ANNUAL REVIEWS

Annual Review of Public Health

Social Media– and Internet-Based Disease Surveillance for Public Health

Allison E. Aiello, Audrey Renson, and Paul N. Zivich

Department of Epidemiology, Carolina Population Center, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27599-7435, USA; email: aaiello@email.unc.edu, arenson@live.unc.edu, pzivich@live.unc.edu

ANNUAL REVIEWS CONNECT

www.annualreviews.org

- Download figures
- Navigate cited references
- Keyword search
- Explore related articles
- Share via email or social media

Keywords

social media, digital health, mhealth, big data, infectious diseases, surveillance

Abstract

Disease surveillance systems are a cornerstone of public health tracking and prevention. This review addresses the use, promise, perils, and ethics of social media– and Internet-based data collection for public health surveillance. Our review highlights untapped opportunities for integrating digital surveillance in public health and current applications that could be improved through better integration, validation, and clarity on rules surrounding ethical considerations. Promising developments include hybrid systems that couple traditional surveillance data with data from search queries, social media posts, and crowdsourcing. In the future, it will be important to identify opportunities for public and private partnerships, train public health experts in data science, reduce biases related to digital data (gathered from Internet use, wearable devices, etc.), and address privacy. We are on the precipice of an unprecedented opportunity to track, predict, and prevent global disease burdens in the population using digital data.

INTRODUCTION

Disease surveillance in the community setting is a cornerstone of public health tracking and pre-
vention. The World Health Organization defines public health surveillance as "the continuous,
systematic collection, analysis, and interpretation of health-related data needed for the planning,
implementation, and evaluation of public health practice" (108; see a related definition from Ref-
erence 76). Public health surveillance acts as a sentinel for identifying trends in disease and emerg-
ing public health concerns and can help to identify potential points of intervention. Furthermore,
surveillance data can provide benchmarks for evaluating intervention measures for curbing disease
spread in populations and allow health experts to set priorities and policies.

Surveillance has undergone numerous changes over the years (20) and will likely continue
to evolve. Recent changes in disease surveillance systems are a result of technological advances
for data collection related to access to the Internet and improved computational power. Digital
disease surveillance can be defined as the use of Internet-based data in the explicit development or
application of systems aimed at nowcasting or forecasting of disease incidence or prevalence (72).
In recent years, Internet-based search tools and social media have created exciting new prospects
for expanding disease surveillance by capturing real-time data and trends for health outcomes.
Figure 1 highlights some of the key developments related to digital public health surveillance,
including the introduction of Google Flu Trends.[1]

Data collected via the Internet and social media sites have typically been used as a complemen-
tary source of surveillance data for existing outpatient, hospital, and laboratory-based systems.
This complement can be extremely useful because traditional surveillance systems rely on track-
ing individuals seeking care and therefore underestimate the total disease burden from a lack of
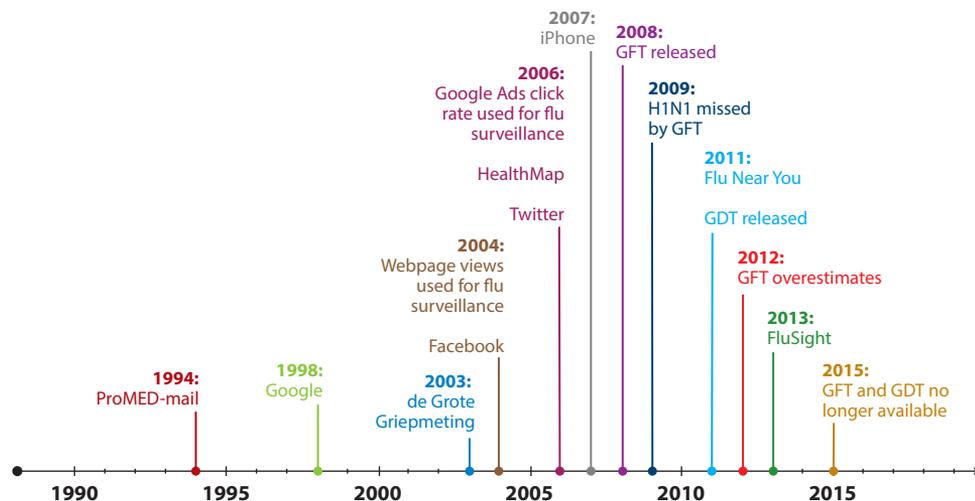
Figure 1

Major events in digital public health surveillance. Data for each timeline entry were taken from the following
sources: ProMED-mail (63), de Grote Griepmeting (58), Webpage views used for flu surveillance (54),
Google Ads click rate used for surveillance (36), HealthMap (11), H1N1 missed by GFT (25), Flu Near You
(96), GFT overestimates (71), FluSight (16), GFT and GDT no longer available (38). Abbreviations: GFT,
Google Flu Trends; GDT, Google Dengue Trends.

---

[1]After failing to predict the 2009 pandemic and overestimating the peak intensity of the 2012–2013 influenza
season, the publicly available Google Flu Trends site was removed (42).

representativeness (99). One example of the complementary approach is the inclusion of electronic health records and historical influenza-like illness data in models built on Google search terms, which increased accuracy over Google search trend data alone (114, 115). However, some digital surveillance arises solely from digital data without necessarily complementing traditional public health surveillance systems. For example, Google searches for "diarrhea," "food poisoning," and other related terms were shown to coincide independently with a peanut butter–associated outbreak of *Salmonella enterica* (10). While this example of digital data surveillance was never directly integrated with traditional passive public health surveillance, it is possible to link these types of data with traditional foodborne illness surveillance data. Examples where the link with traditional public health surveillance is less clear, but may still provide important public health surveillance information, include the Internet collection of data on public attitudes toward vaccination (87), health behaviors such as tracking diet success (39), and smoking cessation (2).

Digital surveillance has been conducted for multiple health events using various sources of search query data, including dengue virus incidence using Google search queries (1), queries from Baidu (a large Chinese search engine similar to Google) (62), and vaccine effectiveness using Google search queries (92). Online restaurant reservation and review logs have been used to identify foodborne disease and influenza-like illness outbreaks (49, 69). While there are a variety of new applications in digital public health surveillance, the most utilized has been within the area of influenza surveillance and tracking.

Several digital surveillance systems have been used for influenza, including Google Flu Trends, Influenzanet, and Flu Near You. These systems broadly aim to provide timely reports of influenza incidence in local areas (15, 23, 96). While these influenza tracking examples and other digital health surveillance tools demonstrate promise for expanding surveillance through digital means, there are also some key trade-offs, such as concerns related to accuracy, privacy, and navigating public/private partnerships for connecting technology companies with government surveillance efforts.

This review addresses the use, promise, potential perils, and ethics of the use of social media and Internet-based data collection for public health surveillance. Within this review, we present a range of examples with a larger focus on influenza, given the high utility of surveillance for this critical public health concern. We also discuss next steps and the potential for future application of surveillance through digital sources.

Digital public health surveillance: public health surveillance with the inclusion of digital data, particularly from social media or other Internet-based sources

Google Flu Trends: publicly available site that used aggregate Google search trend data to forecast influenza-like illness incidence

DIGITAL HEALTH SURVEILLANCE

Public health surveillance is the "systematic and continuous collection, analysis, and interpretation of data, closely integrated with the timely and coherent dissemination of the results and assessment to those who have the right to know so that action can be taken" (76, p. 239). Digital public health surveillance, which we refer to as digital surveillance hereafter, is the inclusion of digital data, particularly from social media or other Internet-based sources, for this same purpose. Others have further distinguished digital surveillance as data collected outside the public health system (84). Moreover, digital health surveillance data are often linked to a nonhealth data source (e.g., pulling health data from a Twitter user who posts on a range of topics, which may also include mentions of their health-related information), unlike more traditional public health surveillance systems.

Since the early 1990s, digital health surveillance has evolved closely in tandem with the Internet itself. Early systems such as ProMED-mail, an expert-moderated list of email messages related to the spread of emerging infectious diseases, helped galvanize interest in the use of the Internet in public health communication and surveillance through the promise of early widespread outbreak notifications (63). Beginning in the early 2000s, digital surveillance efforts have largely

encompassed three major types of Web-based activity: (*a*) aggregate trends derived from searches (e.g., Google search trends, Wikipedia page views), (*b*) social media postings (e.g., Facebook posts, tweets), and (*c*) participatory surveillance efforts (e.g., Flu Near You, Influenzanet). Each of these sources rely on differing online interactions for input. Some require individuals to be actively seeking health information online, thereby providing relevant public health surveillance data indirectly. Conversely, individuals may opt, passively or purposefully, to share health-relevant information on social media for a variety of reasons. Distinguishing between these different types of data collection through digital sources of surveillance data may be important because active versus passive information provides different levels of confidence in the specificity of the data; i.e., there are many reasons to visit the Wikipedia page on influenza, but a tweet describing one's symptoms may be a more accurate reflection of actual illness. Additionally, the population captured by these approaches likely differs, as certain types of individuals may be more likely to engage in providing participatory data on their health or tweet about symptoms to get input about their health from others compared with individuals who utilize a simple query to search the Web about symptoms or a disease.

Digital data have been processed primarily to either forecast or nowcast infectious disease outbreaks. Nowcasting is short-term prediction that attempts to track the present state of incidence in near real time, whereas forecasting aims to predict the future. The way in which the data are prepared for these purposes often requires substantial preprocessing of raw data. To use aggregate trend data from searches, data need to be filtered by keywords that correspond with disease incidence. The selection of search terms to build these prediction algorithms requires careful consideration because the choice of words can have a significant impact on the accuracy of the surveillance predictions (25). With social media data, natural-language processing and image analysis applied to posts may be used to further extract features of users' posts. Owing to the volume and types of data in digital sources, surveillance projects often utilize machine learning algorithms (1, 79, 83). Machine learning is a broad term that refers to approaches that adapt to patterns in data without explicit programming of the prediction task (68). Some examples of algorithms include decision trees, neural networks, and support vector machines. Digital surveillance systems generally use supervised learning, where data patterns are compared with a user-specified outcome. Supervised machine learning has seen success in addressing problems of prediction, particularly with image analysis, including classification of skin lesions (34), identification of early breast cancer indicators (105), and detection of diabetic retinopathy (46). However, having more data with more advanced prediction algorithms does not necessarily imply improvements to prediction (22). The mere use of digital surveillance and associated tools, such as machine learning, does not necessarily mean these systems will be better than traditional surveillance (66).

Examples of Digital Surveillance

Digital surveillance can be used in many ways to monitor trends and detect disease outbreaks, including through enhancing other data sources, identifying geographic spread, and optimizing existing surveillance systems. An example of digital surveillance being used to enhance other data sources is through incorporating Google data as a "virtual provider" to enhance accuracy of an existing influenza-like illness surveillance system based on a network of outpatient providers (91), using tweets to identify restaurants potentially responsible for foodborne infections and subsequently target inspections (48), and using Google to nowcast a plague outbreak in Madagascar, groundtruthed against health care–based statistics (7). Digital data have also been used to identify geographical spread of infectious diseases; in particular, Twitter geolocation data have been used along with air traffic data to track the spread of Chikungunya virus (80) and incorporated into mechanistic models of the flu to forecast peak time and intensity (118). Furthermore, digital data

have optimized traditional surveillance through various means. For influenza surveillance, digital data have been directly integrated with existing surveillance systems on influenza-like illness (16, 18, 91). Possible cases of foodborne illness have been identified through tweets with natural-language processing, and users were provided information on how to report foodborne illnesses (47, 48). Last, digital data have also aided the identification of foodborne outbreak point sources through tweets (48), Google search and location logs (83), and Yelp reviews (70). Other examples of digital data use for surveillance abound, including mosquito-borne infectious diseases (1, 15, 19, 80, 107), foodborne infectious diseases (29, 47, 70, 83), and attitudes/behaviors, including those related to vaccination (2, 39).

Search Query Examples

The most prominent examples of digital surveillance have been related to efforts for tracking influenza, where initial strategies focused on search query-based digital surveillance data. An early example of this approach was to demonstrate the correlation between Google Ads click rates and influenza incidence in Canada (36), and the correlation between Yahoo search trends and influenza incidence (75). Inspired by these approaches, Google Flu Trends was a publicly available platform communicating predicted influenza incidence from a model based on Google search query volumes. The Centers for Disease Control and Prevention (CDC)'s weekly influenza-like illness reports typically experience a 1–2-week lag (8), so the goal of Google Flu Trends was to predict influenza approximately 1 week ahead of the CDC. Google Flu Trends used a linear regression model of 45 unique search queries highly correlated with the influenza time series, manually pruned for feature relevance (e.g., terms related to basketball, which seasonally correlate with flu, were manually removed) (42). After Google Flu Trends's release in 2008, the data were found to correlate highly with traditional surveillance systems in several countries and reliably predict influenza-like illness incidence one to two weeks in advance for the 2007–2008 and 2008–2009 seasons (42, 56, 110). However, failure to detect the 2009 A/H1N1 pandemic led to an initial update to the model fit that correctly predicted the pandemic in retrospect (25). Despite the update, both the original and the updated Google Flu Trends vastly overestimated the peak intensity for the 2012–2013 influenza season (71). Ultimately these failures led to the removal of the public-facing site [although Google still makes its Google Flu Trends data available to researchers who request it directly (38)]. Current efforts to predict infectious diseases using Google typically rely on aggregate search volume data available through Google Trends (time series of relative volumes of specific search terms) and Google Correlate (correlations in Google Trends for different terms and comparative Google Trends between US states).

Another example in the influenza digital surveillance field was the CDC's "Forecast the Influenza Season Collaborative Challenge" (also known as FluSight), an annual competition in which teams of researchers compete to develop the most accurate weekly regional-level influenza-like illness predictions; teams are required to use some form of digital data, whether it is search query, social media, or other Internet-based data (3, 16). Teams were also allowed to incorporate traditional data sources for influenza surveillance, such as ILINet, which provides data on outpatient reports of influenza-like illness rates in a timely manner (8). This effort spurred broad interest in influenza modeling using digital data, with more teams competing every year (3, 4, 65). A major contribution of the FluSight competition has been the development of useful targets for prediction; in the first season, targets included timing of season onset, peak week, peak intensity, and duration (3). Compared with purely statistical targets such as correlation or mean squared error with the observed trend, which are frequently used (21, 113), these targets provide information about the public health impact of a given model (3, 21). Reich et al. (79) recently published a detailed summary of the performance of different FluSight predictive approaches. Performance of

Search query–based digital surveillance: digital public health surveillance systems that use aggregate search query data to monitor disease trends (examples include Google Flu Trends and Google Dengue Trends)

FluSight: the CDC's "Forecast the Influenza Season Collaborative Challenge"; an annual competition in which researchers compete to develop the most accurate weekly influenza-like illness predictions with some form of digital data

algorithms was assessed by onset of the influenza season, peak of the influenza season, and fore-cast incidence at one, two, three, and four weeks in advance. All models were compared with the predictions on the basis of the historical average of past seasons. Most models outperformed the historical average approach. As theory regarding ensemble approaches would suggest (30, 100), an ensemble of all submitted FluSight models outperformed each of the individual models (79). Similar competitions to FluSight have since been developed for other diseases, such as dengue (12), chikungunya (27) and Ebola (103).

Social Media Examples

In addition to search queries, social media–based digital surveillance data have become a vital part of efforts to monitor disease trends, with Twitter being a highly used platform. This high usage is facilitated by Twitter's relatively open data policy, allowing public access to a 1% random sample of raw tweets (**https://developer.twitter.com/en/docs/tweets/sample-realtime/overview**). For this reason, Twitter has become a "model organism" for digital research data (98). Six of nine teams in the first season of FluSight used Twitter alone or in addition to other digital sources (3). The most typical use of Twitter data involves content identification, through either keyword search or natural-language processing, to identify tweets related to health conditions such as the flu. Epidemic levels are then modeled as a function of tweet frequencies (81). However, an additional benefit of Twitter data is the availability of geolocated tweets, which can be used to model disease spread as a function of human geographic movement, potentially offering greater accuracy (80). While Twitter is by far the most frequently used platform in digital surveillance, many others have been used as well. For example, Facebook "like" patterns correlate strongly with a wide range of health conditions and behaviors (43), and Instagram timelines have been used to identify adverse drug reactions (26).

Crowd-Sourced Data

Besides social media and search query data, large-scale, crowd-sourced digital surveillance systems such as Flu Near You (96) and Influenzanet (58) represent major innovations in the digital sphere. These systems recruit users through online and traditional media to participate in repeated Web-based surveys, including detailed symptom reports, and report on observed disease distribution through online maps and newsletters (111). For example, Influenzanet was established in 2009 and includes 10 European countries (although it was built on the Dutch and Belgian platform, De Grote Griepmeting, launched in 2003–2004) (45). The standardized Web survey in Influenzanet collects detailed flu symptom data, which allows for multiple influenza-like illness case definitions used by different European health agencies. Flu Near You is a similar system in the United States (96), and Dengue na Web extended this model to monitor dengue in Salvador, Bahia, Brazil (111). In addition to providing prevalence estimates based on standardized case definitions, a major advantage of participatory surveillance systems is that they provide individual-level demographic and risk factor data, allowing investigators to define the variables of interest and ultimately address research questions of interest (14, 32). Over traditional surveillance systems, clear advantages of these crowd-sourced approaches include lower cost and greater flexibility, as they allow integration of additional questions and varying case definitions (45).

Hybrid Digital/Traditional Public Health Data

Given the complicated biases present in Internet and social media data (covered below), digital data are often best used to supplement rather than replace nondigital public health surveillance

data sources. Indeed, the greatest potential use for digital data has been described as the development of hybrid digital surveillance systems (60, 95, 110). Outside of the FluSight competition (3), a few authors have attempted to formally integrate Web-based surveillance with more traditional sources. Santillana et al. (89) combined Google, Twitter, and Flu Near You data with influenza-like illness percentages from a private health care insurer, showing improvement over Google Flu Trends in terms of root-mean-square error and forecasting horizon (four weeks) but did not report on more public health–relevant performance metrics (e.g., start week, peak week, peak percentage, and duration). Incorporation of humidity (a known weather-related determinant of influenza transmission) data with Google Flu Trends data into mechanistic models has shown promising results (55, 93, 94). Other sources of big data can complement as well. Cloud-based electronic health record data are increasingly available in near real time; a study combining electronic health record influenza-like illness estimates with Google search data reduced the root-mean-square error of predicting the flu intensity four weeks in advance of the CDC relative to HealthMap (114). Another example is the use of air travel volume data combined with Twitter geolocation data to predict the spread of Chikungunya virus (80).

VALIDITY AND BIASES OF SOCIAL MEDIA AND INTERNET DATA

Two key distinctions for digital surveillance with regard to bias are that (*a*) digital data are not owned by the public, and (*b*) except for participatory surveillance, the data are capturing public awareness rather than actual occurrence of disease. An early example of these issues is the inaccurate predictions by Google Flu Trends, which sparked criticism from researchers and led to the removal of the publicly available Google Flu Trends site in 2015 (38). Google Flu Trends was unable to detect the influenza A/H1N1 pandemic in 2009 and greatly overestimated the peak intensity of the 2012–2013 season (71). Proposed explanations for divergent predictions include changes in search behaviors over time; because the 2009 H1N1 pandemic occurred in the spring/summer, the terms used in influenza-related Internet searches may have deviated from the terms more commonly used in the winter (25). For the 2012–2013 season, media coverage is believed to have caused exaggerated public awareness of influenza, leading to a higher volume of searches and inflating predictions (13). The inaccurate predictions by Google Flu Trends in multiple seasons raised awareness of potential biases inherent in digital surveillance (60), including changes to search algorithms (stability), nonindependence of data sources (posting and searching can be influenced by others), confounding (of search terms), representativeness (access to Internet), and lack of case validation (no clinical ascertainment). We highlight each of these potential sources of bias and present some ideas for mitigating them.

Stability of Digital Data Sources

Search engine companies make frequent changes to query algorithms without notifying the public. An almost automatic result of these changes is that predictive models degrade in accuracy over time. Therefore, relying on these sets of data over time can lead to a bias, referred to as "concept drift" (109). Forecasting is especially challenging: Across multiple infectious diseases, prediction accuracy degrades rapidly with forecast horizon, typically extending reasonably to only 2–4 weeks (104). Models with excellent nowcasting skill can have zero forecasting value (77) because digital data may indicate only public awareness and coincide with outbreak peaks (80). However, even with nowcasting, degradation is cumulative. Priedhorsky et al. (77) found that an influenza model trained on Wikipedia traffic showed "staleness" after four months, meaning models were no longer effective and needed to be retrained. This problem was largely unaided by adding more data over a

Hybrid digital surveillance: integration of digital surveillance data with traditional public health surveillance data or multiple sources of digital along with traditional public health surveillance data to monitor disease trends

Cloud-based: describes data that are stored, processed, or analyzed on demand via hosted remote servers made available through the Internet

longer period. Such staleness does not just cause noisy predictions but can create erroneous spikes (77). Furthermore, a website can be deactivated at any time or change owners and, therefore, the persistence of data is not guaranteed. The concerns related to disappearing or changing data on websites, and algorithm changes, make the process of replication of digital surveillance highly tenuous.

Nonindependent Digital Data

Internet data are fundamentally nonindependent and contain self-perpetuating feedback loops of multiple kinds. When predictors are trained on such data, their sensitivity/specificity and predictive values can be expected to change over time, leading to erroneous conclusions. Recommendation systems generate suggestions based in part on the popularity of a particular search term, creating dependence between observations. In social media, trending topics are self-perpetuating and can be manipulated (98). Media attention on an epidemic causes spikes in topic frequency not related to disease rates. For instance, the CDC conducted a press conference in April 2013 regarding the H7N9 "bird flu" pandemic, which led to a spike in flu-related tweets not corresponding to incidence (9).

Digital Search Confounding

Confounding of search terms is also present in correlation-based feature selection, particularly when relying on simple terms without sophisticated filtering. For example, terms related to basketball correlate strongly with influenza due to seasonal overlap (42, 77), and Google trends for the word "cholera" revealed a cholera "epidemic" in the United States in 2007 related to Oprah Winfrey selecting *Love in the Time of Cholera* for her book club (95). Prediction accuracy has also been observed to drop off precipitously during holidays (116). These examples illustrate the necessity of semantic filtering, which was performed by hand in the initial Google Flu Trends algorithm (42), but it is possible to automate the process (77). Related to confounding, there may be social desirability biases related to the use of certain search terms. If individuals are concerned that their searches are being tracked or that they can be identified through their computer searches, they may be less likely to use terms that relate to certain diseases and conditions that may be stigmatized. This bias may reduce the accuracy of digital data surveillance for certain conditions.

Digital Representativeness

Representativeness is a key characteristic of an ideal surveillance system. Well-known sampling biases exist in Internet and social media use. Heavily discussed is the "digital divide": the fact that socioeconomic inequality exists in Internet access and usage (31) and that Internet access is substantially less dense in developing countries (44). While roughly 22% of the US adult population uses Twitter, individuals of higher socioeconomic status, ages 30–44, and those living in urban areas are overrepresented (74), reflecting a notable underrepresentation of groups at highest risk for infectious disease morbidity and mortality. Participatory surveillance systems, though they may mitigate the confounding and nonindependence discussed above, suffer similar representativeness issues, with individuals needing to sign up and manually provide their information. For example, Influenzanet shows underrepresentation of males and the youngest and oldest age groups, as well as a higher influenza vaccination rate among older participants compared with the same age group in the target population (45). As expected, the population participating in Influenzanet may be healthier and more advantaged, showing lower prevalence of asthma and diabetes and higher income and education, which are predictors of lower influenza risk (58). To our knowledge, no study

has specifically aimed to explore the effect of underrepresentation of those most susceptible to flu, the very young and old, on population-level flu estimates based on Internet data. One promise of big data is the potential for highly granular predictions in terms of geography; hence, some Twitter-based studies rely on geolocation data. However, geolocation of tweets is turned off by default, with roughly 1–2% of users turning it on, likely introducing additional sampling biases (61). In particular, geolocation users have measurable differences in language use and are more likely to be men over 40 (73). To our knowledge, no investigation has examined the relationship between the use of geolocation tags and disease, but it might reasonably correlate negatively with influenza risk (particularly by age) and with HIV risk behaviors, many of which are stigmatized or illegal (97).

Digital Validation

Much of traditional public health surveillance relies on validation through clinical case ascertainment, which is not possible with digital surveillance (perhaps except in cases where participatory surveillance is required). To properly evaluate the performance of digital surveillance, appropriate evaluation metrics are needed. A suitable comparator data source for the surveillance system must be chosen because prediction will also replicate any biases present in the groundtruth data. Using influenza as an example, diagnosed case reports (which are used in most studies of the flu) underestimate prevalence, especially among persons with lower access to regular health care. CDC influenza-like illness percentages are the most common ground-truth data used for digital flu surveillance; these are based on weekly reports from approximately 2,200 outpatient providers throughout the United States reflecting the total number of patients meeting the CDC's syndromic case definition (17). Because some providers may exhibit delays in case reporting, influenza-like illness reports are often revised weeks or months after their initial release, which negatively impacts forecasting ability (79). Also, influenza-like illness reports reflect incidence in the general population, which may not accurately reflect the impact of the flu on those who are most physiologically vulnerable, such as the very young and old. To better capture such impacts, digital surveillance efforts should also consider incorporating other routine indicators, such as age-stratified influenza-like illness and influenza-related hospitalization and mortality (5, 55).

Possible Solutions to Biases

Here, we discuss several potential solutions to the biases described above. Addressing both concept drift and nonindependence has been explored previously. Concept drift and resulting model staleness can likely be addressed by including a plan to dynamically retrain models to account for always-changing online behavior and disease dynamics. For example, Santillana et al. (90) were able to dramatically improve Google Flu Trends's accuracy by automatically updating the model when weekly CDC estimates were released, highlighting that language used in searches changes over time and so should independent contributions of search terms used to estimate flu trends. Future work could examine precisely how frequently models need to be recalibrated.

Nonindependence of data is a more challenging problem, since self-perpetuating public awareness potentially impacts all forms of health-related online behavior, but can be addressed at least in part by improving text classification accuracy. For example, Broniatowski et al. (9) used a multistage filtering approach to distinguish media-related chatter from tweets that are true markers of influenza incidence that doubled the accuracy in predicting the weekly direction of change in incidence (i.e., up or down).

Solutions to sampling bias and nonrepresentativeness have been less well explored. A classical approach is to use weighting to adjust samples to be population representative, through

poststratification weights, inverse probability weights, or raking (117). These are standard practices in national surveys with known sampling probabilities but may also be reasonable to apply in a convenience sample such as a selection of Google search queries or tweets. Wang et al. (106) used data from a series of daily voter intention polls conducted on the Xbox gaming platform, a sample not unlike social media or search query data, to predict successfully the 2012 US presidential election results using poststratification weights based on simple demographics such as age, sex, race, and political party affiliation. Although these variables are not often available from Google or Twitter data, they may be straightforward to learn from the available data; reasonably accurate models exist to predict demographics on Twitter from names alone (112). Though population rates of Internet usage have been increasing for all groups, it is worthwhile to note that sampling bias is not necessarily mitigated by high usage rates overall. Ironically, as innovations spread through social networks, holdouts can represent an increasingly more unusual group (82).

ETHICAL CONSIDERATIONS FOR DIGITAL DATA SURVEILLANCE

The primary ethical challenge of public health is appropriately balancing risks and harms to individuals, while protecting and promoting population health (37, 52). This challenge remains for digital surveillance as the primary ethical issue (28, 51, 84), but the nonhealth purpose of the data raises ethical concerns in a different light. To frame our discussion of ethics of digital surveillance, we focus on the following five principles: beneficence, nonmaleficence, respect for autonomy, equity, and efficiency. These principles have been previously used to frame ethical considerations in public health (64) and to examine the ethics of public health surveillance more generally (57). More broadly, consideration of these five principles can help public health researchers with decision making for public health concerns and the ethical concerns related to these decisions (e.g., harm minimization or precautionary principle).

Beneficence

The concept of beneficence is the principle that public health surveillance should improve the health of the target population (64). While digital surveillance has the potential to improve infectious disease surveillance systems, part of this principle is clearly defining the target population of the surveillance system. As such, careful consideration should be directed at identifying who is and who is not captured by digital surveillance systems. Identification of blind spots is necessary to ensure that public health surveillance works toward improving the health of the target population it purports. Digital surveillance is promising because members of the target population who do not come into contact with more traditional medical-based surveillance systems can be captured. In addition, these systems can capture events that may be missed by traditional surveillance. One example is foodborne illness, where only a fraction of cases are captured by traditional surveillance (69). Under this view, the additional coverage of digital surveillance, while still a biased sample, can improve the overall coverage of the target population relative to traditional epidemiologic surveillance approaches alone.

Beneficence also charges us with considering how health would be improved under a digital surveillance system. Merely monitoring a health outcome does not necessarily improve population health. Improving population health depends on effective communication and interventions. Previous work has made the argument that digital surveillance allows for earlier outbreak response (19, 80, 85, 86, 101), since traditional surveillance systems offer insufficient lead time (2–4 weeks in advance of the outbreak) to plan meaningful responses, such as adjustments to hospital surge capacity and vaccine manufacturing (104). However, few digital surveillance systems have actually

described interventions that resulted from the earlier detection via digital surveillance and their effectiveness.

Nonmaleficence

The concept of nonmaleficence is defined by actions that reduce the potential harms and burdens of collecting data and promote the benefits of doing so, to the greatest degree possible. Several threats to nonmaleficence include the use of nonhealth data and stigmatization of risk factors, violation of privacy, and mistrust in public health information and intentions. Digital surveillance relies primarily on systems that were not built for the explicit collection of health-related data, but a substantial amount of nonhealth data are also often collected during queries. If these nonhealth data are subsequently labeled as "risk factors," this approach may attach stigma to certain behaviors or groups that are a proxy for the true underlying risk factors. There are many examples of stigma manifestation that occurred before the Internet—including the HIV epidemic—with groups becoming stigmatized on the basis of sexual orientation, country of origin, and race/ethnicity (41, 88). Collection of nonhealth data for digital surveillance and the speed of nowcasting and social media set up similar misattribution of risk and stigma (78). What is communicated from digital surveillance and how it is presented to the public should be considered thoughtfully and cautiously.

A core tenet of public health is building and maintaining public trust. A major barrier to public trust is false detections and missed outbreaks by surveillance systems. False detections can erode public trust in these surveillance systems, lead to misuse of limited resources, and result in poor risk communication to the public (6, 33). False detections and missed outbreaks can occur for any surveillance system, but the use of nonhealth data, black-box machine learning algorithms, and rapidly declining performance open up more possibilities for false detections to occur. Furthermore, mistrust may result from social media sites themselves through data breaches or sites using public health outbreak detection to market products to users. Last, users may knowingly share information publicly but may not expect its continual collection and analysis. While aggregate data may alleviate some concerns over violations of privacy, the same cannot be said of individual-specific Internet data. Therefore, digital surveillance data may warrant the same privacy protections as do data from more traditional surveillance systems. Public and private data use must be implemented cautiously, and the possibility for misuse should be examined.

Autonomy

The concept of respect for autonomy involves recognizing the right to self-determination of individuals and minimizing subsequent violations. A fundamental concern is informed consent within surveillance systems (57), which is also often cited as a major ethical concern of digital surveillance (50, 95). Informed consent is forgone in surveillance systems to improve the accuracy, with various justifications making the lack of informed consent more acceptable (57). In particular, the anonymization or aggregation of data helps to justify the lack of informed consent for medical institution data (67). For digital surveillance, monitoring of aggregate search trends or page views is similarly defensible. Issues of informed consent for digital surveillance on individual-level data lead to unique complications. Three distinguishing features of digital epidemiology with regard to informed consent are that (*a*) data are not sourced from formal medical institutions, (*b*) data come from proprietary online platforms, and (*c*) data are not limited to health but often include personal attributes (67).

In the clinical setting, there is often a legal mandate to report certain diseases, medical professionals have their own long-standing code of ethics, and data are collected explicitly for health purposes. Health data from medical institutions are covered under the Health Insurance

Portability and Accountability Act. While a patient at a medical institution may reasonably believe that their medical data will be used by health departments to monitor public health, the same is not true for social media. Social media users may not expect their data to be used toward monitoring public health trends because there is no legal mandate for public health reporting of social media data, there are no standard ethics for social media creators, and data are primarily nonhealth related. Social media sites and other Internet platforms have recently come under scrutiny regarding privacy concerns (24). While the user agreements (the Terms and Conditions) cover consent from a legal aspect, the "informed" part has been a concern owing to the complexity of the language used and the volume of conditions. While efforts such as the European Union General Data Protection Regulation (GDPR) seeks to give Internet users greater autonomy over their personal data collected by social media companies and to reduce the complexity of user agreements (35), we remain in a tumultuous time regarding data privacy and consent to online data collection. Because digital data consist of largely nonhealth data, it becomes less clear whether informed consent can be broken for individual-level health data.

As public health surveillance continues to rely on these methods, we need to consider and openly discuss the lack of informed consent and whether there is sufficient justification to warrant digital surveillance without it. Whether the lack of informed consent for digital surveillance is justified depends on the specific scenario and the corresponding risk to public health. With GDPR giving users more control over their data, informed consent could be obtained by directly requesting individuals to share their social media data with public health authorities (84). This approach has the additional benefit of avoiding the issue of social media companies retracting access to data. However, this approach has its own ethical concerns (40).

Equity

Individuals in the target population should have equal opportunity to receive a given public health intervention and a just distribution of benefits. As stated for beneficence, digital surveillance offers the opportunity to identify health problems in individuals who are not in contact with the medical system and thereby may increase opportunities for enhancing equity in who is included in surveillance data. However, as described above in biases, segments of the population still do not have access to the Internet or digital technology, leaving open the potential for loss of equity in digital surveillance. Considering the limitations in representativeness [and, potentially, heterogeneous predictive power (53)] of most Internet data sources, it is unlikely that surveillance data and associated interventions will be justly distributed without taking explicit account of these biases.

Efficiency

The concept of efficiency is related to the cost–benefit analysis of a surveillance system. For digital surveillance to produce benefits, it needs to overcome the biases previously discussed and provide improved tracking of health issues. For these systems to be cost-efficient, they require automated programs to manage and analyze the data. These automated systems may require substantial start-up funds and require regular maintenance to prevent algorithms from becoming stale. Because the data collected by these private companies are proprietary and there is no legal mandate for data provisions to public health, underlying algorithms by the companies require continuous updating, and access could be discontinued at any time without warning (50). If access is revoked, the resources spent on creating the digital surveillance system become poorly spent. Furthermore, proprietary algorithms to detect outbreaks can also be revoked at any time and may not be reproducible (25, 59, 84, 102). An alternative that avoids platforms revoking access is to build legal mandates to allow for access or allow users control over their data and to ask them

for permission to share their data with public health professionals (84). Legal mandates for data provision may violate the autonomy of these companies. Some digital surveillance systems use a variation of the latter approach, like Flu Near You and Influenzanet, where users provide their data directly through Web-based surveys.

In summary, the use of Internet-based data collection offers new opportunities but raises several new ethical concerns compared with traditional public health surveillance. These issues should continue to be addressed and require communication with the public. As a starting point, Mittelstadt et al. (67) have provided conditions to consider and several case studies related to the ethics of digital surveillance.

DISCUSSION

Digital public health surveillance affords the opportunity to revolutionize existing public health surveillance infrastructure. Although some public health officials report monitoring digital data sources such as Google Flu Trends for contextual information (13), our review indicates that public health, in any official capacity, has yet to embrace and build on existing opportunities that have arisen from new digital data sources. While a variety of conditions examined in public health surveillance have been explored to some degree using digital data sources, the main focus has been on influenza. Moreover, road maps describing how public/private partnerships could advance and work together toward creating sustainable, efficient, and ethical digital surveillance systems have yet to be fully developed. In addition, clear standards by which digital public health surveillance can be compared with traditional surveillance systems have yet to be established, making it difficult to assess whether these sources truly amplify the benefits of traditional public health for disease tracking and prevention. While we cover applied uses related to search queries, social media posts, and crowdsourcing, another potentially promising area that has yet to be explored is the use of data from existing digital health collection platforms based on smartphones and wearable devices, such as Apple HealthKit. These sources could offer highly precise information related to both infectious and noninfectious conditions and risk factors at a broad scale, but with many of the biases and representation concerns mentioned above.

Ultimately, digital surveillance systems will need to be developed in ways that avoid the numerous potential pitfalls associated with biases and ethical considerations described in this review. As with all surveillance systems, experts will need to determine whether digital surveillance is necessary and ethically justifiable. In future applications, digital surveillance is likely to have the largest public health impact when integrated with traditional surveillance systems, such as traditional laboratory data, case reports, and electronic health records (95). While digital surveillance offers the ability to build novel surveillance systems where no existing surveillance system exists, appropriate and accurately measured training data sets are needed.

In the future, it will be important to identify the most beneficial ways to use digital data sources through hybrid or completely new independent systems. The prospect of a new system, driven solely by digital technology, seems unlikely at present, but the continued advancement of machine learning, and related technologies for managing and deriving meaning from large sets of data, may bring this idea closer to reality in the future. One challenge will be in the training of public health experts in computer science, big data, and machine learning to harness novel sources of digital data and support innovation in digital surveillance while reducing possible harms. In conclusion, we are on the precipice of an unprecedented opportunity to track, predict, and prevent global population disease burdens using digital data, and it is critical that public health institutions receive the training and resources needed to provide the right input and help to build new systems that move traditional public health surveillance into the future.

DISCLOSURE STATEMENT

ACKNOWLEDGMENTS

LITERATURE CITED

1. Althouse BM, Ng YY, Cummings DAT. 2011. Prediction of dengue incidence using search query surveillance. *PLOS Negl. Trop. Dis.* 5:e1258

2. Aphinyanaphongs Y, Lulejian A, Brown DP, Bonneau R, Krebs P. 2016. Text classification for automatic detection of e-cigarette use and use for smoking cessation from Twitter: a feasibility pilot. *Pac. Symp. Biocomput.* 21:480–91

3. Biggerstaff M, Alper D, Dredze M, Fox S, Fung IC-H, et al. 2016. Results from the Centers for Disease Control and Prevention's predict the 2013–2014 Influenza Season Challenge. *BMC Infect. Dis.* 16:357

4. Biggerstaff M, Johansson M, Alper D, Brooks LC, Chakraborty P, et al. 2018. Results from the second year of a collaborative effort to forecast influenza seasons in the United States. *Epidemics* 24:26–33

5. Biggerstaff M, Kniss K, Jernigan DB, Brammer L, Bresee J, et al. 2018. Systematic assessment of multiple routine and near real-time indicators to classify the severity of influenza seasons and pandemics in the United States, 2003–2004 through 2015–2016. *Am. J. Epidemiol.* 187:1040–50

6. Böl G-F. 2016. Risk communication in times of crisis: pitfalls and challenges in ensuring preparedness instead of hysterics. *EMBO Rep*. 17:1–9

7. Bragazzi NL, Mahroum N. 2019. Google Trends predicts present and future plague cases during the plague outbreak in Madagascar: infodemiological study. *JMIR Public Health Surveill.* 5:e13142

8. Brammer L, Blanton L, Epperson S, Mustaquim D, Bishop A, et al. 2011. Surveillance for influenza during the 2009 influenza A (H1N1) pandemic–United States, April 2009–March 2010. *Clin. Infect. Dis.* 52:S27–35

9. Broniatowski DA, Paul MJ, Dredze M. 2013. National and local influenza surveillance through Twitter: an analysis of the 2012–2013 influenza epidemic. *PLOS ONE* 8:e83672

10. Brownstein JS, Freifeld CC, Madoff LC. 2009. Digital disease detection—harnessing the Web for public health surveillance. *N. Engl. J. Med.* 360:2153–55, 2157

11. Brownstein JS, Freifeld CC, Reis BY, Mandl KD. 2008. Surveillance sans frontières: Internet-based emerging infectious disease intelligence and the HealthMap project. *PLOS Med*. 5:e151

12. Buczak AL, Baugher B, Moniz LJ, Bagley T, Babin SM, Guven E. 2018. Ensemble method for dengue prediction. *PLOS ONE* 13:e0189988

13. Butler D. 2013. When Google got flu wrong. *Nature* 494:155–56

14. Carlson SJ, Durrheim DN, Dalton CB. 2010. Flutracking provides a measure of field influenza vaccine effectiveness, Australia, 2007–2009. *Vaccine* 28:6809–10

15. Carneiro HA, Mylonakis E. 2009. Google trends: a web-based tool for real-time surveillance of disease outbreaks. *Clin. Infect. Dis.* 49:1557–64

16. CDC (Cent. Dis. Control Prev.). 2013. *CDC competition encourages use of social media to predict flu*. Press Release, Nov. 25. **https://www.cdc.gov/flu/news/predict-flu-challenge.htm**

17. CDC (Cent. Dis. Control Prev.). 2019. Overview of influenza surveillance in the United States. *Centers for Disease Control and Prevention, Influenza (Flu)*. **https://www.cdc.gov/flu/weekly/overview.htm**

18. CDC (Cent. Dis. Control Prev.) Epidemic Predict. Initiat. 2019. FluSight 2018–2019. Seasonal influence forecasting at the national and regional level. *Epidemic Prediction Initiative*. **https://predict.cdc.gov/post/5ba1504e5619f003acb7e18f**

19. Chan EH, Sahai V, Conrad C, Brownstein JS. 2011. Using web search query data to monitor dengue epidemics: a new model for neglected tropical disease surveillance. *PLOS Negl. Trop. Dis.* 5:e1206

20. Choi BCK. 2012. The past, present, and future of public health surveillance. *Scientifica* 2012:875253

21. Chretien J-P, George D, Shaman J, Chitale RA, McKenzie FE. 2014. Influenza forecasting in human populations: a scoping review. *PLOS ONE* 9:e94130

22. Christodoulou E, Ma J, Collins GS, Steyerberg EW, Verbakel JY, Van Calster B. 2019. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. *J. Clin. Epidemiol.* 110:12–22

23. Chunara R, Aman S, Smolinski M, Brownstein JS. 2013. Flu Near You: an online self-reported influenza surveillance system in the USA. *Online J. Public Health Inform.* 5:e133

24. Confessore N. 2018. Cambridge Analytica and Facebook: the scandal and the fallout so far. *The New York Times*, April 4. **https://www.nytimes.com/2018/04/04/us/politics/cambridge-analytica-scandal-fallout.html**

25. Cook S, Conrad C, Fowlkes AL, Mohebbi MH. 2011. Assessing Google flu trends performance in the United States during the 2009 influenza virus A (H1N1) pandemic. *PLOS ONE* 6:e23610

26. Correia RB, Li L, Rocha LM. 2016. Monitoring potential drug interactions and reactions via network analysis of Instagram user timelines. *Pac. Symp. Biocomput.* 21:492–503

27. Del Valle SY, McMahon BH, Asher J, Hatchett R, Lega JC, et al. 2018. Summary results of the 2014–2015 DARPA Chikungunya challenge. *BMC Infect. Dis.* 18:245

28. Denecke K. 2017. An ethical assessment model for digital disease detection technologies. *Life Sci. Soc. Policy* 13:16

29. Desai R, Hall AJ, Lopman BA, Shimshoni Y, Rennick M, et al. 2012. Norovirus disease surveillance using Google Internet query share data. *Clin. Infect. Dis.* 55:e75–78

30. Dietterich TG. 2000. Ensemble methods in machine learning. In *Multiple Classifier Systems*, ed. J Kittler, F Roli, pp. 1–15. Berlin/Heidelberg: Springer

31. DiMaggio P, Hargittai E, Neuman WR, Robinson JP. 2001. Social implications of the Internet. *Annu. Rev. Sociol.* 27:307–36

32. Eames KTD, Brooks-Pollock E, Paolotti D, Perosa M, Gioannini C, Edmunds WJ. 2012. Rapid assessment of influenza vaccine effectiveness: analysis of an Internet-based cohort. *Epidemiol. Infect.* 140:1309–15

33. Eckmanns T, Füller H, Roberts SL. 2019. Digital epidemiology and global health security; an interdisciplinary conversation. *Life Sci. Soc. Policy* 15:2

34. Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, et al. 2017. Dermatologist-level classification of skin cancer with deep neural networks. *Nature* 542:115–18

35. Eur. Comm. 2018. *EU data protection rules. European Commission, Data Protection*. **https://ec.europa.eu/commission/priorities/justice-and-fundamental-rights/data-protection/2018-reform-eu-data-protection-rules/eu-data-protection-rules_en**

36. Eysenbach G. 2006. Infodemiology: tracking flu-related searches on the web for syndromic surveillance. *AMIA Annu. Symp. Proc.* 2006:244–48

37. Fairchild AL, Haghdoost AA, Bayer R, Selgelid MJ, Dawson A, et al. 2017. Ethics of public health surveillance: new guidelines. *Lancet Public Health* 2:e348–49

38. Flu Trends Team. 2015. The next chapter for flu trends. *Google AI Blog*, Aug. 20. **https://ai.googleblog.com/2015/08/the-next-chapter-for-flu-trends.html**

39. Fried D, Surdeanu M, Kobourov S, Hingle M, Bell D. 2014. Analyzing the language of food on social media. arXiv:1409.2195v2 [cs.CL]

40. Geneviève LD, Martani A, Wangmo T, Paolotti D, Koppeschaar C, et al. 2019. Participatory disease surveillance systems: ethical framework. *J. Med. Internet Res.* 21:e12273

41. Gilman SL. 1987. AIDS and syphilis: the iconography of disease. *October* 43:87–107

42. Ginsberg J, Mohebbi MH, Patel RS, Brammer L, Smolinski MS, Brilliant L. 2009. Detecting influenza epidemics using search engine query data. *Nature* 457:1012–14

43. Gittelman S, Lange V, Gotway Crawford CA, Okoro CA, Lieb E, et al. 2015. A new source of data for public health surveillance: Facebook likes. *J. Med. Internet Res.* 17:e98

44. Graham M, Hale S, Stephens M. 2012. Featured graphic: Digital divide: the geography of Internet access. *Environ. Plan. A* 44:1009–10

45. Guerrisi C, Turbelin C, Blanchon T, Hanslik T, Bonmarin I, et al. 2016. Participatory syndromic surveillance of influenza in Europe. *J. Infect. Dis.* 214:S386–92

46. Gulshan V, Peng L, Coram M, Stumpe MC, Wu D, et al. 2016. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA* 316:2402–10

47. Harris JK, Hawkins JB, Nguyen L, Nsoesie EO, Tuli G, et al. 2017. Using Twitter to identify and respond to food poisoning: the Food Safety STL Project. *J. Public Health Manag. Pract.* 23:577–80

48. Harris JK, Mansour R, Choucair B, Olson J, Nissen C, et al. 2014. Health department use of social media to identify foodborne illness—Chicago, Illinois, 2013–2014. *MMWR* 63:681–85

49. Harrison C, Jorder M, Stern H, Stavinsky F, Reddy V, et al. 2014. Using online reviews by restaurant patrons to identify unreported cases of foodborne illness—New York City, 2012–2013. *MMWR* 63:441–45

50. Heitmueller A, Henderson S, Warburton W, Elmagarmid A, Pentland AS, Darzi A. 2014. Developing public policy to advance the use of big data in health care. *Health Aff.* 33:1523–30

51. Höhle M. 2017. A statistician's perspective on digital epidemiology. *Life Sci. Soc. Policy* 13:17

52. Holland S. 2015. *Public Health Ethics*. Cambridge, UK: Polity

53. Huang D-C, Wang J-F, Huang J-X, Sui DZ, Zhang H-Y, et al. 2016. Towards identifying and reducing the bias of disease information extracted from search engine data. *PLOS Comput. Biol.* 12:e1004876

54. Johnson HA, Wagner MM, Hogan WR, Chapman W, Olszewski RT, et al. 2004. Analysis of Web access logs for surveillance of influenza. *Stud. Health Technol. Inform.* 107:1202–6

55. Kandula S, Pei S, Shaman J. 2019. Improved forecasts of influenza-associated hospitalization rates with Google Search Trends. *J. R. Soc. Interface* 16:20190080

56. Kelly H, Grant K. 2009. Interim analysis of pandemic influenza (H1N1) 2009 in Australia: surveillance trends, age of infection and effectiveness of seasonal vaccination. *Euro Surveill.* 14. **https://doi.org/10.2807/ese.14.31.19288-en**

57. Klingler C, Silva DS, Schuermann C, Reis AA, Saxena A, Strech D. 2017. Ethical issues in public health surveillance: a systematic qualitative review. *BMC Public Health* 17:295

58. Koppeschaar CE, Colizza V, Guerrisi C, Turbelin C, Duggan J, et al. 2017. Influenzanet: citizens among 10 countries collaborating to monitor influenza in Europe. *JMIR Public Health Surveill.* 3:e66

59. Kostkova P. 2018. Disease surveillance data sharing for public health: the next ethical frontiers. *Life Sci. Soc. Policy* 14:16

60. Lazer D, Kennedy R, King G, Vespignani A. 2014. The parable of Google Flu: traps in big data analysis. *Science* 343:1203–5

61. Leetaru K, Wang S, Cao G, Padmanabhan A, Shook E. 2013. Mapping the global Twitter heartbeat: the geography of Twitter. *First Monday* 18(5). **https://doi.org/10.5210/fm.v18i5.4366**

62. Li Z, Liu T, Zhu G, Lin H, Zhang Y, et al. 2017. Dengue Baidu Search Index data can improve the prediction of local dengue epidemic: a case study in Guangzhou, China. *PLOS Negl. Trop. Dis.* 11:e0005354

63. Madoff LC, Woodall JP. 2005. The Internet and the global monitoring of emerging diseases: lessons from the first 10 years of ProMED-mail. *Arch. Med. Res.* 36:724–30

64. Marckmann G, Schmidt H, Sofaer N, Strech D. 2015. Putting public health ethics into practice: a systematic framework. *Front. Public Health* 3:23

65. McGowan CJ, Biggerstaff M, Johansson M, Apfeldorf KM, Ben-Nun M, et al. 2019. Collaborative efforts to forecast seasonal influenza in the United States, 2015–2016. *Sci. Rep.* 9:683

66. Mignan A, Broccardo M. 2019. One neuron is more informative than a deep neural network for aftershock pattern forecasting. *Nature* 574:E1–3

67. Mittelstadt B, Benzler J, Engelmann L, Prainsack B, Vayena E. 2018. Is there a duty to participate in digital epidemiology? *Life Sci. Soc. Policy* 14:9

68. Mooney SJ, Pejaver V. 2018. Big data in public health: terminology, machine learning, and privacy. *Annu. Rev. Public Health* 39:95–112

69. Nsoesie EO, Buckeridge DL, Brownstein JS. 2014. Guess who's not coming to dinner? Evaluating online restaurant reservations for disease surveillance. *J. Med. Internet Res.* 16:e22

70. Nsoesie EO, Gordon SA, Brownstein JS. 2014. Online reports of foodborne illness capture foods implicated in official foodborne outbreak reports. *Prev. Med.* 67:264–69

71. Olson DR, Konty KJ, Paladini M, Viboud C, Simonsen L. 2013. Reassessing Google Flu Trends data for detection of seasonal and pandemic influenza: a comparative epidemiological study at three geographic scales. *PLOS Comput. Biol.* 9:e1003256

72. Park H-A, Jung H, On J, Park SK, Kang H. 2018. Digital epidemiology: use of digital data collected for non-epidemiological purposes in epidemiological studies. *Healthc. Inform. Res.* 24:253–62

73. Pavalanathan U, Eisenstein J. 2015. Confounds and consequences in geotagged Twitter data. arXiv:1506.02275v2 [cs.CL]

74. Perrin A, Anderson M. 2019. *Share of U.S. adults using social media, including Facebook, is mostly unchanged since 2018*. Pew Res. Cent., FactTank, April 10. **https://www.pewresearch.org/fact-tank/2019/04/10/share-of-u-s-adults-using-social-media-including-facebook-is-mostly-unchanged-since-2018/**

75. Polgreen PM, Chen Y, Pennock DM, Nelson FD. 2008. Using Internet searches for influenza surveillance. *Clin. Infect. Dis.* 47:1443–48

76. Porta M, ed. 2008. *A Dictionary of Epidemiology*. New York: Oxford Univ. Press

77. Priedhorsky R, Osthus D, Daughton AR, Moran KR, Generous N, et al. 2017. Measuring global disease with Wikipedia. *CSCW Conf. Support. Coop. Work.* 2017:1812–34

78. Quinn P. 2018. Crisis communication in public health emergencies: the limits of 'legal control' and the risks for harmful outcomes in a digital age. *Life Sci. Soc. Policy* 14:4

79. Reich NG, Brooks LC, Fox SJ, Kandula S, McGowan CJ, et al. 2019. A collaborative multiyear, multimodel assessment of seasonal influenza forecasting in the United States. *PNAS* 116:3146–54

80. Rocklöv J, Tozan Y, Ramadona A, Sewe MO, Sudre B, et al. 2019. Using big data to monitor the introduction and spread of Chikungunya, Europe, 2017. *Emerg. Infect. Dis. J.* 25:1041

81. Rodríguez-Martínez M, Garzón-Alfonso CC. 2018. Twitter Health Surveillance (THS) system. *Proc. IEEE Int. Conf. Big Data* 2018:1647–54

82. Rogers EM. 2003. *Diffusion of Innovations*. New York: Simon and Schuster. 5th ed.

83. Sadilek A, Caty S, DiPrete L, Mansour R, Schenk T Jr., et al. 2018. Machine-learned epidemiology: real-time detection of foodborne illness at scale. *npj Digit. Med.* 1:36

84. Salathé M. 2018. Digital epidemiology: What is it, and where is it going? *Life Sci. Soc. Policy* 14:1

85. Salathé M, Bengtsson L, Bodnar TJ, Brewer DD, Brownstein JS, et al. 2012. Digital epidemiology. *PLOS Comput. Biol.* 8:e1002616

86. Salathé M, Freifeld CC, Mekaru SR, Tomasulo AF, Brownstein JS. 2013. Influenza A (H7N9) and the importance of digital epidemiology. *N. Engl. J. Med.* 369:401–4

87. Salathé M, Khandelwal S. 2011. Assessing vaccination sentiments with online social media: implications for infectious disease dynamics and control. *PLOS Comput. Biol.* 7:e1002199

88. Santana MA, Dancy BL. 2000. The stigma of being named "AIDS carriers" on Haitian-American women. *Health Care Women Int.* 21:161–71

89. Santillana M, Nguyen AT, Dredze M, Paul MJ, Nsoesie EO, Brownstein JS. 2015. Combining search, social media, and traditional data sources to improve influenza surveillance. *PLOS Comput. Biol.* 11:e1004513

90. Santillana M, Zhang DW, Althouse BM, Ayers JW. 2014. What can digital disease detection learn from (an external revision to) Google Flu Trends? *Am. J. Prev. Med.* 47:341–47

91. Scarpino SV, Dimitrov NB, Meyers LA. 2012. Optimizing provider recruitment for influenza surveillance networks. *PLOS Comput. Biol.* 8:e1002472

92. Shah MP, Lopman BA, Tate JE, Harris J, Esparza-Aguilar M, et al. 2018. Use of Internet search data to monitor rotavirus vaccine impact in the United States, United Kingdom, and Mexico. *J. Pediatr. Infect. Dis. Soc.* 7:56–63

93. Shaman J, Karspeck A. 2012. Forecasting seasonal outbreaks of influenza. *PNAS* 109:20425–30

94. Shaman J, Karspeck A, Yang W, Tamerius J, Lipsitch M. 2013. Real-time influenza forecasts during the 2012–2013 season. *Nat. Commun.* 4:2837

95. Simonsen L, Gog JR, Olson D, Viboud C. 2016. Infectious disease surveillance in the big data era: towards faster and locally relevant systems. *J. Infect. Dis.* 214:S380–85

96. Smolinski MS, Crawley AW, Baltrusaitis K, Chunara R, Olsen JM, et al. 2015. Flu Near You: crowd-sourced symptom reporting spanning 2 influenza seasons. *Am. J. Public Health* 105:2124–30

97. Stoové MA, Pedrana AE. 2014. Making the most of a brave new world: opportunities and considerations for using Twitter as a public health monitoring tool. *Prev. Med.* 63:109–11

98. Tufekci Z. 2014. Big questions for social media big data: representativeness, validity and other methodological pitfalls. arXiv:1403.7400 [cs.SI]

99. Uiters E, Devillé W, Foets M, Spreeuwenberg P, Groenewegen PP. 2009. Differences between immigrant and non-immigrant groups in the use of primary medical care; a systematic review. *BMC Health Serv. Res.* 9:76

100. van der Laan MJ, Polley EC, Hubbard AE. 2007. Super learner. *Stat. Appl. Genet. Mol. Biol.* 6:25

101. Vayena E, Salathé M, Madoff LC, Brownstein JS. 2015. Ethical challenges of big data in public health. *PLOS Comput. Biol.* 11:e1003904

102. Velasco E. 2018. Disease detection, epidemiology and outbreak response: the digital future of public health practice. *Life Sci. Soc. Policy* 14:7

103. Viboud C, Sun K, Gaffey R, Ajelli M, Fumanelli L, et al. 2018. The RAPIDD Ebola forecasting challenge: synthesis and lessons learnt. *Epidemics* 22:13–21

104. Viboud C, Vespignani A. 2019. The future of influenza forecasts. *PNAS* 116:2802–4

105. Wang J, Yang X, Cai H, Tan W, Jin C, Li L. 2016. Discrimination of breast cancer with microcalcifications on mammography by deep learning. *Sci. Rep.* 6:27327

106. Wang W, Rothschild D, Goel S, Gelman A. 2015. Forecasting elections with non-representative polls. *Int. J. Forecast.* 31:980–91

107. Wesolowski A, Eagle N, Tatem AJ, Smith DL, Noor AM, et al. 2012. Quantifying the impact of human mobility on malaria. *Science* 338:267–70

108. WHO (World Health Organ.). 2017. Public health surveillance. *World Health Organization*, *Health Topics*. **https://www.who.int/topics/public_health_surveillance/en/**

109. Widmer G, Kubat M. 1996. Learning in the presence of concept drift and hidden contexts. *Mach. Learn.* 23:69–101

110. Wilson N, Mason K, Tobias M, Peacey M, Huang QS, Baker M. 2009. Interpreting "Google Flu Trends" data for pandemic H1N1 influenza: the New Zealand experience. *Eurosurveillance* 14:19386

111. Wójcik OP, Brownstein JS, Chunara R, Johansson MA. 2014. Public health for the people: participatory infectious disease surveillance in the digital age. *Emerg. Themes Epidemiol.* 11:7

112. Wood-Doughty Z, Andrews N, Marvin R, Dredze M. 2018. Predicting Twitter user demographics from names alone. In *Proceedings of the Second Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media*, pp. 105–11. Stroudsburg, PA: Assoc. Comp. Linguist.

113. Yan SJ, Chughtai AA, Macintyre CR. 2017. Utility and potential of rapid epidemic intelligence from Internet-based sources. *Int. J. Infect. Dis.* 63:77–87

114. Yang S, Santillana M, Brownstein JS, Gray J, Richardson S, Kou SC. 2017. Using electronic health records and Internet search information for accurate influenza forecasting. *BMC Infect. Dis.* 17:332

115. Yang S, Santillana M, Kou SC. 2015. Accurate estimation of influenza epidemics using Google search data via ARGO. *PNAS* 112:14473–78

116. Young SD, Torrone EA, Urata J, Aral SO. 2018. Using search engine data as a tool to predict syphilis. *Epidemiology* 29:574–78

117. Zhang L-C. 2000. Post-stratification and calibration—a synthesis. *Am. Stat.* 54:178–84

118. Zhang Q, Perra N, Perrotta D, Tizzoni M, Paolotti D, Vespignani A. 2017. Forecasting seasonal influenza fusing digital indicators and a mechanistic disease model. In *Proceedings of the 26th International Conference on World Wide Web*, pp. 311–19. New York: ACM