How Could “Real World Data” Help Us Better Understand Pre-hospital Diagnostic Delay?

Elaine O. Nsoesie[1]

1. School of Public Health, Boston University

Abstract

Digital technologies have enabled massive and continuous production of data that can be used to study and improve health. In this paper, we discuss how digital technologies have been used in public health; outline potential opportunities for using these technologies and data in the study of pre-hospital diagnostic delays; discuss critical legal, ethical, and methodological challenges; and provide guidance on what researchers should consider when using these data. This is not meant to be a comprehensive review of the field. We aim to highlight contributions that might be relevant to the topic of pre-hospital diagnostic delay while also providing some history on how these data sources have been used in public health.

Introduction

Public health researchers and practitioners began developing methods that analyzed online media for the surveillance of infectious diseases in the 1990s.[1,2] One of the earliest systems was the biosurveillance system, ProMED - Program for Monitoring Emerging Diseases – introduced in 1994.[3] PROMED collected and shared data on infectious disease outbreaks that affect plants and animals, expert comments on infectious diseases, and moderated news stories using mailing lists and listserv subscriptions. Other systems such as the Global Public Health Intelligence Network (GPHIN) and HealthMap founded in 1997 and 2006 respectively, focused on automated extraction of information from news aggregators, chatroom discussions and blogs.[4] A detailed overview of Internet-based biosurveillance systems including a list of systems can be found in several review articles including, Hartley et al. (2013),[1] O’Shea (2017)[5], Milinovich et al. (2014)[6], and Choi et al. (2016).[7] The goal of these systems was to augment traditional approaches to disease surveillance by providing free and timely information on disease outbreaks even in remote regions. These systems were used by government agencies, and public health organizations and officials including, the World Health Organization (WHO), and were noted in multiple reports to have captured early reports of disease outbreaks on the Internet prior to disclosure from official sources.[8–11]

Due to its initial focus on infectious diseases and biosurveillance, a significant number of studies and platforms using data from the Internet for public health research and practice have focused on infectious diseases, especially influenza and influenza-like diseases. A systematic review on digital health surveillance published in 2021 identified 755 studies, the highest proportion (i.e. 25%) were focused on infectious diseases.[12] Many of the studies included in the review evaluated similar or the same ideas in different contexts (54 countries) and using different sources of data (26 digital platforms). Studies have suggested that these non-traditional sources of data may be best suited for studying specific diseases, such as those with seasonal trends and short incubation periods, an observation that might also apply to the study of other public health topics.[13,14]

Approaches to using Internet and other non-traditional data sources for surveillance have evolved in the last thirty years; dovetailing with novel advances in technology, data generation, and artificial intelligence. Novel data sources have emerged, and existing sources of data have been used in new ways, including consumer reviews of businesses, products or services, crowdsourcing, social media, remote sensing/place data (i.e., satellite images, street view images, point of interest data, geolocated/GIS data), news, wearables/sensor data, mobile phone data and others (see Table 1 for definitions). These data have been used for a variety of public health applications. As noted, some uses include syndromic surveillance of infectious diseases; surveillance of chronic disease risk factors; pharmacovigilance; understanding patient sentiments and healthcare utilization; studying population mobility and response to public health emergencies and interventions; monitoring social determinants of health; and monitoring health misinformation.

Table 1: Definitions

Term	Definition used in this article
Review Websites	Websites that display user opinions on services and institutions.
Crowdsourcing	The collection of data or completion of activities by enlisting a community or large group of people. Crowdsourcing is sometimes referred to broadly as citizen science, or specifically as, participatory disease surveillance in the public health context.
Social Media	Internet networking and information sharing platforms.
Search Data/Queries	Information submitted to Internet search engines.
Remote Sensing	Information about the earth gathered through satellites.
News	Online news media.
Wearables / Sensors	Devices that collect physiological, movement and other types of data for health purposes. Sensors can be embedded in smartphones, fitness trackers, wristbands etc.
Mobile phones	Data from mobile phones including, call data records (CDR), and Global Positioning System (GPS) location data. Other data, including, Bluetooth used in contact tracing will be mentioned but not discussed in-depth.

In this paper, we provide an overview of digital technologies and data, their applications and highlight studies published between 2011 and 2022 that capture important ideas that could be relevant for studying diagnostic delays. Specifically, we present the following: (1) a review of digital technologies and data and their use in public health; (2) potential opportunities for using digital technologies and data sources to understand pre-hospital diagnostic delay; (3) challenges with using digital technologies and data for public health research and application; and (4) recommendations on what researchers should know about digital technologies and data sources before or while designing a research study.

USES OF DIGITAL TECHNOLOGIES AND DATA SOURCES IN PUBLIC HEALTH

We provide an overview of the applications of various digital technologies and data sources.

Search Data

The first studies exploring the use of search queries for disease surveillance used data from Yahoo and Google and were published in 2008 and 2009, respectively.[15,16] Using a list of influenza-related terms and additional queries that correlated with prevalence of influenza and influenza-like illnesses, researchers developed models for forecasting temporal trends of the same. The model estimates were compared to and shown to be significantly correlated with official reports from the United States Centers for Disease Control and Prevention (CDC). The system developed by Google was named Google Flu Trends and was deployed for public use. Many issues were later raised about the Google Flu Trends system, including its deviation from CDC reports of influenza-like illness and the lack of precision in the selection of terms that were used in creating the model.[17,18] Researchers proposed solutions that focused on both addressing the challenges associated with the data and the methods.[19] Similar methods have been used in forecasting other infectious diseases including, MERS; cholera; dengue; malaria; hand, foot, and mouth disease; Zika; and chicken pox.[20–22] In general, early warning systems using search data have been shown to be most suitable for vector-borne and vaccine preventable diseases and have been useful for predicting disease incidence or prevalence several weeks in advance.[13,23]

In most cases, search data is easier to access and process than the other data sources mentioned in this paper so there is a plethora of studies using these data. In addition to infectious disease surveillance and forecasting, web searches from a variety of platforms have been used in a range of public health applications. These applications have taken advantage of the global use of search engines for seeking health information and the assumption by users that these platforms lend some degree of privacy when compared to social media.[24] Studies have even used these data to study information seeking for sensitive or stigmatized health conditions and topics including, HIV/AIDS, and mental health.[25,26]

To illustrate the broad applicability of search data for public health, we provide examples. In one study, researchers focused on search behaviors of guardians to children with biliary atresia or hypertrophic pyloric stenosis to show that data on guardian behaviors can be used for detecting childhood diseases.[27] Some studies have characterized general online health seeking behavior,[28] with some focusing on describing health information needs and misconceptions in a particular region.[29] Studies on eye diseases have focused on mapping search patterns across regions or looking at the association between searches for terms such as, cataract, glaucoma, and diabetic retinopathy, and the prevalence of each condition.[30,31] Temporal, spatial and diurnal variations in chest pain have been shown and also correlated with data from the CDC.[32] Others have looked at searches for risk factors associated with obesity and other chronic diseases.[33] Cyclical trends have been noted for searches for suicide and depression related terms.[34] Search data has also been used for forecasting demand for medical devices.[35] Use of search data during the Covid-19 pandemic have included monitoring of interest in self-medications,[36,37] misinformation,[38] insomnia,[39] assessing associations between symptom searches and Covid-19 cases,[40] and forecasting trends in COVID-19 dynamics.[41]

Though not classified as a search engine, access logs of the online encyclopedia, Wikipedia, is another data source that has been used in disease surveillance and forecasting. Studies have accessed useability for Cholera, Dengue, Ebola, HIV/AIDS, Influenza, Plague and Tuberculosis.[14,42]

Analytic Approach

Studies using search queries generally start with a generation of terms or keywords. Depending on the focus of the study, terms could include disease names (e.g., cholera, dengue), etiological agents (e.g., brucella), symptoms (e.g., cough, vomiting), medications (e.g., chloroquine) or general treatment options (e.g., malaria treatment), colloquialisms (e.g., TB), effects of disease on different populations (e.g., Covid and children) etc. This step can be highly subjective and requires context, culture, language, and subject-matter expertise.[43,44] Most studies generate keywords based on expert opinion. However, there are other approaches (i.e., data mining, and machine learning) that have been explored, and this will be further discussed in the bias section since it applies to other data sources. The wrong set of keywords could lead to spurious or no correlations.

Search data are available at different time scales (e.g., days, weeks, months) depending on the volume and frequency with which the data is being generated. The data structure varies across search engine platforms, from absolute counts to normalized values. For example, Google Trends (Table 2) does not report the total search volume, instead it provides values normalized on a scale from one to one hundred, relative to the popularity of a specific search term across time and geography.

Next the search terms are filtered to identify relevant features for machine learning or variables for a regression model. Methods as simple as Pearson correlation and as complex as Artificial Intelligence approaches have been used to establish associations between search data and forecast future trends, respectively. In infectious disease forecasting, these methods have been used to predict different aspects of disease dynamics including, when an epidemic will peak, how many people will be infected at the peak, when it will end and the incidence/prevalence of the disease in a population. These data and datasets from other digital platforms were used to address the delay (usually in weeks) between when disease epidemic data was collected and when it was shared by official public health surveillance systems.

To validate trends observed in the search data, researchers have used data from official sources (such as, Ministries of Health, Centers for Disease Control, or the World Health Organization), hospitalization records and surveys.

Individual-level search data is typically not publicly available but could be obtained through a collaboration with the company that owns the data. Since this data is at the individual level, predictions and inferred diagnosis can also happen at that level. However, there are many ethical issues with individual level diagnosis given its potential emotional and psychological impact. In some cases, a study begun with individual level data and was then aggregated to the population level prior to analysis (e.g., the study by Sadilek et al. (2020) on forecasting Lyme disease[45]). Additional information on individual-level analysis is provided in the section on
Opportunities for Using Digital Technologies and Data Sources to Study Pre-Hospital Diagnostic Delay.

Social Media

Social media platforms have allowed the sharing of opinions, preferences, and behaviors in real-time and across geographic regions. Opinions expressed by users in one part of the world can easily impact decisions made by users in another part of the world as has been shown for web-based content on anti-vaccine sentiments.[46] Some of the earliest studies on the use of social media platforms for public health surveillance focused on monitoring influenza,[47] cholera,[48] and vaccine sentiments.[49] Researchers were interested in how behavioral information shared on platforms such as, Twitter, could be used in tracking the spread of infectious diseases and sentiments towards public health interventions.

Similar to search data, many studies have explored the use of social media data from diverse platforms for public health research and practice. Twitter is overrepresented in the field because it made its data available to academics for research almost since its launch in 2006. A few examples on the use of social media data for public health follow. Social media data has been used to characterize sleep issues based on usage patterns and the inclusion of keywords such as, insomnia and Ambien or sleep aid mentions in tweets.[50] Patient reviews or discussions of experiences in healthcare settings on social media have been used as a measure of quality of care.[51] Tweets have been analyzed to understand disease and vaccine sentiments during outbreaks.[49,52–56] Geolocated data aggregated to the neighborhood level has been analyzed to characterize happiness, diet, and engagement in physical activity, and to study spatial and gender disparities in these health outcomes.[57–60] Systems have been developed to support local departments' of health foodborne illness surveillance efforts by mining reports on social media.[61–64] In studies linking social media data to medical records, researchers have demonstrated that users' language can be used for disease screening, identifying indicators of disease risk, and extracting information on disease epidemiology.[65–67] A number of studies have also focused on pharmacovigilance – the study of adverse medicine/vaccine effects.[68–71] Social media data has also been used to study sensitive and stigmatized topics.[72,73]

Analytic Approach

Similar to search data, the first step in the analytic process for social media data usually involves developing a set of keywords for data extraction. Depending on the social media platform, extracted data can include the following attributes: username, the user's unique platform identifier, the message, timestamp of posted message, and geographical location such as, latitude and longitude (if the user has opted to share), number of followers, number following, and biographical information associated with the account. Platforms, such as, Twitter, only share information from public accounts.[74] Researchers collect data from an Application Programming Interface (API) or through third-party apps created to process and provide data to businesses.

Unlike search data that is usually processed (i.e., aggregated) prior to being made publicly available for research and can be easily managed in a single file, data from social media platforms can be enormous requiring pre-processing and processing to extract meaningful information. The ratio of noise to signal can be significant depending on the specificity of the keywords used in extracting the data and the topic of study. Data analysis steps can include, manual and/or machine annotation of relevant and irrelevant content to produce a corpus for machine learning classification; data mining to create data tables (e.g., weekly counts of individuals reporting engagement in physical activity) or natural language processing to extract data themes; and interpretation of findings. If the study involves spatial analysis, geospatial methods will be needed to map data coordinates to geographic regions prior to statistical analysis.

To further illustrate, analytical steps are presented for two studies. The study by Nikfarjam et al. (2015),[68] on mining adverse drug reactions with word embedding features involved the following steps: data collection from an API, annotation, developing/expanding on an adverse drug reaction lexicon, a conditional random fields classifier was used to label user sentences, learn word embeddings, and embed cluster features. In another study by Hernandez et al. (2022)[75] which focused on understanding changes in diet during the Covid-19 pandemic, the analytic steps involved obtaining geolocated data from Twitter API, use of data mining to filter the dataset to focus on tweets containing mentions of food items, a sample of the data was annotated into negative and positive classes, machine learning methods were developed to classify the remainder of the tweets, the data was further classified into three classes (fast food, healthy food and alcohol) using data mining techniques, geospatial methods were used to map the tweets to US regions and regression methods were used to study changes in diet.

Data used in external validation come from the same sources as those used for search data studies, namely, official sources, hospital records and surveys. Studies have tried to match patient medical records with their social media postings to demonstrate the validity of using social media platforms for public health surveillance.[76] Data is usually provided at an individual-level, however, due to privacy concerns and potential for harm, individual-level analysis is discouraged. Findings from most studies are presented at the population level.

Although examples are limited, surveillance tools developed using data from social media platforms have been integrated into public health practice; one example is foodborne illness surveillance.[62,63]

Reviews

Review websites such as, Yelp and RealSelf, were borne out of a lack of information on the quality of services provided by medical providers and institutions on the Internet.[77–79] Reviews on the Internet have been described as an online version of "word-of-mouth testimonials".[80] Reviews can provide businesses with useful input to improve and change customer service practices. Because reviews are unstructured, they are not limited to biased predefined categories usually included in surveys (e.g., post-hospital-visit surveys).[80,81] Most studies using review data focus on a specific business (e.g., hospitals) or specific product (e.g., food products). Data from review websites have been used in several public health applications, though not as widely as search or social media data likely due to data access challenges and the structure of the data, which limits the types of questions that can be investigated.

Broadly, studies using Internet reviews can be classified into patient or user experiences at healthcare institutions, experiences of products that affect health, and reviews of businesses that have a health impact (e.g., vaping shops). We present a few examples. Seltzer et al. (2022) used data from Yelp to characterize patients' experiences of obstetric care in hospitals.[82] Agarwal et al. (2022) used machine learning to extract themes associated with negative and positive reviews of substance use disorder treatment facilities in the USA. [83] Donnally et al. (2018)[84] and Furnas et al. (2020)[79] compared reviews of spine surgeons and plastic surgeons, respectively across multiple platforms. Tong et al. (2022) manually coded reviews on Yelp to document examples of institutional racism in healthcare facilities and describe recurrent themes.[85] Other businesses that have been studied include, vaping shops, hookah bars and tobacco vendors,[86–88] and mental health treatment facilities.[89]

Reviews are one data type that have been integrated into public health practice. For example, Yelp and the Los Angeles County Department of Public Health collaborated to display restaurant hygiene ratings on restaurant profiles, giving users access to this information when making a choice about where to dine out.[90] The New York City Department of Health and Mental Hygiene used reviews to identify foodborne illness and outbreak reports that were not reported through their standard systems.[91,92] Other examples of food safety studies include, the application of an artificial intelligence algorithm to Amazon reviews to identify unsafe food products,[93] assessing restaurant quality and sanitation,[94] and comparing foods mentioned in foodborne illness reviews to those implicated in outbreaks reported to the CDC.[95]

Analytic Approach

The process used to analyze online reviews for public health applications and research usually includes data mining, manual or automated labeling, natural language processing or quantitative modeling. Data access and gathering might require a collaboration with the company since the data is not usually available through an API. Companies such as, Yelp, make a limited subset of their data available for academic research (Table 2).

Qualitative analysis of these data is typically focused on extracting themes or clusters of information. Manual coding is sometimes required prior to training and applying machine learning methods. However, some studies completely rely on manual coding if machine learning or artificial intelligence methods cannot capture the nuances present in the data. Quantitative analysis might involve using correlations or regression to measure the association between ratings, volume of reviews, and other quantitative measures, and health outcomes (such as, mortality) or rankings from official websites such as, the Hospital Consumer Assessment of Healthcare Providers and Systems (HCAHPS).[51,96]

There are concerns about the quality and validity of reviews. Studies suggest negative review fraud is likely committed by competitors.[97] To integrate reviews into traditional surveillance systems, public health authorities contact customers who have submitted a review for additional information. This is one approach for verifying the validity of reviews.[64] However, this process can be cumbersome and costly when handling large volumes of reviews.

Other Data Sources and Applications

Sensors, Wearables and Apps

A quick search of "health apps" on the Apple store retrieves thousands of apps; from those measuring heart rates, to meditation apps, all promising improvements in health. However, few of these apps have been thoroughly tested or are based on methods that have been scientifically evaluated. Apps can be used for patient engagement, facilitating treatment and sustaining treatment gains.[98,99] Applications for patient engagement might be the most relevant for the topic of pre-diagnostic delay. We also discuss possible applications in the next section.

Wearable sensors can be used to detect and collect physiological and movement data in a variety of applications. Wearable technology is widely used and the smartphone is the most widely used passive sensor.[100,101] According to reports, there were 600 million wearable devices in use in 2020. Based on current trends, this number is expected to rise to 1100 million in 2022.[102] The majority of smartphones can also detect voice, touchscreen pressure, ambient light, ambient pressure, and geographic position. Wearable gadgets, like gloves, insoles, hats, and smartwatches, are often attached to the wearer's body directly, inside of clothing, or in semi-rigid structures.[101,103] Well known wearable devices such as the Apple Watch can track fitness activities, sleep patterns, blood oxygen level, ECG and step count. Another example is Fitbit Sense, which can track fitness activity, skin temperature, ECG, electrodermal activity and stress management.[101] The amount of data collected by wearable sensors over long periods of time can be massive.

Studies have discussed both the useability and accuracy of wearable sensors.[104–107] A notable example is the pulse oximeter, which has been shown to perform poorly on darker skin, which can lead to delayed treatment and possibly death.[108,109] The accuracy and reliability of wearable sensor especially in clinical settings is extremely important given its possible impact on health outcomes.

Sound

Clinicians frequently use audio signals produced by the human body, such as digestive sounds, breathing, vibrations, and heartbeats, to diagnose or track the progression of diseases.[110] During the Covid-19 pandemic, researchers used crowdsourcing approaches to develop datasets of audio recordings including cough, speech, voice, breathing, and other sounds produced by Covid and non-Covid patients.[110–121] The data was collected from social media sites (such as, YouTube, Instagram, Twitter), website recordings and using existing recording apps on mobile phones (such as, Wechat App). Examples of apps and crowdsourced datasets included the COVID-19 Sounds App,[122] Corona Voice Detector,[123] Coswara [124]and COUGHVID.[114] The datasets included a diversity of participants across geographic locations, gender, and age. Speech processing, artificial intelligence, and other approaches were applied to study how these data could be used in Covid-19 surveillance and diagnosis. Some studies recommended using artificial intelligence tools built on these data for Covid screening at school and work, and for more efficient testing.

Mobile Phone Data

Human mobility and activity patterns can be extracted from mobile phone call records, social media check-in (defined as the act of reporting or registering one's presence at a particular location using a mobile device) and other types of location data.[125–130] Movement information extracted from mobile phone call records have been aggregated and analyzed for modeling of disease epidemics.[131–135] Call Records Data usually includes, a time stamp, the GPS coordinates of a nearby cell tower, and a unique identity for each subscriber.[136] GPS location data, which was widely used during the pandemic typically includes, a time stamp, the GPS coordinates of the phone, and a unique identity for each user.[136] Prior to the Covid-19 pandemic, these data was mostly available to select research groups who had relationships with telecom companies. The public release of data by companies such as, SafeGraph, and Cuebiq during the pandemic created new opportunities for studying movement and its association to disease spread and the impact of public health interventions during the pandemic. For example, researchers used mobile phone location data to study inequities in Covid burden and how different socioeconomic groups responded to social distancing policies.[137,138] Mobile phone Bluetooth data was also used in contact tracing applications.[136]

Remote Sensing Data

Remote sensing and place data from satellites, street view images, and geographic information systems have been used recently to study the association between the built environment and health outcomes.[139–141] The availability of efficient artificial intelligence methods that can be applied to classify and extract information from images has created new opportunities for studying the presence of resources in neighborhoods. Crosswalks, buildings, greenery, and streets, have been identified using images, and the data used to examine their relationship with chronic health outcomes and socioeconomic factors.[140,141,145–147] High levels of agreement between neighborhood characteristics extracted from images and field assessments have been used to demonstrate the dependability of employing images to extract data on physical features.[142–144]

Opportunities for Using Digital Technologies and Data Sources to Study Pre-Hospital Diagnostic Delay

Pre-hospital diagnostic delay is defined in this paper as the time period prior to a patient reaching the care environment in which a diagnosis takes place. Here, we make five recommendations with accompanying examples that may provide insights into how digital tools and data sources can be used in studying and addressing pre-hospital diagnostic delay.

First, both individual-level and aggregated data from search engines can be used for screening patients, "diagnosing" or predicting health outcomes. Similar approaches have been developed using data from social media, mobile apps, and wearables.[100,148]

Studies have used individual-level search data to screen or "diagnose" Type I diabetes, neurodegenerative disorders, cancers, and mental health disorders. Hochberg et al. (2019) developed models for diagnosing Type I diabetes early based on the search behavior of 11,050 search engine users with diabetes and a control group of 11.5 million users.[149] The patient group included users whose searches stated a disease diagnosis. White et al. (2018) developed classifiers for detecting neurodegenerative disorders, specifically, Parkinson's disease (PD) and Alzheimer's disease (AD) from individual search logs.[150] Similar to the previous study, cohorts of "patients" and "controls" were developed. The analysis included 31,321,773 search engine users. Evidence used in the developed models are not available to clinicians, including, longitudinal query repetition, and mouse cursor activity. White et al. (2017) focused on how to use search logs for screening patients for lung carcinoma.[151] Gold standard data was unavailable. The presence of disease was determined based on a "landmark query" after several months of symptom searches. Searches after the "landmark query" focused on medications and treatments. Positive cases included 5443 users and negative cases included 4,808,542. Yom-Tov et al.[152] identified individuals with mood disorders by using search queries about medications used to treat the condition as well as changes in their behavior close to the occurrence of mood disorder events. Eichstaedt at al. (2018), showed that they could use patients' Facebook statuses to diagnose depression.[66] Ofran et al.[153] identified people with a likely cancer diagnoses and then tracked their information seeking over time by setting a threshold on the quantity of cancer-specific queries. Lastly, Paparrizos et al.[154] predicted the diagnosis of people who self-identified as having pancreatic cancer prior to diagnosis. In all these studies, the authors state that individual search logs could be useful in the early screening and diagnosis of these diseases.

Second, apps can be used to increase access to information that can encourage healthcare seeking behavior and increase diagnosis. Specifically, apps developed for patient engagement could (1) focus on education about disease conditions, risks, and treatment; (2) be used to overcome structural barriers regarding distance to medical facilities, cost of consultation, stigma or fear associated with some health conditions by providing access to information; (3) provide information on how to seek financial assistance to cover treatment cost; and (4) find locations of nearest health centers for diagnosis and treatment. For example, the HealthMap vaccine finder was developed to help patients locate the nearest facility offering a specific type of vaccine.[155] Several apps with a user-centered design have been developed to educate patients about different medical conditions and treatment options. See examples in Alberts et al. (2020), Birkhoff et al. (2017) and Roger et al. (2018).[156–158]

Third, Internet reviews and social media can be used to understand patients' perceptions of care at healthcare institutions including experiences of discrimination that could be deterrents to seeking care. There are many studies showing that patient reviews of healthcare institutions on Internet platforms correlate with reviews from official sources. Examples of such studies include research by Hawkins et al. (2016) using data from Twitter,[51] Ranard et al. (2016) using data from Yelp,[81] and Greaves et al. (2012) using data from a government owned hospital review website.[159] Additionally, at least one study has shown that specific types of discrimination could be identified from healthcare reviews.[85] These information could be useful in addressing the broader systemic issues and policies that impact certain populations from seeking diagnostic care.

Fourth, mobile phone and remote sensing data can be used to study access to healthcare locations for populations most impacted by diagnostic delays for the specific diseases of interest. Mobile phone data has been widely used to study movement and its relation to disease spread and public health interventions especially during the Covid pandemic.[128,136,138,160] Similar approaches can be used to map locations of healthcare facilities and to study variability in access to healthcare resources.[161] Nguyen et al. (2018) and Maharana et al. (2018) showed that AI can be applied to neighborhood satellite and street images to study access to neighborhood-level built environment indicators that have been associated with health and well-being. [139,140,162,163]

Fifth, all of the aforementioned data types can be used to study other social determinants of health. These datasets can be combined with neighborhood economic, demographic and health data to answer the following questions. Who is affected by diagnostic delays? What role does geography, race/ethnicity, socioeconomic class, immigrant status, gender, sexual orientation, or culture have in observed disparities? Are there specific policies that hinder access to diagnostic resources? How does disparities in Internet access impede diagnostic care? What role can community health workers play in reducing diagnostic delays?

Challenges with Using Digital Technologies and Data for Public Health

There are many ethical and legal challenges associated with the use of digital data and tools for public health. These include ensuring public benefit; ensuring scientific validity and accuracy; protecting privacy; preserving autonomy; avoiding discrimination; validation; misinformation; the cost of false detections or predictions; mental and emotional impact of unsolicited diagnosis; and preventing digital inequality.[164–167] Methodological challenges include, linking these digital data sources to traditional data, such as, electronic health records, and bias in the analytical process.

Ethical and Legal Challenges

Studies using digital data sources can be harmful to certain groups or individuals. It is therefore important to weigh the risk versus benefits before pursuing a research question. If a research study is likely to exacerbate existing health inequity, marginalization, or general disadvantage for a specific population, then it shouldn't be pursued.

Furthermore, ensuing validity or accuracy can be challenging in many of these data sources. For example, online reviews can be fake or edited to make a business look better than it actually is. Also, not everyone searching for a particular illness on a search engine is sick, there are many other reasons (including, education and curiosity) why an individual might be interested in researching information on an illness.

However, compared to surveys, reports on social media, which are generally voluntary, are less likely to be affected by recall bias or social desirability bias.[168–170]

Also, individuals who use online platforms might not always be aware that their data can be used for research and there are no Institutional Review Board (IRB) standards for informed consent, unless the research involves contacting users to collect additional information beyond what is considered publicly available. There have been several publications recommending ethical guidelines that should be followed by researchers, including, not linking individual information from multiple sources since this can lead to user identification and the disclosure of sensitive personal information, not publishing social media postings since this can be easily linked to an individual with a simple search on Google or other search engines.[164,165,167]

Furthermore, there are ethical concerns regarding the impact of research. A lack of representation or "missingness" of certain populations from research data implies that they are less likely to benefit from the findings, technology and policies resulting from a research study.[171,172] Also, analysis and predictions at the individual level can cause unnecessary emotional and psychological harm especially if unsolicited. Lastly, there is unequal access to the Internet and digital health technologies, and disparities in digital literacy, implying that solely relying on digital tools and data might not be beneficial to poor, marginalized and vulnerable populations.[173] Addressing this challenge might require combining digital data and tools with data from traditional systems to achieve equity across populations.

Methodological Concerns

There are various methodological challenges that have been associated with the use of digital data and technology for public health. These challenges, which include data linking, representation bias, algorithmic bias, keyword bias, and platform bias, might apply to a single data source or multiple data sources.

Data Linking. Several studies have shown that it is possible to link social media to medical records provided that patients consent to have their records linked by providing access to their social media accounts.[76,174,175] However, this process is not easy to implement when dealing with millions of records from a social media site or when user information are unavailable. There are also privacy concerns especially when working with sensitive health topics, for example mental health.

Representation Bias. Lack of demographic data in some social media sources makes it difficult to assess and quantify demographic representation in these data. Studies have emphasized the need for demographic data to better understand who is included and who isn't, and to make more targeted public health assessments and recommendations.[57,58,176,177] In an effort to obtain accurate evaluations of users' overall wellbeing using Twitter data, Jaidka et al. (2020) and Iacus et al. (2020) tried to reduce demographic bias by stratifying Twitter users based on their geographic distributions.[178,179] Cesare et al. (2019) uncovered disparities in reported physical activity prevalence after inferring and stratifying the data by gender.[57] Furthermore, Weeg et al. (2015) reported that after stratifying Twitter users by demographics, the association between findings from social media data and those from a nationwide survey was greatly increased.[180]

Algorithmic Bias. Bias in machine learning and artificial intelligence algorithms has been widely discussed in recent years.[171,181,182] Analysis of data from the aforementioned digital data sources, usually require the application of machine learning algorithms. It is therefore important for researchers to be aware of current concerns about machine learning algorithms and ensure that application of these algorithms do not perpetuate existing biases. There are many papers discussing fairness and ethical frameworks for machine learning algorithms that should be referenced prior to research.[183–187]

Keyword Bias. Most studies select keywords for extracting data from the aforementioned sources manually; either based on expert opinion, previous studies, or social context. A few studies have highlighted the potential impact of culture and context in the selection of keywords and how research findings can be significantly altered if the wrong keywords are selected.[43,44] Furthermore, selected keywords may be inadequate due to the exclusions of misspellings or slangs. To address bias in the selection of keywords for extracting data, a few methods have been proposed, including, machine learning filtering and rule-based filtering approaches.[178,188–194] For example, some researchers have manually evaluated a sample of tweets after filtering using keywords to ensure that tweets capture the intent of the study (for example, in studying influenza, are the tweets indicating that the user is experiencing symptoms or does in mention symptoms in reference to something else). Researchers also manually label thousands of tweets to develop "positive" and "negative" samples, which are then used to train a machine learning or artificial intelligence classifier to identify relevant content.

Platform Bias. Different platforms might capture different population samples implying that the data might not be representative. Also, the nature of the platforms (i.e., image, video or text or a mixture) might be better suited or adaptive to different types of information. Studies have compared different platforms to show differences in content[195] and also combined data from various platforms to address bias that might be represented in individual platforms.[196]

Conclusion

There are many factors to consider when making decisions about using digital data sources for research on diagnostic delays. First, it is important to clearly define the research question. The research question will determine what data platform would be most useful. For example, to understand user's comments about access to diagnostic tools, social media (such as, Twitter or Facebook) or online forums where a particular diagnostic tool is being discussed might be the best options.

Second, consider whether you need a digital data or nontraditional data source to answer the research question or if there are other existing datasets that might answer the same question. If a dataset from a digital source is appropriate, do you need another dataset to supplement or fill in gaps in the selected dataset? Also, can you get access to the dataset, or would you need to establish a collaboration with the company to obtain data?

Third, assess whether there are limitations and challenges with using the selected dataset that need a plan prior to download and analysis. Also, consider the computational needs of the data. Where will it be stored? Do you need special computational tools to analyze the data? Also, consider the ethical challenges with using the data. Research using public datasets from the Internet does not usually require an IRB review. However, if there are plans to combine these data with another dataset (e.g., electronic health records or individual-level survey data), then an IRB might be needed. Demographic representation is another important issue to consider. If the data does not represent Census demographics for the region of study, that should be addressed with statistical methods or acknowledged in the study. The population impacted by the study's findings must be clearly articulated.

Fourth, appropriate methods should be selected. Sometimes multiple methods are needed throughout the analysis pipeline. It is important to consider what methods have been used in similar studies and invite individuals with the necessary expertise to join the team. For example, in a paper analyzing text data from Twitter, data mining, natural language processing, and statistical modeling might be needed. Also, if the study is focused on miscarriages, then clinicians with the appropriate expertise should also be included.

Lastly, it is important to think carefully about appropriate methods for quantifying and communicating measures of uncertainty and generalizability. How uncertainty is communicated might depend on the data, the methods, and the audience. As previously noted, these digital technologies and data sources might not be the solution to addressing certain questions about diagnostic delays. There are disparities in smartphone access, and Internet and broadband infrastructure that lead to underrepresentation of low-income communities in some of these datasets. In cases where the use of data from digital sources might create or exacerbate existing disparities, alternative solutions including, using more representative datasets or integrating with other datasets must be considered to ensure equitable solutions are proposed.

Acknowledgement: I thank Nina Cesare, PhD for reading and providing comments on an early draft of this paper

Table 2: Data Access for Frequently Used Platforms

Type	Source	URL	Access Type	Academic Program
Social Media	Twitter	https://developer.twitter.com/en/products/twitter-api/academic-research	Free	Yes
	Instagram	https://help.crowdtangle.com/en/articles/4302208-crowdtangle-for-academics-and-researchers	Free	Yes
	Facebook	https://research.facebook.com/data/ https://help.crowdtangle.com/en/articles/4302208-crowdtangle-for-academics-and-researchers	Free	Yes
	YouTube	https://research.youtube/how-it-works/	Free	Yes
	Foursquare	https://foursquare.com/products/for-academics/	Free	Yes (select datasets with limited subscription)
	Reddit	Several scraping tools available online	Free/paid	No
Business/Product/ Service Review	Yelp	https://www.yelp.com/dataset	Free	Yes (single dataset)
	Vitals	Vitals.com	Unavailable	No
	Amazon	Some data available from secondary sources	Unavailable	No
	Healthgrades	Healthgrades.com	Unavailable	No
	Google	Google.com	Unavailable	No
Place/Location/ Remote Sensing	Google Satellite/Maps	https://www.google.com/intl/en-GB_ALL/permissions/geoguidelines/	Free (with restrictions)	Yes
Search	Bing	https://www.bing.com/	Unavailable (dataset released during pandemic)	No
	Wikipedia	https://dumps.wikimedia.org/other/pageviews/readme.html	Free	No
	Google	https://trends.google.com	Free	No

*Some platforms are open to collaborating with academics to provide data that is not publicly available.

References

1. Hartley DM, Nelson NP, Arthur RR, et al. An overview of internet biosurveillance. *Clin Microbiol Infect*. 2013;19(11):1006-1013. doi:10.1111/1469-0691.12273

2. Salathe M, Bengtsson L, Bodnar TJ, et al. Digital epidemiology. *PLoS computational biology*. 2012;8:e1002616. doi:10.1371/journal.pcbi.1002616

3. Madoff LC, Woodall JP. The internet and the global monitoring of emerging diseases: lessons from the first 10 years of ProMED-mail. *Arch Med Res*. 2005;36(6):724-730. doi:10.1016/j.arcmed.2005.06.005

4. Brownstein JS, Freifeld CC, Madoff LC. Digital disease detection—harnessing the Web for public health surveillance. *New England Journal of Medicine*. 2009;360(21):2153-2157.

5. O'Shea J. Digital disease detection: A systematic review of event-based internet biosurveillance systems. *International journal of medical informatics*. 2017;101:15-22.

6. Milinovich GJ, Williams GM, Clements AC, Hu W. Internet-based surveillance systems for monitoring emerging infectious diseases. *The Lancet infectious diseases*. 2014;14(2):160-168.

7. Choi J, Cho Y, Shim E, Woo H. Web-based infectious disease surveillance systems and public health perspectives: a systematic review. *BMC public health*. 2016;16(1):1-10.

8. Anema A, Kluberg S, Wilson K, et al. Digital surveillance for enhanced detection and response to outbreaks. *The Lancet Infectious Diseases*. 2014;14(11):1035-1037.

9. Salathé M, Freifeld CC, Mekaru SR, Tomasulo AF, Brownstein JS. Influenza A (H7N9) and the Importance of Digital Epidemiology. *N Engl J Med*. 2013;369(5):401-404. doi:10.1056/NEJMp1307752

10. Brownstein JS, Freifeld CC, Madoff LC. Digital disease detection--harnessing the Web for public health surveillance. *The New England journal of medicine*. 2009;360:2153-2155, 2157. doi:10.1056/NEJMp0900702

11. Brownstein JS, Freifeld CC, Madoff LC. Influenza A (H1N1) virus, 2009—online monitoring. *New England Journal of Medicine*. 2009;360(21):2156-2156.

12. Shakeri Hossein Abad Z, Kline A, Sultana M, et al. Digital public health surveillance: a systematic scoping review. *NPJ digital medicine*. 2021;4(1):1-13.

13. Milinovich GJ, Avril SM, Clements AC, Brownstein JS, Tong S, Hu W. Using internet search queries for infectious disease surveillance: screening diseases for suitability. *BMC infectious diseases*. 2014;14(1):1-9.

14. Generous N, Fairchild G, Deshpande A, Del Valle SY, Priedhorsky R. Global disease monitoring and forecasting with Wikipedia. *PLoS computational biology*. 2014;10(11):e1003892.

15. Polgreen PM, Chen Y, Pennock DM, Nelson FD, Weinstein RA. Using Internet Searches for Influenza Surveillance. *Clinical Infectious Diseases*. 2008;47(11):1443-1448. doi:10.1086/593098

16. Ginsberg J, Mohebbi MH, Patel RS, Brammer L, Smolinski MS, Brilliant L. Detecting influenza epidemics using search engine query data. *Nature*. 2009;457:1012-1014. doi:10.1038/nature07634

17. Butler D. When Google got flu wrong. *Nature*. 2013;494:155-156. doi:10.1038/494155a

18. Lazer D, Kennedy R, King G, Vespignani A. The Parable of Google Flu: Traps in Big Data Analysis. *Science*. 2014;343(6176):1203-1205. doi:10.1126/science.1248506

19. Santillana M, Zhang DW, Althouse BM, Ayers JW. What can digital disease detection learn from (an external revision to) Google Flu Trends? *American journal of preventive medicine*. 2014;47(3):341-347.

20. Ocampo AJ, Chunara R, Brownstein JS. Using search queries for malaria surveillance, Thailand. *Malaria journal*. 2013;12(1):390.

21. Seo DW, Shin SY. Methods using social media and search queries to predict infectious disease outbreaks. *Healthcare informatics research*. 2017;23(4):343-348.

22. Bakker KM, Martinez-Bakker ME, Helm B, Stevenson TJ. Digital epidemiology reveals global childhood disease seasonality and the effects of immunization. *Proceedings of the National Academy of Sciences*. 2016;113(24):6689-6694. doi:10.1073/pnas.1523941113

23. Nsoesie E, Mararthe M, Brownstein J. Forecasting peaks of seasonal influenza epidemics. *PLoS currents*. 2013;5. doi:10.1371/currents.outbreaks.bb1e879a23137022ea79a8c508b030bc

24. De Choudhury M, Morris MR, White RW. Seeking and sharing health information online: comparing search engines and social media. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM; 2014:1365-1376.

25. Vaidyanathan U, Sun Y, Shekel T, et al. An evaluation of Internet searches as a marker of trends in population mental health in the US. *Scientific Reports*. 2022;12(1):1-9.

26. Allem JP, Leas EC, Caputi TL, et al. The Charlie Sheen effect on rapid in-home human immunodeficiency virus test sales. *Prevention Science*. 2017;18(5):541-544.

27. Yamaguchi S, Hinoki A, Tsubouchi K, Amano H, Tajima A, Uchida H. Usefulness of web search queries for early detection of diseases in infants. *Nagoya Journal of Medical Science*. 2021;83(1):107.

28. Li X, Tang K. Effects of Online Health Information-Seeking Behavior on Sexually Transmitted Disease in China: An Infodemiology Study Based on the Baidu Index. *Available at SSRN 4190263*.

29. Abebe R, Hill S, Vaughan JW, Small PM, Schwartz HA. Using search queries to understand health information needs in africa. In: *Proceedings of the International AAAI Conference on Web and Social Media*. Vol 13. ; 2019:3-14.

30. Azzam DB, Nag N, Tran J, et al. A novel epidemiological approach to geographically mapping population dry eye disease in the United States through Google Trends. *Cornea*. 2021;40(3):282-291.

31. Hom GL, Chen AX, Greenlee TE, Singh RP. Internet search engine queries of common causes of blindness and low vision in the United States. *American Journal of Ophthalmology*. 2021;222:373-381.

32. Senecal C, Widmer RJ, Lerman LO, Lerman A. Association of search engine queries for chest pain with coronary heart disease epidemiology. *JAMA cardiology*. 2018;3(12):1218-1221.

33. Oladeji O, Zhang C, Moradi T, et al. Monitoring Information-Seeking Patterns and Obesity Prevalence in Africa With Internet Search Data: Observational Study. *JMIR public health and surveillance*. 2021;7(4):e24348.

34. Arora VS, Stuckler D, Mckee M. Tracking search engine queries for suicide in the United Kingdom, 2004–2013. *Public health*. 2016;137:147-153.

35. Xu S, Chan HK. Forecasting medical device demand with online search queries: a big data and machine learning approach. *Procedia Manufacturing*. 2019;39:32-39.

36. Rivera JM, Gupta S, Ramjee D, et al. Evaluating interest in off-label use of disinfectants for COVID-19. *The Lancet Digital Health*. 2020;2(11):e564-e566.

37. Onchonga D. A Google Trends study on the interest in self-medication during the 2019 novel coronavirus (COVID-19) disease pandemic. *Saudi Pharmaceutical Journal: SPJ*. 2020;28(7):903.

38. Nsoesie EO, Cesare N, Müller M, Ozonoff A. COVID-19 Misinformation Spread in Eight Countries: Exponential Growth Modeling Study. *J Med Internet Res*. 2020;22(12):e24425. doi:10.2196/24425

39. Zitting KM, Lammers-van der Holst HM, Yuan RK, Wang W, Quan SF, Duffy JF. Google Trends reveals increases in internet searches for insomnia during the 2019 coronavirus disease (COVID-19) global pandemic. *Journal of Clinical Sleep Medicine*. 2021;17(2):177-184.

40. Rajan A, Sharaf R, Brown RS, Sharaiha RZ, Lebwohl B, Mahadev S. Association of search query interest in gastrointestinal symptoms with COVID-19 diagnosis in the United States: infodemiology study. *JMIR public health and surveillance*. 2020;6(3):e19354.

41. Ayyoubzadeh SM, Ayyoubzadeh SM, Zahedi H, Ahmadi M, Kalhori SRN. Predicting COVID-19 incidence through analysis of google trends data in Iran: data mining and deep learning pilot study. *JMIR public health and surveillance*. 2020;6(2):e18828.

42. McIver DJ, Brownstein JS. Wikipedia Usage Estimates Prevalence of Influenza-Like Illness in the United States in Near Real-Time. *PLoS Comput Biol*. 2014;10(4):e1003581. doi:10.1371/journal.pcbi.1003581

43. Nsoesie EO, Sy KTL, Oladeji O, Sefala R, Nichols BE. Nowcasting and forecasting provincial-level SARS-CoV-2 case positivity using google search data in South Africa. *medRxiv*. Published online 2020.

44. Nsoesie EO, Oladeji O, Abah ASA, Ndeffo-Mbah ML. Forecasting influenza-like illness trends in Cameroon using Google Search Data. *Scientific Reports*. 2021;11(1):1-11.

45. Sadilek A, Hswen Y, Bavadekar S, Shekel T, Brownstein JS, Gabrilovich E. Lymelight: forecasting Lyme disease risk using web search data. *NPJ digital medicine*. 2020;3(1):1-12.

46. Burnett RJ, van Gogh LJ, Moloi MH, François G. A profile of anti-vaccination lobbying on the South African internet, 2011-2013. *South African Medical Journal*. 2015;105(11):922-926.

47. Signorini A, Segre AM, Polgreen PM. The use of Twitter to track levels of disease activity and public concern in the US during the influenza A H1N1 pandemic. *PloS one*. 2011;6(5):e19467.

48. Chunara R, Andrews JR, Brownstein JS. Social and news media enable estimation of epidemiological patterns early in the 2010 Haitian cholera outbreak. *The American journal of tropical medicine and hygiene*. 2012;86(1):39.

49. Salathé M, Khandelwal S. Assessing Vaccination Sentiments with Online Social Media: Implications for Infectious Disease Dynamics and Control. *PLoS Comput Biol*. 2011;7(10):e1002199. doi:10.1371/journal.pcbi.1002199

50. McIver DJ, Hawkins JB, Chunara R, et al. Characterizing sleep issues using Twitter. *Journal of medical Internet research*. 2015;17(6).

51. Hawkins JB, Brownstein JS, Tuli G, et al. Measuring patient-perceived quality of care in US hospitals using Twitter. *BMJ Qual Saf*. 2016;25(6):404. doi:10.1136/bmjqs-2015-004309

52. Meadows CZ, Tang L, Liu W. Twitter message types, health beliefs, and vaccine attitudes during the 2015 measles outbreak in California. *American journal of infection control*. 2019;47(11):1314-1318.

53. Blankenship EB, Goff ME, Yin J, et al. Sentiment, contents, and retweets: a study of two vaccine-related twitter datasets. *The Permanente Journal*. 2018;22.

54. Yousefinaghani S, Dara R, Mubareka S, Papadopoulos A, Sharif S. An analysis of COVID-19 vaccine sentiments and opinions on Twitter. *International Journal of Infectious Diseases*. 2021;108:256-262.

55. Broniatowski DA, Jamison AM, Qi S, et al. Weaponized health communication: Twitter bots and Russian trolls amplify the vaccine debate. *American journal of public health*. 2018;108(10):1378-1384.

56. Seltzer EK, Horst-Martz E, Lu M, Merchant RM. Public sentiment and discourse about Zika virus on Instagram. *Public Health*. 2017;150:170-175.

57. Cesare N, Nguyen QC, Grant C, Nsoesie EO. Social media captures demographic and regional physical activity. *BMJ open sport & exercise medicine*. 2019;5(1):e000567.

58. Cesare N, Dwivedi P, Nguyen QC, Nsoesie EO. Use of social media, search queries, and demographic data to assess obesity prevalence in the United States. *Palgrave Communications*. 2019;5(1):1-9.

59. Nguyen QC, McCullough M, Meng H wen, et al. Geotagged US Tweets as Predictors of County-Level Health Outcomes, 2015–2016. *Am J Public Health*. 2017;107(11):1776-1782. doi:10.2105/AJPH.2017.303993

60. Nguyen QC, Li D, Meng HW, et al. Building a National Neighborhood Dataset From Geotagged Twitter Data for Indicators of Happiness, Diet, and Physical Activity. Eysenbach G, ed. *JMIR Public Health and Surveillance*. 2016;2(2):e158. doi:10.2196/publichealth.5869

61. HealthMap Foodborne Dashboard. Published online September 29, 2016. Accessed September 29, 2016. http://www.healthmap.org/foodborne/

62. Harris JK, Hinyard L, Beatty K, et al. Evaluating the implementation of a Twitter-based foodborne illness reporting tool in the city of St. Louis Department of Health. *International journal of environmental research and public health*. 2018;15(5):833.

63. Harris JK, Mansour R, Choucair B, Olson J, Nissen C, Bhatt J. Health department use of social media to identify foodborne illness-Chicago, Illinois, 2013-2014. *Morbidity and Mortality Weekly Report*. 2014;63(32):681-685.

64. Hawkins JB, Tuli G, Kluberg S, Harris J, Brownstein JS, Nsoesie E. A digital platform for local foodborne illness and outbreak surveillance. *Online Journal of Public Health Informatics*. 2016;8(1).

65. Guntuku SC, Schwartz HA, Kashyap A, et al. Variability in language used on social media prior to hospital visits. *Scientific reports*. 2020;10(1):1-9.

66. Eichstaedt JC, Smith RJ, Merchant RM, et al. Facebook language predicts depression in medical records. *Proceedings of the National Academy of Sciences*. 2018;115(44):11203-11208.

67. Merchant RM, Asch DA, Crutchley P, et al. Evaluating the predictability of medical conditions from social media posts. *PloS one*. 2019;14(6):e0215476.

68. Nikfarjam A, Sarker A, O'connor K, Ginn R, Gonzalez G. Pharmacovigilance from social media: mining adverse drug reaction mentions using sequence labeling with word embedding cluster features. *Journal of the American Medical Informatics Association*. 2015;22(3):671-681.

69. Graves RL, Perrone J, Al-Garadi MA, et al. Thematic analysis of reddit content about buprenorphine-naloxone using manual annotation and natural language processing techniques. *Journal of Addiction Medicine*. 2022;16(4):454.

70. Sarker A, Ginn R, Nikfarjam A, et al. Utilizing social media data for pharmacovigilance: a review. *Journal of biomedical informatics*. 2015;54:202-212.

71. Salathé M. Digital Pharmacovigilance and Disease Surveillance: Combining Traditional and Big-Data Systems for Better Public Health. *Journal of Infectious Diseases*. 2016;214(suppl 4):S399-S403. doi:10.1093/infdis/jiw281

72. Cesare N, Oladeji O, Ferryman K, et al. Discussions of miscarriage and preterm births on Twitter. *Paediatric and Perinatal Epidemiology*. Published online 2020.

73. De Choudhury M, Counts S, Horvitz EJ, Hoff A. Characterizing and predicting postpartum depression from shared facebook data. In: *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing*. ACM; 2014:626-638.

74. McCormick TH, Lee H, Cesare N, Shojaie A, Spiro ES. Using Twitter for Demographic and Social Science Research: Tools for Data Collection and Processing. *Sociological Methods & Research*. Published online October 9, 2015. doi:10.1177/0049124115605339

75. Hernandez MA, Modi S, Mittal K, et al. Diet during the COVID-19 pandemic: An analysis of Twitter data. *Patterns*. 2022;3(8):100547.

76. Bodnar T, Barclay VC, Ram N, Tucker CS, Salathé M. On the ground validation of online diagnosis with Twitter and medical records. In: *Proceedings of the 23rd International Conference on World Wide Web*. ; 2014:651-656.

77. Rexly Penaflorida II. Doctor Review Sites Where You Should Have A Profile. ReviewTrackers Blog. Published June 1, 2021. Accessed September 20, 2022. https://www.reviewtrackers.com/blog/doctor-review-sites/

78. Jayne O'Donnell. New doctors site rates for experience, quality. USA Today. Published October 14, 2019. Accessed September 20, 2022. https://www. usatoday.com/story/news/nation/2014/10/19/doctors- ratings-open-enrollment-quality-price/17371575/

79. Furnas HJ, Korman JM, Canales FL, Pence LD. Patient Reviews: Yelp, Google, Healthgrades, Vitals, and RealSelf. *Plastic and Reconstructive Surgery*. 2020;146(6):1419-1431.

80. Merchant RM, Volpp KG, Asch DA. Learning by listening—improving health Care in the era of yelp. *Jama*. 2016;316(23):2483-2484.

81. Ranard BL, Werner RM, Antanavicius T, et al. Yelp Reviews Of Hospital Care Can Supplement And Inform Traditional Surveys Of The Patient Experience Of Care. *Health Affairs*. 2016;35(4):697-705. doi:10.1377/hlthaff.2015.1030

82. Seltzer EK, Guntuku SC, Lanza AL, et al. Patient Experience and Satisfaction in Online Reviews of Obstetric Care: Observational Study. *JMIR Formative Research*. 2022;6(3):e28379.

83. Agarwal AK, Guntuku SC, Meisel ZF, Pelullo A, Kinkle B, Merchant RM. Analyzing Online Reviews of Substance Use Disorder Treatment Facilities in the USA Using Machine Learning. *Journal of General Internal Medicine*. 2022;37(4):977-980.

84. Donnally III CJ, Li DJ, Maguire Jr JA, et al. How social media, training, and demographics influence online reviews across three leading review websites for spine surgeons. *The spine journal*. 2018;18(11):2081-2090.

85. Tong JK, Akpek E, Naik A, et al. Reporting of Discrimination by Health Care Consumers Through Online Consumer Reviews. *JAMA network open*. 2022;5(2):e220715-e220715.

86. Nali MC, Purushothaman V, Li J, Mackey TK. Characterizing California licensure status and tobacco user experience with adverse events using Yelp data. *Preventive medicine reports*. 2022;26:101720.

87. Sussman S, Garcia R, Cruz TB, Baezconde-Garbanati L, Pentz MA, Unger JB. Consumers' perceptions of vape shops in Southern California: an analysis of online Yelp reviews. *Tobacco induced diseases*. 2014;12(1):1-9.

88. Cawkwell PB, Lee L, Weitzman M, Sherman SE. Tracking Hookah Bars in New York: Utilizing Yelp as a Powerful Public Health Tool. Eysenbach G, ed. *JMIR Public Health and Surveillance*. 2015;1(2):e19. doi:10.2196/publichealth.4809

89. Stokes DC, Kishton R, McCalpin HJ, et al. Online reviews of mental health treatment facilities: narrative themes associated with positive and negative ratings. *Psychiatric services*. 2021;72(7):776-783.

90. Loew L. *Los Angeles County Restaurant Hygiene Grades Now on Yelp.*; 2016.

91. Harrison C, Jorder M, Stern H, et al. Using online reviews by restaurant patrons to identify unreported cases of foodborne illness-New York City, 2012-2013. *MMWR Morbidity and mortality weekly report*. 2014;63(20):441-445.

92. Effland T, Lawson A, Balter S, et al. Discovering foodborne illness in online restaurant reviews. *Journal of the American Medical Informatics Association*. Published online 2018:ocx093. doi:10.1093/jamia/ocx093

93. Maharana A, Cai K, Hellerstein J, et al. Detecting reports of unsafe foods in consumer product reviews. *JAMIA Open*. 2019;2(3):330-338. doi:10.1093/jamiaopen/ooz030

94. Park H, Kim J, Almanza B. Yelp versus inspection reports: is quality correlated with sanitation in retail food facilities? *Journal of Environmental Health*. 2016;78(10):8-13.

95. Elaine O Nsoesie, Sheryl A Gordon, John S Brownstein. Online Reports of Foodborne Illness Capture Foods Implicated in Official Foodborne Outbreak Reports. *Prev Med*. 2014;In Press.

96. Stokes DC, Pelullo AP, Mitra N, et al. Association between crowdsourced health care facility ratings and mortality in US counties. *JAMA network open*. 2021;4(10):e2127799-e2127799.

97. Luca M, Zervas G. Fake it till you make it: Reputation, competition, and Yelp review fraud. *Management Science*. 2016;62(12):3412-3427.

98. Anthes E. Pocket psychiatry: mobile mental-health apps have exploded onto the market, but few have been thoroughly tested. *Nature*. 2016;532(7597):20-24.

99. Price M, Yuen EK, Goetter EM, et al. mHealth: a mechanism to deliver more accessible, more effective mental health care. *Clinical psychology & psychotherapy*. 2014;21(5):427-436.

100. Sim I. Mobile devices and health. *New England Journal of Medicine*. 2019;381(10):956-968.

101. Vijayan V, Connolly JP, Condell J, McKelvey N, Gardiner P. Review of wearable devices and data collection considerations for connected health. *Sensors*. 2021;21(16):5589.

102. Global Connected Wearable Devices 2016–2022 Statista. Statista. Published 2020. www.statista.com/statistics/487291/global-connected-wearable-devices/

103. Perez AJ, Zeadally S. Recent advances in wearable sensing technologies. *Sensors*. 2021;21(20):6828.

104. Case MA, Burwick HA, Volpp KG, Patel MS. Accuracy of smartphone applications and wearable devices for tracking physical activity data. *Jama*. 2015;313(6):625-626.

105. Althoff T, Sosič R, Hicks JL, King AC, Delp SL, Leskovec J. Large-scale physical activity data reveal worldwide activity inequality. *Nature*. 2017;547(7663):336-339.

106. Piwek L, Ellis DA, Andrews S, Joinson A. The rise of consumer health wearables: promises and barriers. *PLoS medicine*. 2016;13(2):e1001953.

107. Feehan LM, Geldman J, Sayre EC, et al. Accuracy of Fitbit devices: systematic review and narrative syntheses of quantitative data. *JMIR mHealth and uHealth*. 2018;6(8):e10527.

108. Feiner JR, Severinghaus JW, Bickler PE. Dark skin decreases the accuracy of pulse oximeters at low oxygen saturation: the effects of oximeter probe type and gender. *Anesthesia & Analgesia*. 2007;105(6):S18-S23.

109. Fawzy A, Wu TD, Wang K, et al. Racial and Ethnic Discrepancy in Pulse Oximetry and Delayed Identification of Treatment Eligibility Among Patients With COVID-19. *JAMA Internal Medicine*. Published online 2022.

110. Brown C, Chauhan J, Grammenos A, et al. Exploring automatic diagnosis of COVID-19 from crowdsourced respiratory sound data. *arXiv preprint arXiv:200605919*. Published online 2020.

111. Lella KK, PJA A. A literature review on COVID-19 disease diagnosis from respiratory sound data. *arXiv preprint arXiv:211207670*. Published online 2021.

112. Imran A, Posokhova I, Qureshi HN, et al. AI4COVID-19: AI enabled preliminary diagnosis for COVID-19 from cough samples via an app. *Informatics in Medicine Unlocked*. 2020;20:100378.

113. Hassan A, Shahin I, Alsabek MB. Covid-19 detection system using recurrent neural networks. In: *2020 International Conference on Communications, Computing, Cybersecurity, and Informatics (CCCI)*. IEEE; 2020:1-5.

114. Orlandic L, Teijeiro T, Atienza D. The COUGHVID crowdsourcing dataset, a corpus for the study of large-scale cough analysis algorithms. *Scientific Data*. 2021;8(1):1-10.

115. Alsabek MB, Shahin I, Hassan A. Studying the Similarity of COVID-19 Sounds based on Correlation Analysis of MFCC. In: *2020 International Conference on Communications, Computing, Cybersecurity, and Informatics (CCCI)*. IEEE; 2020:1-5.

116. Al Ismail M, Deshmukh S, Singh R. Detection of COVID-19 through the analysis of vocal fold oscillations. In: *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE; 2021:1035-1039.

117. Chaudhari G, Jiang X, Fakhry A, et al. Virufy: Global applicability of crowdsourced and clinical datasets for AI detection of COVID-19 from cough. *arXiv preprint arXiv:201113320*. Published online 2020.

118. Laguarta J, Hueto F, Subirana B. COVID-19 artificial intelligence diagnosis using only cough recordings. *IEEE Open Journal of Engineering in Medicine and Biology*. 2020;1:275-281.

119. Quatieri TF, Talkar T, Palmer JS. A framework for biomarkers of COVID-19 based on coordination of speech-production subsystems. *IEEE Open Journal of Engineering in Medicine and Biology*. 2020;1:203-206.

120. Han J, Qian K, Song M, et al. An early study on intelligent analysis of speech under COVID-19: Severity, sleep quality, fatigue, and anxiety. *arXiv preprint arXiv:200500096*. Published online 2020.

121. Ritwik KVS, Kalluri SB, Vijayasenan D. COVID-19 patient detection from telephone quality speech data. *arXiv preprint arXiv:201104299*. Published online 2020.

122. Accessed September 22, 2022. www.covid-19-sounds.org

123. Accessed September 22, 2022. https://cvd.lti.cmu.edu/

124. Sharma N, Krishnan P, Kumar R, et al. Coswara--a database of breathing, cough, and voice sounds for COVID-19 diagnosis. *arXiv preprint arXiv:200510548*. Published online 2020.

125. Wesolowski A, Qureshi T, Boni MF, et al. Impact of human mobility on the emergence of dengue epidemics in Pakistan. *Proceedings of the National Academy of Sciences*. 2015;112(38):11887-11892. doi:10.1073/pnas.1504964112

126. Buckee CO, Balsari S, Chan J, et al. Aggregated mobility data could help fight COVID-19. *Science (New York, NY)*. Published online 2020.

127. Zufiria PJ, Pastor-Escuredo D, Úbeda-Medina L, et al. Identifying seasonal mobility profiles from anonymized and aggregated mobile phone data. Application in food security. *PloS one*. 2018;13(4):e0195714.

128. Wesolowski A, Buckee CO, Engø-Monsen K, Metcalf CJE. Connecting mobility to infectious diseases: the promise and limits of mobile phone data. *The Journal of infectious diseases*. 2016;214(suppl_4):S414-S420.

129. Gonzalez MC, Hidalgo CA, Barabasi AL. Understanding individual human mobility patterns. *Nature*. 2008;453:779-782.

130. Wesolowski A, Eagle N, Tatem AJ, et al. Quantifying the Impact of Human Mobility on Malaria. *Science*. 2012;338(6104):267-270. doi:10.1126/science.1223467

131. Tatem A, Qiu Y, Smith D, Sabot O, Ali A, Moonen B. The use of mobile phone data for the estimation of the travel patterns and imported Plasmodium falciparum rates among Zanzibar residents. *Malaria Journal*. 2009;8(1):287.

132. Tizzoni M, Bajardi P, Decuyper A, et al. On the use of human mobility proxies for modeling epidemics. *PLoS Comput Biol*. 2014;10. doi:10.1371/journal.pcbi.1003716

133. Bengtsson L, Gaudart J, Lu X, et al. Using mobile phone data to predict the spatial spread of cholera. *Sci Rep*. 2015;5. doi:10.1038/srep08923

134. Tatem A, Huang Z, Narib C, et al. Integrating rapid risk mapping and mobile phone call record data for strategic malaria elimination planning. *Malaria Journal*. 2014;13(1):52.

135. Wesolowski A, Eagle B, Tatem AJ, et al. Quantifying the impact of human mobility on malaria. *Science*. 2012;338. doi:10.1126/science.1223467

136. Grantz KH, Meredith HR, Cummings DA, et al. The use of mobile phone data to inform analysis of COVID-19 pandemic epidemiology. *Nature communications*. 2020;11(1):1-8.

137. Chang S, Pierson E, Koh PW, et al. Mobility network models of COVID-19 explain inequities and inform reopening. *Nature*. 2021;589(7840):82-87.

138. Jay J, Bor J, Nsoesie EO, et al. Neighbourhood income and physical distancing during the COVID-19 pandemic in the United States. *Nature human behaviour*. 2020;4(12):1294-1302.

139. Nguyen TT, Nguyen QC, Rubinsky AD, et al. Google Street View-Derived Neighborhood Characteristics in California Associated with Coronary Heart Disease, Hypertension, Diabetes. *IJERPH*. 2021;18(19):10428. doi:10.3390/ijerph181910428

140. Nguyen QC, Sajjadi M, McCullough M, et al. Neighbourhood looking glass: 360° automated characterisation of the built environment for neighbourhood effects research. *Journal of Epidemiology & Community Health*. Published online 2018. doi:10.1136/jech-2017-209456

141. Maharana A, Nsoesie EO. Using Deep Learning to Examine the Association between the Built Environment and Neighborhood Adult Obesity Prevalence. *arXiv preprint arXiv:171100885*. Published online 2017.

142. Rundle AG, Bader MDM, Richards CA, Neckerman KM, Teitler JO. Using Google Street View to Audit Neighborhood Environments. *American Journal of Preventive Medicine*. 2011;40(1):94-100. doi:10.1016/j.amepre.2010.09.034

143. Silva V, Grande A, Rech C, Peccin M. Geoprocessing via google maps for assessing obesogenic built environments related to physical activity and chronic noncommunicable diseases: validity and reliability. *Journal of healthcare engineering*. 2015;6(1):41-54.

144. Kelly CM, Wilson JS, Baker EA, Miller DK, Schootman M. Using Google Street View to audit the built environment: inter-rater reliability results. *Annals of Behavioral Medicine*. 2013;45(suppl_1):S108-S112.

145. Nguyen QC, Khanna S, Dwivedi P, et al. Using Google Street View to examine associations between built environment characteristics and US health outcomes. *Preventive medicine reports*. 2019;14:100859.

146. Albert A, Kaur J, Gonzalez M. Using convolutional networks and satellite imagery to identify patterns in urban environments at a large scale. *arXiv preprint arXiv:170402965*. Published online 2017.

147. Jean N, Burke M, Xie M, Davis WM, Lobell DB, Ermon S. Combining satellite imagery and machine learning to predict poverty. *Science*. 2016;353(6301):790-794.

148. Guntuku SC, Yaden DB, Kern ML, Ungar LH, Eichstaedt JC. Detecting depression and mental illness on social media: an integrative review. *Current Opinion in Behavioral Sciences*. 2017;18:43-49.

149. Hochberg I, Daoud D, Shehadeh N, Yom-Tov E. Can internet search engine queries be used to diagnose diabetes? Analysis of archival search data. *Acta diabetologica*. 2019;56(10):1149-1154.

150. White RW, Doraiswamy PM, Horvitz E. Detecting neurodegenerative disorders from web search signals. *NPJ digital medicine*. 2018;1(1):1-4.

151. White RW, Horvitz E. Evaluation of the feasibility of screening patients for early signs of lung carcinoma in web search logs. *JAMA oncology*. 2017;3(3):398-401.

152. Yom-Tov E, White RW, Horvitz E. Seeking insights about cycling mood disorders via anonymized search logs. *Journal of medical Internet research*. 2014;16(2):e2664.

153. Ofran Y, Paltiel O, Pelleg D, Rowe JM, Yom-Tov E. Patterns of information-seeking for cancer on the internet: an analysis of real world data. Published online 2012.

154. Paparrizos J, White RW, Horvitz E. Screening for pancreatic adenocarcinoma using signals from web search logs: Feasibility study and results. *Journal of Oncology Practice*. 2016;12(8):737-744.

155. Huston JE, Mekaru SR, Kluberg S, Brownstein JS. Searching the web for influenza vaccines: Healthmap vaccine finder. *American journal of public health*. 2015;105(8):e134-e139.

156. Birkhoff SD, Smeltzer SC. Perceptions of smartphone user-centered mobile health tracking apps across various chronic illness populations: an integrative review. *Journal of nursing scholarship*. 2017;49(4):371-378.

157. Vilardaga R, Rizo J, Zeng E, et al. User-centered design of learn to quit, a smoking cessation smartphone app for people with serious mental illness. *JMIR serious games*. 2018;6(1):e8881.

158. Alberts NM, Badawy SM, Hodges J, et al. Development of the InCharge Health Mobile App to improve adherence to hydroxyurea in patients with sickle cell disease: user-centered design approach. *JMIR mHealth and uHealth*. 2020;8(5):e14884.

159. Greaves F, Pape UJ, King D, et al. Associations between Internet-based patient ratings and conventional surveys of patient experience in the English NHS: an observational study. *BMJ quality & safety*. 2012;21(7):600-605.

160. Pepe E, Bajardi P, Gauvin L, et al. COVID-19 outbreak response, a dataset to assess mobility changes in Italy following national lockdown. *Scientific data*. 2020;7(1):1-7.

161. Rader B, Astley CM, Sy KTL, et al. Geographic access to United States SARS-CoV-2 testing sites highlights healthcare disparities and may bias transmission estimates. *Journal of travel medicine*. Published online 2020.

162. Maharana A, Nsoesie EO. Use of deep learning to examine the association of the built environment with prevalence of neighborhood adult obesity. *JAMA network open*. 2018;1(4):e181535-e181535.

163. Nguyen QC, Belnap T, Dwivedi P, et al. Google Street View Images as Predictors of Patient Health Outcomes, 2017–2019. *BDCC*. 2022;6(1):15. doi:10.3390/bdcc6010015

164. Gasser U, Ienca M, Scheibner J, Sleigh J, Vayena E. Digital tools against COVID-19: taxonomy, ethical challenges, and navigation aid. *The Lancet Digital Health*. 2020;2(8):e425-e434.

165. Vayena E, Salathé M, Madoff LC, Brownstein JS. *Ethical Challenges of Big Data in Public Health*. Public Library of Science; 2015.

166. Metcalf J, Crawford K. Where are human subjects in big data research? The emerging ethics divide. *Big Data & Society*. 2016;3(1):2053951716650211.

167. Brownstein JS, Cassa CA, Mandl KD. No place to hide--reverse identification of patients from published maps. *The New England journal of medicine*. 2006;355:1741-1742. doi:10.1056/NEJMc061891

168. Van de Mortel TF. Faking it: social desirability response bias in self-report research. *Australian Journal of Advanced Nursing, The*. 2008;25(4):40-48.

169. Hassan E. Recall bias can be a threat to retrospective and prospective research designs. *The Internet Journal of Epidemiology*. 2006;3(2):339-412.

170. Althubaiti A. Information bias in health research: definition, pitfalls, and adjustment methods. *Journal of multidisciplinary healthcare*. 2016;9:211.

171. Kuhlman C, Jackson L, Chunara R. No computation without representation: Avoiding data and algorithm biases through diversity. *arXiv preprint arXiv:200211836*. Published online 2020.

172. Henly S, Tuli G, Kluberg SA, et al. Disparities in digital reporting of illness: A demographic and socioeconomic assessment. *Preventive Medicine*. 2017;101:18-22.

173. Thomasian NM, Eickhoff C, Adashi EY. Advancing health equity with artificial intelligence. *Journal of public health policy*. Published online 2021:1-10.

174. Merchant RM, Asch DA, Crutchley P, et al. Evaluating the predictability of medical conditions from social media posts. *PloS one*. 2019;14(6):e0215476.

175. Padrez KA, Ungar L, Schwartz HA, et al. Linking social media and medical record data: a study of adults presenting to an academic, urban emergency department. *BMJ quality & safety*. 2016;25(6):414-423.

176. Cesare N, Grant C, Nsoesie EO. Detection of User Demographics on Social Media: A Review of Methods and Recommendations for Best Practices. *CoRR*. 2017;abs/1702.01807. http://arxiv.org/abs/1702.01807

177. Cesare N, Grant C, Hawkins JB, Brownstein JS, Nsoesie EO. Demographics in Social Media Data for Public Health Research: Does it matter? *arXiv preprint arXiv:171011048*. Published online 2017.

178. Jaidka K, Giorgi S, Schwartz HA, Kern ML, Ungar LH, Eichstaedt JC. Estimating geographic subjective well-being from Twitter: A comparison of dictionary and data-driven language methods. *Proceedings of the National Academy of Sciences*. 2020;117(19):10165-10171.

179. Iacus SM, Porro G, Salini S, Siletti E. An Italian composite subjective well-being index: The voice of Twitter users from 2012 to 2017. *Social Indicators Research*. Published online 2020:1-19.

180. Weeg C, Schwartz HA, Hill S, Merchant RM, Arango C, Ungar L. Using Twitter to measure public discussion of diseases: a case study. *JMIR public health and surveillance*. 2015;1(1):e3953.

181. Bias M. *There's Software Used across the Country to Predict Future Criminals. And It's Biased against Blacks. by Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica, May 23, 2016.*

182. West SM, Whittaker M, Crawford K. Discriminating systems. *AI Now*. Published online 2019.

183. Mitchell S, Potash E, Barocas S, D'Amour A, Lum K. Algorithmic fairness: Choices, assumptions, and definitions. *Annual Review of Statistics and Its Application*. 2021;8:141-163.

184. Fletcher RR, Nakeshimana A, Olubeko O. *Addressing Fairness, Bias, and Appropriate Use of Artificial Intelligence and Machine Learning in Global Health*. Vol 3. Frontiers Media SA; 2021:561802.

185. Celedonia KL, Corrales Compagnucci M, Minssen T, Lowery Wilson M. Legal, ethical, and wider implications of suicide risk detection systems in social media platforms. *Journal of Law and the Biosciences*. 2021;8(1):lsab021.

186. Xafis V, Schaefer GO, Labude MK, et al. An ethics framework for big data in health and research. *Asian Bioethics Review*. 2019;11(3):227-254.

187. Floridi L, Cowls J, Beltrametti M, et al. AI4People—An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. *Minds and machines*. 2018;28(4):689-707.

188. McCosker A, Gerrard Y. Hashtagging depression on Instagram: Towards a more inclusive mental health research methodology. *new media & society*. 2021;23(7):1899-1919.

189. Yang JA, Tsou MH, Jung CT, et al. Social media analytics and research testbed (SMART): Exploring spatiotemporal patterns of human dynamics with geo-targeted social media messages. *Big Data & Society*. 2016;3(1):2053951716652914.

190. Aslam AA, Tsou MH, Spitzberg BH, et al. The reliability of tweets as a supplementary method of seasonal influenza surveillance. *Journal of medical Internet research*. 2014;16(11):e3532.

191. Culotta A. Lightweight methods to estimate influenza rates and alcohol sales volume from Twitter messages. *Language resources and evaluation*. 2013;47(1):217-238.

192. Tufts C, Polsky D, Volpp KG, et al. Characterizing tweet volume and content about common health conditions across Pennsylvania: retrospective analysis. *JMIR Public Health and Surveillance*. 2018;4(4):e10834.

193. Mowery D, Smith H, Cheney T, et al. Understanding depressive symptoms and psychosocial stressors on Twitter: a corpus-based study. *Journal of medical Internet research*. 2017;19(2):e6895.

194. Gattepaille LM, Hedfors Vidlin S, Bergvall T, Pierce CE, Ellenius J. Prospective evaluation of adverse event recognition systems in Twitter: Results from the Web-RADR Project. *Drug safety*. 2020;43(8):797-808.

195. Elkin LE, Pullon SR, Stubbe MH. 'Should I vaccinate my child?'Comparing the displayed stances of vaccine information retrieved from Google, Facebook and YouTube. *Vaccine*. 2020;38(13):2771-2778.

196. Santillana M, Nguyen AT, Dredze M, Paul MJ, Nsoesie EO, Brownstein JS. Combining Search, Social Media, and Traditional Data Sources to Improve Influenza Surveillance. Salathé M, ed. *PLoS Computational Biology*. 2015;11(10):e1004513. doi:10.1371/journal.pcbi.1004513