Title
Risks and Opportunities to Ensure Equity in the Application of Big Data Research in Public Health

Journal
Annual Review of Public Health, 43(1)

ISSN
0163-7525

Authors
Wesson, Paul
Hswen, Yulin
Valdes, Gilmer
et al.

Publication Date
2022-04-05

DOI
10.1146/annurev-publhealth-051920-110928

Peer reviewed

HHS Public Access

Author manuscript

Annu Rev Public Health. Author manuscript; available in PMC 2023 April 05.

Published in final edited form as:

Annu Rev Public Health. 2022 April 05; 43: 59–78. doi:10.1146/annurev-publhealth-051920-110928.

Risks and Opportunities to Ensure Equity in the Application of Big Data Research in Public Health

Paul Wesson[1,2], **Yulin Hswen**[1,2], **Gilmer Valdes**[1,3], **Kristefer Stojonowski**[4], **Margaret A Handley**[1,5,6]

[1.]Department of Epidemiology and Biostatistics, UCSF, San Francisco

[2.]Bakar Computational Health Sciences Institute, UCSF, San Francisco

[3.]Department of Radiation Oncology, UCSF, San Francisco

[4.]Department of Health Behavior and Health Education, School of Public Health, University of Michigan

[5.]Department of Medicine, UCSF, Zuckerberg San Francisco General Hospital and Trauma Center, San Francisco

[6.]PRISE Center: Partnerships for Research in Implementation Science for Equity, University of California San Francisco

Abstract

The “big data” revolution presents an exciting frontier to expand public health research, broadening the scope of research, and increasing the precision of answers. Despite these advances, scientists must be vigilant against also advancing potential harms towards marginalized communities. In this review we provide examples in which big data applications have (unintentionally) perpetuated discriminatory practices, while also highlighting opportunities for big data applications to advance equity in public health. Here, big data is framed in the context of the five Vs (Volume, Velocity, Veracity, Variety and Value), and we propose a sixth V, Virtuosity, which incorporates equity and justice frameworks. Analytic approaches to improve equity are presented using social computational big data, fairness in machine learning algorithms, medical claims data, and data augmentation approaches as illustrations. Throughout, we emphasize the biasing influence of data absenteeism and positionality and conclude with recommendations to incorporate an equity lens into big data research.

Keywords

health equity; multi-level models; Computational Epidemiology; Digital Phenotype; machine learning; multiple systems estimation

Corresponding Author: Margaret A Handley PhD MPH, Department of Epidemiology and Biostatistics, University of California, San Francisco, Mission Hall, 550 16th Street, Suite 300, San Francisco California, Margaret.Handley@ucsf.edu.

INTRODUCTION

In the context of health, big data is considered “health data from multiple sources at scale”(84) and is considered ‘big’ when, in fact, it is complex. Big data encompasses: data from biospecimens; health records; medical and other imaging; individual, community, or satellite sensors; administrative records; policies, laws, and human rights records; environmental indicators; behavioral data; internet, social media, and mobile phone data; stock market trends, public opinion and more. A commonality for big data research in public health is that these data are almost always evaluated using observational rather than experimental methodologies.

Since a 2018 review of big data and public health (99), the risks to misinterpretations of big data findings for vulnerable populations that may exacerbate health inequity have become more apparent[1]. Equity-focused critiques suggest that incomplete conceptualizations of big data may solidify biases, which further marginalize vulnerable populations (37). For example, concerns about racial bias in algorithms (112), a digital divide that only includes data for certain populations (84), and the absence of big data in health and social systems in several lower income countries (36), are widely expressed.

There are also incredible opportunities to leverage big data for health equity, through bringing together structural determinants data previously not included in epidemiologic research, integrating new forms of data, and bridging policy analysis with health planning -- all of which can aid in directing public health action to ensure more equitable outcomes (28,78) (Table 1). Advocates highlight the importance of big data for the uptake of evidence-based interventions (the purview of implementation science) as well as for, “delivery of the right intervention to the right population at the right time, and includes consideration of social and environmental determinants” (78). Research linking health data with new forms of structural determinants data, for example with measures of racially directed violence or discriminatory housing policies (such as redlining) (104), can help expand the scope of public health (28). By de-silo-ing and amalgamating data from seemingly unconnected sources (“mash ups”) advocates hope to create new actionable knowledge (34).

In this paper we: (1) summarize concerns about big data and health equity (2) present a series of analytic approaches to explore and address inequities using big data, and (3) discuss data augmentation methods to embed a health equity lens into big data research. We use the 5 V’s framing (6) encompassing: Volume, Velocity, Veracity, Variety and Value (Box 1, Figure 1) adding a 6th V, for Virtuosity, to re-conceptualize big data research to explicitly focus on equity.

1. DATA DO NOT MAKE THEMSELVES: RISKS FROM BIG DATA CONCEPTUALIZATION

Many critiques of big data concern the veracity and variety of data and risks to health equity resulting from how data are created. Precision public health, for example, is criticized for relying on datasets that don’t reflect underlying structural relationships, missing

[1]Health equity, in the context of this paper can be considered not as the absence of health disparities but a view that prioritizes and pursues positive change to achieve equity for all (13).

upstream socio-economic determinants (77). For example, early in the COVID-19 pandemic, cell phone-based mobility data was misinterpreted to suggest that non-compliance with restrictions was widespread in low-income communities in the U.S (compared to higher income communities) (20,106). A fuller exploration demonstrated that the privileges of working from home were not available to essential workers who had to continue commuting for work (20). Without a deeper understanding of upstream drivers, there is the risk that big data-based policies will further exacerbate inequities.

The potential to address health equity through big data research rests in large part, on the inclusiveness and accuracy of data for all types of individuals and communities (veracity and variety). Although there are numerous national population-based surveys in the US and other countries which have engaged methods to include more diverse samples, declining survey participation rates over time have contributed to challenges in using these national surveys for some equity-focused research (8). Widespread concerns have emerged about how little attention to inclusion and representation exist in big data formation, such that those who identify as racial ethnic or gender minorities, individuals from resource poor settings with limited cyberinfrastructures to capture data, and from environments for which providing personal data can result in discrimination and underrepresentation (30). One of the most important challenges associated with gaps in data for vulnerable populations can be thought of as ‘data absenteeism’ (84) whereby some groups are ‘absent’ from data. Paramount to any scientific endeavor is a need to understand how data were collected (or generated) and the gap between the study population (study participant data) and the target population. Barring a true census in which everyone is included, people will be missed; and it is incumbent on the scientist to assess if certain types of people or communities are systematically missed or misrepresented. Equity in who benefits from scientific knowledge and resource allocation fundamentally cannot be achieved if groups are not represented in the scientific studies that generate the knowledge that informs the distribution of resources.

Additionally, data are not reflective of the real world but instead represent protocols designed to collect or produce data, and the conscious or unconscious biases of investigators which may systematically miss or undercount certain communities (30). The United States census, for example, tasked with counting every resident of the United States every ten years, has systematically undercounted racial and ethnic minorities (and minority language groups) in communities because census workers did not feel comfortable visiting homes in low-income neighborhoods, where racial/ethnic minorities disproportionately live (4). As a result, these communities do not benefit from resources that are allocated according to population size.

There is a well-recognized digital divide in which many people do not have access to digital tools used to collect data (e.g., smartphones, internet portals). Data absenteeism is particularly relevant to global health as many data do not exist for populations who do not have widespread internet use. Assumptions that individuals are the sole owners of cell phones or computer devices and that data captured reflect these same unique individuals also generates biases in data, as demonstrated in a study of Ebola transmission risk in which multiple individuals shared a phone (36). Conceptualizing virtuosity in the formation of big data requires asking: *Who is included and not included here and why (variety)? Can we*

address possible data gaps with supplementation methods or datasets (variety)? How could discrimination through lack of inclusion bias interpretation, and have a detrimental impact (value)? Are there theories related to marginalization we can apply to the research questions that enable a more thorough understanding of the data inputs and their meaning (veracity)? What assumptions are we making about historical, upstream or socioecological factors, racialization or discrimination risk in regard to the questions we are exploring (veracity)? How are we making assumptions about variables in our data and who they speak for, and how can we address them, if biases may be present (veracity)?

Data absenteeism has been linked to inaccurate algorithms, as when limited diversity among individuals included in machine learning-based risk models, results in model preferences that unfairly characterize what is 'normal', deviating from 'normal' or excluded altogether (9,24,30,112,118). Big data critiques also refer to uneven power structures embedded in the data's origins (referred to as the positionality of data and its creators), resulting in calls for the democratization of data and data augmentation (37). Theory can play a role in how one approaches big data and equity (39), not only a socioecological theory to help understand different layers of information, but with theories presenting a perspective on *What shapes inequities and for whom?* and *Who's in the position of power in the data structure, development of data, analysis?*

Machine learning programs are intended to improve health outcomes, reduce expenditures, and improve service delivery in clinical and public health programs. Machine learning involves developing algorithms and applying them to synthesized data (41). Recent studies show how discrimination can occur when algorithms are applied in the absence of complete data, in the presence of biased data and/or when biased assumptions are used to characterize data inputs (35, 41, 112, 117, 133). In a highly publicized study, electronic health record algorithms that determined which patients were referred for extra care, Obermeyer et al identified that the algorithm excluded Black patients who were equally as sick as White patients (112). The algorithm developers had inappropriately applied a variable capturing total health care expenditures as a proxy for unmet health needs in establishing risk scores. This example illustrates the inverse prevention law, in which those with greatest need, are least likely to receive resources (14) because the cost variable did not capture the unmet needs of Black patients (value). In the following sections, we present various approaches to enmesh equity approaches into big data research to mitigate these challenges.

2. ANALYTIC APPROACHES TO IMPROVE EQUITY FOR BIG DATA

Equity in Social Computational Big Data—Social computational big data includes data from technology companies such as Google, Facebook and Twitter (139) which are collected and stored from application programing interfaces (APIs), and contain data restrictions. (126). Moment by moment oscillations in behavior that are depicted on these online platforms leave a digital footprint which can be aggregated to uncover emerging population trends across a wide range of health topics (54,56,57,65,69–70, 82–83,111,114,120–122,,127,136,139,148). Social computational data is frequently validated against gold standard metrics, such as governmental sources, to corroborate it's use for predicting real world activities (11,55,56,62,87,128). However, ethical challenges exist and

this section examines the sixth “V” of Virtuosity in the context of using social computational big data focusing on: 1) data access and availability, 2) inclusion and representativeness 3) advantages, 4) methods of analyses, and 5) future goals for the field as related to equity.

As of 2021, on average 500 million tweets are tweeted on Twitter, daily, with 1.88 billion daily active users on Facebook, 500 million on Instagram, and 3.5 billion searches on Google (17,38,97). This massive amount of data volume and velocity provides an archive of human behavior. Data must be captured through the application programming interface (APIs) that technology giants provide. However, the API has access limitations based on industry controls (124), which reduce the accessibility for scientists to capturing the full volume and velocity of big data on these platforms. Originally, Facebook provided a public API to collect their data for research related purposes (51,85), but it is no longer publicly accessible. For Twitter, it is estimated that Twitter’s free streaming API enables the data collection of 1% of all tweets (100–101) significantly less than the full dataset of tweets from “Twitter’s Firehose,” a prohibitively expensive service (72). Consequently, social computational big data is highly volatile as it is based on the ever-changing nature of the industry and public opinion, such as personal privacy and security demands (2,68). Academic scientists are not the gatekeepers of this data but must work with the industry’s positionality as to what and how much data can be retrieved.

Data absenteeism is an important concern in social computational big data research (59,81). These online data sources do not bridge the digital divide but may in fact widen it and cannot be considered a replacement for methods that capture the needs of digitally excluded populations, such as those who cannot afford access, rural areas with limited connectivity, and resource constrained nations with political and economic barriers to access (138). Social computational big data does not undergo population-based sampling and thus, is not representative of the general population (100). Social media platforms are skewed towards younger users (18–29 years old) and those in urban areas. (32,95). Both reduce the generalizability of findings to estimate populations trends (32,45,93). Finally, the ability to identify the demographics of online users is a challenge, and even though new computational methods can detect user demographics on social media (18–19), data anonymity and privacy set forth by the industry often limit the collection of meta-data from their users profiles.

Social computational big data can provide a richer picture of experiences, compared with traditional data capturing methods, which has important equity implications for public health research. The origins of these platforms were not intended for research but recall and social desirability bias may be less present. Internet search queries and discussions on social platforms are organic, making them more authentic, unfiltered and genuine compared to data collected explicitly for research related purposes 12,33, 42,88). Lastly, without the barriers of social desirability bias these data can present more truthful and accurate view about beliefs and behaviors (15,22,53, 60, 64,89,131,140,147) For instance, people tend to lie when it comes to racially charged topics. An empirical example explored how racist searches on Google were a robust negative predictor of Obama’s voting share while national survey estimates about being racist were not (135).

A validation method for social computational big data compares results to governmental sources (gold standard) as with the example of Yelp reviews being validated for capturing foodborne illness through correlating Yelp results with data from Centers for Disease Control foodborne outbreak reports (110). This research led to the development of a supplemental foodborne illness surveillance system combining Twitter and Yelp data for public health departments' foodborne illness tracking.(50) Such an approach is more sensitive and responsive to real-time signals compared to federal surveys of health conditions and behaviors, which can be limited in data lag and representativeness. (8)

Another method of validation is to establish a relationship between online beliefs and behaviors and offline consequences. An example of this online-to-offline relationship is how online racism can be a signal of the perpetuation of real-world hate crimes (58,105). A strong time series lag time correlation has been found between Anti-Islam related hashtags on Twitter and anti-Muslim hate crimes (102) and evidence of a 1-week earlier lag time between the negative sentiment towards Mexicans and Hispanics on Twitter and worse mental health outcomes in this population (58) suggests the connection between the online and offline world. Therefore, data from online sources can be used to reveal subversive feelings and beliefs that traditional epidemiologic methods may not fully contend with.

Finally, the racialization of diseases and the long history of abuse conducted in medical research (40) has prevented scientists from capturing data on stigmatized and marginalized populations (5,12,43,91). For instance, a review of government funded cancer studies found that all racial and ethnic minorities are considerably underrepresented in cancer clinical trials with fewer than 2% of studies focusing on minority health needs (21,80). Even in studies of environmentally-related diseases that disproportionally affect minority communities, Black, Brown, and Indigenous populations are less likely than their White counterparts to be represented.(16,80). Data from online social media platforms may help reduce this gap in recruitment, as it has been shown that a greater proportion of racial and ethnic minorities use social networks (32,45).

Fairness in Machine Learning—Technically, the machine learning community approaches equity through the concept of fairness. In recent years there have been substantial efforts to formalize this concept in machine learning algorithms. Three formal definitions have gained recognition: 1) anti-classification, 2) classification parity and 3) calibration (23). In the anti-classification approach, protected attributes like race, gender or age, and their proxies, are not explicitly included in analyses. The idea is that these protected attributes should not be considered in the final model, which by definition would make predictions unequally across all protected groups. Therefore, by excluding them the model is forced to predict the expected value of the outcome by using variables that are not considered to lead to unfair models. Here, no guarantees about the accuracy of the model for different groups are made. With classification parity, common definitions of model accuracy or performance (e.g. false negative rate) are forced to be equal across different groups defined by the protected variables. As such, in the case the model is used to redistribute limited resources (i.e. access to health care), and then protected groups are equally considered. Although this guarantees the model performs the same for all groups, it usually comes at the expense of a minimum common denominator where at least one group

performs worse than it otherwise would. In calibration, it is required that conditional on the risk-estimate provided by the model, outcomes are independent of the protected attributes. Although this last definition is always desirable, in practice it falls short of protecting all groups.

The definitions of fairness seem intuitive to an extent. For instance, it is natural to think that if one is expected to build a fair model, protected variables should not be used to decrease or increase the probability of certain prediction (anti-classification). Equally, if the final model does not have the same accuracy across all groups defined by the protected variables one might think that the model is not fair towards that group (classification parity). Finally, it is logical to assume that if a model is on average representing the true underlying risk of the specific group this model is fair (calibration). More importantly, given that a formal definition of fairness is agreed upon, algorithms can be designed to satisfy it (3). For instance, a recent method constructs decision rules that ensure true positive rates and false positive rates are equal among pre-specified groups (48). Equally, techniques that use pre- and post-processing of the input variables and regularization to lead to parsimonious models have been devised to guarantee parity across different demographic variables (103,150). Unfortunately, no notion of fairness comes for free without affecting model performance for some or all groups (96).

Let us take the concept of anti-classification. Sometimes, the variables left out have highly predictive values and not including them results in a less accurate model for one of the groups. For example, measurements of oxygen saturation level has been shown to be less accurate for Black patients than White patients (133). As such, an algorithm that leaves out protected variables would not correct for the inherent bias of the technique. Similarly, although male patients with breast cancer have higher mortality as compared to female patients (while the latter have substantially higher incidence) (143), any model trying to predict mortality of breast cancer patients or incidence must consider sex to accurately predict risk levels. Additionally, identifying all proxy variables is extremely difficult and if they are left in the model the influence of the protected variables on the outcome can be ‘learned’ even when they are not present. For instance, the difference in intensity of the pixels for computer tomography images is enough for an algorithm to ‘learn’ that the images come from different hospital systems (149). If hospital systems see patients with different underlying risks, then the model would use this information without accurately learning the true underlying features that would lead to a correct interpretation of the image.

Expanding on the male breast cancer example, due to the stark differences, any dataset built to predict mortality of breast cancer patients will be comprised of many more female than male patients, and requiring a false negative rate to be the same across all protected groups (classification parity) will be detrimental for female patients (143). Finally, requiring proper calibration (outcomes need to be independent of the protected attributes conditional on the risk estimate) can be extremely misleading if the outcome is not well explained by the collected variables or a proxy outcome is analyzed. In the Obermeyer study described earlier, total health care expenditures was used as a proxy for unmet health needs in establishing risk scores (112). Requiring calibration of a model built with these data would have been problematic for Black patients as it would be reinforce the root problem; the

fact that the outcomes collected (expenditures) did not represent the true need for medical attention.

Although debiasing machine learning models have gained a lot of attention in the recent years, solutions to apply all scenarios have not yet been found, underscoring the difficult nature of this problem. Therefore, depending on the specific problem at hand, scientists should investigate the three current accepted definitions of fairness and evaluate which are most relevant to the specific problem being tackled.

Equity and Clustered Big (claims) Data—Claims data from the Centers for Medicare & Medicaid or private health insurers are rich sources of big data that are used to collect habitually unchanging patient health information and other demographics. Claims data can be considered as high in Volume given how much data are collected during healthcare visits, of which there are a Variety of data, such as health conditions, diagnostic codes, billing, and payment and the data are produced at a high Velocity—frequently updating given how many health services are provided daily. But, what about the Veracity or the trustworthiness of the data? Incorporating the 6th V—Virtuosity—supports an equity lens in which to interrogate the data to properly understand the patterning of health outcomes uncovered in claims data. In this section, we consider large administrative data sets that reflect routinely collected patient information.

To analyze claims data, scientists use AI to develop machine learning algorithms that "learn" to detect patterns in the data. These data are being used to examine relationships of health outcomes, for example mortality among Veterans Health Administration patients and predicting Type II diabetes (125,144). However, elucidating health patterns using claims data is challenging given the initial purpose for collecting the data are for billing rather than monitoring of patient and population health. There are numerous challenges with the use of algorithms to detect patterns in claims data.

First, many patterns detected are correlational, and are limited in their ability to explain why certain patterns are uncovered (137,146). Moreover, the validity of methods and algorithm transparency need to be addressed. One of the main biases in algorithms is omitted variable bias, which arises from having limited information on other potentially relevant factors that influence the outcome (146). The lack of information is particularly true for datapoints that measure the structural and social determinants of health. For example, in a study that predicted Type II diabetes, having limited information about the patient's diet or access to healthy and nutritious food creates omitted variable bias. Another challenge is data absenteeism. When states did not expand Medicaid entire swaths of populations were missing within claim's records. Given the various challenges with AI to analyze big (claims) data, a reorientation toward ensuring Veracity and Virtuosity are evaluated is paramount. Attention to biases including measurement error, sampling bias, and ascertainment bias, can mitigate equity risks from claims data (75). Equally important, such methods may best be viewed as hypothesis generating, rather than hypothesis testing.

To help explain "why" the patterns are seen in claims data, analysis approaches would benefit from explicitly leveraging the inherent data structures. Clustering of populations

within different environments (e.g. counties) are intrinsic data structures. One statistical technique that supports incorporating health equity lens to big data analysis are multi-level (hierarchical) models. Multi-level models (MLM) concern themself with data that are nested within clusters of higher orders of influence (e.g., years, months, classrooms, countries, etc.) (52). In typical MLM analyses, the units of analysis are individuals (at a lower level) that are grouped within contextual (higher level) units. The use of MLM, and other similarly situated analyses has been growing, particularly in research aimed at understanding how policy and other structural determinants may shape the outcomes of interest (49,79,92,115–116). For example, in a 2011 study by Klawitter the results indicated that gay men who lived in states with anti-discrimination policy protections in employment earned 8% more per year, compared to gay men in states with no anti-discrimination legislation-- portraying the importance of leveraging the clustered structure of the data to explain the patterns (79).

In health insurance claims, the data points (patient data and outcomes) are clustered within geographic regions (e.g., Census tracts, Counties, States) with diverse policy and structural environments. The diverse geographies the data are clustered within have varied policies, such as minimum wage, labor rights, anti-discrimination protections, and poverty alleviation measures. How might the policy and environmental contexts in which the patients live shape the patterning of the health outcomes that the AI analyses might find in the claims data? Scholarship is beginning to grapple with these questions as applied to big data. A study by Davis et al., found that 42% of Oregon Medicaid and commercial insurance patients that were eligible for colorectal screenings had a completed screening over a four year period (29). Interestingly, with the use of MLM, the authors found that the percentage varied by county, such that counties with higher rates of socioeconomic deprivation (e.g., lower high school graduation rates, higher unemployment rates) and lower rates of endoscopy specialists had the lowest colorectal screening rates (29). Studies that intentionally leverage the data structure can enhance equity approaches to claims data analysis, and help ensure that proper algorithm “learning” is taking place to support accurate interpretation of the results.

3. DATA AUGMENTATION TO IMPROVE BIG DATA EQUITY

Much of the appeal of ‘big data’ comes from its high Volume attribute, which carries the aspiration of statistical flexibility to support complex analyses and protect against Type I and II error. High volume, referring to the number of participants, may even bring the aspiration of generalizability, in which identified statistical associations are broadly applicable to other populations. However, a high volume of participants cannot replace the representativeness of participants, and no amount of data can confer generalizability if representativeness is lacking (71).

Data augmentation methods such as multiple systems estimation (MSE), also known as capture-recapture, can be used to assess and correct undercounting of populations, thereby improving representativeness and moving closer to equity (Virtuosity). With origins in wildlife biology, MSE has common applications in public health to estimate the size of “hidden” and “hard-to-reach” populations as well as evaluating the completeness of surveillance systems and identifying patterns of differential underreporting (47,145). MSE

estimates the total size of a population based on the degree of overlap of two or more incomplete lists (or samples) of that population (i.e., how many unique individuals are observed on multiple lists). The greater the overlap, the smaller the unobserved population (i.e., the number of people not already observed on any of the lists); conversely, the smaller the overlap, the greater the unobserved population. Several assumptions are necessary for valid estimation, yet the 'list independence' assumption often receives the most attention in public health applications (67). This assumption states that the lists used for MSE must be statistically independent from one another; being on one list does not increase or decrease one's probability of being on another list. A variety of design and statistical approaches are available to satisfy or relax this assumption, most notably regression modeling to account for list dependency (145).

MSE has been applied in countless settings, using various types of lists, to provide context and scope for populations of public health interest (10, 47,66,107,113,145). For example, Hu et al. implemented a two-sample MSE of men who have sex with men (MSM) in mainland China using high volume data from social networking sites (66). Simulating mobile phone positions covering the country, the authors recorded profile IDs from Blued, a popular MSM social networking app, over two 2-week periods seven months apart, as two independent samples for the MSE study. Each sample included nearly 2.5 million profiles, with an overlap of 1.3 million profiles. They estimated 8,288,536 MSM in mainland China. Subnational estimates by cities and provinces served as denominators to calculate the burden of HIV and sexually transmitted infections among MSM in different areas.

Incorporating at least three lists provides additional statistical flexibility for MSE. In an example of incorporating high variety data, Min et al. linked six data sets covering the spectrum of medical touchpoints to estimate the prevalence of and trends in opioid use disorder (OUD) in British Columbia, Canada from 1996 to 2017 (94). Stratified analyses identified the greatest increase in OUD prevalence among males 12–30 years and 31–44 years, from 2013 to 2017. Stratified analyses, as illustrated in the example, are especially important for the advancement of equity and highlight whether subgroups are systematically underrepresented in data (and differentially deprived of resources) or, in the case of human rights abuses, systematically targeted in instances of mass killings (44,86). As noted by Barocas, commenting on the need to include stratified estimates of race/ethnicity in applications of MSE, "inaccurate counting is another form of structural racism and leads to widening disparities." (7).

Potential Challenges and Opportunities in Record Linkage—Lists and data sources are almost always incomplete, and oftentimes systematically incomplete. Failing to acknowledge this fact risks perpetuating inequities by continuing to underrepresent vulnerable populations in scientific studies. Populations that are of key interest due to their unique vulnerabilities to health outcomes are often not represented in high Volume data sets. These populations may be missed because they only comprise a small proportion of the overall population, or because of the stigma and discrimination that often follows from disclosing their identity.

Beyond the MSE applications, record linkage of overlapping and complementary data sets can be used for other data augmentation purposes. Just as administrative records may be linked for MSE, administrative records, as well as federal population-based surveys (8), are increasingly being linked to generate complex data sets mapping individuals' contacts with institutional touchpoints. Some of these data sources may come from institutions that explicitly serve specific marginalized populations (e.g., housing services, social welfare programs, health clinics providing services to specific groups such as female sex workers or transgender women), ensuring that these populations are included (by design) in the creation of the data. Other data sources may come from hospital records, arrest records, and social services (e.g., housing, child welfare, substance use treatment). Each data source collects overlapping information with a different focus and, in aggregate, provides more context through which to view the lived experience of the individual, albeit keeping in mind the limitations of data positionality described earlier. In order to not perpetuate inequities and discriminatory practices, scientists must exercise caution in which data sets are selected for data augmentation and be vigilant against any biases that may be in place in the construction of those data sets (37).

Record linkage, which was not an aim when creating these siloed data systems, poses additional challenges. Investigators must think creatively about how to accurately match the same individual across multiple systems based on limited personal identifiers while protecting everyone's privacy. A potential solution may be found in hashing (31) and biometric scanning (e.g., fingerprint scanners, iris scanners) (1,130,141); emerging, low-cost methods, in which algorithms generate complex alphanumeric IDs based on identifiable information. The ID uniquely identifies the client's record but cannot be reverse coded to identify the client. The algorithm can be applied internally within each institution to preserve client privacy while facilitating the linkage of records across institutional touch points, clinics, and other data sources. However, in relation to biometric scanning, fingerprint scanners may be associated with the criminal justice system. Even though the code generated from the scan cannot be linked back to criminal or immigration databases, this approach to record linkage may not be considered acceptable to certain groups or be met with skepticism (1,76).

DISCUSSION

Our paper examines important equity and justice considerations in the conceptualization and application of big data methods. For users of big data, it is paramount to challenge our assumptions and develop new or improved frameworks to ensure equitable big data research practices. Williams et al., states "In particular, creators of algorithmic systems have three general classes of approaches to prevent discrimination: they can make the data less biased beforehand, build fairness criteria into the algorithm (discussed earlier), or alter the application of the rules after the algorithm runs."(146). However, it would be more ideal to prevent discrimination in all three classes. One strategic way to enshrine Virtuosity within big data practice is to foreground the experiences of particular groups that experience marginalization. Incorporating frameworks such as intersectionality theory, which recognizes that when social identity categories intersect (reflecting interlocking systems of privilege and oppression and they may result in unique and intensified forms

of discrimination), can improve the way we approach uses of big data in public health (25). At a minimum, this requires practitioners of big data to recognize and challenge their positionality to the data itself and its interpretations. For example, use of big data sources through a health equity framework mandates the grappling of issues, such as racism, sexism, homophobia, transphobia, classism, etc. The explicit application of theoretical and analytic frameworks that contend with the structural and contextual factors that shape our lives is paramount in research using big data.

To improve equity in big data research, we provide several suggestions to critically incorporate the 6th V of Virtuosity. The first is to include social epidemiologists in the research and prioritize social epidemiology training beyond programs in epidemiology and other public health disciplines. Scientists who study social epidemiology have a deeper knowledge of the structural and systemic forces that have generated a distribution of advantages and disadvantages in society (73). The second suggestion is to increase the level of diversity in researchers across disciplines pursuing big data and equity. Discriminatory biases can be prevented through the addition of a wide range of perspectives as this can reduce the likelihood of generating biases based on singular viewpoints (46). Thirdly, generate parterships between industry and academia (128). Big tech should work with social epidemiologists to generate more ethical and virtuous research. Fourthly, federal and state policies are needed to safeguard against biased and discriminatory production of big data. Fifth, as scientists it is important for us to evaluate our own biases and understand that we do not have the breadth of experience to know what is fully needed to improve equity.

Coupling these recommendations with a focus on training, it is critical that schools of public health, departments of epidemiology and biostatistics, and other data science training programs have diverse representation among students, staff, and faculty. These programs should emphasize core competencies in sampling theory and designs, participatory engagement and working with data and collaborators spanning diverse disciplines (e.g., social welfare, criminal justice), as well as understanding the role of historical and structural racism and inequality. When teaching focuses on biases, especially as they relate to study design and analysis, competencies should include a historical lens highlighting structural inequalities and interrogating its influence on data, data absenteeism, and the positionality of data. We also encourage the discussion of methodological approaches to reduce the impact of such biases. As an example, the Department of Epidemiology and Biostatistics at the University of California, San Francisco, in partnership with the Bakar Computational Health Sciences Institute and the Center for Health and Community, has developed a pre-doctoral training program that integrates the theoretical frameworks and methodological tools of behavioral and social scientiests with the rapidly evolving technical repertoire of computational health scientists. Graduates of the Data Science Training to Advance Behavioral and Social Science Expertise for Health Research (DaTABASE) program (NIH/NIMHD T32MD015070) are trained to apply analytic tools to novel data sets for research on behavioral and social processes underlying health disparities. Finally, public health practitioners must honor data augmentation and community involvement in it, and as such must listen and acknowledge the lived experience of often ‘absent’ communities and ensure that they are at the forefront of the design and development of equitable research (15, 22, 60, 89, 131,140, 147).

Acknowledgments

Research reported in this review was supported by National Institutes of Health (NIH)/National Institute of Allergy and Infectious Disease Career Development Award 5K01MH119910-02 (to P.W.); the National Institute of Biomedical Imaging and Bioengineering of the NIH under award K08EB026500 (to G.V.); NIH grants R01DK115492, 7N91020C00039, and R56AR063705 (to Y.H.) and PRISE (Partnerships for Research in Implementation Science for Equity) at the University of California, San Francisco (to M.H.).

REFERENCES

1. Abrams MP, Torres FE, Little SJ. Biometric Registration to an HIV Research Study may Deter Participation. AIDS Behav [Internet]. 2021;25(5):1552–9. Available from: 10.1007/s10461-020-02995-y [PubMed: 32767155]
2. Acquisti A, & Gross R. (2006). Imagined communities: Awareness, information sharing, and privacy on the Facebook. Paper presented at the International workshop on privacy enhancing technologies.
3. Agarwal A, Beygelzimer A, Dud´ık M, Langford J, and Wallach H. (2018). A reductions approach to fair classification. In International Conference on Machine Learning
4. Anderson MJ, Fienberg SE. Who Counts? The politics of census-taking in contemporary America. New York: Russell Sage Foundation; 1999. 119–127 p.
5. Ayhan CHB, Bilgin H, Uluman OT, Sukut O, Yilmaz S, & Buzlu S. (2020). A systematic review of the discrimination against sexual and gender minority in health care settings. International Journal of Health Services, 50(1), 44–61. [PubMed: 31684808]
6. Baro E, Degoul S, Beuscart R, Chazard E. Toward a Literature-Driven Definition of Big Data in Healthcare. Biomed Res Int. 2015;2015:639021. doi: 10.1155/2015/639021. Epub 2015 Jun 2. PMID: 26137488; PMCID: PMC4468280.
7. Barocas JA. Commentary on Jones et al. (2020): Using indirect estimation methods of drug use prevalence to address racial and ethnic health disparities. Addiction. 2020;115(12):2405–6. [PubMed: 32822524]
8. Blewett Lynn A., Call Kathleen Thiede, Turner Joanna, Robert Hest Data Resources for Conducting Health Services and Policy Research" by Blewett et al. (2018) Annual Review of Public Health 2018 39:1, 437–452
9. Biller-Andorno N and Biller A. 2019. Algorithm-Aided Prediction of Patient Preferences — An Ethics Sneak Peek. N Engl J Med. 381: 15.
10. Böhning D, Rocchetti I, Maruotti A, Holling H. Estimating the undetected infections in the Covid-19 outbreak by harnessing capture – recapture methods. Int J Infect Dis. 2020;97:197–201 [PubMed: 32534143]
11. Bollen J, Mao H, & Zeng X. (2011). Twitter mood predicts the stock market. Journal of computational science, 2(1), 1–8.
12. Brandon DT, Isaac LA, & LaVeist TA (2005). The legacy of Tuskegee and trust in medical care: is Tuskegee responsible for race differences in mistrust of medical care? Journal of the National Medical Association, 97(7), 951. [PubMed: 16080664]
13. Braveman PA, Kumanyika S, Fielding J, Laveist T, Borrell LN, Manderscheid R, Troutman A. Health disparities and health equity: the issue is justice. Am J Public Health. 2011 Dec;101 Suppl 1(Suppl 1):S149–55. doi: 10.2105/AJPH.2010.300062. Epub 2011 May 6. PMID: 21551385; PMCID: PMC3222512. [PubMed: 21551385]
14. Brownson RC, Kumanyika SK, Kreuter MW, Haire-Joshu D. Implementation science should give higher priority to health equity. Implement Sci. 2021 Mar 19;16(1):28. doi: 10.1186/s13012-021-01097-0. PMID: 33740999; PMCID: PMC7977499. [PubMed: 33740999]
15. Buntain C, & Golbeck J. (2015). This is your Twitter on drugs: Any questions? Paper presented at the Proceedings of the 24th international conference on World Wide Web.
16. Burchard EG, Oh SS, Foreman MG, & Celedón JC (2015). Moving toward true inclusion of racial/ethnic minorities in federally funded studies. A key step for achieving respiratory health equality in the United States. American journal of respiratory and critical care medicine, 191(5), 514–521. [PubMed: 25584658]

17. Carvalho JP, Rosa H, Brogueira G, & Batista F. (2017). MISNIS: An intelligent platform for twitter topic mining. Expert Systems with Applications, 89, 374–388.
18. Cesare N, Grant C, Nguyen Q, Lee H, & Nsoesie EO (2017). How well can machine learning predict demographics of social media users? arXiv preprint arXiv:1702.01807.
19. Cesare N, Grant C, & Nsoesie EO (2017). Detection of user demographics on social media: A review of methods and recommendations for best practices. arXiv preprint arXiv:1702.01807.
20. Chang S, Pierson E, Koh PW et al. Mobility network models of COVID-19 explain inequities and inform reopening. Nature 589, 82–87 (2021). 10.1038/s41586-020-2923-3 [PubMed: 33171481]
21. Chen MS Jr, Lara PN, Dang JH, Paterniti DA, & Kelly K. (2014). Twenty years post-NIH Revitalization Act: enhancing minority participation in clinical trials (EMPaCT): laying the groundwork for improving minority clinical trial accrual: renewing the case for enhancing minority participation in cancer clinical trials. Cancer, 120, 1091–1096. [PubMed: 24643646]
22. Comenetz J. (2016). Frequently occurring surnames in the 2010 Census. United States Census Bureau.
23. Corbett-Davies Sam, and Goel Sharad. “The measure and mis measure of fairness: A critical review of fair machine learning.” arXiv preprint arXiv:1808.00023 (2018).)
24. Costanza-Chock Sasha. 2018. “Design Justice, A.I., and Escape from the Matrix of Domination.” Journal of Design and Science (JoDS).
25. Crenshaw K. (1989). Demarginalizing the Intersection of Race and Sex: A Black Feminist Critique of Antidiscrimination Doctrine, Feminist Theory and Antiracist Politics. University of Chicago Legal Forum, 1989(1), 139–167. https://chicagounbound.uchicago.edu/uclf/vol1989/iss1/8/?utm_source=chicagounbound.uchicago.edu%2Fuclf%2Fvol1989%2Fiss1%2F8&utm_medium=PDF&utm_campaign=PDFCoverPages
26. Curtis A, Ajayakumar J, Curtis J, Mihalik S, Purohit M, Scott Z, Muisyo J, Labadorf J, Vijitakula S, Yax J, Goldberg DW. Geographic monitoring for early disease detection (GeoMEDD). Sci Rep. 2020 Dec 10;10(1):21753. doi: 10.1038/s41598-020-78704-5. PMID: 33303896; PMCID: PMC7728804. [PubMed: 33303896]
27. Curtis DS, Washburn T, Lee H, Smith KR, Kim J, Martz CD, Kramer MR, Chae DH. Highly public anti-Black violence is associated with poor mental health days for Black Americans. Proc Natl Acad Sci U S A. 2021 Apr 27;118(17):e2019624118. doi: 10.1073/pnas.2019624118. PMID: 33875593; PMCID: PMC8092615.
28. Dankwa-Mullan I Zhang X, Le PT, Riley WT., et al. 2021. The Science of Health Disparities Research, First Edition. Edited by Dankwa-Mullan Irene, Pérez-Stable Eliseo J., Gardner Kevin L., Zhang Xinzhi, and Rosario Adelaida M.. 2021 John Wiley & Sons, Inc. Published 2021 by John Wiley & Sons, Inc. Chapter 14 Applications of Big Data Science and Analytic Techniques for Health Disparities Research
29. Davis MM, Renfro S, Pham R, Hassmiller Lich K, Shannon J, Coronado GD, & Wheeler SB (2017). Geographic and population-level disparities in colorectal cancer testing: A multilevel analysis of Medicaid and commercial claims data. Preventive Medicine, 101(2017), 44–52. 10.1016/j.ypmed.2017.05.001 [PubMed: 28506715]
30. D’Ignazzio C and Klein LF. 2020. Data Feminism MIT Press
31. Dombrowski K, Khan B, Wendel T, Mclean K, Misshula E, Curtis R. Estimating the Size of the Methamphetamine-Using Population in New York City Using Network Sampling Techniques. Adv Appl Sociol. 2012;2(4):245–52. [PubMed: 24672746]
32. Duggan M, & Brenner J. 2013. The demographics of social media users, 2012 (Vol. 14): Pew Research Center’s Internet & American Life Project Washington, DC.
33. Dula A. 1994. African American suspicion of the healthcare system is justified: what do we do about it? Cambridge Q. Healthcare Ethics, 3, 347.
34. Dumbill E. Making Sense of Big Data. Big Data. 2013 Mar;1(1):1–2. doi: 10.1089/big.2012.1503. Epub 2012 Nov 7. PMID: 27447028. [PubMed: 27447028]
35. Eneanya ND, Yang W and Reese PP. Reconsidering the Consequences of Using Race to Estimate Kidney Function JAMA July 9, 2019 Volume 322, Number 2

36. Erikson SL. Cell Phones & Self and Other Problems with Big Data Detection and Containment during Epidemics MEDICAL ANTHROPOLOGY QUARTERLY, Vol. 32, Issue 3, pp. 315–339 [PubMed: 29520829]
37. Eubanks V. Automating Inequality 2018. St Martin's Press. 175 Fifth Avenue New York NY. US
38. Facebook Statistics 2020. (2021). Retrieved from https://www.statista.com/statistics/346167/facebook-global-dau/
39. Ford CL, Takahashi LM, Chandanabhumma PP, Ruiz ME, Cunningham WE. Anti-Racism Methods for Big Data Research: Lessons Learned from the HIV Testing, Linkage, & Retention in Care (HIV TLR) Study. Ethn Dis. 2018 Aug 9;28(Suppl 1):261–266. doi: 10.18865/ed.28.S1.261. PMID: 30116096; PMCID: PMC6092168. [PubMed: 30116096]
40. Freimuth VS, Quinn SC, Thomas SB, Cole G, Zook E, & Duncan T. (2001). African Americans' views on research and the Tuskegee Syphilis Study. Social Science & Medicine, 52(5), 797–808. [PubMed: 11218181]
41. Gianfranceso MA Tamang S Yazdany J et al. , Potential Biases in Machine Learning Algorithms Using Electronic Health Record Data JAMA Intern Med. 2018 November 01; 178(11): 1544–1547. doi:10.1001/jamainternmed.2018.3763
42. Giordano LA, Elliott MN, Goldstein E, Lehrman WG, & Spencer PA (2010). Development, implementation, and public reporting of the HCAHPS survey. Medical Care Research and Review, 67(1), 27–37. [PubMed: 19638641]
43. Göçmen İ, & Yılmaz V. (2017). Exploring perceived discrimination among LGBT individuals in Turkey in education, employment, and health care: Results of an online survey. Journal of Homosexuality, 64(8), 1052–1068. [PubMed: 27645489]
44. Green AH & Ball P(2019). Civilian killings and disappearances during civil war in El Salvador (1980–1992). Demographic Research 41:27, 781–814
45. Greenwood S, Perrin A, & Duggan M. 2016. Social media update 2016. Pew Research Center, 11(2), 1–18.
46. Hajian S, Bonchi F, & Castillo C. 2016. Algorithmic bias: From discrimination discovery to fairness-aware data mining. Paper presented at the Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining.
47. Hall HI, Song R, Gerstle JE, Lee LM. Assessing the completeness of reporting of human immunodeficiency virus diagnoses in 2002–2003: capture-recapture methods. Am J Epidemiol [Internet]. 2006 Aug 15 [cited 2014 Feb 3];164(4):391–7. Available from: http://www.ncbi.nlm.nih.gov/pubmed/16772373 [PubMed: 16772373]
48. Hardt M, Price E, and Srebro N. (2016). Equality of opportunity in supervised learning. In Advances In Neural Information Processing Systems, pages 3315–3323.).
49. Hatzenbuehler ML (2011). The social environment and suicide attempts in lesbian, gay, and bisexual youth. Pediatrics, 127(5), 896–903. 10.1542/peds.2010-3020 [PubMed: 21502225]
50. Hawkins JB, Tuli G, Kluberg S, Harris J, Brownstein JS, & Nsoesie E. (2016). A digital platform for local foodborne illness and outbreak surveillance. Online Journal of Public Health Informatics, 8(1).
51. Hogan B. (2008). A comparison of on and offline networks through the Facebook API. Available at SSRN 1331029.
52. Hox JJ, Moerbeek M, & van de Schoot R. (2018). Multilevel Analysis (3rd ed.). Routledge.Hulsen T, Jamuar SS, Moody AR, Karnes JH, Varga O, Hedensted S, Spreafico R, Hafler DA, & McKinney EF (2019). From big data to precision medicine. Frontiers in Medicine, 6(34), 1–14. 10.3389/fmed.2019.00034 [PubMed: 30723716]
53. Hswen Y, & Brownstein JS (2019). Real-time digital surveillance of vaping-induced pulmonary disease. New England Journal of Medicine, 381(18), 1778–1780.
54. Hswen Y, Brownstein JS, Liu J, & Hawkins JB (2017). Use of a digital health application for influenza surveillance in China. American journal of public health, 107(7), 1130–1136. [PubMed: 28520492]
55. Hswen Y, Brownstein JS, Xu X, & Yom-Tov E. (2020). Early detection of COVID-19 in China and the USA: summary of the implementation of a digital decision-support and disease surveillance tool. BMJ open, 10(12), e041004.

56. Hswen Y, Gopaluni A, Brownstein JS, & Hawkins JB (2019). Using twitter to detect psychological characteristics of self-identified persons with autism Spectrum disorder: a feasibility study. JMIR mHealth and uHealth, 7(2), e12264.
57. Hswen Y, Naslund JA, Brownstein JS, & Hawkins JB (2018). Online communication about depression and anxiety among twitter users with schizophrenia: preliminary findings to inform a digital phenotype using social media. Psychiatric Quarterly, 89(3), 569–580.
58. Hswen Y, Qin Q, Williams DR, Viswanath K, Subramanian S, & Brownstein JS (2020). Online negative sentiment towards Mexicans and Hispanics and impact on mental well-being: A time-series analysis of social media data during the 2016 United States presidential election. Heliyon, 6(9), e04910.
59. Hswen Y, & Viswanath K. (2015). Beyond the hype: mobile technologies and opportunities to address health disparities. Journal Of Mobile Technology In Medicine, 4(1), 39–40.
60. Hswen Y, Williams DR, Tuli G, Sewalk K, Hawkins JB, Viswanath K,... Brownstein JS (2020). Racial and Ethnic Disparities in Patient Experiences in the United States: 4-Year Content Analysis of Twitter. Journal of medical Internet research, 22(8), e17048.
61. Hswen Y, Xu X, Hing A, Hawkins JB, Brownstein JS, Gee GC. Association of "#COVID19" versus "#Chinesevirus" with anti-Asian sentiments on Twitter: March 9–23, 2020. Am J Public Health. 2021;111(5):956–964. [PubMed: 33734838]
62. Hswen Y, Zhang A, & Brownstein JS (2020). Leveraging black-market street buprenorphine pricing to increase capacity to treat opioid addiction, 2010–2018. Preventive medicine, 137, 106105.
63. Hswen Y, Zhang A, Freifeld C, & Brownstein JS (2020). Evaluation of Volume of News Reporting and Opioid-Related Deaths in the United States: Comparative Analysis Study of Geographic and Socioeconomic Differences. Journal of medical Internet research, 22(7), e17693.
64. Hswen Y, Zhang A, Sewalk K, Tuli G, Brownstein JS, & Hawkins JB (2020a). Use of social media to assess the impact of equitable state policies on LGBTQ patient experiences: An exploratory study. Paper presented at the Healthcare.
65. Hswen Y, Zhang A, Sewalk KC, Tuli G, Brownstein JS, & Hawkins JB (2020b). Investigation of Geographic and Macrolevel Variations in LGBTQ Patient Experiences: Longitudinal Social Media Analysis. Journal of medical Internet research, 22(7), e17087.
66. Hu M, Xu C, Wang J. Spatiotemporal Analysis of Men Who Have Sex With Men in Mainland China: Social App Capture-Recapture Method. JMIR MHealth UHealth. 2020; 8(1):1–13.
67. International Working Group for Disease Monitoring and Forcasting. Capture-recapture and multiple-record systems estimation. I: History and theoretical development. Am J Epidemiol [Internet]. 1995 [cited 2013 Apr 28];142(10):1047–58. Available from: http://hub.hku.hk/handle/10722/82976 [PubMed: 7485050]
68. Isaak J, & Hanna MJ (2018). User data privacy: Facebook, Cambridge Analytica, and privacy protection. Computer, 51(8), 56–59.
69. Jain SH, Powers BW, Hawkins JB, & Brownstein JS (2015). The digital phenotype. Nature biotechnology, 33(5), 462–463.
70. Jha A, & Mamidi R. (2017). When does a compliment become sexist? analysis and classification of ambivalent sexism using twitter data. Paper presented at the Proceedings of the second workshop on NLP and computational social science.
71. Johnson CL, Paulose-Ram R, Ogden CL, Carroll MD, Kruszan-Moran D, Dohrmann SM, & Curtin LR (2013). National health and nutrition examination survey. Analytic guidelines, 1999–2010.
72. Joseph K, Landwehr PM, & Carley KM (2014). Two 1% s don't make a whole: Comparing simultaneous samples from Twitter's streaming API. Paper presented at the International conference on social computing, behavioral-cultural modeling, and prediction.
73. Kawachi I, & Subramanian S. (2018). Social epidemiology for the 21st century. Social Science & Medicine, 196, 240–245. [PubMed: 29113687]
74. Kamishima T, Akaho S, Asoh H, and Sakuma J. 2012. Fairness-aware classifier with prejudice remover regularizer. In Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Springer, 35–50.

75. Kaplan RM, Chambers DA, & Glasgow RE (2014). Big data and large sample size: A cautionary note on the potential for bias. Clinical and Translational Science, 7(4), 342–346. 10.1111/cts.12178 [PubMed: 25043853]
76. Kavanagh MM, Baral SD, Milanga M, Sugarman J. Biometrics and public health surveillance in criminalised and key populations : policy, ethics, and human rights considerations. Lancet HIV. 2019;6(1):e51–9.
77. Kenney M, Mamo L. Med Humanit 2020;46:192–203. doi:10.1136/medhum-2018-011597 [PubMed: 31420373]
78. Khoury MJ, Engelgau M, Chambers DA, Mensah GA. Beyond Public Health Genomics: Can Big Data and Predictive Analytics Deliver Precision Public Health? Public Health Genomics. 2018;21(5–6):244–250. doi: 10.1159/000501465. Epub 2019 Jul 17. PMID: 31315115; PMCID: PMC6687519. [PubMed: 31315115]
79. Klawitter M. (2011). Multilevel analysis of the effects of antidiscrimination policies on earnings by sexual orientation. Journal of Policy Analysis and Management, 30(2), 334–358. 10.1002/pam.20563
80. Konkel L. (2015). Racial and ethnic disparities in research studies: the challenge of creating more diverse cohorts. In: National Institute of Environmental Health Sciences.
81. Kontos EZ, Emmons KM, Puleo E, & Viswanath K. (2010). Communication inequalities and public health implications of adult social networking site use in the United States. Journal of health communication, 15(sup3), 216–235. [PubMed: 21154095]
82. Kristoufek L. (2013). BitCoin meets Google Trends and Wikipedia: Quantifying the relationship between phenomena of the Internet era. Scientific reports, 3(1), 1–7.
83. Lamb MR, Kandula S, & Shaman J. (2021). Differential COVID-19 case positivity in New York City neighborhoods: Socioeconomic factors and mobility. Influenza and Other Respiratory Viruses, 15(2), 209–217. [PubMed: 33280263]
84. Lee EW, Viswanath K. Big Data in Context: Addressing the Twin Perils of Data Absenteeism and Chauvinism in the Context of Health Disparities Research (2020). J Med Internet Res 2020;22(1):e16377) doi: 10.2196/16377 [PubMed: 31909724]
85. Lewis K, Kaufman J, Gonzalez M, Wimmer A, & Christakis N. (2008). Tastes, ties, and time: A new social network dataset using Facebook. com. Social networks, 30(4), 330–342.
86. Lum K, Meg Price EP, and Banks D (2013). Applications of Multiple Systems Estimation in Human Rights Research. The American Statistician, 67:4, 191–200.
87. Maharana A, Cai K, Hellerstein J, Hswen Y, Munsell M, Staneva V,... Nsoesie EO (2019). Detecting reports of unsafe foods in consumer product reviews. JAMIA Open, 2(3), 330–338. [PubMed: 31984365]
88. Malebranche DJ, Peterson JL, Fullilove RE, & Stackhouse RW (2004). Race and sexual identity: perceptions about medical culture and healthcare among Black men who have sex with men. Journal of the National Medical Association, 96(1), 97. [PubMed: 14746359]
89. Mateos P. 2007. A review of name-based ethnicity classification methods and their potential in population studies. Population, Space and Place, 13(4), 243–263.
90. McKenna B, Myers MD, & Newman M. (2017). Social media in qualitative research: Challenges and recommendations. Information and Organization, 27(2), 87–99.
91. McMurtry CL, Findling MG, Casey LS, Blendon RJ, Benson JM, Sayde JM, & Miller C. (2019). Discrimination in the United States: Experiences of Asian Americans. Health services research, 54, 1419–1430. [PubMed: 31657465]
92. Medeiros V, Ribeiro RSM, & Amaral PVM do. (2021). Infrastructure and household poverty in Brazil: A regional approach using multilevel models. World Development, 137(105118), 1–14. 10.1016/j.worlddev.2020.105118
93. Mellon J, & 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), 2053168017720008.
94. Min JE, Pearce LA, Homayra F, Dale LM, Barocas JA, Irvine MA, et al. Estimates of opioid use disorder prevalence from a regression-based multi-sample stratified capture-recapture analysis. Drug Alcohol Depend. 2020; 217(July).

95. Mislove A, Lehmann S, Ahn Y-Y, Onnela J-P, & Rosenquist J. (2011). Understanding the demographics of Twitter users. Paper presented at the Proceedings of the International AAAI Conference on Web and Social Media.

96. Mitchell S, Potash E, Barocas S, D'Amour A, & Lum K. (2018). Prediction-Based Decisions and Fairness: A Catalogue of Choices, Assumptions, and Definitions. arXiv e-prints, arXiv-1811

97. Mohsin M. (2020). 10 Google Search Statistics You Need to Know Retrieved from https://www.oberlo.com/blog/google-search-statistics

98. Mondria J, & Wu T. (2013). Imperfect financial integration and asymmetric information: competing explanations of the home bias puzzle? Canadian Journal of Economics/Revue canadienne d'économique, 46(1), 310–337.

99. Mooney SJ, Pejaver V. Big Data in Public Health: Terminology, Machine Learning, and Privacy. Annu Rev Public Health. 2018 Apr 1;39:95–112. doi: 10.1146/annurev-publhealth-040617-014208. Epub 2017 Dec 20. PMID: 29261408; PMCID: PMC6394411. [PubMed: 29261408]

100. Morstatter F, Pfeffer J, & Liu H. (2014). When is it biased? Assessing the representativeness of twitter's streaming API. Paper presented at the Proceedings of the 23rd international conference on world wide web.

101. Morstatter F, Pfeffer J, Liu H, & Carley K. (2013). Is the sample good enough? comparing data from twitter's streaming api with twitter's firehose. Paper presented at the Proceedings of the International AAAI Conference on Web and Social Media.

102. Müller K, & Schwarz C. (2020). From hashtag to hate crime: Twitter and anti-minority sentiment. Available at SSRN 3149103.

103. Goel Naman, Yaghini Mohammad, and Faltings Boi. 2018. Non-discriminatory machine learning through convex fairness criteria. In Thirty-Second AAAI Conference on Artificial Intelligence

104. Nardone A, Casey JA, Morello-Frosch R, Mujahid M, Balmes JR, Thakur N. Associations between historical residential redlining and current age-adjusted rates of emergency department visits due to asthma across eight cities in California: an ecological study. Lancet Planet Health. 2020 Jan;4(1):e24–e31. doi: 10.1016/S2542-5196(19)30241-4. PMID: 31999951. [PubMed: 31999951]

105. Nayak A. (2010). Race, affect, and emotion: young people, racism, and graffiti in the postcolonial English suburbs. Environment and Planning A, 42(10), 2370–2392.

106. New York Times https://www.nytimes.com/interactive/2020/04/03/us/coronavirus-stay-home-rich-poor.html?auth=link-dismiss-google1tap

107. Nielsen S, Hansen JF, Hay G, Cowan S, Jepsen P, Omland LH, et al. Hepatitis C prevalence in Denmark in 2016 — An updated estimate using multiple national registers. PLoS One. 2020;15(9):1–12.

108. Nguyen TT, Criss S, Dwivedi P, Huang D, Keralis J, Hsu E, Phan L, Nguyen LH, Yardi I, Glymour MM, Allen AM, Chae DH, Gee GC, Nguyen QC. Exploring U.S. Shifts in Anti-Asian Sentiment with the Emergence of COVID-19. Int J Environ Res Public Health. 2020 Sep 25;17(19):7032. doi: 10.3390/ijerph17197032. PMID: 32993005; PMCID: PMC7579565.

109. Nguyen TT, Adams N, Huang D, Glymour MM, Allen AM, Nguyen QC. The Association Between State-Level Racial Attitudes Assessed From Twitter Data and Adverse Birth Outcomes: Observational Study. JMIR Public Health Surveill. 2020 Jul 6;6(3):e17103. doi: 10.2196/17103. PMID: 32298232; PMCID: PMC7381033.

110. Nsoesie EO, Kluberg SA, & Brownstein JS (2014). Online reports of foodborne illness capture foods implicated in official foodborne outbreak reports. Preventive medicine, 67, 264–269. [PubMed: 25124281]

111. Nuti SV, Wayda B, Ranasinghe I, Wang S, Dreyer RP, Chen SI, & Murugiah K. (2014). The use of google trends in health care research: a systematic review. PloS one, 9(10), e109583.

112. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. Science. 2019 Oct 25;366(6464):447–453. doi: 10.1126/science.aax2342. PMID: 31649194. [PubMed: 31649194]

113. Orellana C, Kreshpaj B, Burstrom B, Davis L, Frumento P, Hemmingsson T, et al. Organisational factors and under-reporting of occupational injuries in Sweden : a population-based study using capture-recapture methodology. Occup Env Med. 2021;1–8.
114. Oum S, Chandramohan D, & Cairncross S. (2005). Community-based surveillance: a pilot study from rural Cambodia. Tropical medicine & international health, 10(7), 689–697. [PubMed: 15960708]
115. Pachankis JE, Hatzenbuehler ML, Berg RC, Fernández-Dávila P, Mirandola M, Marcus U, Weatherburn P, & Schmidt AJ (2017). Anti-LGBT and Anti-Immigrant Structural Stigma: An Intersectional Analysis of Sexual Minority Men's HIV Risk When Migrating to or Within Europe. Journal of Acquired Immune Deficiency Syndromes, 76(4), 356–366. 10.1097/QAI.0000000000001519 [PubMed: 28787329]
116. Pachankis JE, Hatzenbuehler ML, Mirandola M, Weatherburn P, Berg RC, Marcus U, & Schmidt AJ (2017). The Geography of Sexual Orientation: Structural Stigma and Sexual Attraction, Behavior, and Identity Among Men Who Have Sex with Men Across 38 European Countries. Archives of Sexual Behavior, 46(5), 1491–1502. 10.1007/s10508-016-0819-y [PubMed: 27620320]
117. Parikh RB, Teeple S, Navathe AS. Addressing Bias in Artificial Intelligence in Health Care. JAMA. 2019;322(24):2377–2378. doi:10.1001/jama.2019.18058 [PubMed: 31755905]
118. Paulus JK, Kent DM. Predictably unequal: understanding and addressing concerns that algorithmic clinical prediction may increase health disparities. NPJ Digit Med. 2020 Jul 30;3:99. doi: 10.1038/s41746-020-0304-9. PMID: 32821854; PMCID: PMC7393367. [PubMed: 32821854]
119. Potash E, Ghani R, Walsh J, Jorgensen E, Lohff C, Prachand N, Mansour R. Validation of a Machine Learning Model to Predict Childhood Lead Poisoning. JAMA Netw Open. 2020 Sep 1;3(9):e2012734. doi: 10.1001/jamanetworkopen.2020.12734. PMID: 32936296; PMCID: PMC7495240.
120. Pourebrahim N, Sultana S, Edwards J, Gochanour A, & Mohanty S. (2019). Understanding communication dynamics on Twitter during natural disasters: A case study of Hurricane Sandy. International journal of disaster risk reduction, 37, 101176.
121. Preis T, Moat HS, & Stanley HE (2013). Quantifying trading behavior in financial markets using Google Trends. Scientific reports, 3(1), 1–6.
122. Preis T, Moat HS, Stanley HE, & Bishop SR (2012). Quantifying the advantage of looking forward. Scientific reports, 2(1), 1–2.
123. Pruss D, Fujinuma Y, Daughton AR, Paul MJ, Arnot B, Albers Szafir D, Boyd-Graber J. Zika discourse in the Americas: A multilingual topic analysis of Twitter. PLoS One. 2019 May 23;14(5):e0216922. doi: 10.1371/journal.pone.0216922. PMID: 31120935; PMCID: PMC6532961.
124. Rate Limits. (2021). Retrieved from https://developer.twitter.com/en/docs/rate-limits
125. Razavian N, Blecker S, Schmidt AM, Smith-Mclallen A, Nigam S, & Sontag D. (2015). Population-level prediction of type 2 diabetes from claims data and analysis of risk factors. Big Data, 3(4), 277–287. 10.1089/big.2015.0020 [PubMed: 27441408]
126. Reips U-D, & Matzat U. (2014). Mining "Big Data" using big data services. International Journal of Internet Science, 9(1), 1–8.
127. Runge-Ranzinger S, Horstick O, Marx M, & Kroeger A. (2008). What does dengue disease surveillance contribute to predicting and detecting outbreaks and describing trends? Tropical medicine & international health, 13(8), 1022–1041. [PubMed: 18768080]
128. Sadilek A, Hswen Y, Bavadekar S, Shekel T, Brownstein JS, & Gabrilovich E. (2020). Lymelight: Forecasting Lyme disease risk using web search data. NPJ digital medicine, 3(1), 1–12. [PubMed: 31934645]
129. Sandifer PA, Knapp LC, Lichtveld MY, Manley RE, Abramson D, Caffey R,... Engel L. (2020). Framework for a community health observing system for the Gulf of Mexico region: preparing for future disasters. Frontiers in public health, 8, 588.

130. Serwaa-bonsu A, Herbst A, Reniers G, Ijaa W, Clark B, Kabudula C, et al. First experiences in the implementation of biometric technology to link data from Health and Demographic Surveillance Systems with health facility data. Glob Health Action. 2010;3.

131. Sewalk KC, Tuli G, Hswen Y, Brownstein JS, & Hawkins JB (2018). Using Twitter to examine Web-based patient experience sentiments in the United States: Longitudinal study. Journal of medical Internet research, 20(10), e10043.

132. Silverman B. Modern Slavery : an application of Multiple Systems Estimation [Internet]. 2014. Available from: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/386841/Modern_Slavery_an_application_of_MSE_revised.pdf

133. Sjoding MW, Dickson RP, Iwashyna TJ, Gay SE, Valley TS. Racial Bias in Pulse Oximetry Measurement. N Engl J Med. 2020 Dec 17;383(25):2477–2478. doi: 10.1056/NEJMc2029240. PMID: 33326721; PMCID: PMC7808260. [PubMed: 33326721]

134. Song TM, Song J, An JY, Hayman LL, Woo JM. Psychological and social factors affecting internet searches on suicide in Korea: A big data analysis of google search trends. Yonsei Med J. 2014;55(1):254–63. [PubMed: 24339315]

135. Stephens-Davidowitz S. (2014). The cost of racial animus on a black candidate: Evidence using Google search data. Journal of Public Economics, 118, 26–40.

136. Tavoschi L, Quattrone F, D'Andrea E, Ducange P, Vabanesi M, Marcelloni F, & Lopalco PL (2020). Twitter as a sentinel tool to monitor public opinion on vaccination: an opinion mining analysis from September 2016 to August 2017 in Italy. Human vaccines & immunotherapeutics, 16(5), 1062–1069. [PubMed: 32118519]

137. Thesmar D, Sraer D, Pinheiro L, Dadson N, Veliche R, & Greenberg P. (2019). Combining the Power of Artificial Intelligence with the Richness of Healthcare Claims Data: Opportunities and Challenges. PharmacoEconomics, 37(6), 745–752. 10.1007/s40273-019-00777-6 [PubMed: 30848452]

138. Thompson C. (2006). Google's China Problem. The Power of Information.

139. Tufekci Z. (2014). Big questions for social media big data: Representativeness, validity and other methodological pitfalls. Paper presented at the Proceedings of the International AAAI Conference on Web and Social Media.

140. Tzioumis K. (2018). Demographic aspects of first names. Scientific data, 5(1), 1–9. [PubMed: 30482902] Vosen S, & Schmidt T. (2011). Forecasting private consumption: survey-based indicators vs. Google trends. Journal of forecasting, 30(6), 565–578.

141. Wall KM, Kilembe W, Inambao M, Chen YN, Mchoongo M, Kimaru L, et al. Implementation of an electronic fingerprint-linked data collection system : a feasibility and acceptability study among Zambian female sex workers. Global Health [Internet]. 2015;11(27):1–11. Available from: 10.1186/s12992-015-0114-z [PubMed: 25889826]

142. Wang B, & Zhuang J. (2017). Crisis information distribution on Twitter: a content analysis of tweets during Hurricane Sandy. Natural hazards, 89(1), 161–181.

143. Wang F, Shu X, Meszoely I, Pal T, Mayer IA, Yu Z, Zheng W, Bailey CE, Shu XO. Overall Mortality After Diagnosis of Breast Cancer in Men vs Women. JAMA Oncol. 2019 Nov 1;5(11):1589–1596. doi: 10.1001/jamaoncol.2019.2803. PMID: 31536134; PMCID: PMC6753503. [PubMed: 31536134]

144. Wang L, Porter B, Maynard C, Evans G, Bryson C, Sun H, Gupta I, Lowy E, McDonell M, Frisbee K, Nielson C, Kirkland F, & Fihn S. (2013). Predicting risk of hospitalization or death among patients receiving primary care in the veterans health administration. Medical Care, 51(4), 368–373. 10.1097/MLR.0b013e31827da95a [PubMed: 23269113]

145. Wesson P, Murgai N. Evaluating the Completeness of HIV Surveillance Using Capture – Recapture Models, Alameda County, California. AIDS Behav. 2017;

146. Williams Brooks, & Shmargad. (2018). How Algorithms Discriminate Based on Data They Lack: Challenges, Solutions, and Policy Implications. Journal of Information Policy, 8, 78–115. 10.5325/jinfopoli.8.2018.0078

147. Wong KO, Zaïane OR, Davis FG, & Yasui Y. (2020). A machine learning approach to predict ethnicity using personal name and census location in Canada. PloS one, 15(11), e0241239.

148. Xiong J, Hswen Y, & Naslund JA (2020). Digital Surveillance for Monitoring Environmental Health Threats: A Case Study Capturing Public Opinion from Twitter about the 2019 Chennai Water Crisis. International journal of environmental research and public health, 17(14), 5077.

149. Zech JR, Badgeley MA, Liu M, Costa AB, Titano JJ, Oermann EK. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study. PLoS Med. 2018 Nov 6;15(11):e1002683. doi: 10.1371/journal.pmed.1002683. PMID: 30399157; PMCID: PMC6219764.

150. Zemel R, Wu Y, Swersky K, Pitassi T, and Dwork C. (2013). Learning fair representations. In International Conference on Machine Learning, pages 325–333.).

BOX 1:

THE 6 Vs

Volume:

The breadth and depth of data.

Velocity:

The speed that data is accumulated - often close to the time of data collection.

Variety:

Types of data that can include unstructured (e.g. video, images, free text) as well as structured formats (rows and columns of data).

Veracity:

Refers to the trustworthiness of data for the questions being posed.

Value:

The decision-making potential derived from big data analyses.

Virtuosity:

The obligation of scientists to: (1) incorporate equity and justice frameworks into big data, for each of the V's and (2) develop solutions for dismantling the -isms and -phobias that perpetuate biases in the data and interpretation.

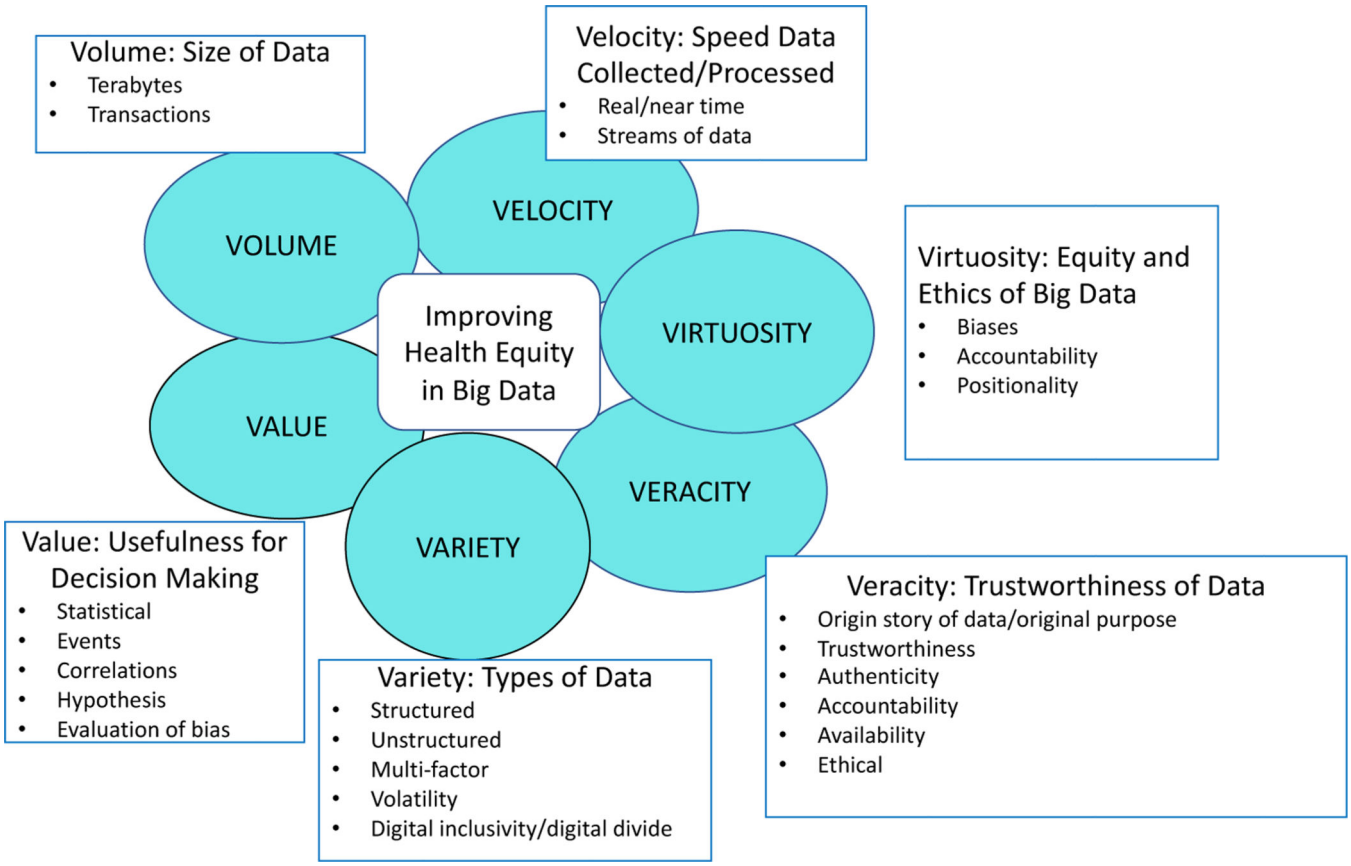

Figure 1:
The 6 Vs of Big Data Adapted for conceptualizing Big Data with Foregrounding of Health Equity Considerations

Table 1:

Examples of Recent Uses of Big Data for Selected Core Public Health Activities and Topics

PUBLIC HEALTH AREA	EXAMPLES			
1. **Descriptive Epidemiology** Focus: Description, exploratory analysis, prediction	Analysis of an algorithm based on multiple data sources can applied to predict families at risk for lead poisoning, compared with a validated sources (119)	Examines associations between Twitter-derived sentiments related to racial or ethnic minorities and birth outcomes, building on construct development for 'sentiments' as an area-level measure of social context (108–109)	Explores relationship between highly public anti-Black violence, responses to public violence based on Internet activity following events, and patterns of mental health distress among Black Americans in the US (26)	Two-sample capture-recapture study building on profiles for MSM (men who have sex with men) through a social networking app to estimate the total number of MSM in mainland China and migration (66)
2. Social Behavior Monitoring (often in conjunction with sentinel events) Focus: Detection of events or patterns, monitoring of behavior signals, often with hypothesis testing	Application of multi-level models to study the relationship between searches for "stress" and "스트레스"; "drinking" and "음주"; "exercise" and "운동"; and "suicide" and "자살"in South Korea at a primary level (month) and Korea's suicide rate, unemployment rate, and after reports of a celebrity suicide or not at a secondary level (year). Results showed that stress-related searches increase suicide search volume on Google (134)	Multiple Systems Estimation (MSE) to estimate the number of extra-judicial killings committed by the Guatemalan government during a massive counter-insurgency campaign from 1978–1996. Stratified analyses revealed that, in some regions, indigenous people were killed at a rate 5–8 times greater than nonindigenous people (86)	Application of Multiple Systems Estimation (MSE) methods, using a combination of log-linear and Bayesian model averaging, to estimate human rights abuses focusing on estimating the total number of civilian killings and disappearances in El Salvador during 1980–1992 (44)	Exploration of the extent to which the phrases, "COVID-19" and "Chinese virus" were associated with anti-Asian sentiments in the context of WHO recommendations to avoid discriminatory language (e.g. location, ethnicity) and tweets by the U.S. President Trump containing racially discriminatory ant-Asian language (61)
3.Disease Surveillance Focus: Detection of disease/symptom status changes (infectious diseases); modelling trends in behaviors and outcomes	Regression-based, multi-sample stratified capture-recapture analysis to estimate the prevalence of opioid use disorder in British Columbia, Canada over time (94)	Exploration of Zika-related Twitter conversations around the time of North and South American outbreak, finding showed geographic rates of Zika-related discussion are moderately correlated with Zika incidence (123)	Infectious-disease modeling for predicting new COVID-19 infections, with inclusion of human-mobility data based on mobile-phone records in the US, presenting limitations of the method when data were absent from key sectors of society, such as those in prison and in nursing homes and helped explain disparities in mobility patters based on economic drivers of mobility (2)	Development of a local geography tool for identification of different types of COVID-19 clusters, combining multiple sources of data (27)
4. Evaluate Impact of Programs/ Interventions/Policies Focus: Impact of programs and policies on health outcomes, behavior, care processes, systems of care/ public health	Assessment of a national health system risk prediction tool/algorithm for population health management which identified inherent biases towards Black patients when cost (instead of illness) was evaluated – (112)	Evaluation of capture-recapture methods to estimate undetected COVID-19 cases, as compared with independent prevalence estimation using epidemiologic techniques, reported similar findings and less uncertainty (10)	Evaluation of black-market pricing of buprenorphine to better understand supply and demand for opioid addiction treatment using and substance use program needs (62).	Oregon Medicaid and commercial colorectal screening clustered by County socioeconomic and health statistics found that low-income areas and those with limited providers had the lowest rates of screening (29)