate web-accessible information: news aggregators, eyewitness reports, expert-curated discussions, and validated official reports. The algorithms pull data from these sources through an automated process, constantly updating the system. HealthMap's apps are used by public health departments and government agencies, including the CDC, Department of Defense, and World Health Organization. Outbreaks Near Me[12] is among the most prominent of HealthMap's apps, providing real-time information on reports of disease outbreaks and mapping their locations through GPS data from users.

HealthMap is a Linux/Apache/MySQL/PHP application relying on open-source products and APIs for mapping locations of reports and aggregating information from across the web. A Bayesian machine learning approach is used to automatically tag and separate breaking news stories (Robinson 2003). Duplicate reports are automatically filtered by the algorithm.

Challenges with drawing statistical inferences using HealthMap data include limitations in coverage of news sources, timeliness of the reporting of the sources HealthMap draws from, the limited availability of human reviewers to conduct quality checks on the findings, and questions about the effectiveness of the automated algorithms (Freifeld et al. 2010). Further, HealthMap is limited in its ability to corroborate or verify submitted information. Users can help review and correct submitted data, similar to how Wikipedia is maintained, but challenges remain in understanding the veracity of HealthMap data.

One notable use of HealthMap that emerged for disease outbreak surveillance is Flu Near You (FNY) (Smolinski et al. 2015).[13] Unlike HealthMap's other applications, which aggregate information from across the web, FNY relies on crowdsourced app-based mobile reporting, collecting locations of individuals making reports. Participation is completely voluntary. After a user signs up, the user is prompted weekly to report any symptoms related to the flu. Then, the user is classified as either having influenza-like illness or not. Demographic data are collected upon registration to participate. The 2013-2014 flu season included more than 300,000 reports of flu via FNY.

Smolinski et al. (2015) compared estimates of flu prevalence from FNY to the CDC's official benchmark estimate and found that FNY compared favorably. The researchers found that the estimates improved when first-time users were excluded, to avoid analyzing non-serious reporting, and by using noise-filtering to avoid extreme changes in estimates of flu prevalence. The authors posited that noise-filtering could prevent sharp changes in increased reporting of the flu due to external events, e.g., increased interest in FNY when news stations report on flu outbreaks.

The research noted speed, sensitivity, and scalability as advantages of FNY. The crowdsourced reporting allowed real-time updating of estimates and quick tracking in changes in patterns of flu prevalence. Reporting allows geographic granularity, tracking trends at the zip code level. However, the authors recognize that FNY relies on a convenience sample and is not representative of the U.S. population. Therefore, good performance in the past does not necessarily mean the estimates will perform well in the future. The authors also recognize the possibility of multiple user accounts. Further, the reliance on crowdsourcing could allow malicious users to corrupt the estimates. In general, NORC found that Big

[12] http://www.healthmap.org/outbreaksnearme/

[13] https://flunearyou.org

Data uses like FNY lacked the development of standards, as well as mature thinking about measuring data quality, to assure the veracity of the data.

Related Developments

One other notable use of Big Data that relies on voluntary participants is the NIH Precision Medicine Initiative's All of Us Research Program[14] (Hudson, Lifton, and Patrick-Lake 2015). All of Us seeks to recruit more than 1 million volunteers to contribute their health data and biospecimens to a centralized national database to support research on a range of medical and health questions. The All of Us database will constitute a hybrid system that includes self-reported measures, data from electronic health records, sensor-based observations through phones and wearable devices, geospatial and environment data, and data from social media use, among many other data sources. As compared with FNY, the analysis of the user-generated data included in the All of Us initiative is less challenging as each data source can be directly connected to a specific user. However, All of Us faces the challenges of the representativeness of the volunteer cohort and of the completeness of data sources accumulated in the centralized database.

Social Media for Pharmacovigilance and Adverse Drug Reaction Monitoring

The text mining of social media data can be a promising tool for monitoring adverse drug reactions (ADR) and related events, but like other uses of user-generated data, is also subject to some substantial challenges. Researchers have noted the potential to improve upon traditional ADR monitoring, which can be slow and patchy, often assessed by drug and pharmaceutical companies (Salathé 2016). Further, many are not aware of the FDA's Adverse Event Reporting System (FAERS) for reporting a possible adverse event for further review. Thus, researchers have recognized the potential for Big Data sources to enhance pharmacovigilance, including the development of FDA's Sentinel Initiative.

The recognition of social media platforms as a place where people may share possible ADRs led to investigating using text mining of social media data for digital pharmacovigilance. Freifeld et al. (2014) studied 6.9 million tweets from Twitter and, using a combination of manual and semi-automated techniques, found 4,401 possible ADRs. Although assessing the validity of the findings was difficult, the researchers compared their findings to those from FAERS and found similarities in patterns between the two data sources.

The performance of machine learning and text mining algorithms for analyzing ADRs in social media data can be critical. Yang et al. (2015) describe methods for classifying large volumes of social media messages as either related or unrelated to ADRs. Their approach uses Latent Dirichlet Allocation, a largely unsupervised learning approach that uses a probabilistic model to construct a topic space, assigning messages to topics identified in the messages. Since the posts related to ADRs have similar focuses, while the irrelevant, non-ADR messages discuss diverse topics, the authors advocate using a partially supervised approach using a small number of examples of posts known to be related to ADRs to train the model.

While in the United States, the use of social media as an early warning system for ADRs has largely been explored in academic literature and is less used by government agencies, the European Union's

[14] https://allofus.nih.gov/

Innovative Medicines Initiative has launched a system for Web-Recognizing Adverse Drug Reactions (WEB-RADR)[15] (Lengsavath et al. 2017) through a public-private partnership. WEB-RADR aims to identify new data sources for pharmacovigilance and optimize the aggregation of information on possible ADRs. The program will include development of an EU-wide mobile phone app for reporting adverse events and the development of text mining techniques for publicly available data on social media sites. However, this effort is in an early stage. The WEB-RADR system will involve quality checks on the data collected, including efforts to verify the contact information and identity of online reporters of adverse events.

Related Developments

Social media hold promise for an array of phenomena requiring early detection for response, including for natural disasters (Houston et al. 2015). Activity may be reported in social media, and sometimes earlier than government agencies may otherwise be aware of the event. Social media may also alert to developments at the site of the emergency as events evolve and provide the specific location of individuals needing immediate assistance from emergency responders

Other Uses of User-Generated Data for Public Health Policy: Activity Monitors

Activity monitors and sensors are a promising development for health research. Compared with self-reported health and activity measures, activity monitors in some instances may provide more accurate measurements, as they are not affected by the recall of the participant (Matthews et al. 2012). Activity monitors can be worn most of the day and are not intrusive, allowing a rich set of information about a user's behavior and activities to be collected. However, users of activity monitors are far from representative of the U.S. population.

Wearable devices can be used in research studies and provided to participants in the study to support statistical inferences. Huisingh-Scheetz et al. (2016) discuss research on disability in older adults and using wrist accelerometry. They used a representative sample for the study to support the external validity of findings. However, even after extensive work to identify the right device for the study, questions remained about the quality of the measurements—including measurements' construct validity and the comparability of device measurements across participants. There were also differences in compliance in using the device across participants. In a separate development, an evaluation is underway to evaluate whether data collected from Apple Watches can be used to identify irregular heart rhythms (ClinicalTrials.gov 2017). That research remains in an early stage.

Skills and Capabilities Needed to Manage Big Data

This section identifies technical skills and technological capabilities needed to harness the power of Big Data for health policy research.

[15] http://web-radr.eua

Technical Skills

Managing Big Data requires a unique blend of domain knowledge, technical skills, behavioral attributes, and personality traits from an organization's staff (Booz Allen Hamilton 2018). Instead of looking for elusive candidates who possess all of these qualities, a more successful strategy builds data science teams with diverse capabilities that can complement each other (Olavsrud 2015, Andrade 2017). Such teams require at least four set of roles (Japec et al. 2015):

Domain expert: Member with deep subject matter expertise related to the data, their appropriate use, and their limitations.

Research methodologist: Member trained in formal research methods including survey methodology and statistics.

Computer scientist: Member possessing computer programming and database management skills.

System administrator: Member responsible for managing technology infrastructure that enables large-scale computation.

Domain expertise is required in each step of validating, cleaning, and interpreting the data (Leetaru 2016). As phenomena captured through data are highly fluid and rapidly changing, extensive domain knowledge is required to understand how the data was compiled and to consider the nuances in interpretation. Domain expertise is especially important in relatively new data sources such as social media data, where understanding of data structure and data quality is not yet mature.

Foundational research skills such as survey methodology and statistics are necessary as there is an increased risk of untrained users finding false associations in Big Data. The sheer volume of data can lead to an increased number of patterns emerging that are not meaningful for policy (see Section II.B). Foundational research skills are particularly important in policy as research results often have high impact on the public.

Basic *computer science skills* include solid knowledge of an object-oriented programming language (e.g., Python), ability to work with databases and database languages (e.g., Hadoop), and capacity for quickly learning new skills in a fast-changing landscape. The importance of database management skills increases as the size of data grows.

System administration involves managing computing environments for storing, analyzing, and visualizing Big Data. Many complex decisions need to be made such as determining whether to set up internal compute clusters, outsource it to "utility computing" service providers (e.g., Google Cloud Platform, Amazon Web Service), or to implement a hybrid model. Long-term benefits and costs should also be considered for deciding whether to use open-source or proprietary tools for data analysis and visualization.

In addition to forming such data management teams, agencies should provide continuous training to their staff and judiciously outsource key technical steps to meet their Big Data needs. Most agencies have many staff with extensive backgrounds and strong technical skills in dealing with a variety of data. However, continuous training is required to help them keep pace with the constant update in new methodologies and technologies in the Big Data realm (NAS 2017a).

Outsourcing specialized technical skills have also resulted in successful management of Big Data projects (NAS 2017a). One such project was a pilot to explore new sources of spending data by the Federal Reserve Board. In this project, data from BLS, the Census Bureau, the Energy Information Agency, and the Bureau of Economic Analysis, along with credit card transaction information from the First Data database was combined to produce a timely, independent measure of spending data that tracked the indicators of retail trade as measured by the Monthly Retail Trade Surveys. In another example, a private American software company was contracted to support a variety of specialized technical capabilities. The company provided flexible data management configuration, reliable and easily parsed auditing tools, a highly secure environment, and a wide variety of collaboration options to the project, which would have been difficult to carry out in-house.

Technological Capabilities

A system with 10 computation units (10 Central Processing Units and 10 hard drives) theoretically processes data 10 times faster than a system with one computation unit. However, use of Apache Hadoop to simplify parallel computations can drastically reduce computation time. At the bare minimum, a Big Data analysis system consists of (Japec et al. 2015):

Disk-based storage media

Active computation components: Central Processing Unit, Random Access Memory.

Additional infrastructure: Server farm 32achine32s, cooling costs, network access, security, etc.

Although non-specialized Information Technology (IT) staff could handle lower data sizes, large-scale computing environments need high-powered computing stacks of hardware and software that require specialized IT training. Most projects that require Big Data computation platforms choose one of three strategies (Japec et al. 2015):

Internal compute cluster: This strategy has advantages when long-term storage of sensitive data is needed. An Apache Hadoop cluster of networked servers is created within the organization's internal network for a secure and low-cost option.

External compute cluster: This strategy allows the fastest and smoothest set-up as it rents pre-existing computation infrastructure from "utility computing" service providers such as Amazon Web Services or Google Cloud Platform (McNulty 2014). However, this option may be more expensive than setting up an internal compute cluster in the long run.

Hybrid compute cluster: This strategy mixes the two above options by outsourcing on-demand Big Data analysis tasks to external compute clusters, while creating an internal compute cluster for long-term data storage.

Conclusion

Uses of Big Data for health policymaking are diverse both in the kinds of health policy areas studied and the data types used. In our review of the literature, NORC found that understanding the data quality of a Big Data source is critical to successful application of the data to support statistical inferences. Whether a data source is a survey or a Big Data source, many of the same considerations about data quality apply. Thinking through the aspects of data quality presented in Table 3, as they apply to a specific data source, can help with determining what benefits and limitations that data source has. Further, Big Data sources can be subject to additional concerns due to the technological challenges of using such data sources. It is important when data is acquired from one or more organization(s) to establish standards for transparency of the curation of that data source, with as complete of documentation as possible. Algorithm dynamics and violations of the ideal user assumption make certain Big Data sources particularly challenging for making statistical inferences.

The benefits and challenges of using different data sources can vary greatly by data type. NORC grouped Big Data sources into three categories, finding that the attributes and issues of the data sources are more similar within the three categories:

Data maintained primarily by the public sector: Such data sources include administrative data, some insurance claims, Medicare and Medicaid enrollment data, and registries. Among Big Data sources reviewed, these in general have among the highest data quality, although the importance of assessing and verifying data quality remains important. There are several successful examples of combining administrative data sources with surveys to benefit policymaking, including record linkage and use of administrative data as auxiliary information for the survey.

Data maintained primarily by private sector organizations, including hybrid systems: These include electronic health records, some insurance claims, e-pharmacy data, and consumer purchase data. In general, there are more questions about using data from private sector organizations than about data from the public sector. Close coordination with data providers and requiring transparency of the data curation process is critical for policymakers to have an adequate understanding of data quality. There are several successful uses of this data type, particularly electronic health records in hybrid systems, like HCUP, CDC NSSP, and FDA's Sentinel Initiative. These successful uses have been supported by the HL7 standard and the Common Data Model, reflecting a fairly mature understanding of how to maintain high data quality for electronic health records. The volume and velocity of these data can be a strength, making these data sources promising for monitoring and surveillance.

User-generated data: These include social media, data from mobile phone use, GPS, and data from environmental and health sensors. Uses of these data sources tend to be challenging and are in much more of a developing phase. Technological challenges are most acute for these data sources, particularly the difficulty of verifying the truthfulness and identity of users providing data and possible changes in algorithms by a website or platform. Volume and velocity of data may be even greater than from private sector data sources, so the exploration of use of these data sources for surveillance and monitoring should continue in spite of these challenges.

Summary Observations

Based on this literature review, NORC makes the following observations to help guide future statistical uses of Big Data sources for evidence-building:

1. The data quality needed from a data source depends on how the data are used. Any application of a Big Data source should be evaluated in the context of what aspects of data quality are critical for the successful use of that data source.
2. The use of administrative records to support surveys has many examples of proven success. Careful guidance should be developed for evaluating and maintaining data quality of government administrative data sources. DHHS should continue to seek opportunities to use administrative data sources for health policymaking. Administrative data may be particularly useful when they have low measurement error and can be used to replace survey questions and reduce respondent burden.
3. When data from one or more public or private sector organizations are used, the establishment of standards for data cleaning, data maintenance, documentation, and quality checks is critical to make these data usable for policymaking. HL7 and the Common Data Model, which are emerging standards for electronic medical records that are championed by federal agencies including the National Coordinator for Health Information Technology, are useful examples of the kinds of standards that should be developed for other Big Data sources of interest.
4. The volume, velocity, and low cost of data of some private sector and user-generated data sources make these data particularly promising for monitoring and surveillance purposes. In these applications, where development of an early warning system is more critical than the data's veracity, Big Data offer particular promise. These data can offer a real-time signal that an emerging issue requires response, as well as geographic granularity to monitor where an issue is emerging.
    a. In the case of electronic health records, Big Data applications are fairly mature, and new opportunities to use these data for surveillance and monitoring should be pursued.
    b. For user-generated data, many questions about the quality of these sources remain, but they can be useful for generating hypotheses to investigate with higher-quality data sources.

References

Andrade, R. (2017). Big data and its need for unicorns. CIAT Blog. Retrieved January 12, 2018, from http://blog.ciat.cgiar.org/big-data-and-its-need-for-unicorns/

Agency for Healthcare Research and Quality (AHRQ). Accessed January 2018 at https://www.hcup-us.ahrq.gov/

Banks, D., Ruddle, P., and Kennedy, E. (2016). Arrest-Related Deaths Program Redesign Study, 2015-16: Preliminary Findings. Accessed January 2018 at https://www.bjs.gov/index.cfm?ty=pbdetail&iid=5864

Beyer, M. (2011). Gartner says solving big data challenge involves more than just managing volumes of data. *Gartner*. Retrieved January 12, 2018, from http://www.gartner.com/newsroom/id/1731916

Biemer, P.P. (2016). Errors and inference. In Foster, I., Ghani, R., Jarmin, R. S., Kreuter, F., and Lane, J. (eds.). *Big Data and Social Science: A Practical Guide to Methods and Tools* (pp. 265-297). CRC Press.

Binder, H., and Blettner, M. (2015). Big data in medical science—a biostatistical view: part 21 of a series on evaluation of scientific publications. *Deutsches Ärzteblatt International, 112*(9), 137.

Booz Allen Hamilton. (2018). *Tips for Building a Data Science Capability*. Retrieved January 12, 2018, from https://www.boozallen.com/content/dam/boozallen_site/legacy/pdf/ds-capability-handbook.pdf

Brownstein, J.S., Freifeld, C.C., Reis, B.Y., and Mandl, K.D. (2008). Surveillance sans 35achine35s: internet-based emerging infectious disease intelligence and the HealthMap project. *PloS Medicine*, *5*(7), e151.

Cavallo, A. (2017). Are online and offline prices similar? Evidence from large multi-channel retailers. *American Economic Review*, *107*(1), 283-303.

Centers for Disease Control and Prevention (CDC). (2018). National Syndromic Surveillance Program (NSSP). Retrieved January 12, 2018, from https://www.cdc.gov/nssp/index.html

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

ClinicalTrials.gov. (2017). Identifier NCT03335800, Apple heart study: assessment of wristwatch-based photoplethysmography to identify cardiac arrhythmias. National Library of Medicine. Retrieved January 12, 2018, from https://clinicaltrials.gov/ct2/show/NCT03335800

Commission on Evidenced-Based Policymaking. (2017). *The Promise of Evidence-Based Policymaking*. Retrieved January 12, 2018, from https://www.cep.gov/report/cep-final-report.pdf

Dahlhamer, J.M., and Cox, C.S. (2007). Respondent consent to link survey data with administrative records: results from a split-ballot field test with the 2007 National Health Interview survey. *In Proceedings of the Federal Committee on Statistical Methodology Research Conference.*

English, R. (2017). NSSP Grantee Meeting [PowerPoint slides]. Retrieved January 12, 2018, from https://www.cdc.gov/nssp/events/documents/Web-DQ_Grantee_2017.pdf

Eppig, F.J., and Chulis, G.S. (1997). Matching MCBS and Medicare data: the best of both worlds. *Health Care Financing Review*, *18*(3), 211.

Fan, J., Han, F., and Liu, H. (2014). Challenges of big data analysis. *National Science Review*, *1*(2), 293-314.

Flu Near You. (2018). Retrieved January 12, 2018, from https://flunearyou.org

Foundations for Evidence-Based Policymaking Act, H.R.4174, 115th Cong. (2017).

Freifeld, C.C., Brownstein, J.S., Menone, C.M., Bao, W., Filice, R., Kass-Hout, T., and Dasgupta, N. (2014). Digital drug safety surveillance: monitoring pharmaceutical products in twitter. *Drug Safety*, *37*(5), 343-350.

Freifeld, C.C., Chunara, R., Mekaru, S.R., Chan, E.H., Kass-Hout, T., Iacucci, A.A., and Brownstein, J.S. (2010). Participatory epidemiology: use of mobile phones for community-based health reporting. *PloS Medicine*, *7*(12), e1000376.

Friedman, J., Hastie, T., and Tibshirani, R. (2001). *The Elements of Statistical Learning* (vol. 1, pp. 241-249). New York: Springer.

Groves, R.M. (2011). Three eras of survey research. *Public Opinion Quarterly, 75*(5): 861–871.

Hansen, S.E., Benson, G., Bowers, A., Pennell, B.E., Lin, Y., and Duffey, B. (2010). Survey quality. University of Michigan Institute for Social Research Cross-Cultural Survey Guidelines. Retrieved January 12, 2018, from http://projects.isr.umich.edu/csdi/quality.cfm

HealthMap. (2018). Retrieved January 12, 2018, from http://www.healthmap.org/en/

Hinkson, I.V., Davidsen, T.M., Klemm, J.D., Kerlavage, A.R., and Kibbe, W.A. (2017). A comprehensive infrastructure for big data in cancer research: accelerating cancer research and precision medicine. *Frontiers in Cell and Developmental Biology*, *5*, 83.

Horrigan, M. (2013). *Big Data and Official Statistics*. Retrieved January 12, 2018 https://www.bls.gov/osmr/symp2013_horrigan.pdf

Houston, J.B., Hawthorne, J., Perreault, M.F., Park, E.H., Goldstein Hode, M., Halliwell, M.R., Turner McGowen, S.E., Davis, R., Vaid, S., McElderry, J.A., and Griffith, S.A. (2015). Social media and disasters: a functional framework for social media use in disaster planning, response, and research. *Disasters*, *39*(1), 1-22.

Hudson, K., Lifton, R., and Patrick-Lake, B. (2015). The precision medicine initiative cohort program—building a research foundation for 21st century medicine. Precision Medicine Initiative (PMI) Working Group Report to the Advisory Committee to the Director, NIH. Retrieved January 12, 2018, from https://acd.od.nih.gov/documents/reports/DRAFT-PMI-WG-Report-9-11-2015-508.pdf

Huisingh-Scheetz, M.J., Kocherginsky, M., Magett, E., Rush, P., Dale, W., and Waite, L. (2016). Relating wrist accelerometry measures to disability in older adults. *Archives of Gerontology and Geriatrics*, *62*, 68-74.

IBM. (2017). The Four V's of Big Data. Retrieved January 12, 2018, from http://www.ibmbigdatahub.com/infographic/four-vs-big-data

Japec, L., Kreuter, F., Berg, M., Biemer, P., Decker, P., Lampe, C., Lane, J., O'Neil, C., and Usher, A. (2015). Big data in survey research: AAPOR task force report. *Public Opinion Quarterly, 79*(4), 839-880.

Jarmin, R.S., and O'Hara, A.B. (2016). Big data and the transformation of public policy analysis. *Journal of Policy Analysis and Management, 35*(3), 715-721.

Kaiser Family Foundation. (2017). Accessed January 15, 2018 at https://www.kff.org/medicare/

Johnson, D.S., Massey, C., and O'Hara, A. (2015). The opportunities and challenges of using administrative data linkages to evaluate mobility. *ANNALS of the American Academy of Political and Social Science*, *657*(1), 247-264.

Kim, Y., Huang, J., and Emery, S. (2016). Garbage in, garbage out: data collection, quality assessment and reporting standards for social media data use in health research, infodemiology and digital disease detection. *Journal of Medical Internet Research, 18*(2).

Khera, R., and Krumholz, H.M. (2017). With great power comes great responsibility: big data research from the National Inpatient Sample. *Circulation: Cardiovascular Quality and Outcomes*, *10*(7), e003846.

Kitchin, R. (2015). The opportunities, challenges and risks of big data for official statistics. *Statistical Journal of the IAOS, 31*(3), 471-481.

Lane, J. (2016). Big data for public policy: the quadruple helix. *Journal of Policy Analysis and Management, 35*(3), 708-715.

Lazer, D., Kennedy, R., King, G., and Vespignani, A. (2014). The parable of Google Flu: traps in big data analysis. *Science*, *343*(6176), 1203-1205.

Lazer, D., and Radford, J. (2017). Data ex 37achine: introduction to big data. *Annual Review of Sociology, 43*, 19-39.

Leetaru, K. (2016). Why we need more domain experts in the data sciences. *Forbes*. Retrieved January 12, 2018, from https://www.forbes.com/sites/kalevleetaru/2016/06/12/why-we-need-more-domain-experts-in-the-data-sciences/

Lengsavath, M., Dal Pra, A., de Ferran, A.M., Brosch, S., Härmark, L., Newbould, V., and Goncalves, S. (2017). Social media monitoring and adverse drug reaction reporting in pharmacovigilance: an overview of the regulatory landscape. *Therapeutic Innovation & Regulatory Science*, *51*(1), 125-131.

Lohr, S.L., and Raghunathan, T.E. (2017). Combining survey data with other data sources. *Statistical Science, 32*(2), 293-312.

Margolies, L.R., Pandey, G., Horowitz, E.R., and Mendelson, D.S. (2016). Breast imaging in the era of big data: structured reporting and data mining. *American Journal of Roentgenology, 206*(2), 259-264.

Matthews, C.E., Hagströmer, M., Pober, D.M., and Bowles, H.R. (2012). Best practices for using physical activity monitors in population-based research. *Medicine and Science in Sports and Exercise*, *44*(1 S1), S68.

McNulty, E. (2014). Understanding big data: cross infrastructure and analytics. *Dataconomy*. Retrieved January 12, 2018, from http://dataconomy.com/2014/07/understanding-big-data-cross-infrastructure-analytics/

Meng, X.L. (2014). A trio of inference problems that could win you a Nobel prize in statistics (if you help fund it). *Past, Present, and Future of Statistical Science*.

Mooney, S.J., Westreich, D.J., and El-Sayed, A.M. (2015). Epidemiology in the era of big data. *Epidemiology (Cambridge, MA), 26*(3), 390.

National Academies of Sciences, Engineering, and Medicine (NAS). (2017a). *Federal Statistics, Multiple Data Sources, and Privacy Protection: Next Steps*. Washington, DC: National Academies Press.

National Academies of Sciences, Engineering, and Medicine (NAS). (2017b). *Innovations in Federal Statistics: Combining Data Sources While Protecting Privacy*. Washington, DC: National Academies Press.

National Academies of Sciences, Engineering, and Medicine (NAS). (2017c). *Refining the Concept of Scientific Inference When Working with Big Data: Proceedings of a Workshop*. Washington, DC: National Academies Press.

National Center for Health Statistics, Office of Analysis and Epidemiology (NCHS). (2017). The Linkage of National Center for Health Statistics Surveys to Medicare Enrollment and Claims Data - Methodology and Analytic Considerations. Hyattsville, Maryland.

National Institutes of Health (NIH). (2018). All of Us Research Program. Retrieved January 12, 2018, from https://allofus.nih.gov/

NIST Big Data Public Working Group. (2017). *Draft NIST Big Data Interoperability Framework: Volume 1, Definitions*. NIST Special Publications 1500-1, Version 2, Draft 2. Retrieved January 12, 2018, from https://bigdatawg.nist.gov/_uploadfiles/M0613_v1_3911475184.docx

Office of the National Coordinator for Health Information Technology (ONC). (2016). *2016 Report To Congress on Health IT Progress*. Accessed January 2018 at https://www.healthit.gov/sites/default/files/2016_report_to_congress_on_healthit_progress.pdf

Olavsrud, T. (2015). Don't look for unicorns, building a data science team. *CIO*. Retrieved January 12, 2018, from https://www.cio.com/article/3011648/analytics/dont-look-for-unicorns-build-a-data-science-team.html

Outbreaks Near Me. (2018). HealthMap. Retrieved January 12, 2018, from http://www.healthmap.org/outbreaksnearme/

Park, S., Basu, A., Coe, N., and Khalil, F. (2017). *Service-Level Selection: Strategic Risk Selection in Medicare Advantage in Response to Risk Adjustment* (No. w24038). National Bureau of Economic Research.

Popovic, J.R. (2017). Distributed data networks: a blueprint for big data sharing and healthcare analytics. *Annals of the New York Academy of Sciences, 1387*(1), 105-111.

Puts, M.J.H., Tennekes, M., Daas, P.J.H., and de Blois, C. (2016). Using huge amounts of road sensor data for official statistics. In *Proceedings of the European Conference on Quality in Official Statistics* (Q2016), Madrid, Spain. Accessed January 2018 at http://www.pietdaas.nl/beta/pubs/pubs/q2016Final00177.pdf

Robb, M.A., Racoosin, J.A., Sherman, R.E., Gross, T.P., Ball, R., Reichman, M.E., Midthun, K., and Woodcock, J. (2012). The U.S. Food and Drug Administration's Sentinel Initiative: expanding the horizons of medical product safety. *Pharmacoepidemiology and Drug Safety*, *21*(S1), 9-11.

Robinson, G. (2003). A statistical approach to the spam problem. *Linux Journal*, *2003*(107), 3.

Roski, J., Bo-Linn, G.W., and Andrews, T.A. (2014). Creating value in health care through big data: opportunities and policy implications. *Health Affairs, 33*(7), 1115-1122.

Ryan, C., and Lewis, J.M. (2017). *Computer and Internet Use in the United States: 2015*. American Community Survey Reports: ACS-37. U.S. Census Bureau. Retrieved January 12, 2018, from https://www.census.gov/content/dam/Census/library/publications/2017/acs/acs-37.pdf

Salathé, M. (2016). Digital pharmacovigilance and disease surveillance: combining traditional and big-data systems for better public health. *Journal of Infectious Diseases*, *214*(S4), S399-S403.

Sankar, P.L., and Parker, L.S. (2017). The Precision Medicine Initiative's All of Us Research Program: an agenda for research on its ethical, legal, and social issues. *Genetics in Medicine*, *19*(7), 743.

Shadish, W.R., Cook, T.D., and Campbell, D.T. (2002). *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Boston, MA: Houghton Mifflin.

Shameer, K., Johnson, K.W., Yahi, A., Miotto, R., Li, L.I., Ricks, D., Jebakaran, J., Kovatch, P., Sengupta, P. P., Gelijns, A., Moskovitz, A., Darrow, B., Reich, D. L., Kasarskis, A., Tatonetti, N. P., Pinney, S., and Dudley, J. T. (2017). Predictive modeling of hospital readmission rates using electronic medical record-wide machine learning: a case-study using Mount Sinai Heart Failure Cohort. In *Pacific Symposium on Biocomputing 2017* (pp. 276-287).

Simonsen, L., Gog, J.R., Olson, D., and Viboud, C. (2016). Infectious disease surveillance in the big data era: towards faster and locally relevant systems. *Journal of Infectious Diseases*, *214*(S4), S380-S385.

Smolinski, M.S., Crawley, A.W., Baltrusaitis, K., Chunara, R., Olsen, J.M., Wójcik, O., Santillana, M., Nguyen, A., and Brownstein, J.S. (2015). Flu Near You: crowdsourced symptom reporting spanning two influenza seasons. *American Journal of Public Health*, *105*(10), 2124-2130.

Swain, M., Charles, D., Patel, V., and Searcy, T. (2015). *Health Information Exchange among U.S. Non-Federal Acute Care Hospitals: 2008-2014*. ONC Data Brief, no.24. Washington, DC: Office of the National Coordinator for Health Information Technology.

U.S. Food and Drug Administration (FDA). (2018). FDA's Sentinel Initiative. Retrieved January 12, 2018, from https://www.fda.gov/safety/fdas-sentinel-initiative

WEB-RADR. (2018). Retrieved January 12, 2018, from http://web-radr.eu

Yang, M., Kiang, M., and Shang, W. (2015). Filtering big data from social media – building an early warning system for adverse drug reactions. *Journal of Biomedical Informatics*, *54*, 230-240.

Zhang, G., Parker, J.D., and Schenker, N. (2016). Multiple imputation for missingness due to nonlinkage and program characteristics: a case study of the National Health Interview Survey linked to Medicare claims. *Journal of Survey Statistics and Methodology, 4*(3), 319-338.

Appendix

Appendix Table 1. Notable Use Cases from Literature Review

Data Category	Data Type	Organization	Topic	Summary	Citation
Public Sector	Medicare/Medicaid enrollment Insurance claims	Centers for Disease Control and Prevention	Survey linkage	The National Health Interview Survey (NHIS) is linked to Medicare enrollment and claims data under an interagency agreement with the Centers for Medicare and Medicaid Services (CMS). Previous linkage for NHIS was facilitated by the Office of The Assistant Secretary for Planning and Evaluation (ASPE) and the Social Security Administration. The linkage provides individual record matches between 1994-2013 NHIS survey respondents to a variety of Medicare eligibility and claims files including demographic and enrollment information of beneficiaries.	National Health Interview Survey. (2018). Retrieved January 29, 2018, from https://www.cms.gov/About-CMS/Agency-Information/OMH/resource-center/hcps-and-researchers/data-tools/sgm-clearinghouse/nhis.html
Public Sector	Medicare/Medicaid enrollment Insurance claims	Centers for Medicare & Medicaid Services NORC at the University of Chicago	Survey linkage	The Medicare Current Beneficiary Survey (MCBS) is linked to Medicare enrollment and claims. Master Beneficiary Enrollment File from Medicare claims serves as the sampling frame and MCBS respondents are matched to claims files by design. The original design was premised on a full partnership between survey data collection and administrative records. MCBS provides important information on beneficiaries that is not available in CMS administrative data and plays an essential role in monitoring and evaluating beneficiary health status and health care policy.	Medicare Current Beneficiary Survey. (2018). Retrieved January 29, 2018, from https://www.cms.gov/Research-Statistics-Data-and-Systems/Research/MCBS/

Data Category	Data Type	Organization	Topic	Summary	Citation
Public Sector	Medicare/Medicaid enrollment Public health registries Medication treatments County-level determinants of health	Centers for Medicare & Medicaid Services	Research database	HealthData.gov is an open data community and data navigator created by CMS. The platform integrates Medicare and Medicaid cost reports, public health registries, medication treatments, and county-level determinants of health. It also houses nearly 1000 valuable data sets and gives the users the ability filter the data sets by categories such as subject, agency, sub-agency, date, and geography.	HealthData.gov. (2018). Retrieved January 29, 2018, from https://www.healthdata.gov/content/about
Public Sector	Government registry	Centers for Disease Control and Prevention	Influenza-related deaths	Mortality Surveillance Data from the National Center for Health Statistics (NCHS) is in pilot use by the Center for Disease Control (CDC) for Pneumonia and Influenza (P&I) mortality surveillance. NCHS has recently improved its reporting and statistical infrastructure to be able to provide near real-time surveillance of mortality. CDC's pilot program which monitors influenza-related deaths based on real-time electronic samples of US death certificates will replace the older 122 Cities Mortality Reporting System of manually evaluated death certificates.	Simonsen, L., Gog, J. R., Olson, D., & Viboud, C. (2016). Infectious disease surveillance in the Big Data era: Towards faster and locally relevant systems. The Journal of Infectious Diseases, 214(suppl_4), S380-S385.
Public Sector	Government registry	Academic research	Cardiology research	The Thrombus Aspiration in ST-Elevation Myocardial Infarction in Scandinavia successfully carried out a registry-based randomized trial comparing the use of thrombus aspiration with no aspiration before percutaneous coronary intervention. The Swedish Coronary Angiography and Angioplasty Registry and the Swedish Web System for Enhancement and Development of Evidence-based Care in Heart Disease Evaluated According to Recommended Therapies were used.	Zannad, F., Pfeffer, M. A., Bhatt, D. L., Bonds, D. E., Borer, J. S., Calvo-Rojas, G., Fiore, L., Lund, L. H., Madigan, D., Maggioni, A. P., Meyers, C. M., Rosenberg, Y., Simon, T., Gattis Stough, W., Zalewski, A., Zariffa, N., & Temple, R. (2017). Streamlining cardiovascular clinical trials to improve efficiency and generalisability. Heart, heartjnl-2017.

Data Category	Data Type	Organization	Topic	Summary	Citation
Public Sector	Government registry	Academic research	Cardiology research	Study of Access Site for Enhancement of Percutaneous Coronary Intervention for Women tried to determine the outcome of radial access on women receiving percutaneous coronary intervention. Subjects were randomized to a treatment using an online randomization module within the existing CathPCI Registry database through the National Institute of Health's National Cardiovascular Research Infrastructure, which allowed for efficiency in the design of the study.	Zannad, F., Pfeffer, M. A., Bhatt, D. L., Bonds, D. E., Borer, J. S., Calvo-Rojas, G., Fiore, L., Lund, L. H., Madigan, D., Maggioni, A. P., Meyers, C. M., Rosenberg, Y., Simon, T., Gattis Stough, W., Zalewski, A., Zariffa, N., & Temple, R. (2017). Streamlining cardiovascular clinical trials to improve efficiency and generalisability. Heart, heartjnl-2017.
Public Sector	Medicare/Medicaid enrollment Insurance Claims Assessment data	Centers for Medicare & Medicaid Services	Research database	Chronic Conditions Data Warehouse (CCW) is a research database launched by CMS with the purpose of making Medicare, Medicaid, Assessments, and Part D Prescription Drug Events data readily available for research. Medicare and Medicaid beneficiary, claims, and assessment data are linked by beneficiary across the continuum of care and saves data users from huge data wrangling efforts.	Chronic Conditions Data Warehouse. (2018). Retrieved January 29, 2018, from https://www.ccwdata.org
Public Sector	Insurance Claims	Centers for Medicare & Medicaid Services	Research database	Medicare Claims Synthetic Public Use Files (SynPUFs) has been created by CMS to allow interested users to gain familiarity with claims data without going through the procedure needed to require restricted access. SynPUFs were created with the aim of lowering the barrier-to-use for data users and software developers looking to work with claims data. Users will be much more informed on which CMS data product they need after engaging with SynPUFs.	Medicare Claims Synthetic Public Use Files. (2018). Retrieved January 29, 2018 from https://www.cms.gov/Research-Statistics-Data-and-Systems/Downloadable-Public-Use-Files/SynPUFs/
Public Sector	Insurance Claims	Academic research	Vaccination estimates	Using medical claims to track vaccine uptake has been demonstrated by researchers in Germany as a promising low cost approach where vaccination is largely administered through the private sector. They found that systemic overestimation of coverage due to children never seeing a physician and, thus not being entered into the database, was small.	Kalies, H., Redel, R., Varga, R., Tauscher, M., & von Kries, R. (2008). Vaccination coverage in children can be estimated from health insurance data. BMC Public Health, 8, 82.

Data Category	Data Type	Organization	Topic	Summary	Citation
Public Sector	Satellite imagery data	Academic research	Measles transmission	Nighttime satellite imagery from the Defense of Meteorological Satellite Program (DMSP) Operational Linescan System (OLS) was used by researchers to quantify migration patterns and relative population density. Researchers found that population density and measles transmission were highly correlated in three cities in Niger.	Bharti, N., Tatem, A. J., Ferrari, M. J., Grais, R. F., Djibo, A., & Grenfell, B. T. (2011). Explaining Seasonal Fluctuations of Measles in Niger Using Nighttime Lights Imagery. Science (New York, N.Y.), 334(6061), 1424–1427.
Public/Private	Medicare/Medicaid enrollment Public health registries Hospital records	Health Resources and Services Administration	Research database	Area Health Resource File (AHRF) is provided by Health Resources and Services Administration and contains over 6,000 variables related to health care access at the county, state, and national-level. AHRF integrates data from over 50 sources including: the American Hospital Association, the American Medical Association, the US Census Bureau, CMS, Bureau of Labor Statistics, InterStudy, and the Veteran's Administration.	Area Health Resource File. (2018). Retrieved January 29, 2018 from https://www.healthypeople.gov/2020/data-source/area-health-resource-file
Public/Private	Administrative EHR/EMR	Agency for Health Care Research and Quality	Research database	The Health Care Cost and Utilization Project (HCUP) is a long term successful collaboration between the Agency for Health Care Research and Quality, states, hospitals and private organizations to provide individual level encounter data from hospitals in almost every state in the nation. HCUP uses administrative data, hospital discharge record, demographic information, services provided, disease status, and the cost of services and payers. States, municipalities, and private organizations receive this information through voluntary donations or state mandates.	Health Care Cost and Utilization Project. (2018). Retrieved January 29, 2018 from https://www.hcup-us.ahrq.gov/
Public/Private	Government registry Insurance claims EHR/EMR	Food and Drug Administration	Monitoring and surveillance	The Sentinel Initiative from the Food and Drug Administration combines electronic health records, insurance claims data, and registries for adverse event monitoring to ensure safety of drugs and other regulated medical products. A distributed data infrastructure allows for rapid analysis across the database of more than 193 million patients. The use of the Common Data Model helps ensure standardization and maintain data quality.	U.S. Food & Drug Administration (2018). FDA's Sentinel Initiative. Retrieved January 12, 2018, from https://www.fda.gov/safety/fdas-sentinel-initiative

Data Category	Data Type	Organization	Topic	Summary	Citation
Public/Private	EHR/EMR	Center for Disease Control	Monitoring and surveillance	The National Syndromic Surveillance Program (NSSP) by Center for Disease Control (CDC) integrates electronic health information for emergency departments, urgent care, ambulatory care, inpatient care, pharmacy data, and lab data, with standardized analytic tools to support detection of and rapid response to hazardous events and disease outbreaks. The sheer volume of data help support surveillance with high spatial and temporal resolution. The BioSense platform allows for cloud-based sharing of health information with tools to capture, store, and analyze data.	Centers for Disease Control and Prevention (2018). National Syndromic Surveillance Program (NSSP). Retrieved January 12, 2018, from https://www.cdc.gov/nssp/index.html
Public/Private	EHR/EMR	China Stroke Prevention Committee Sanofi China	Stroke screening	The China Stroke Data Center is a nationwide stroke screening platform that has been built in 2011 to support national stroke prevention programs and stroke research. The data integration system collects information on stroke patients' risk factors, diagnosis history, treatment, socio-demographic characteristics, and EMR.	Yu, J., Mao, H., Li, M., Ye, D., & Zhao, D. (2016, August). CSDC—A nationwide screening platform for stroke control and prevention in China. In Engineering in Medicine and Biology Society (EMBC), 2016 IEEE 38th Annual International Conference of the (pp. 2974-2977). IEEE.
Public/Private	Web data	Bureau of Justice Statistics	Arrest related deaths	In the redesign of Census of Arrest Related Deaths, the Bureau of Justice Statistics began reviewing open information sources such as web-scraped news articles and official agency documents to collect data about deaths to persons arrested or in custody more rigorously.	Banks, D., Ruddle, P., Kennedy, E., & Planty, M. G. (2016). Arrest-related deaths program redesign study, 2015-16: preliminary findings (U.S. Department of Justice, Office of Justice Programs).
Private Sector	Transaction data	MIT	Inflation trends	The Billion Prices Project is an academic initiative at MIT to track price shifts in 22 countries using daily web-scraped prices for 5 million items. Changes in inflation trends can be spotted in a much timelier manner compared to the monthly Consumer Price Index (CPI). These statistics are considered to be a source of CPI by many statistical agencies world-wide. Validation exercises and ongoing assessment of data quality and cleaning are currently being used to assess fitness of use.	The Billion Prices Project. (2018). Retrieved January 29, 2018, from http://www.thebillionpricesproject.com

Data Category	Data Type	Organization	Topic	Summary	Citation
Private Sector	Transaction data	Bureau of Labor Statistics	Consumer price index	The Consumer Price Index produced by the Bureau of Labor Statistics uses data from retail scanners, web-based price scrapping, and JD Power car prices for adjustment and calculation. The use of retail scanner data to augment traditional price gather mechanisms began in the late 1990s, but has expanded as access to retail scanner data has been routinized by several private market research firms.	Consumer Price Index. (2018). Retrieved January 29, 2018, from https://www.bls.gov/opub/ted/2019/consumer-price-index-2018-in-review.htm
Private Sector	Transaction data	Academic research	Grocery purchase quality	The Grocery Purchase Quality Index-2016 (GPQI-2016) is a system for evaluating the quality of household grocery purchases, which has been developed and validated by researchers. GPQI-2016 used a grocery sales data set provided by a national grocery chain by drawing a sample of 4000 household in each four geographic locations. Construct validity of the index was established through confirming that households that never purchased tobacco had higher median total quality scores than households that purchased tobacco, as well as scoring higher in every component of Department of Agriculture's grouped Food Plan market baskets.	Brewster, P. J., Guenther, P. M., Jordan, K. C., & Hurdle, J. F. (2017). The Grocery Purchase Quality Index-2016: An innovative approach to assessing grocery food purchases. Journal of Food Composition and Analysis, 64, 119-126.
Private Sector	Mobile phone data	Academic research	Dengue outbreak	Climate and mobility data from around 40 million mobile phone subscribers were used by Wesolowski et al. (2015) to examine the outbreak of 2013 dengue outbreak in Pakistan. Spatially explicit dengue case data were compared to an epidemiological model of dengue virus transmission based on mobile phone data. The researchers found "that mobile phone-based mobility estimates predict the geographic spread and timing of epidemics."	Wesolowski, A., Qureshi, T., Boni, M. F., Sundsøy, P. R., Johansson, M. A., Rasheed, S. B., Engø-Monsen, K., & Buckee, C. O. (2015). Impact of human mobility on dengue epidemics. Proceedings of the National Academy of Sciences, 112(38), 11887-11892.

Data Category	Data Type	Organization	Topic	Summary	Citation
Private Sector	Web data	Health Canada World Health Organization	Monitoring and surveillance	The Global Public Health Intelligence Network (GPHIN) is developed by Health Canada in collaboration with the World Health Organization and gathers epidemic intelligence from informal sources. The network is a multilingual early-warning tool that continuously scours global media sources for disease outbreaks and public health concerns such as communicable disease, food and water safety, and chemical events.	Global Public Health Intelligence Network. (2018). Retrieved January 29, 2018, from http://www.who.int/csr/alertresponse/epidemicintelligence/en/
Private Sector	Sensor data	Statistics of Netherlands	Traffic and road conditions	National traffic and road conditions are provided in real time by Statistics of Netherlands by capturing traffic sensor data, which has become ubiquitous enough to provide nearly complete coverage of national roadways.	Daas, P. J.H., Puts, M. J. H., Buelens, B., & Hurk, P. A. M. (2013). *Big Data and Official Statistics.* Presented at the 2013 New Techniques and Technologies for Statistics conference (NTTS).
Public/Private/User-Generated	Internet data	Boston Children's Hospital	Monitoring and surveillance	HealthMap has been developed by the Boston Children's Hospital's Computational Epidemiology Group to support applications for monitoring and surveillance of disease outbreaks and emerging public health threats. HealthMap's applications primarily use algorithms to accumulate web-accessible information: news aggregators, eyewitness reports, expert-curated discussions, and validated official reports. HealthMap's apps are used by public health departments and government agencies, including Center for Disease Control, the Department of Defense, and the World Health Organization.	HealthMap. (2018). Retrieved January 12, 2018, from http://www.healthmap.org/en/
Private/User-Generated	Internet data	European Union	Adverse drug reactions	Web-Recognizing Adverse Drug Reactions (WEB-RADR) has been launched by the European Union's Innovative Medicines Initiative through a public-private partnership. WEB-RADR aims to identify new data sources for pharmacovigilance and optimize the aggregation of information on possible adverse drug reactions. The effort is in early stages and will include development of an EU-wide mobile phone app for reporting adverse events and the development of text mining techniques for publicly available data on social media sites.	WEB-RADR. (2018). Retrieved January 12, 2018, from http://web-radr.eu

Data Category	Data Type	Organization	Topic	Summary	Citation
User-Generated	Social media data	Academic research	Adverse drug reaction	Freifeld et al. (2014) studied 6.9 million tweets from Twitter using a combination of manual and semi-automated techniques. They found 4,401 possible adverse drug reactions. Although assessing the validity of the findings was difficult, the researchers compared their findings to those from FDA's Adverse Event Reporting System and found similarities in patterns between the two data sources.	Freifeld, C. C., Brownstein, J. S., Menone, C. M., Bao, W., Filice, R., Kass-Hout, T., & Dasgupta, N. (2014). Digital drug safety surveillance: monitoring pharmaceutical products in twitter. Drug safety, 37(5), 343-350.
User-Generated	Biospecimen Self-reported data Social media data Sensor data	National Institutes of Health	Research database	The All of Us Research Program has been spearheaded by National Institutes of Health's Precision Medicine Initiative. All of Us seeks to recruit more than one million volunteer participants that contribute health data and biospecimens to a centralized national database to support research on a range of medical and health questions. The All of Us database will constitute a hybrid system that includes self-reported measures, EHR, sensor-based observations through phones and wearable devices, geospatial and environment data, and social media data.	National Institutes of Health. (2018). All of Us Research Program. Retrieved January 12, 2018, from https://allofus.nih.gov/
User-Generated	Sensor data	Academic research	Accelerometer	Wrist accelerometry has been explored as a tool in disability research in older adults by Husingh-Scheetz et al. (2016). They used a representative sample for the study to support the external validity of findings. However, even after extensive work to identify the right device for the study, questions remained about the quality of the measurements - including measurements' construct validity and the comparability of device measurements across participants.	Huisingh-Scheetz, M. J., Kocherginsky, M., Magett, E., Rush, P., Dale, W., & Waite, L. (2016). Relating wrist accelerometry measures to disability in older adults. Archives of gerontology and geriatrics, 62, 68-74.

Data Category	Data Type	Organization	Topic	Summary	Citation
User-Generated	Sensor data	Computation Institute Argonne National Laboratory University of Chicago School of the Art Institute of Chicago Urban Center for Computation and Data City of Chicago	Interactive sensors	The Array of Things is a collaborative project of scientists, universities, the city of Chicago, and local citizens to collect real-time data on the city's environment for public use and research. It consists of a network of interactive sensors that are installed around Chicago that collect real-time data on livability factors such as climate, air quality, and noise. The project aims to provide granular data of the city for scientists, policy-makers, and citizens to use in improving the livability and efficiency of Chicago.	Array of Things. (2018). Retrieved January 29, 2018, from http://arrayofthings.github.io/
User-Generated	Sensor data	Apple Stanford	Cardiology research	The Apple Heart Study is a collaborative research project between Apple and Stanford Medicine to assess whether Apple Watches can be used to identify irregular heart rhythms. The study launched in late 2017 and is still in its early stages of recruiting voluntary participants.	ClinicalTrials.gov. (2017). Identifier NCT03335800, Apple heart study: Assessment of wristwatch-based photoplethysmography to identify cardiac arrhythmias. National Library of Medicine. Retrieved January 12, 2018, from https://clinicaltrials.gov/ct2/show/NCT03335800

APPENDIX B:
Interview Protocol

Interview Protocol

Interview Questions

1. Please describe your professional background.

2. Please describe your current organization and your role there.
   a. How long have you worked at your current organization?
   b. Do you have control over how data are produced in your organization?

3. Where do your interests in "big" data stem from?

4. Why is "big" data important for research purposes today?

5. How has "big" data evolved throughout your professional career?

6. What type or types of "big" data do you use to make important decisions for your organization?

Please tell us anything else about your experience related to data quality, fitness for use, and the type of big data, you view as relevant.