Uses of Alternative Data Sources for Public Health Statistics and Policymaking: Challenges and Opportunities

Zachary H. Seeskin[1], Felicia LeClere[1], Jaehoon Ahn[1], Joshua A. Williams[2]
[1]NORC at the University of Chicago, 55 E. Monroe Street, 31st Floor, Chicago, IL 60603
[2]Office of the Assistant Secretary for Planning and Evaluation, U.S. Department of Health and Human Services, 200 Independence Avenue SW, Washington, DC 20201

Abstract
Alternative data sources beyond surveys and censuses are increasingly seen as a potential resource for health statistics and policy analysis. These non-traditional data sources can have advantages due to, among other factors, providing novel information, the speed of data collection, and increased geographic granularity. However, such data sources must be evaluated carefully to ensure that they meet the quality standards needed for policy analysis. While administrative records have been used successfully for some data products and analyses, the understanding of how to evaluate the quality of other alternative data sources, from electronic health records to data from environmental and health sensors to social media, is still maturing. The strengths and weaknesses of different alternative sources should guide whether and how they are applied. Researchers and analysts may look to alternative health data sources, for example, for new methods of analysis because the strength of timely data collection is more important than weaknesses of accuracy and reliability. In this paper, we examine the challenges and opportunities of using alternative data sources to answer policy-relevant questions in the context of public health policy by reviewing recent examples of such uses.

Key Words: Data quality; Combining data from multiple sources; Administrative records; Electronic health records; Sensor data; Social media

1. Introduction

There is a substantial opportunity for government agencies and researchers to use new data sources to inform policymaking and evidence-building, including in the area of public health. Diverse data sources such as government administrative records, electronic health records (EHR), environmental and health sensors, and social media are being studied for use for health statistics and research. These data sources differ from those traditionally used for evidence-building, specifically surveys, censuses, and randomized experiments, in that they were not originally collected for statistical purposes. For federal decision-making and policymaking purposes, evidence "can be defined broadly as information that aids the generation of a conclusion," but traditionally with an emphasis on "information produced by 'statistical activities' with a 'statistical purpose' that is potentially useful when evaluating government programs and policies;" this can be distinguished from data used for non-statistical purposes, such as those used to determine individual benefits (Commission on Evidenced-Based Policymaking 2017, 8-12).

This article reviews different uses of alternative data sources for policy analysis, for both health and public health policy, to demonstrate the promise and challenges of these data sources for future evidence-building. While distinct concepts with different communities, this article refers to both health policy (which is typically inclusive of health care financing and delivery issues) and public health policy (which is inclusive of more population-level issues such as disease burden, prevalence, and risk factors) simply as public health policy. As evidence-based policymaking is a major focus of the federal government, this article looks to

help understand how analysis of alternative data sources may inform decision-making in the public health policy field, which in turn may lead to a more informed and effective policymaking process.

Among the 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. We distinguish among three main categories of data types upon finding that different kinds of issues tend to affect these data types, public sector data including administrative records, private sector data as well as combined public and private data, and user-generated data including such sources as social media and sensors. Some of these alternative data sources, particularly public sector data and to a lesser extent private sector data, have been used for evidence-building and decision-making activities for a long while, while others, mostly from the user-generated data category, are just starting to be explored for their potential usability and suitability in public health.

In this review, we highlight important use cases chosen to represent the promise and challenges of alternative data sources for evidence-building. Because of the many kinds of applications being explored, we cannot discuss all use cases or even all prominent successful use cases. 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. We find that understanding the quality of alternative data sources is critical to the potential for their successful use. These alternative sources come with additional technical concerns that must be considered and mitigated. Both the challenges and the potential benefits of using alternative data sources can vary based on the type of data and the desired application.

This article is organized as follows. Section 2 discusses characteristics of different alternative data sources investigated for this literature review and describes major themes from the literature. This section provides the framework for describing the benefits and challenges of the use cases discussed in the subsequent sections. Section 3 presents the literature review plan and strategy. Sections 4 to 6 investigate use cases for the three data types described earlier: Section 4 discussing public sector data, Section 5 discussing private sector data and combined public and private data, and Section 6 discussing user-generated data. Section 7 then concludes and summarizes observations to guide uses of alternative data sources for policy analysis to inform evidence-based decision-making.

2. Background

2.1 Characteristics of Alternative Data Sources

For this literature review, we define alternative data sources using Groves's (2011) description of "organic data," *data that are not originally collected for statistical purposes*. This definition differentiates the alternative data sources of interest from surveys, censuses, and randomized experiments, or "designed" data, as the data are not originally collected for the purposes of statistical use or evidence-building. By focusing on "organic data," we focus on the benefits and challenges resulting from how the data were created.

"Organic data" encompasses a wide range of datasets that can be used for evidence-building, all facing the challenge that the data collection was not designed to support statistical uses. Among the data sources reviewed here include private and public health insurance claims, Medicare and Medicaid enrollment records, state health registries, EHR, consumer purchase data from businesses, electronic prescription data, social media, data generated by mobile phone use, and data produced by electronic and health sensors. Many of these data sources fall under what the federal government defines as "administrative data," which the federal agencies are already attempting to use for evidence-building in various capacities. However, federal administrative data, as defined by OMB Memorandum 14-06 (Guidance for Providing and Using

Administrative Data for Statistical Purposes), does not cover certain forms of organic data like those derived from social media. M-14-06 specifically states 'administrative data' refers to "administrative, regulatory, law enforcement, adjudicatory, financial, or other data held by agencies and offices of the government or their contractors or grantees (including States or other units of government) and collected for other than statistical purposes. Administrative data are typically collected to carry out the basic administration of a program, such as processing benefit applications or tracking services received. These data relate to individuals, businesses, and other institutions."

While we are defining all of these data sources as "organic data", some of the data sources of interest for this article may be also be described as 'Big Data' sources. Big data sources are often described by their characteristics, referred to as the "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 alternative data sources. The volume, velocity, and variety of some data sources suggest how these data sources may mitigate weaknesses of and/or add value to surveys and censuses. These attributes, however, also make these data sources 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 alternative data sources, extensive data cleaning may be required to make the data usable.

It is important that statistical uses of data ensure fairness so that different groups are treated equitably. Researchers and analysts must also understand the limitations of the data sources they use in order to avoid doing harm to those affected by potential policies. Government agencies rely greatly on public trust in the information they provide, so understanding when and how alternative data sources can be used for policy research based on their level of veracity is of the utmost importance.

The origin of alternative data sources is critical to our evaluation of data quality. Lazer and Radford (2017) distinguish two kinds of alternative data sources that yield different issues when those sources are analyzed using this approach's perspective.

- **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 alternative data sources can usually be classified as either reflecting digital trace or digital life. Digital trace data 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 policy analysis.

2.2 Data Quality and Alternative Data Sources

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, alternative 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. Here, we present a data quality framework that can be applied to assess alternative data sources' fitness for statistical uses.

Data quality is multidimensional, with elements reflecting different aspects needed to support valid statistical inferences. Table 1 describes different aspects of data quality that can be applied to either traditional or alternative data sources to support policymaking, grouped into five categories: 1) accuracy, 2) relevance, 3) timeliness, 4) accessibility, clarity, and transparency, and 5) coherence and comparability (Hansen et al. 2010, Japec et al. 2015, NAS 2017a).

Table 1: 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 are available when expected and in time for policy purposes.
Accessibility **Clarity** **Transparency**	Data can be readily obtained and analyzed by users. Data and statistics are presented in clear, understandable format. Methodologies for data preparation and statistical processes are available.
Coherence **Comparability**	Enough metadata is available to understand data structure and allow for combination with other statistical information. Data over time and from different records or sets of records reflect the same concept.

Data accuracy is directly related to veracity in the context of alternative data sources and is often a concern for an alternative data source. Since the data are 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 interest. Further, the variables available from an alternative data source may not directly correspond to those of interest for policymaking. On the other hand, alternative 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.

Related to accuracy, one common issue with the veracity of alternative 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 (Mellon and Prosser 2017), 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 an alternative data source for evidence-building 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.

A possible strength of alternative data sources is the novel measures they can provide for research that may be more relevant to the policy topic of interest than what can be measured via a survey questionnaire. However, other challenges emerge because the data are not collected for statistical purposes. 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 can bring into question the resulting statistics and their use to inform policy.

Alternative data sources, however, often have the advantage of timeliness or velocity. Traditional data collection can take time to intake, process, and review before estimates are available to policymakers. Some alternative data sources can provide data at enough speed to allow for rapid estimation. When the validity of these inferences can be ensured, alternative data sources can offer new opportunities for more timely and evidence-based decision-making.

2.3 Fitness for Use

There are a variety of uses of alternative data sources, and different uses require different strengths from the data source. Therefore, any evaluation of a data source for 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 using alternative data sources for policy analysis 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.

2.4 Alternative Data Sources and a Fast-Changing Technological World

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

2.4.1 Business Processes and Alternative Data Sources

It is critical for the organizations maintaining alternative 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 an alternative data source for evidence-building. Businesses typically do not have expertise in policy research and can be unaware of what is needed from data sources to be trustworthy for federal use. Further, a business or other organization may have a financial interest in keeping their methodologies proprietary, compromising the clarity or coherence of their data for use as an alternative 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 evidence-building.

One theme among the successful use cases reviewed in this study is the use of standardization to improve the quality of alternative data sources. When data come from different organizations that are not collecting the data for evidence-building purposes, having requirements for data maintenance and 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 an alternative data source. Standardization can be critical for making alternative data sources usable for public health policy research.

2.4.2 Algorithm Dynamics

Additional challenges can arise with using alternative 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, and sometimes not even fully understood by the company themselves due to the complexity of underlying systems, 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 alternative 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 well-documented overestimate of flu prevalence in 2013 after multiple years of apparent success. Thus, algorithm dynamics can greatly harm statistical validity.

2.4.3 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 for the data source of interest. Further, much traffic on the internet is generated by bots, automated programs which may or may not have been created or run by malicious actors (Zeifman 2017), and their activities may be inadvertently attributed to human users. There is even a possibility for users with some understanding of how an alternative data source is being used for decision-making to intentionally corrupt the data. As discussed later, the 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.

2.5 Summary for Three Types of Alternative Data Sources

Different kinds of alternative data sources have different strengths and face different challenges. Thus, we have grouped our following review into three categories: 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 is user-generated, e.g., from online platforms. We found that different kinds of issues tend to affect these data types. As will be discussed, the first two typically represent digital trace data whereas user-generated data typically represent digital life.

Different data types have different readiness levels for statistical uses and decision-making. The maturation of standards and requirements for processing and documentation of alternative data sources can be critical to assure strong data quality and to guide the successful use of alternative data sources for decision-making. A summary of our observations on the different data types is provided in Table 2.

Table 2: 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) Varies 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

3. Literature Review Strategy

For this literature review, we first identified and reviewed a set of papers providing overviews regarding alternative data for statistical uses and decision-making. Then, we identified papers that use alternative data sources for policymaking or population studies focusing on applications both in and outside of government and preferring articles released since 2015. 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, we used our judgment to identify papers that demonstrated the most promising, successful uses of

alternative data, as well as a set of uses that could demonstrate the breadth of data types, benefits, and challenges of using alternative data sources. The review includes both published articles and grey literature, and we explored both social science and medical literature. Table 3 presents a list of many of the search terms used in the literature review, including different public health policy topics, terminology related to alternative data sources and data science, and different data types of interest.

Table 3: Partial List of Search Terms Used for Literature Review

Public 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

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

4. Public Sector Data

4.1 Considerations for Public Sector Data

Public sector alternative data sources have long been used alone or in conjunction with sample surveys to support decision-making and evaluation. Recently, it has become more straightforward to use them as primary or ancillary sources of data due to improvements in timeliness and data quality. Administrative 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 alternative data sources. Two reports from the National Academies of Sciences, Engineering, and Medicine (2017a, 2017b) and the report of the Commission on Evidence-Based Policymaking (2017) 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.

The U.S. Census Bureau has long used administrative records to improve and expand the federal data systems. The Census Bureau received federal funding in 2016 to build upon their well-developed Federal Statistical Research Data Center 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).

Administrative data, by and large, are collected for administrative, regulatory, law enforcement, adjudicatory, or financial 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 by-products or digital traces 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, such as for notifiable diseases such as HIV (e.g., the Enhanced HIV/AIDS Reporting System), or voluntary state-level reporting, such as of childhood immunization (e.g., 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 with the intent to 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, administrative data have become more robust and timely for policy analyses due to increased automation, improved data quality checks, and harmonization efforts. 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 to validate or supplement extant surveys. Program enrollment and eligibility rosters are often used as sampling frames, either as the source of complete 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 nonrespondents. 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 systematically 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 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 a survey sample member. 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 way of linking 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 way administrative data and survey data are linked 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, in part, because of their organic nature. Using the data quality framework described in Table 1, 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.

4.2 Uses of Administrative Data for Public 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.

4.2.1 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 the U.S. Department of Health and Human Services 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 the Office of the Assistant Secretary for Planning and Evaluation and the Social Security Administration.

The most recent linkage for the NHIS provided individual record matches of 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 needed for efficient linking, 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 example, the percentage of the total sample age 65 and over in the NHIS linked to the Medicare administrative data dropped from 67.0 percent in 1994 to 43.6 percent in 2005. Those figures rose to 44.3 percent in 2007 and then hovered between 51.0 and 59.0 percent between 2008 and 2013, in part as a function of the change in methods of gathering PII and for informed consent (NCHS 2017).

Like other record linkage attempts, the NHIS Medicare administrative data linkage creates data files that are enriched by the linkage, but also subject to error due to a variety of factors that include (as noted earlier): 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 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 reports the 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 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 data files for statistically sound estimates.

4.2.2 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 (CMS 2018). 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 different age groups.

As part of data collection, respondents are asked detailed questions that focus on use of medical services and the resulting costs, and are asked questions 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 data files 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 (MA) 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 MA 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.

5. Private Sector Data

5.1 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 two recent reports from the Committee on National Statistics (NAS 2017a, 2017b), the authors lay 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 decision-making. 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 five 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 for 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 and decision-making. 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, agreed upon characteristics, include 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 complementary 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, are 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 alternative data sources 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 analysis and decision-making. 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 EHR 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, informally known as the Meaningful Use 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. The Incentive Program, in the early years of the Affordable Care Act, successfully encouraged 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 data 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 data mining (Margolies et al. 2016) and predictive modeling with machine learning of hospital readmission rates for heart failure (Shameer et al. 2017). Some additional collaborations, which rely heavily on the maturation of this private-public partnership, including the National Institutes of Health's *Cancer Moonshot* and the *All of Us* Research Program, under the *Precision Medicine Initiative*. However, these initiatives will be heavily dependent on the ongoing efforts at standardization of data items and exchange.

5.2 Uses of Private Sector Data

We discuss two examples representing benefits and challenges with different kinds of private sector data sources, conducting nutrition research using structured data from retail food purchases and using data integration to provide more timely detection of adverse drug reactions (ADRs).

5.2.1 Commercial Scanner Data for Retail Food Purchases

The Department of Agriculture's Economic Research Service is using commercial scanner data on retail food purchases from market research firm IRI for research on food economics, nutrition, and health behavior (Muth et al. 2016). Two major data products from IRI have been discussed in the literature. First, the Consumer Network household scanner data links survey data on product purchases from a sample of 120,000 households to product characteristics and nutrition data to provide a rich picture on household nutritional consumption. Second, InfoScan retail scanner data are aggregated directly from different retailers in the U.S, providing a resource with billions of transactions from an array of outlet types across the country.

The primary benefit of these data sources is the novel information that they provide for policy analysis and research. The level of detail on purchases and nutrition is not possible to obtain through a survey data alone, providing a new source of information for conducting research. However, the two data sources have different kinds of challenges.

For the InfoScan retail scanner data, there are restrictions from retailers and from IRI on what data can be released. This causes challenges to the representativity of the information available, and survey weights are not available to help adjust for this issue. Second, retailers vary in the level of aggregation at which information is available. While some retailers make data available at the store-level, others provide data for different geographic levels of aggregation corresponding to marketing areas. As the geographic aggregations vary by retailer, it is difficult to make geographic comparisons among retailers. Limited product information is available for random-weight items, such as perishable and private-label products. Finally, only some Universal Purchase Codes have nutrition data available to link, leaving 19 percent of the total sales in the data without nutrition information, creating challenges to generalizability and limiting some information available for nutrition research.

The Consumer Network household scanner data overcomes some of the issues with generalizability of inferences by developing estimates based on a random sample of households linked to product purchases. While valuable data is added to the survey, it is difficult to adequately capture information on unpackaged and random-weight items, like meats, cheeses, fruits, vegetables, and bakery items, limiting the ability to study some important food categories for nutrition research (Sweitzer et al. 2017).

5.2.2 FDA's Sentinel Initiative for Monitoring Adverse Drug Reactions

The U.S. Food and Drug Administration's (FDA's) Sentinel Initiative (https://www.fda.gov/Safety/FDAsSentinelInitiative/) combines EHR, public and private insurance claims, registries, and other data sources to ensure the safety of drugs and other regulated medical products (Robb et al. 2012). Recognition of the need for a system like the Sentinel Initiative arose due to low awareness of the FDA's Adverse Event Reporting System for reporting a possible ADRs and an over-reliance on pharmaceutical companies to monitor their own products.

For timely intervention to remove potentially dangerous drugs from the market, quick detection of possible issues can be important relative to the accuracy of initial estimates. Among the elements that allow for rapid analysis and detection of issues across a database of 193 million patients are a distributed data infrastructure and the application of the Common Data Model (Popovic 2017). Combined, these systems help provide a standard data structure and coding of fields across different sites that allow for the standardized computer programs to run identically at different sites for analysis.

While the Sentinel Initiative is primarily used for initial, timely detection of adverse events, the system can also be used for further analysis to verify the initial signal. Once a possible issue is detected, the Sentinel Initiative allows for prospective monitoring at certain time intervals, or signal refinement, to verify the issue initially detected. Further, findings may also inform full-scale epidemiological studies, or signal evaluation.

Overall, the Sentinel Initiative demonstrates the promise of using private-sector data for monitoring when hypothesis generation is of interest and the timeliness dimension of data quality is most critical. The Sentinel Initiative's success has been paved by progress in data standardization and the use of the Common Data Model.

6. User-Generated Data

6.1 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 policy analysis and decision-making 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 policy analysis. 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 about all processes affecting the data source, which limits transparency. Inferences relying on the ideal user assumption may be misleading 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 analyses often use 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 (APIs), 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 policy analysis 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 policy analysis and research. The examples we highlight are just a subset of the potential uses of user-generated data for public health policy analysis, but were chosen to reflect the range data types, applications, benefits, and challenges of these data sources.

6.2 Uses of User-Generated Data for Public Health Policy Analysis

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 a further example of monitoring ADRs, but using social media reports. The choice to highlight these cases reflects our 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.

6.2.1 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 (Brownstein et al. 2008, http://www.healthmap.org/) 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. Outbreaks Near Me (http://www.healthmap.org/outbreaksnearme/) 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, but challenges remain in understanding the veracity of HealthMap data.

One notable use of HealthMap is Flu Near You (FNY) for disease outbreak surveillance (Smolinski et al. 2015, https://flunearyou.org). 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 Centers for Disease Control and Prevention'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, we found that alternative data sources uses like FNY lacked the development of standards, as well as mature thinking about measuring data quality, to assure the veracity of the data.

6.2.2 Social Media for Adverse Drug Reaction Monitoring

The text mining of social media data can be a promising tool for monitoring ADRs and related events, but like other uses of user-generated data, is also subject to some substantial challenges. Just as FDA's Sentinel Initiative emerged due to low awareness of traditional reporting systems, there was also a realization that social media could be valuable for this monitoring.

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 monitoring. 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's traditional ADR reporting system 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 applies 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 Innovative Medicines Initiative has launched a system for Web-Recognizing Adverse Drug Reactions (WEB-RADR, http://web-radr.eu, Lengsavath et al. 2017) through a public-private partnership. WEB-RADR aims to identify new data sources for monitoring ADRs and to optimize the aggregation of information. The program includes deployment of an EU-wide mobile phone app to providers and patients 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. To address violations of the ideal user assumption, 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.

7. Conclusion

Uses of alternative data sources for policy analysis are diverse both in the kinds of public health policy areas studied and the data types used. In our review of the literature, we found that understanding the data quality of an alternative data source is critical to successful use of the data to support statistical inferences. Whether a data source is a survey or an alternative data source, many of the same considerations about data quality apply. Thinking through the aspects of data quality presented in Table 1, as they apply to a specific data source, can help with determining what benefits and limitations that data source has. Further, alternative data sources can be subject to additional concerns due to the technological challenges of using such data sources. It is important to establish standards for transparency of the curation of data sources when data are acquired from third party organizations, with as complete of documentation as possible. Algorithm dynamics and violations of the ideal user assumption make digital life data particularly challenging for making statistical inferences.

The benefits and challenges of using different data sources can vary greatly by data type. The data quality needed from a data source depends on how the data are used. Any application of an alternative data source should be evaluated in the context of what aspects of data quality are critical for the successful use of that data source. We grouped alternative data sources into three categories, finding that attributes of and issues with the data sources are more similar within the three categories.

Public Sector Data Sources

Among alternative data sources reviewed, public sector data sources in general have higher 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 policy analysis, including record linkage and use of administrative data as auxiliary information for the survey. 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. Administrative data may be particularly useful when they have low measurement error and can be used to replace survey questions and reduce respondent burden.

Private Sector and Combined Data Sources

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 researchers and analysts to have an adequate understanding of data quality. There are several successful uses of this data type, including EHR combined with other public and private sector data

for the FDA's Sentinel Initiative. These successful uses have been supported by the HL7 standards and the Common Data Model, reflecting a fairly mature understanding of how to standardize EHR for analysis. These standards can serve as model to be applied to other uses of alternative data sources. The volume and velocity of these data can be strengths, making these data sources promising for monitoring and surveillance. These data can offer a real-time signal that an emerging issue requires action, as well as geographic granularity to monitor where an issue is emerging.

User-Generated Data

Uses of user-generated or digital life 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 for private sector data sources, so the exploration of use of these data sources for surveillance and monitoring is promising in spite of these challenges. User-generated data can be useful for generating hypotheses to investigate with higher quality data sources.

Acknowledgments

This work was supported through federal contract number HHSP233201500048I. 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 thank the following individuals for the support and comments which have greatly improved this research: Sharon Arnold, Martin Barron, Lynette Bertsche, Teresa Zayas Cabán, Amanda Cash, Mike Cohen, Rashida Dorsey, Prashila Dullabh, Sherry Emery, Don Jang, Daniel Lawrence, Jeongsoo Kim, Yoonsang Kim, Katie O'Doherty, Megha Ravanam, Samantha Rosner, Ben Skalland, Jim Sorace, and Ernie Tani.

References

Agency for Healthcare Research and Quality (AHRQ). Retrieved 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. Retrieved September 26, 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.

Brownstein, J.S., Freifeld, C.C., Reis, B.Y., and Mandl, K.D. (2008). Surveillance sans frontières: 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.

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

Center for Medicare and Medicaid Services (CMS) (2018). Medicare Current Beneficiary Survey 2015 methodology report. Retrieved September 26, 2018 at https://www.cms.gov/Research-Statistics-Data-and-Systems/Research/MCBS/Downloads/MCBS2015MethodReport508.pdf

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

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.

Flu Near You. (2018). Retrieved August 14, 2018, from https://flunearyou.org

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.

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. http://projects.isr.umich.edu/csdi/quality.cfm

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

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

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

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.

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.

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

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).

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

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

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.

Mellon, J., and Prosser, C., 2017. Twitter and Facebook are not representative of the general population: Political attitudes and demographics of British social media users. *Research & Politics*, *4*(3).

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

Muth, M. K., Sweitzer, M., Brown, D., Capogrossi, K., Karns, S., Levin, D., Okrent, A., Siegel, P., and Zhen, C. (2016). Understanding IRI household-based and store-based scanner data. United States Department of Agriculture, Economic Research Service.

National Academies of Sciences, Engineering, and Medicine (NAS). (2017a). *Federal Statistics, Multiple Data Sources, and Privacy Protection: Next Steps*. Division of Behavioral and Social Sciences and Education; Committee on National Statistics; Panel on Improving Federal Statistics for Policy and Social Science Research Using Multiple Data Sources and State-of-the-Art Estimation Methods. Washington, DC: National Academies Press.

National Academies of Sciences, Engineering, and Medicine (NAS). (2017b). *Innovations in Federal Statistics: Combining Data Sources While Protecting Privacy*. Division of Behavioral and Social Sciences and Education; Committee on National Statistics; Panel on Improving Federal Statistics for Policy and Social Science Research Using Multiple Data Sources and State-of-the-Art Estimation Methods. 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. Retrieved September 26, 2018 from https://www.cdc.gov/nchs/data-linkage/cms/nchs_medicare_linkage_methodology_and_analytic_considerations.pdf

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 Management and Budget. (2014). *M-14-06: Guidance for Providing and Using Administrative Data for Statistical Purposes*. Retrieved September 24, 2018 from https://obamawhitehouse.archives.gov/sites/default/files/omb/memoranda/2014/m-14-06.pdf

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

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. Retrieved 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.

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

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.

Sweitzer, M., Brown, D., Karns, S., Muth, M. K., Siegel, P., and Zhen, C. (2017). Food-at-home expenditures: Comparing commercial household scanner data From IRI and government survey data. United States Department of Agriculture, Economic Research Service.

U.S. Food and Drug Administration (FDA). (2018). FDA's Sentinel Initiative. Retrieved January 12, 2018, from https://www.fda.gov/Safety/FDAsSentinelInitiative/ucm2007250.htm

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.

Zeifman, I. (2017). Bot traffic report 2016. *Imperva*. Retrieved August 17, 2018, from https://www.incapsula.com/blog/bot-traffic-report-2016.html

Appendix

Appendix Table 1: Notable Use Cases from Literature Review

Data Category	Data Type	Organization	Topic	Summary	Citation(s)
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 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 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.

Appendix Table 1 (continued): Notable Use Cases from Literature Review

Data Category	Data Type	Organization	Topic	Summary	Citation(s)
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.	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.
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 to 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

Appendix Table 1 (continued): Notable Use Cases from Literature Review

Data Category	Data Type	Organization	Topic	Summary	Citation(s)
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 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., 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.
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.

Appendix Table 1 (continued): Notable Use Cases from Literature Review

Data Category	Data Type	Organization	Topic	Summary	Citation(s)
Public Sector	Government registry	Academic research	Cardiology research	The 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 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

Appendix Table 1 (continued): Notable Use Cases from Literature Review

Data Category	Data Type	Organization	Topic	Summary	Citation(s)
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., and von Kries, R. (2008). Vaccination coverage in children can be estimated from health insurance data. BMC Public Health, 8, 82.

Appendix Table 1 (continued): Notable Use Cases from Literature Review

Data Category	Data Type	Organization	Topic	Summary	Citation(s)
Public Sector	Satellite imagery data	Academic research	Measles transmission	Nighttime satellite imagery from the Defense of Meteorological Satellite Program Operational Linescan System 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

Appendix Table 1 (continued): Notable Use Cases from Literature Review

Data Category	Data Type	Organization	Topic	Summary	Citation(s)
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 EHR, 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/FDAsSentinelInitiative/ucm2007250.htm

Appendix Table 1 (continued): Notable Use Cases from Literature Review

Data Category	Data Type	Organization	Topic	Summary	Citation(s)
Public/Private	EHR/EMR	Centers for Disease Control and Prevention	Monitoring and surveillance	The National Syndromic Surveillance Program from the 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. 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.

Appendix Table 1 (continued): Notable Use Cases from Literature Review

Data Category	Data Type	Organization	Topic	Summary	Citation(s)
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 five million items. Changes in inflation trends can be spotted in a much timelier manner compared to the monthly Consumer Price Index. These statistics are considered to be an inflation measure 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/

Appendix Table 1 (continued): Notable Use Cases from Literature Review

Data Category	Data Type	Organization	Topic	Summary	Citation(s)
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/cpi/overview

Appendix Table 1 (continued): Notable Use Cases from Literature Review

Data Category	Data Type	Organization	Topic	Summary	Citation(s)
Private Sector	Transaction Data	USDA Economic Research Service	Grocery retails sales for nutrition research	Billions of transactions from on grocery retail sales from an array of outlet types provide a unique resource for conducting nutrition research. Two products have been developed, the InfoScan retail scanner data as a standalone data source of retail sales transactions and the Consumer Network household scanner data which links survey data from 120,000 households to their scanner data.	Muth, M. K., Sweitzer, M., Brown, D., Capogrossi, K., Karns, S., Levin, D., Okrent, A., Siegel, P., and Zhen, C. (2016). *Understanding IRI household-based and store-based scanner data*. United States Department of Agriculture, Economic Research Service; Sweitzer, M., Brown, D., Karns, S., Muth, M. K., Siegel, P., and Zhen, C. (2017). Food-at-Home Expenditures: Comparing Commercial Household Scanner Data From IRI and Government Survey Data. United States Department of Agriculture, Economic Research Service.

Appendix Table 1 (continued): Notable Use Cases from Literature Review

Data Category	Data Type	Organization	Topic	Summary	Citation(s)
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 households 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.

Appendix Table 1 (continued): Notable Use Cases from Literature Review

Data Category	Data Type	Organization	Topic	Summary	Citation(s)
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.
Private Sector	Web data	Health Canada World Health Organization	Monitoring and surveillance	The Global Public Health Intelligence Network 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.

Appendix Table 1 (continued): Notable Use Cases from Literature Review

Data Category	Data Type	Organization	Topic	Summary	Citation(s)
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, including data from news aggregators, eyewitness reports, expert-curated discussions, and validated official reports. HealthMap's apps are used by public health departments and government agencies, including CDC, the Department of Defense, and the World Health Organization.	HealthMap. (2018). Retrieved January 12, 2018, from http://www.healthmap.org/en/

Appendix Table 1 (continued): Notable Use Cases from Literature Review

Data Category	Data Type	Organization	Topic	Summary	Citation(s)
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 includes deployment 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
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.

Appendix Table 1 (continued): Notable Use Cases from Literature Review

Data Category	Data Type	Organization	Topic	Summary	Citation(s)
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 include 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 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.

Appendix Table 1 (continued): Notable Use Cases from Literature Review

Data Category	Data Type	Organization	Topic	Summary	Citation(s)
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 residents 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