Review

The Application of Internet-Based Sources for Public Health Surveillance (Infoveillance): Systematic Review

Joana M Barros[1,2], MSc; Jim Duggan[2], PhD; Dietrich Rebholz-Schuhmann[3], PhD

[1]Insight Centre for Data Analytics, National University of Ireland Galway, Galway, Ireland

[2]School of Computer Science, National University of Ireland Galway, Galway, Ireland

[3]ZB MED - Information Centre for Life Sciences, University Cologne, Cologne, Germany

Corresponding Author:
Joana M Barros, MSc
Insight Centre for Data Analytics
National University of Ireland Galway
The DERI Building
IDA Business Park, Lower Dangan
Galway, H91 AEX4
Ireland
Phone: 353 0838327458
Email: joana.barros@insight-centre.org

Abstract

Background: Public health surveillance is based on the continuous and systematic collection, analysis, and interpretation of data. This informs the development of early warning systems to monitor epidemics and documents the impact of intervention measures. The introduction of digital data sources, and specifically sources available on the internet, has impacted the field of public health surveillance. New opportunities enabled by the underlying availability and scale of internet-based sources (IBSs) have paved the way for novel approaches for disease surveillance, exploration of health communities, and the study of epidemic dynamics. This field and approach is also known as infodemiology or infoveillance.

Objective: This review aimed to assess research findings regarding the application of IBSs for public health surveillance (infodemiology or infoveillance). To achieve this, we have presented a comprehensive systematic literature review with a focus on these sources and their limitations, the diseases targeted, and commonly applied methods.

Methods: A systematic literature review was conducted targeting publications between 2012 and 2018 that leveraged IBSs for public health surveillance, outbreak forecasting, disease characterization, diagnosis prediction, content analysis, and health-topic identification. The search results were filtered according to previously defined inclusion and exclusion criteria.

Results: Spanning a total of 162 publications, we determined infectious diseases to be the preferred case study (108/162, 66.7%). Of the eight categories of IBSs (search queries, social media, news, discussion forums, websites, web encyclopedia, and online obituaries), search queries and social media were applied in 95.1% (154/162) of the reviewed publications. We also identified limitations in representativeness and biased user age groups, as well as high susceptibility to media events by search queries, social media, and web encyclopedias.

Conclusions: IBSs are a valuable proxy to study illnesses affecting the general population; however, it is important to characterize which diseases are best suited for the available sources; the literature shows that the level of engagement among online platforms can be a potential indicator. There is a necessity to understand the population's online behavior; in addition, the exploration of health information dissemination and its content is significantly unexplored. With this information, we can understand how the population communicates about illnesses online and, in the process, benefit public health.

(J Med Internet Res 2020;22(3):e13680) doi: 10.2196/13680

KEYWORDS

medical informatics; public health informatics; public health; infectious diseases; chronic diseases; infodemiology; infoveillance

Introduction

Background

Public health is "the art and science of preventing disease, prolonging life and promoting health through the organized efforts of society" [1]. As a research and political field, it is focused on improving the quality of life of the population by identifying, suggesting, and applying prevention measures (eg, through the promotion of healthy behaviors) and health-related treatments [2]. Monitoring health is one important contribution to public health measures and involves the systematic collection, analysis, and interpretation of large amounts of health-related data. The key aim of public health surveillance is to design and guide interventions; in particular, (1) it serves as an early warning system for health emergencies (*epidemics*, ie, acute events), (2) it documents public health interventions and tracks their progress (ie, *monitoring health*), and (3) it monitors and clarifies the epidemiology of health problems, enabling the prioritization of information necessary for the formulation of health policy (ie, targeting chronic events) [3].

In the past, surveillance has been based on reports from health care workers constituting an active surveillance system when consistent and standardized reporting is in place [3,4]. However, this architecture is costly to maintain and involves significant delays between the moment of data capture to the time point of the first diagnosis, thus hampering any rapid or even immediate detection of outbreaks [5]. Instead of attempting to gather surveillance data from a network of health facilities and laboratories, health entities can employ a passive surveillance system in which hospitals, clinics, or other similar sources submit their respective health reports. This system provides an inexpensive way to monitor the community's health; however, data quality is an issue owing to nonuniform reporting standards, and timeliness remains difficult to achieve [4]. To further complement these systems, syndromic surveillance was created to deal with the timeliness issue by using clinical (eg, emergency department admissions) and nonclinical sources (eg, over-the-counter drug sales), which are available before a diagnosis is confirmed [4]. This type of surveillance is based on the assumption that an outbreak would manifest itself as an anomaly in normal behavior [5]. In line with syndromic surveillance and with the growth of the internet, new opportunities for the detection of health-related information have arisen, with the potential to capture the patient's input directly from the source. This leads to the ambitious endeavor of being able to monitor the health of a significant portion of the population at any point in time and at any geographical location, with the ultimate goal of monitoring public health.

The abovementioned technological advancements have enabled unofficial informal sources to currently provide more than 60% of epidemic reports [6]. Data analytics based on these data sources can provide near real-time outbreak information in various formats (independently from official governmental output) and have been successfully tested for health-related purposes. Furthermore, these sources offer the unique advantage of providing firsthand evidence for occurrences of health-related events (eg, through social media channels) and real-time informal reports (eg, news), which can be immediately investigated. Any analysis can be focused only on continuous monitoring, or by contrast to the identification of specific events (ie, single disease focus). In the latter case, it can be targeted to identify isolated hints (eg, mentions of flu) or to determine significant changes in public reporting; it can be further extended to consider the location of the population at risk or to monitor the distribution or extension of an epidemic (eg, influenced by the travelling population). The potential of data analytics applied to public data for health-related developments is ever more far-reaching in our increasingly digitally equipped society; thus, these approaches have an important role in the improvement of timeliness and sensitivity (ie, rapidly and correctly identifying health mentions) in public health surveillance [7].

Internet-Based Sources for Public Health Surveillance

IBSs are characterized by providing unstructured information from multiple origins and have proven to detect the first evidence of an outbreak, which is particularly beneficial for locations with a limited capacity for public health surveillance. The use of these sources for public health is also known as infodemiology or infoveillance. With the evidence provided by these sources, health agents are capable of mobilizing rapid response, reducing morbidity and mortality [8,9]. Some examples of IBSs include search queries, web encyclopedias, microblogs, and other social media.

Infectious diseases became the initial case study for the application of IBSs for disease surveillance. These continue to be a major cause of death in low-income countries [10], with research initially focusing on dengue, and are responsible for recurrent threats in the rest of the world (eg, *swine flu* and *bird flu*). Furthermore, these diseases are continuously monitored by official sources through laboratory tests or sentinel systems over many years and such information now forms the ground-truth data used to validate the findings from IBSs [11].

ProMED-mail is one of the first applications of such sources. This system is currently used for communication, via email and reports, among the infectious disease community [12]. Other systems include aggregators such as Global Public Health Intelligence Network, BioCaster, and HealthMap. These initially targeted a variety of sources including emails, Really Simple Syndication feeds, and PDF documents to extract information referring to an increased number of clusters of infected people at a specified time, period, or location, which could indicate a threat. The aggregator systems still in operation also include additional sources such as social media [7,13]. Moving to other sources, influenza-like illnesses (ILIs) served as a prototypical case study owing to being seasonal, worldwide, and well-reported diseases and initiated the monitoring of web-based queries. In particular, one of the first studies utilized Google search volumes to estimate the percentage of ILI-related physician visits [14]. This source was further adapted to the surveillance of other diseases such as dengue [15], gastroenteritis, and chickenpox [16]. This initial success led Google to develop targeted tools for the monitoring of influenza (Google Flu Trends) and dengue (Google Dengue Trends) in 2008 and 2011, respectively, which were later discontinued. Research continued and aimed to identify the

most appropriate search terms to utilize as well as other search services (eg, Yahoo [17]) and other languages (eg, Vårdguiden [18]). Following search queries, microblogs, in particular, Twitter, showed to be another source of health information characterized by providing more descriptive information and potentially containing semistructured metadata (eg, location and gender) [19,20]. By filtering messages containing disease-related keywords, the frequency of disease mentions can be tracked and outbreaks can be identified as unusual spikes in the message frequency [21]. Similar to search queries, subsequent research focused on the identification of adequate keywords, as well as the identification of personal health messages, ie, containing a keyword relevant to the disease and describing a first-person infection case, among others [22,23]. With the use of more descriptive albeit semantically ambiguous data, the focus shifted to detecting true signals, ie, first-person occurrences of diseases. The application of IBSs continued to grow [24] in tandem with the addition of new sources such as Facebook [25], Instagram [26], and discussion forums [27]. Noncommunicable diseases (NCDs) are the cause of 71% of deaths globally, ranging from 37% in low-income countries to 88% in high-income countries, hence, internet-based surveillance focus has begun to also include NCDs [10]. In this case, emphasis was given to the online behavior of affected individuals, as well as to the content of the information present in the sources [28], with the goal to establish or improve health practices and support the dissemination of health information to address the needs of the population [29,30].

Owing to the unstructured nature and to the large volumes of data provided by these IBSs, tailoring of solutions, applications, and even tools for retrieving and filtering the content becomes vital for success. Subsequent automatic use of these methods then becomes the key step to monitor the internet sources continuously, and eventually to identify potential public health risks or, even better, risks to individual patients [31]. However, disease surveillance based on online sources must be used with caution. Automatic identification of disease events from web-based data streams has to cope with inherent biases, ie, false-positive events, introduced through geographic or cultural variability in language and reporting when compared with reliable traditional surveillance methods [32]. Furthermore, traditional epidemiological parameters (eg, attack rate) are often not available as a gold standard and thus limit the proper assessment of the applied methods [31].

Objectives

Our objective was to provide a discriminative assessment of the applications of IBSs for disease surveillance and their use as ground truth for future research. To achieve this, we have presented a comprehensive systematic literature review with a focus on IBSs and their limitations, the diseases targeted, and methods commonly applied for disease surveillance. Our research questions (RQs) were as follows:

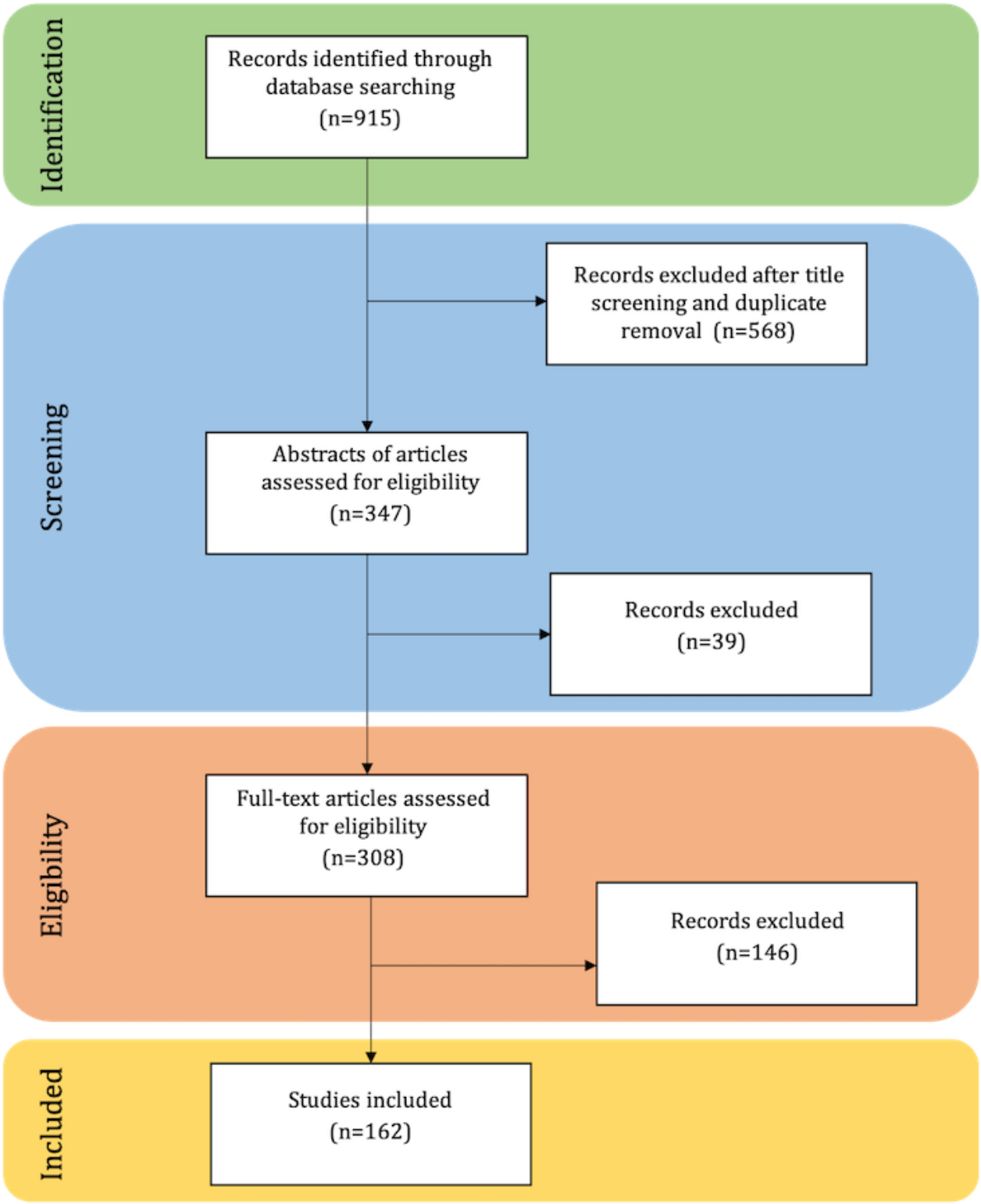
1. What internet-based sources are utilized for infoveillance and infodemiology?
2. What is the aim of the research conducted using these sources?
3. How are internet-based sources applied to generate knowledge?
4. Which sources have shown a preference for studying communicable and noncommunicable diseases?
5. What are the common limitations of internet-based sources?

Methods

Search Strategy

This review was conducted through several stages based on the Preferred Reporting Items for Systematic Reviews and Meta-Analyses process (Figure 1). To be the most inclusive, eight mutually exclusive research libraries that contain a variety of journals in the fields of Informatics and Biomedical Sciences were selected. The libraries were Europe PubMed Central, Institute of Electrical and Electronics Engineers Xplore Digital Library, Association for Computing Machinery Digital Library, SpringerLink, EBSCO Host, PubMed, Scopus, and Web of Science. Keyword generation was focused on IBSs of public health data, infoveillance, infodemiology, and disease outbreak and surveillance. We considered all conference and journal articles identified in this process.

Figure 1. Flowchart applied for the literature search.



Article Selection

The keywords were generated taking into consideration a preliminary assessment of the literature through a manual screening of relevant studies to ensure the list was complete. The complete list of these keywords can be found in Multimedia Appendix 1. The literature search was initially performed from October 10 to October 31, 2017, on the abovementioned repositories, focusing on the publication period of 2012 to 2017. This literature search was later augmented to include additional search terms and to extend the publication period until 2018. In total, the literature search had a duration of 2 months, excluding the screening and eligibility steps. Our review focused on the literature published after 2012 to cover a wider variety of sources, given the time lag between their popularity peaks; furthermore, by analyzing the literature published a few years after the first studies (eg, 2009 for search queries [14] and 2010 for social media [19]), we focused on research with finer-grained

and adapted methodologies (eg, improved keyword selection and relevancy filters). To select relevant articles, a multiphase process was implemented. First, the article title was screened for relevance and duplicates were removed; subsequently the abstract was screened, and only articles that passed both steps were considered for the eligibility phase. The inclusion and exclusion criteria for the articles were decided according to a modified PICOTS. The criteria are specified in Multimedia Appendix 1. The first author performed the screening process and retrieved the data. When doubts were raised regarding the inclusion of certain publications, the remaining authors were consulted.

Quality Assessment

To address the quality of the studies, we implemented a set of criteria to evaluate the publications retrieved. This assessment was based on a set of questions with regard to the purpose of research, contextualization, methodology, study design, the results obtained, and findings. The quality criteria are based on the work by Kofod-Petersen [33] and are present in Multimedia Appendix 1.

Data Extraction

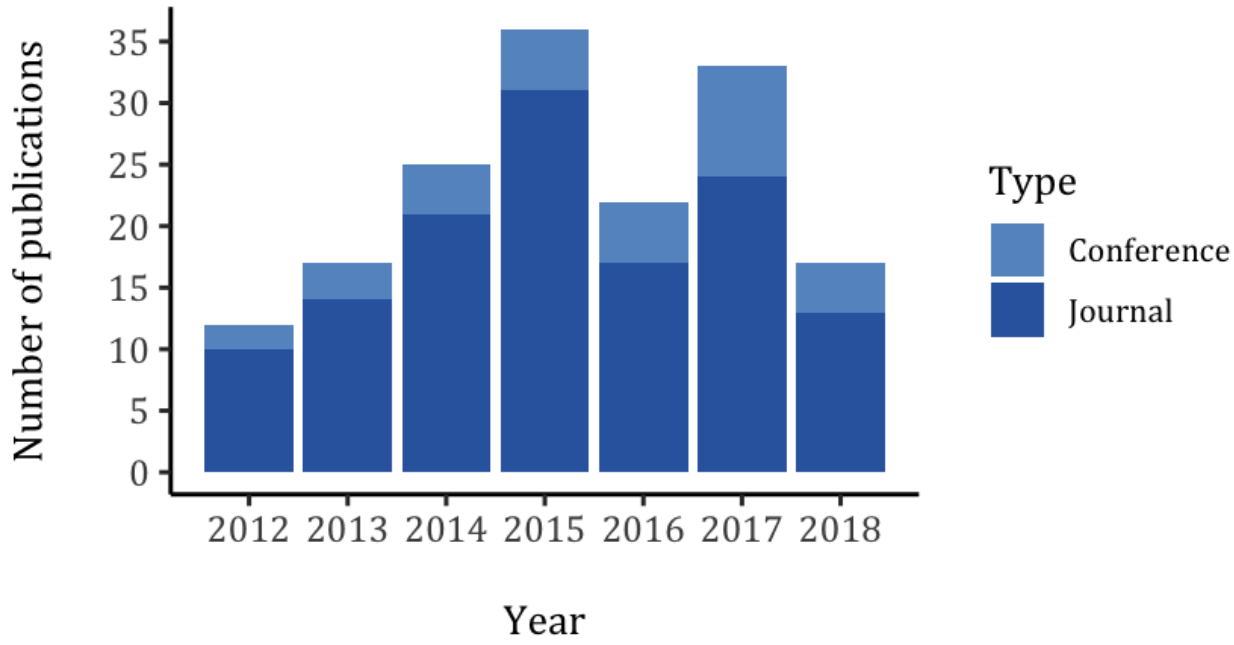
We also developed an extraction form to gather information about the studies allowing us to understand how the issues related to the proposed RQs have been addressed. This step was performed using the NVivo version 11 qualitative software database (QSR International Pty Ltd), nested *cases* were used to annotate each item of the extraction form. The extraction guidelines are available in Multimedia Appendix 1. For each checklist item in the guidelines, we created a classification that has been detailed in the following sections. Each paper was classified as *journal* or *conference*, in accordance with the inclusion criteria. Regarding the targeted diseases, we divided this into *chronic*, *infectious*, *medical conditions*, and *health topics*. The first three categories have been further specified. For the Goal/Objective item, a paper was classified as *outbreak forecasting* if it explicitly stated that the research was aimed at forecasting; else, it was assigned to *surveillance* (ie, when the purpose is only to identify the degree of correlation with ground-truth data and there is no mention of forecasting); *disease characterisation* was assigned when the aim was to determine identifiable characteristics related to a disease, eg, patient search behavior and commonly mentioned treatments, or when the aim was to classify a text as related to a disease; *content analysis* was assigned to the study of the sources' content (eg, news presence and expressed sentiment) referring to a disease or medical condition; *personal health mention classification* focused on separating general mentions of a disease or medical conditions (eg, news) from first-person mentions; and *diagnosis prediction* was assigned when the purpose was to attribute a disease or medical condition to a text and its creator by proxy. The *Internet-based data source* can be classified into search queries, social media (including microblogs), websites, news, discussion forums, web encyclopedia, and media monitoring systems. We also considered the use of data sources external to the IBSs, which can be classified as demographic, socioeconomic, and climate statistics, as well as data from governmental and laboratory sources, among others. To address the study design/methodology, we devised the following criteria: *topic analysis* corresponds to when topic modelling and similar approaches are used; *regression models* encompass all regression and autoregression models (eg, linear regression and autoregressive moving average); *statistical models* was assigned to more complex models (eg, Hidden Markov Chain); *correlation analysis* was used when correlation scores are calculated (eg, Pearson); *rule-based techniques* and *ranking techniques* are self-explanatory; *manual analysis* was assigned when no specific techniques are used other than a manual assessment; *epidemiology theory* refers to the use of techniques and measures commonly used in epidemiology (eg, Susceptible, Exposed, Infectious, and Recovered models); *linguistic analysis* was assigned when sentiment analysis and lexicons, among others, were used; and finally, we split *machine learning* and *deep learning*. Finally, we did not add a classification to the findings and limitations; we chose to keep this as an open field and manually analyzed the outcomes.

Results

Overview

The results from the search strategy are shown in Figure 1; in total, 162 papers were considered for this systematic literature review. The summary of the review results according to the data extraction guidelines is presented in Multimedia Appendix 1 [34-188]. The year with the highest number of publications is 2015 (n=36), followed by 2017 (n=33), 2014 (n=25), 2016 (n=22), 2013 (n=17), 2018 (n=17), and 2012 (n=12; Figure 2). Journal articles accounted for 130 of all publications and the remaining 32, for conferences. The remaining results were split into subsections correspondent to the extraction guidelines followed. The complete summary of the literature analysis is provided in Multimedia Appendix 1.

Figure 2. Distribution of the selected literature per year and type.



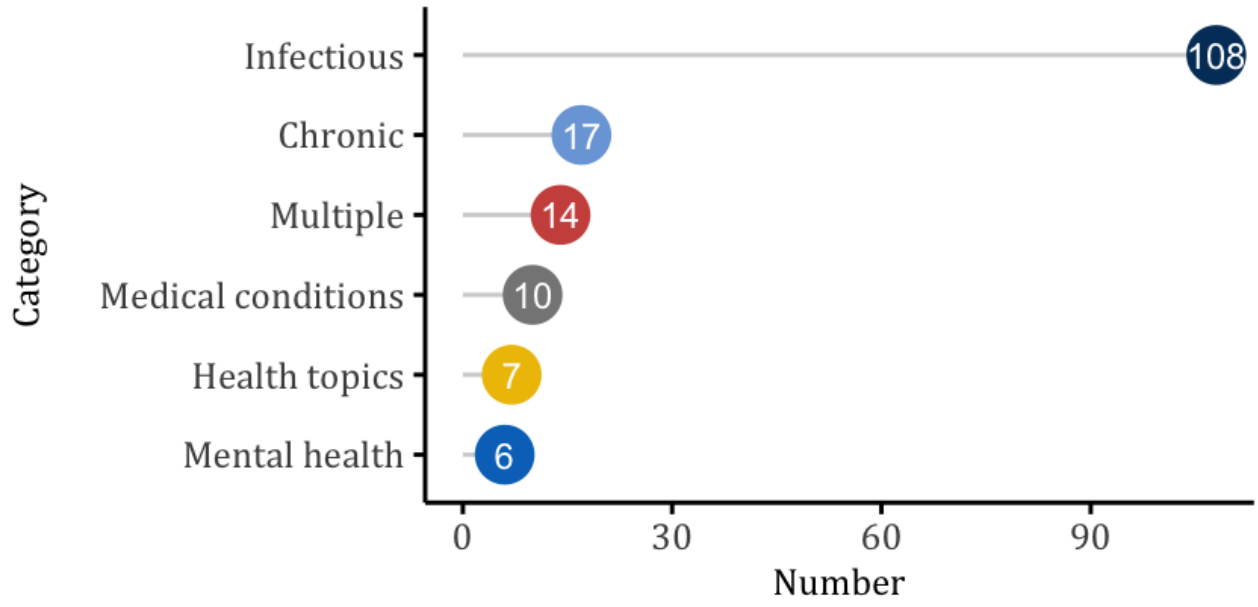
Goal

The analyzed papers mostly focused on surveillance (n=90), content analysis (n=46), and outbreak forecasting (n=45); other goals included personal health mention classification (n=10), disease characterization (n=5), and diagnosis prediction (n=4), with 36 papers having multiple targets.

Diseases, Medical Conditions, and Health Topics

Infectious diseases are markedly the most researched cases, with a total of 108 papers assigned. Chronic diseases are the focus of 17 publications, followed by medical conditions with 10 publications, health topics with 7 mentions, and mental health with 6 assigned articles (Figure 3). A set of 14 publications target multiple diseases from all the previously mentioned categories. Among the infectious diseases, ILIs, dengue, and infectious intestinal diseases are the top choices with 57, 7, and 7 assigned publications, respectively. In terms of chronic illnesses, cancer is the most researched disease (n=3). Excluding the publications focusing on multiple cases, 78% of the diseases appear in less than two articles.

Figure 3. Distribution of the case studies in the literature.



Internet-Based and External Data Sources

The 162 analyzed papers can be classified into eight distinct categories: search queries, social media, news, discussion forums, websites, media monitoring systems, web encyclopedia, and online obituaries. Social media (n=80) and search queries (n=79) are the most utilized IBSs, followed by web encyclopedias (n=13) that, in the selected papers, corresponded solely to Wikipedia. The remaining are utilized in the following decreasing order, forums (n=9), news (n=8), media monitoring systems (n=2), online obituaries (n=2), and websites not related to newspapers (n=1). A total of 29 papers utilize combinations of these sources, with the majority (n=11) combining search queries with social media. For social media, Twitter is mostly used (n=71) with the remaining sources marginally appearing. For search queries, the same behavior is seen with Google Trends; it is present in 42 publications, and when aggregating with Google Flu Trends and Google Dengue Trends, this value rises to 61.

Regarding sources external to IBSs, governmental, or laboratory surveillance statistics are the most utilized and used as ground-truth data (n=107), the second most used external source is hospital and emergency department visits (n=16), which are also used as ground-truth data. Climate or temperature statistics are applied in 8 papers, and socioeconomic statistics in 5 publications followed by health records in 4 publications, demographic or population statistics in 3 papers, and pharmaceutical sales in 2 publications. Scientific search engines, Flu Near You [189], and telephone triage are used individually in only 1 publication. In total, 45 publications do not use any external data source, and 26 publications share multiple external sources.

Study Design

Regarding the methodologies used, correlation analysis (n=59) was predominant and closely followed by regression models (n=46). Machine learning was used in 27 of the analyzed articles, statistical models are preferred in 20 publications, manual analysis was used in 18 of the articles, topic analysis is used in 12 publications, and deep learning and linguistic analysis were used in 10 articles each. Regarding the remaining methodology, rule-based techniques (n=7), epidemiology theory (n=6), surveys (n=3), and ranking techniques (n=1) were used in less than 10 papers.

Findings and Limitations

Qualitatively, the studies reported positive results (n=125), mentioning high or improved correlations with ground-truth data, as well as the outbreak predictive power, and high accuracy when the goal was surveillance, outbreak forecasting, personal health mention classification, disease characterization, and diagnosis prediction. The studies by Olson et al [55], Alicino et al [85], and Yom-tov [109] report negative results caused by questionable reliability with search query data and media influence affecting social networks and web encyclopedias. For the publications solely focused on content analysis (n=30), the findings were reported without a negative or positive association. A total of four publications [54,91,115,130] mention positive and negative results related to a large variation in the correlation score with social network data, surveillance inaccuracies for different age groups and the lack of specificity for search query data, and media influence when applying both social network and search query data.

In terms of limitations, these can be divided into gold standard (n=22); representativeness (n=76); general bias, eg, change in search behavior, symptom variability (n=9), and media effect (n=17); dataset size (n=7); methodology, eg, computational cost, keywords, and spelling errors (n=63); language (n=11); geographical restriction (n=33); and timeframe restriction (n=20).

Discussion

In this study, we aimed to provide a discriminative assessment on the application of IBSs in public health. To achieve this, we focused on the literature published in the last 6 years and applied systematic selection criteria to determine the appropriate studies to include. As a result, we proposed a taxonomy and identified the gaps to be addressed in future research, represented by the identified limitations of IBSs. Hence, this section addresses each RQ stated in the objectives.

Research Question 1: What Internet-Based Sources Are Utilized for Infoveillance and Infodemiology?

As reported in the Results, search queries, social media, discussion forums, news, web encyclopedia, online obituaries, media monitoring systems, and websites constitute the general categories of the IBSs present in the analyzed literature.

For search queries and social media, there is a large variation in the sources, which is mostly caused by geographical differences. The sources include platforms that are only available to certain countries. In the case of search queries, this potentially brings benefits in representativeness as it is possible to estimate the country-wide disease surveillance data from online search behavior. We argue that using a worldwide search engine, cultural differences that shape online search behavior could be diluted further complicating the surveillance task. Google Trends and its variations are the most common and widely represented; Bing also has an extensive geographical representation but a lower market share [190]. Also included are Baidu, Naver, Yandex, Vårdguiden and Websök, and Sapo, which cater to different countries, namely China, South Korea, Russia, Sweden, and Portugal, respectively. Nonetheless, the use of country-specific search engines can be limited by their market share, as is the case for Sapo [191], and their fine-grained geographical representation, eg, Vårdguiden is mostly used by people in Stockholm [68]. These limitations are further discussed in the following subsections.

Regarding social media, the sources differ in content richness. For example, while Twitter is a microblogging service, Weibo incorporates functionalities that can also be found on Facebook [192]. Nevertheless, the same reasoning applies, country-specific platforms can potentially bring benefits in representativeness and more closely estimate the country-wide health-related statistics.

The remaining sources, web encyclopedia and online obituaries, are used without a defined geographical restriction, and only English data were considered. Discussion forums and websites are an exception as they were utilized in different language-specific scenarios. Media monitoring systems also work on data from multiple sources and languages.

Research Question 2: What Is the Aim of the Research Conducted Using These Sources?

With IBSs of health information, the approaches are mostly based on monitoring the internet search and information-sharing behavior; the underlying assumption is that people actively seek and share information on diseases they develop.

In terms of surveillance and outbreak forecasting, estimates of disease activity within a community can be expressed by monitoring the frequency of related internet searches, disease mentions on social and news media, and page views in a web encyclopedia, among others. In addition, these sources also provide complementary information to the ground truth, eg, by targeting sick people who might not go to the hospital. When dealing with outbreak detection, an early and fast response is essential. Traditional surveillance is slower to transmit information across its different channels; therefore, IBSs complement the traditional mechanism when dealing with outbreaks [5,193,194].

Sources that go beyond single keywords pose a challenge as the occurrence of disease mentions does not correspond to the assumption that the text/health report in consideration is referring to the user suffering from the mentioned disease. For example, the microblog “Don’t forget to get your flu shot” is not as valuable as the microblog “I have the flu”; the latter corresponds to a personal health mention that has the potential to more closely correlate with gold standard data. Hence, personal health mention classification is based on the application of classification techniques that aim to filter false-positives, ie, a text containing a disease mention but not stating the user is carrying the disease, from true-positives, ie, a first person mentions of a disease by the affected user [37,72]. This is an important step that has been introduced when dealing with microblogs and online forums, as it has shown improvements for surveillance and outbreak forecasting.

Diagnosis prediction was not a common aim of the analyzed studies as it is difficult to validate owing to the lack of available gold standard data and owing to privacy concerns. The studies by De Choudhury et al [47] and Bodnar et al [64] include a prior user selection process for whom the authors have diagnosis information. In these cases, the source utilized was social media as it provides more contextual information and the potential for sentiment analysis, which is particularly valuable for mental health infodemiology studies. In contrast, the work by Karmen et al [97] targets the diagnosis of a health report (in the form of a forum post) and not the user itself (as not all information for the user is available) utilizing a similar methodology. Yom-tov et al [113], in their study, identify risk markers that correlate with a set of diseases based on the search behavior of assumed affected users.

When considering long-term patients, they also seek the internet for health information but also for online support through the connection with other patients [36,53,86,134]. This corresponds to the task of content analysis and disease characterization. IBSs are not only useful to perform disease surveillance but also to understand the information that is being shared online, which directly relates to public health tasks. The literature also points to the preference of sources for particular user groups, namely, users who seek support groups or connection with other patients and who suffer from chronic illnesses. In this situation, forums and social media, namely, microblogs, provide a suitable medium for the discussion of examination results, symptoms, treatments, and support, offering insights into how diseases are discussed online [36,41,72,86].

Research Question 3: How Are Internet-Based Sources Applied to Generate Knowledge?

As most of the publications aim to perform disease surveillance and outbreak forecasting, correlation analysis is regularly applied to determine the relationship between the IBSs and ground truth. Surveillance data are also commonly incorporated into surveillance and forecasting using regression models, which can also include autoregression, ie, past values of the ground-truth data. However, these methods make several assumptions regarding the distribution of the data, which might not be correct and overly simplistic.

Studies that utilize multiple sources of external data tend to apply more complex statistical methods which attempt to address the assumptions made by regression and autoregression models in trade of higher complexity.

The techniques mentioned earlier are applied to time-series data that can be obtained from the search query volume, page views from a web encyclopedia, and message/health report frequency in social media. In the latter case, to obtain the frequency, machine learning is commonly used to filter messages that are considered nonrelevant for the disease or medical condition in question. Thus, most of the machine learning approaches focus on social media and are reliant on annotated datasets, ie, a set of messages previously labelled as relevant or nonrelevant, which carry an added cost as this is necessary to train the models, as well as the lack of generalization as the labelled dataset targets a specific disease/medical condition.

Deep learning approaches improve generalization, they are capable of modelling complex nonlinear relationships, and do not impose restrictions on the data, eg, distribution; however, they have much higher complexity and can act as a black box owing to the high number of *tunable* hyperparameters. Furthermore, they require large sets of data that might not be available for diseases with a lower prevalence.

Topic analysis is mostly used for content analysis and it provides added benefits to manual analysis and surveys as it is unsupervised, ie, it does not require human input to perform the analysis. However, the topics identified might not be clearly related to a subject, which can lead to subjective interpretations; furthermore, it also carries high computational costs.

Linguistic analysis, in particular, sentiment analysis, can provide insights regarding negative and positive word use, among others,

and how it associates with diseases. For example, this is used for mental health research as the sentiment expressed in words can be fundamental to detect the mental state of a user [82,157]. In the same category, named entity recognition aids in the detection of locations and disease names, among others [144]. Furthermore, the use of lexical and syntactic features and the use of lexicons have shown to improve classification tasks, eg, self-mentions of disease and disease-related categories [81,162].

The use of the epidemiology theory is not common as it can require data that are not available through the use of IBSs owing to its limits in terms of user information (eg, age and location); however, some studies have implemented various epidemiological models [96,121,140,170], as well as epidemiological parameters [34,119].

Rule-based techniques are manually created and are specific to the disease/medical condition studied; hence, they suffer from lack of generalization.

Ranking techniques were only used in 1 of the analyzed papers [195], and it was used to rank the topics generated from a topic modelling approach, suggesting that these can be used to facilitate the interpretation of the topic analysis results.

Research Question 4: Which Sources Have Shown a Preference for Studying Communicable and Noncommunicable Diseases, Health Topics, and Mental Health?

The nature of infectious diseases, ie, fast moving and with easier measurable effects, makes these a preferential case study for outbreak detection and surveillance. In tandem, the sources commonly applied for these tasks are search queries and social media, both combined and with other sources. These are the preferred sources as their output can be transformed into time-series data and compared with a gold standard for evaluation. With regard to search queries, a variety has been used to provide the most representative search behavior for the countries and languages targeted. A similar methodology was applied with social media, although mostly restricted to microblogs. Another important task in studying communicable illnesses is to explore what type of information is being shared and when; this is vital to identify the spread of misinformation and to analyze how far-reaching the counteractions are from health care agencies. Hence, content analysis is also performed by mainly utilizing social media as it provides more contextual information than search queries. Forums and news are utilized for the same reason; the higher contextual value allows for more insights into the study of information dissemination.

When discussing NCDs, monitoring and content analysis are the major approaches taken on the papers analyzed. Collecting epidemiological data for NCDs is a labor-intensive process [57,196]; hence, monitoring through digital sources aims primarily to estimate the number of affected individuals, given that official statistics are released with a significant delay [47,57,58,95,136]. To perform such a task, the commonly used sources are social media, search queries, and online obituaries. Content analysis focuses on determining the behavior and characteristics of users who actively mention a disease (eg, through a forum or social media), and the content and dissemination of health information. This is relevant as it allows to explore the information that is spread within these communities, such as personal medical advice. Additionally, past research has shown that online communities can provide a more convenient environment, for some patients, than traditional face-to-face interactions with health providers [36,41,43,82]. For content analysis, the sources applied are social media, forums, and web encyclopedia. As stated earlier, data with greater contextual content could provide more detailed information regarding online behavior and information exchange. With regard to forecasting, Gu et al [95] and Zhang et al [138] mention the task of predicting erythromelalgia-related hospital visits one week ahead and detecting early signals of diabetes, both through the use of search query data.

The research on mental health, medical conditions, and health topics focuses on content analysis; thus, the sources utilized in the analyzed papers refer to social media, in particular, microblogs, and forums.

Overall, the choice of the source of data is significantly related to the health topics, aim of the study, and the data available for evaluation. Infectious diseases have a large incidence variation; hence, they tend to have surveillance data available and most of the approaches focus on surveillance and outbreak forecasting which in itself requires sources that immediately show changes in the online behavior of users. Thus, search queries and microblogs are preferred for an analysis requiring a timely response. Regarding the remaining health topics, the population affected does not fluctuate as is the case with infectious diseases; hence, the focus is on the discussion that occurs online. It is more valuable to determine the information being disseminated, the questions raised online, and the needs of the patients so that the health agencies can cater to this segment of the population.

Research Question 5: What Are the Common Limitations of Internet-Based Sources?

Most of the studies analyzed report on positive outcomes when utilizing IBSs for public health applications; however, some limitations are frequently cited and only a few authors have given these greater importance. Although recent statistics illustrate the growth in search for health information online [197], internet access is neither equally distributed among countries nor equally penetrating in all regions within a country, which significantly affects the application of IBSs of health information [129,169]. In all the sources, common limitations refer to the lack of representativeness and bias caused by internet penetration and access, and preference to certain user age groups. For example, in the case of Twitter, 62% of its 330 million users are aged between 18 and 49 years [198]; around 53% of American internet users look up information on Wikipedia, with these users being mostly highly educated (69%) and under the age of 30 years (62%) [199]. This type of information elucidates on the potential bias caused by the nongeneralized use of these sources, in particular, when a given age group is more susceptible to a disease (eg, elderly and children).

Another common limitation is related to precision issues caused by the inherent nature of diseases. These tend to share symptoms and treatments that are commonly used as keywords to detect

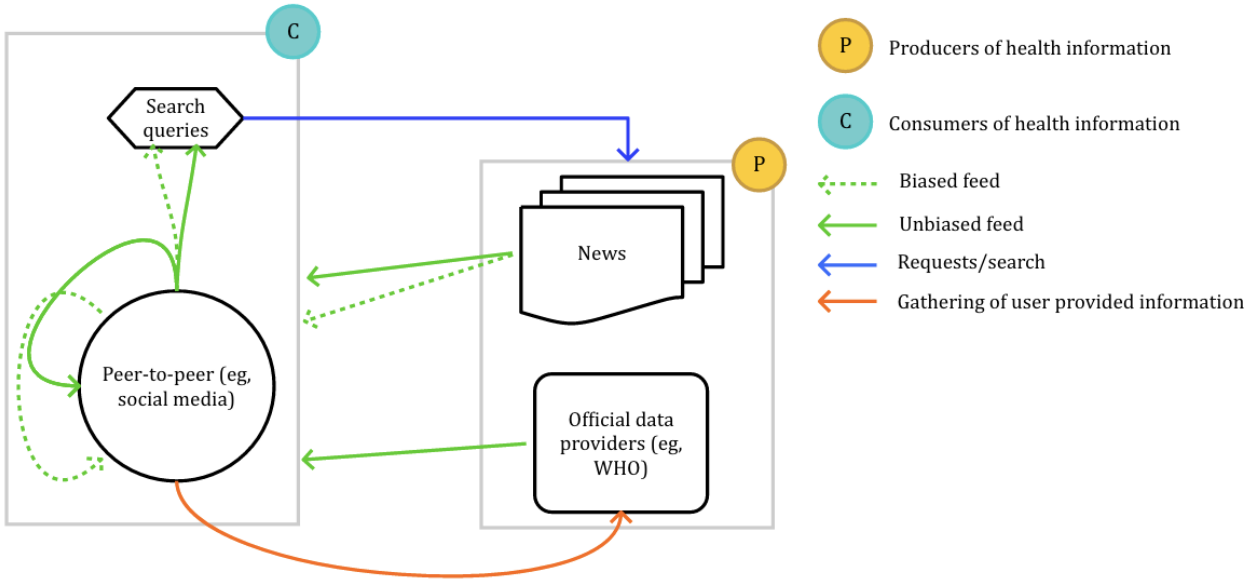
relevant health reports, furthermore, the use of unspecific health-related terminology is also common. The layman language used in IBSs is a challenge, given that most approaches in the field rely on the use of keywords selected from specialized medical vocabulary, as summarized by Dai et al [142]. However, the evolution of learning-based, lexicon-based, and embedding approaches has started to mitigate the language specificity effect.

When ground-truth data are available, some studies question their quality as they can be updated after the initial publication; other mentioned limitations concern the amount of data as well as their timespan, and geographical coverage. Language restriction is also a common limitation of the studies, as most are performed only in English.

In particular to search queries and social media, the lack or limited geographical resolution is also cited as a limitation. Using Google Trends, various studies refer to the lack of transparency on how the search volumes are obtained, especially since forecasting systems based on Google Trends (ie, Google Flu Trends and Google Dengue Trends) have shown significant algorithmic problems that led to their termination [200,201]. The need for costly, manually annotated datasets is a common issue when the goal is to perform classification, and it mainly occurs with social media data.

In addition, a common limitation to search queries, social media, and web encyclopedia is the effect caused by media events. A media event is an event or activity conducted for media publicity. In this definition, we include examples such as panic-inducing news [109] and celebrities being diagnosed with medical conditions [173]. Media events have shown to significantly affect the reliability of these sources. The results of studies by Yom-tov [109] and Mollema et al [100] demonstrate how these sources show a higher correlation with media events than actual surveillance data. The study by Alicino et al [85] also reveals that the presence of news strongly influences the search volume in locations where an outbreak is not occurring. In light of this, we present an interaction schema in Figure 4; the sources of public health–related data predominantly comprise search engine queries, which target public sources, and peer-to-peer (P2P) social media networks; we can further distinguish primary consumers of health information, eg, members of P2P networks, from producers, eg, biased and unbiased news and unbiased official data providers (eg, governmental sources). News and official data providers deliver biased and unbiased information to the consumers. Consumers receive this information and spread it, affecting their search and share behavior, namely, in search queries, social media, and web encyclopedias.

Figure 4. Data sources interaction cycle. WHO: World Health Organization.



Conclusions

IBSs of health information are a valuable proxy to study illnesses affecting the population. Their benefits and applications are far-reaching and continue to evolve as a potential asset to public health. The knowledge gathered from this review suggests that search queries and social media provide useful data to monitor infectious diseases. With regard to studying chronic illnesses, discussion forums and social media are preferred.

The methods used to select relevant keywords or messages target specific illnesses, thus requiring constant updates to reflect the population's changing search behavior as well as emerging trends. Here, we identify the first research gap; disease outbreaks outside of the targeted disease names will not be identified, and new terminologies crucial for the detection of previously targeted diseases will be missed. Future approaches must focus on the ever-changing nature of diseases. For example, new related keywords could be identified through services such as Google Trends' *related topics*. To identify emerging illnesses, more emphasis must be given to the structure and syntax of messages describing a first-person mention of a disease, as this could be applicable to other illnesses.

The strong susceptibility to media events and the absence of approaches dealing with this issue constitute the second research gap. As shown in Figure 4, the interactions between the different sources and the type of information (biased and unbiased) reach

the consumers and affect their search and share behavior, namely, in search queries, social media, and web encyclopedias. Such an effect must be mitigated to ensure improved reliability when utilizing IBSs.

The third research gap relates to the absence of consistent training and test periods, which impedes the appropriate comparison among the different methodologies. To address this, we suggest the creation of standard datasets, allowing to quantify the improvements of distinct methodologies. We also found a significant lack of interaction with public health officials, which would be the entities receiving the information from these models.

As a final recommendation, we suggest the use of an alternative strategy to better harness the information provided by IBSs. Namely, a proactive approach where the users are asked to report on their health state requesting the user to anonymously publish this information while avoiding the inclination to only publish positive messages. Such implementations can potentially make IBSs more comprehensible and a more valuable asset for disease monitoring.

Systematic Literature Review Limitations

This study makes use of eight databases, aiming to achieve a high coverage of the scientific literature. However, these databases do not guarantee full coverage and, hence, the inclusion of all relevant publications in our systematic methodology. In addition, we only considered articles in English as it is the predominant language of the scientific literature; thus, some contributions were potentially missed. The publication period is restricted to the last 6 years to allow for a focus on recent trends; earlier studies were referenced in the Introduction; however, not in an exhaustive way. We included a variety of keywords for the literature search although we understand that these might not cover all relevant publications.

Given that the authors followed a rigorous and systematic methodology when including and excluding publications for this literature review, selection bias was minimized. However, we cannot guarantee the absence of a bias when qualitatively presenting the findings; some categories and articles might be over- or under-represented.

Acknowledgments

This publication has emanated from research conducted with the financial support of Science Foundation Ireland (SFI) under Grant Number SFI/12/RC/2289_P2, cofunded by the European Regional Development Fund.

Authors' Contributions

JB originated the study, collected and analyzed data, and drafted the paper. JD and DR aided in the conceptualization of the study, article selection, and drafting of the paper.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Supporting information for the systematic literature review (keywords, inclusion and exclusion criteria, quality assessment, data extraction checklist, literature synthesis/summary).
[DOCX File , 63 KB-Multimedia Appendix 1]

References

1. Acheson D. Public Health in England: The Report of the Committee of Inquiry into the Future Development to the Public Health Function. London: London Her Majesty's Stationery Office; 1988.
2. World Health Organization. 2018. Public Health Services URL: http://www.euro.who.int/en/health-topics/Health-systems/public-health-services [accessed 2018-08-27]
3. World Health Organization. 2018. Public Health Surveillance URL: http://www.who.int/topics/public_health_surveillance/en/ [accessed 2018-08-27]
4. Nsubuga P, White ME, Thacker SB, Anderson MA, Blount SB, Broome CV, et al. Public health surveillance: a tool for targeting and monitoring interventions. In: Jamison DT, Breman JG, Measham AR, Alleyne G, Claeson M, Evans DV, et al, editors. Disease Control Priorities in Developing Countries. Second Edition. New York: Oxford University Press; 2006:997-1018.
5. Hope K, Durrheim DN, d'Espaignet ET, Dalton C. Syndromic Surveillance: is it a useful tool for local outbreak detection? J Epidemiol Community Health 2006 May;60(5):374-375 [FREE Full text] [doi: 10.1136/jech.2005.035337] [Medline: 16680907]
6. World Health Organization. 2018. Epidemic intelligence - systematic event detection URL: https://www.who.int/csr/alertresponse/epidemicintelligence/en/ [accessed 2018-08-27]
7. O'Shea J. Digital disease detection: a systematic review of event-based internet biosurveillance systems. Int J Med Inform 2017 May;101:15-22. [doi: 10.1016/j.ijmedinf.2017.01.019] [Medline: 28347443]

8. Wilson K, Brownstein JS. Early detection of disease outbreaks using the internet. Can Med Assoc J 2009 Apr 14;180(8):829-831 [FREE Full text] [doi: 10.1503/cmaj.090215] [Medline: 19364791]
9. Center for Disease Control and Prevention. 2003. Syndromic Surveillance: Reports from a National Conference URL: https://www.cdc.gov/mmwr/preview/su5301toc.htm [accessed 2020-01-23]
10. World Health Organization. 2018 May 24. The Top 10 Causes of Death URL: https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death [accessed 2019-08-20]
11. Achrekar H, Gandhe A, Lazarus R, Yu S, Liu B, Ssu-Hsin Y, et al. Predicting Flu Trends using Twitter Data. In: Proceedings of the 2011 IEEE Conference on Computer Communications Workshops. 2011 Presented at: INFOCOM WKSHPS'11; April 10-15, 2011; Shanghai, China p. 702-707. [doi: 10.1109/infcomw.2011.5928903]
12. Madoff LC. ProMED-mail: an early warning system for emerging diseases. Clin Infect Dis 2004 Jul 15;39(2):227-232. [doi: 10.1086/422003] [Medline: 15307032]
13. Hartley DM, Nelson NP, Arthur RR, Barboza P, Collier N, Lightfoot N, et al. An overview of internet biosurveillance. Clin Microbiol Infect 2013 Nov;19(11):1006-1013 [FREE Full text] [doi: 10.1111/1469-0691.12273] [Medline: 23789639]
14. Ginsberg J, Mohebbi MH, Patel RS, Brammer L, Smolinski MS, Brilliant L. Detecting influenza epidemics using search engine query data. Nature 2009 Feb 19;457(7232):1012-1014. [doi: 10.1038/nature07634] [Medline: 19020500]
15. Runge-Ranzinger S, McCall PJ, Kroeger A, Horstick O. Dengue disease surveillance: an updated systematic literature review. Trop Med Int Health 2014 Sep;19(9):1116-1160 [FREE Full text] [doi: 10.1111/tmi.12333] [Medline: 24889501]
16. Pelat C, Turbelin C, Bar-Hen A, Flahault A, Valleron AJ. More diseases tracked by using Google Trends. Emerg Infect Dis 2009 Aug;15(8):1327-1328 [FREE Full text] [doi: 10.3201/eid1508.090299] [Medline: 19751610]
17. Polgreen PM, Chen Y, Pennock DM, Nelson FD. Using internet searches for influenza surveillance. Clin Infect Dis 2008 Dec 1;47(11):1443-1448. [doi: 10.1086/593098] [Medline: 18954267]
18. Hulth A, Rydevik G, Linde A. Web queries as a source for syndromic surveillance. PLoS One 2009;4(2):e4378 [FREE Full text] [doi: 10.1371/journal.pone.0004378] [Medline: 19197389]
19. Culotta A. Towards Detecting Influenza Epidemics by Analyzing Twitter Messages. In: Proceedings of the First Workshop on Social Media Analytics. USA: ACM; 2010 Presented at: SOMA'10; July 25 - 28, 2010; Washington DC p. 115-122. [doi: 10.1145/1964858.1964874]
20. Bernardo TM, Rajic A, Young I, Robiadek K, Pham MT, Funk JA. Scoping review on search queries and social media for disease surveillance: a chronology of innovation. J Med Internet Res 2013 Jul 18;15(7):e147 [FREE Full text] [doi: 10.2196/jmir.2740] [Medline: 23896182]
21. Kostkova P, Szomszor M, St Louis C. #swineflu: the use of Twitter as an early warning and risk communication tool in the 2009 swine flu pandemic. ACM Trans Manage Inf Syst 2014;5(2):1-25. [doi: 10.1145/2597892]
22. Aramaki E, Maskawa S, Morita M. Twitter Catches the Flu: Detecting Influenza Epidemics Using Twitter. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. 2011 Presented at: EMNLP'11; July 27–31, 2011; Edinburgh, United Kingdom p. 1568-1576.
23. Lamb A, Paul MJ, Dredze M. Separating Fact from Fear: Tracking Flu Infections on Twitter. In: Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2013 Presented at: NAACL'13; June 9-14, 2013; Atlanta, Georgia p. 789-795.
24. Gianfredi V, Bragazzi NL, Nucci D, Martini M, Rosselli R, Minelli L, et al. Harnessing big data for communicable tropical and sub-tropical disorders: implications from a systematic review of the literature. Front Public Health 2018;6:90 [FREE Full text] [doi: 10.3389/fpubh.2018.00090] [Medline: 29619364]
25. Gittelman S, Lange V, Crawford CA, Okoro CA, Lieb E, Dhingra SS, et al. A new source of data for public health surveillance: Facebook likes. J Med Internet Res 2015 Apr 20;17(4):e98 [FREE Full text] [doi: 10.2196/jmir.3970] [Medline: 25895907]
26. Seltzer EK, Jean NS, Kramer-Golinkoff E, Asch DA, Merchant RM. The content of social media's shared images about Ebola: a retrospective study. Public Health 2015 Sep;129(9):1273-1277. [doi: 10.1016/j.puhe.2015.07.025] [Medline: 26285825]
27. Marcus MA, Westra HA, Eastwood JD, Barnes KL, Mobilizing Minds Research Group. What are young adults saying about mental health? An analysis of Internet blogs. J Med Internet Res 2012 Jan 30;14(1):e17 [FREE Full text] [doi: 10.2196/jmir.1868] [Medline: 22569642]
28. Paul MJ, Dredze M. You Are What You Tweet: Analyzing Twitter for Public Health. In: Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media. 2011 Presented at: ICWSM'11; July 17-21, 2011; Barcelona, Spain. [doi: 10.1145/2405716.2405728]
29. Roccetti M, Casari A, Marfia G. Inside Chronic Autoimmune Disease Communities: A Social Networks Perspective to Crohn's Patient Behavior and Medical Information. In: Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015. 2015 Presented at: ASONAM'15; August 25 - 28, 2015; Paris, France p. 1089-1096. [doi: 10.1145/2808797.2808813]
30. Sciascia S, Radin M. What can Google and Wikipedia can tell us about a disease? Big Data trends analysis in Systemic Lupus Erythematosus. Int J Med Inform 2017 Nov;107:65-69. [doi: 10.1016/j.ijmedinf.2017.09.002] [Medline: 29029693]

31. Barboza P, Vaillant L, Mawudeku A, Nelson NP, Hartley DM, Madoff LC, Early Alerting Reporting Project Of The Global Health Security Initiative. Evaluation of epidemic intelligence systems integrated in the early alerting and reporting project for the detection of A/H5N1 influenza events. PLoS One 2013;8(3):e57252 [FREE Full text] [doi: 10.1371/journal.pone.0057252] [Medline: 23472077]
32. Yang S, Santillana M, Brownstein JS, Gray J, Richardson S, Kou SC. Using electronic health records and internet search information for accurate influenza forecasting. BMC Infect Dis 2017 May 8;17(1):332 [FREE Full text] [doi: 10.1186/s12879-017-2424-7] [Medline: 28482810]
33. Kofod-Petersen A. Research Gate. 2014. How to Do a Structured Literature Review in Computer Science URL: https://www.researchgate.net/profile/Anders_Kofod-Petersen/publication/265158913_How_to_do_a_Structured_Literature_Review_in_computer_science/links/599a00350f7e9b3edb17cda2/How-to-do-a-Structured-Literature-Review-in-computer-science.pdf [accessed 2020-01-23]
34. Chunara R, Andrews JR, Brownstein JS. Social and news media enable estimation of epidemiological patterns early in the 2010 Haitian cholera outbreak. Am J Trop Med Hyg 2012 Jan;86(1):39-45 [FREE Full text] [doi: 10.4269/ajtmh.2012.11-0597] [Medline: 22232449]
35. Dugas AF, Hsieh Y, Levin SR, Pines JM, Mareiniss DP, Mohareb A, et al. Google Flu Trends: correlation with emergency department influenza rates and crowding metrics. Clin Infect Dis 2012 Feb 15;54(4):463-469 [FREE Full text] [doi: 10.1093/cid/cir883] [Medline: 22230244]
36. Gold KJ, Boggs ME, Mugisha E, Palladino CL. Internet message boards for pregnancy loss: who's on-line and why? Womens Health Issues 2012;22(1):e67-e72 [FREE Full text] [doi: 10.1016/j.whi.2011.07.006] [Medline: 21907592]
37. Khan MA, Iwai M, Sezaki K. A Robust and Scalable Framework for Detecting Self-reported Illness from Twitter. In: 2012 IEEE 14th International Conference on e-Health Networking, Applications and Services. 2012 Presented at: Healthcom'12; October 10-13, 2012; Beijing, China p. 303-308. [doi: 10.1109/healthcom.2012.6379425]
38. Patwardhan A, Bilkovski R. Comparison: flu prescription sales data from a retail pharmacy in the US with Google Flu trends and US ILINet (CDC) data as flu activity indicator. PLoS One 2012;7(8):e43611 [FREE Full text] [doi: 10.1371/journal.pone.0043611] [Medline: 22952719]
39. Pervaiz F, Pervaiz M, Rehman NA, Saif U. FluBreaks: early epidemic detection from Google flu trends. J Med Internet Res 2012 Oct 4;14(5):e125 [FREE Full text] [doi: 10.2196/jmir.2102] [Medline: 23037553]
40. Samaras L, García-Barriocanal E, Sicilia M. Syndromic surveillance models using web data: the case of scarlet fever in the UK. Inform Health Soc Care 2012 Mar;37(2):106-124. [doi: 10.3109/17538157.2011.647934] [Medline: 22360741]
41. Sugawara Y, Narimatsu H, Hozawa A, Shao L, Otani K, Fukao A. Cancer patients on Twitter: a novel patient community on social media. BMC Res Notes 2012 Dec 27;5:699 [FREE Full text] [doi: 10.1186/1756-0500-5-699] [Medline: 23270426]
42. van Velsen L, van Gemert-Pijnen JE, Beaujean DJ, Wentzel J, van Steenbergen JE. Should health organizations use web 2.0 media in times of an infectious disease crisis? An in-depth qualitative study of citizens' information behavior during an EHEC outbreak. J Med Internet Res 2012 Dec 20;14(6):e181 [FREE Full text] [doi: 10.2196/jmir.2123] [Medline: 23257066]
43. Bosley JC, Zhao NW, Hill S, Shofer FS, Asch DA, Becker LB, et al. Decoding twitter: surveillance and trends for cardiac arrest and resuscitation communication. Resuscitation 2013 Feb;84(2):206-212 [FREE Full text] [doi: 10.1016/j.resuscitation.2012.10.017] [Medline: 23108239]
44. Bragazzi NL. Infodemiology and infoveillance of multiple sclerosis in Italy. Mult Scler Int 2013;2013:924029 [FREE Full text] [doi: 10.1155/2013/924029] [Medline: 24027636]
45. Broniatowski DA, Paul MJ, Dredze M. National and local influenza surveillance through Twitter: an analysis of the 2012-2013 influenza epidemic. PLoS One 2013;8(12):e83672 [FREE Full text] [doi: 10.1371/journal.pone.0083672] [Medline: 24349542]
46. Cho S, Sohn CH, Jo MW, Shin S, Lee JH, Ryoo SM, et al. Correlation between national influenza surveillance data and google trends in South Korea. PLoS One 2013;8(12):e81422 [FREE Full text] [doi: 10.1371/journal.pone.0081422] [Medline: 24339927]
47. de Choudhury M, Counts S, Horvitz E. Social Media as a Measurement Tool of Depression in Populations. In: Proceedings of the 5th Annual ACM Web Science Conference. 2013 Presented at: WebSci'13; May 2 - 4, 2013; Paris, France p. 47-56. [doi: 10.1145/2464464.2464480]
48. Dugas AF, Jalalpour M, Gel Y, Levin S, Torcaso F, Igusa T, et al. Influenza forecasting with Google Flu Trends. PLoS One 2013;8(2):e56176 [FREE Full text] [doi: 10.1371/journal.pone.0056176] [Medline: 23457520]
49. Gesualdo F, Stilo G, Agricola E, Gonfiantini MV, Pandolfi E, Velardi P, et al. Influenza-like illness surveillance on Twitter through automated learning of naïve language. PLoS One 2013;8(12):e82489 [FREE Full text] [doi: 10.1371/journal.pone.0082489] [Medline: 24324799]
50. Ghosh DD, Guha R. What are we 'tweeting' about obesity? Mapping tweets with Topic Modeling and Geographic Information System. Cartogr Geogr Inf Sci 2013;40(2):90-102 [FREE Full text] [doi: 10.1080/15230406.2013.776210] [Medline: 25126022]
51. Kang M, Zhong H, He J, Rutherford S, Yang F. Using Google Trends for influenza surveillance in South China. PLoS One 2013;8(1):e55205 [FREE Full text] [doi: 10.1371/journal.pone.0055205] [Medline: 23372837]

52. Kim E, Seok JH, Oh JS, Lee HW, Kim KH. Use of Hangeul Twitter to track and predict human influenza infection. PLoS One 2013;8(7):e69305 [FREE Full text] [doi: 10.1371/journal.pone.0069305] [Medline: 23894447]
53. Lu Y, Zhang P, Liu J, Li J, Deng S. Health-related hot topic detection in online communities using text clustering. PLoS One 2013;8(2):e56221 [FREE Full text] [doi: 10.1371/journal.pone.0056221] [Medline: 23457530]
54. Nagel AC, Tsou M, Spitzberg BH, An L, Gawron JM, Gupta DK, et al. The complex relationship of realspace events and messages in cyberspace: case study of influenza and pertussis using tweets. J Med Internet Res 2013 Oct 24;15(10):e237 [FREE Full text] [doi: 10.2196/jmir.2705] [Medline: 24158773]
55. Olson DR, Konty KJ, Paladini M, Viboud C, Simonsen L. Reassessing Google Flu Trends data for detection of seasonal and pandemic influenza: a comparative epidemiological study at three geographic scales. PLoS Comput Biol 2013;9(10):e1003256 [FREE Full text] [doi: 10.1371/journal.pcbi.1003256] [Medline: 24146603]
56. Parker J, Wei Y, Yates A, Frieder O, Goharian N. A Framework for Detecting Public Health Trends with Twitter. In: Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining. 2013 Presented at: ASONAM'13; August 25 - 28, 2013; Niagara Falls, ON, Canada p. 556-563. [doi: 10.1145/2492517.2492544]
57. Willard SD, Nguyen MM. Internet search trends analysis tools can provide real-time data on kidney stone disease in the United States. Urology 2013 Jan;81(1):37-42. [doi: 10.1016/j.urology.2011.04.024] [Medline: 21676450]
58. Won HH, Myung W, Song GY, Lee W, Kim JW, Carroll BJ, et al. Predicting national suicide numbers with social media data. PLoS One 2013;8(4):e61809 [FREE Full text] [doi: 10.1371/journal.pone.0061809] [Medline: 23630615]
59. Zheluk A, Quinn C, Hercz D, Gillespie JA. Internet search patterns of human immunodeficiency virus and the digital divide in the Russian Federation: infoveillance study. J Med Internet Res 2013 Nov 12;15(11):e256 [FREE Full text] [doi: 10.2196/jmir.2936] [Medline: 24220250]
60. Zhou X, Li Q, Zhu Z, Zhao H, Tang H, Feng Y. Monitoring epidemic alert levels by analyzing internet search volume. IEEE Trans Biomed Eng 2013 Feb;60(2):446-452. [doi: 10.1109/TBME.2012.2228264] [Medline: 23192470]
61. Andersson T, Bjelkmar P, Hulth A, Lindh J, Stenmark S, Widerström M. Syndromic surveillance for local outbreak detection and awareness: evaluating outbreak signals of acute gastroenteritis in telephone triage, web-based queries and over-the-counter pharmacy sales. Epidemiol Infect 2014 Feb;142(2):303-313 [FREE Full text] [doi: 10.1017/S0950268813001088] [Medline: 23672877]
62. Araz OM, Bentley D, Muelleman RL. Using Google Flu Trends data in forecasting influenza-like-illness related ED visits in Omaha, Nebraska. Am J Emerg Med 2014 Sep;32(9):1016-1023. [doi: 10.1016/j.ajem.2014.05.052] [Medline: 25037278]
63. Aslam AA, Tsou M, Spitzberg BH, An L, Gawron JM, Gupta DK, et al. The reliability of tweets as a supplementary method of seasonal influenza surveillance. J Med Internet Res 2014 Nov 14;16(11):e250 [FREE Full text] [doi: 10.2196/jmir.3532] [Medline: 25406040]
64. 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 Presented at: WWW'14; April 7 - 11, 2014; Seoul, Korea p. 651-656. [doi: 10.1145/2567948.2579272]
65. Chen L, Hossain KS, Butler P, Ramakrishnan N, Prakash BA. Flu Gone Viral: Syndromic Surveillance of Flu on Twitter Using Temporal Topic Models. In: 2014 IEEE International Conference on Data Mining. 2014 Presented at: ICDM'14; December 14-17. 2014; Shenzhen, China p. 755-760. [doi: 10.1109/icdm.2014.137]
66. Culotta A. Estimating County Health Statistics with Twitter. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. 2014 Presented at: CHI '14; April 26 - May 1, 2014; Toronto, Ontario p. 1335-1344. [doi: 10.1145/2556288.2557139]
67. Dubey D, Amritphale A, Sawhney A, Dubey D, Srivastav N. Analysis of YouTube as a source of information for West Nile virus infection. Clin Med Res 2014 Dec;12(3-4):129-132 [FREE Full text] [doi: 10.3121/cmr.2013.1194] [Medline: 24573700]
68. Edelstein M, Wallensten A, Zetterqvist I, Hulth A. Detecting the norovirus season in Sweden using search engine data--meeting the needs of hospital infection control teams. PLoS One 2014;9(6):e100309 [FREE Full text] [doi: 10.1371/journal.pone.0100309] [Medline: 24955857]
69. Generous N, Fairchild G, Deshpande A, del Valle SY, Priedhorsky R. Global disease monitoring and forecasting with Wikipedia. PLoS Comput Biol 2014 Nov;10(11):e1003892 [FREE Full text] [doi: 10.1371/journal.pcbi.1003892] [Medline: 25392913]
70. Gluskin RT, Johansson MA, Santillana M, Brownstein JS. Evaluation of internet-based dengue query data: Google Dengue Trends. PLoS Negl Trop Dis 2014 Feb;8(2):e2713 [FREE Full text] [doi: 10.1371/journal.pntd.0002713] [Medline: 24587465]
71. Gu H, Chen B, Zhu H, Jiang T, Wang X, Chen L, et al. Importance of internet surveillance in public health emergency control and prevention: evidence from a digital epidemiologic study during avian influenza A H7N9 outbreaks. J Med Internet Res 2014 Jan 17;16(1):e20 [FREE Full text] [doi: 10.2196/jmir.2911] [Medline: 24440770]
72. Ku Y, Chiu C, Zhang Y, Chen H, Su H. Text mining self-disclosing health information for public health service. J Assoc Inf Sci Technol 2014;65(5):928-947. [doi: 10.1002/asi.23025]

73. McIver DJ, Brownstein JS. Wikipedia usage estimates prevalence of influenza-like illness in the United States in near real-time. PLoS Comput Biol 2014 Apr;10(4):e1003581 [FREE Full text] [doi: 10.1371/journal.pcbi.1003581] [Medline: 24743682]
74. 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 Infect Dis 2014 Dec 31;14:690 [FREE Full text] [doi: 10.1186/s12879-014-0690-1] [Medline: 25551277]
75. Nagar R, Yuan Q, Freifeld CC, Santillana M, Nojima A, Chunara R, et al. A case study of the New York City 2012-2013 influenza season with daily geocoded Twitter data from temporal and spatiotemporal perspectives. J Med Internet Res 2014 Oct 20;16(10):e236 [FREE Full text] [doi: 10.2196/jmir.3416] [Medline: 25331122]
76. Seo D, Jo M, Sohn CH, Shin S, Lee J, Yu M, et al. Cumulative query method for influenza surveillance using search engine data. J Med Internet Res 2014 Dec 16;16(12):e289 [FREE Full text] [doi: 10.2196/jmir.3680] [Medline: 25517353]
77. Paul MJ, Dredze M. Discovering health topics in social media using topic models. PLoS One 2014;9(8):e103408 [FREE Full text] [doi: 10.1371/journal.pone.0103408] [Medline: 25084530]
78. Prieto VM, Matos S, Álvarez M, Cacheda F, Oliveira JL. Twitter: a good place to detect health conditions. PLoS One 2014;9(1):e86191 [FREE Full text] [doi: 10.1371/journal.pone.0086191] [Medline: 24489699]
79. Santos JC, Matos S. Analysing Twitter and web queries for flu trend prediction. Theor Biol Med Model 2014 May 7;11(Suppl 1):S6 [FREE Full text] [doi: 10.1186/1742-4682-11-S1-S6] [Medline: 25077431]
80. Thompson LH, Malik MT, Gumel A, Strome T, Mahmud SM. Emergency department and 'Google flu trends' data as syndromic surveillance indicators for seasonal influenza. Epidemiol Infect 2014 Nov;142(11):2397-2405. [doi: 10.1017/S0950268813003464] [Medline: 24480399]
81. Velardi P, Stilo G, Tozzi AE, Gesualdo F. Twitter mining for fine-grained syndromic surveillance. Artif Intell Med 2014 Jul;61(3):153-163. [doi: 10.1016/j.artmed.2014.01.002] [Medline: 24613716]
82. Wilson ML, Ali S, Valstar MF. Finding Information about Mental Health in Microblogging Platforms: A Case Study of Depression. In: Proceedings of the 5th Information Interaction in Context Symposium. 2014 Presented at: IIiX'14; August 26 - 30, 2014; Regensburg, Germany p. 8-17. [doi: 10.1145/2637002.2637006]
83. Yom-Tov E, Borsa D, Cox IJ, McKendry RA. Detecting disease outbreaks in mass gatherings using internet data. J Med Internet Res 2014 Jun 18;16(6):e154 [FREE Full text] [doi: 10.2196/jmir.3156] [Medline: 24943128]
84. Young SD, Rivers C, Lewis B. Methods of using real-time social media technologies for detection and remote monitoring of HIV outcomes. Prev Med 2014 Jun;63:112-115 [FREE Full text] [doi: 10.1016/j.ypmed.2014.01.024] [Medline: 24513169]
85. Alicino C, Bragazzi NL, Faccio V, Amicizia D, Panatto D, Gasparini R, et al. Assessing Ebola-related web search behaviour: insights and implications from an analytical study of Google Trends-based query volumes. Infect Dis Poverty 2015 Dec 10;4:54 [FREE Full text] [doi: 10.1186/s40249-015-0090-9] [Medline: 26654247]
86. Betts D, Dahlen HG, Smith CA. A search for hope and understanding: an analysis of threatened miscarriage internet forums. Midwifery 2014 Jun;30(6):650-656. [doi: 10.1016/j.midw.2013.12.011] [Medline: 24439850]
87. Broniatowski DA, Dredze M, Paul MJ, Dugas A. Using social media to perform local influenza surveillance in an inner-city hospital: a retrospective observational study. JMIR Public Health Surveill 2015;1(1):e5 [FREE Full text] [doi: 10.2196/publichealth.4472] [Medline: 27014744]
88. Cleaton JM, Viboud C, Simonsen L, Hurtado AM, Chowell G. Characterizing Ebola transmission patterns based on internet news reports. Clin Infect Dis 2016 Jan 1;62(1):24-31 [FREE Full text] [doi: 10.1093/cid/civ748] [Medline: 26338786]
89. Cui X, Yang N, Wang Z, Hu C, Zhu W, Li H, et al. Chinese social media analysis for disease surveillance. Pers Ubiquit Comput 2015;19(7):1125-1132. [doi: 10.1007/s00779-015-0877-5]
90. Davidson MW, Haim DA, Radin JM. Using networks to combine 'big data' and traditional surveillance to improve influenza predictions. Sci Rep 2015 Jan 29;5:8154 [FREE Full text] [doi: 10.1038/srep08154] [Medline: 25634021]
91. Domnich A, Panatto D, Signori A, Lai PL, Gasparini R, Amicizia D. Age-related differences in the accuracy of web query-based predictions of influenza-like illness. PLoS One 2015;10(5):e0127754 [FREE Full text] [doi: 10.1371/journal.pone.0127754] [Medline: 26011418]
92. Eichstaedt JC, Schwartz HA, Kern ML, Park G, Labarthe DR, Merchant RM, et al. Psychological language on Twitter predicts county-level heart disease mortality. Psychol Sci 2015 Feb;26(2):159-169 [FREE Full text] [doi: 10.1177/0956797614557867] [Medline: 25605707]
93. Fung IC, Hao Y, Cai J, Ying Y, Schaible BJ, Yu CM, et al. Chinese social media reaction to information about 42 notifiable infectious diseases. PLoS One 2015;10(5):e0126092 [FREE Full text] [doi: 10.1371/journal.pone.0126092] [Medline: 25946020]
94. Gesualdo F, Stilo G, D'Ambrosio A, Carloni E, Pandolfi E, Velardi P, et al. Can Twitter be a source of information on allergy? Correlation of pollen counts with tweets reporting symptoms of allergic rhinoconjunctivitis and names of antihistamine drugs. PLoS One 2015;10(7):e0133706 [FREE Full text] [doi: 10.1371/journal.pone.0133706] [Medline: 26197474]
95. Gu Y, Chen F, Liu T, Lv X, Shao Z, Lin H, et al. Early detection of an epidemic erythromelalgia outbreak using Baidu search data. Sci Rep 2015 Jul 28;5:12649 [FREE Full text] [doi: 10.1038/srep12649] [Medline: 26218589]

96. Hickmann KS, Fairchild G, Priedhorsky R, Generous N, Hyman JM, Deshpande A, et al. Forecasting the 2013-2014 influenza season using Wikipedia. PLoS Comput Biol 2015 May;11(5):e1004239 [FREE Full text] [doi: 10.1371/journal.pcbi.1004239] [Medline: 25974758]
97. Karmen C, Hsiung RC, Wetter T. Screening internet forum participants for depression symptoms by assembling and enhancing multiple NLP methods. Comput Methods Programs Biomed 2015 Jun;120(1):27-36. [doi: 10.1016/j.cmpb.2015.03.008] [Medline: 25891366]
98. Koopman B, Karimi S, Nguyen A, McGuire R, Muscatello D, Kemp M, et al. Automatic classification of diseases from free-text death certificates for real-time surveillance. BMC Med Inform Decis Mak 2015 Jul 15;15:53 [FREE Full text] [doi: 10.1186/s12911-015-0174-2] [Medline: 26174442]
99. Lee K, Agrawal A, Choudhary A. Mining Social Media Streams to Improve Public Health Allergy Surveillance. In: Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015. 2015 Presented at: ASONAM'15; August 25 - 28, 2015; Paris, France p. 815-822. [doi: 10.1145/2808797.2808896]
100. Mollema L, Harmsen IA, Broekhuizen E, Clijnk R, de Melker H, Paulussen T, et al. Disease detection or public opinion reflection? Content analysis of tweets, other social media, and online newspapers during the measles outbreak in The Netherlands in 2013. J Med Internet Res 2015 May 26;17(5):e128 [FREE Full text] [doi: 10.2196/jmir.3863] [Medline: 26013683]
101. Odlum M, Yoon S. What can we learn about the Ebola outbreak from tweets? Am J Infect Control 2015 Jun;43(6):563-571 [FREE Full text] [doi: 10.1016/j.ajic.2015.02.023] [Medline: 26042846]
102. Parker J, Yates A, Goharian N, Frieder O. Health-related hypothesis generation using social media data. Soc Netw Anal Min 2015;5:7. [doi: 10.1007/s13278-014-0239-8]
103. Pathak R, Poudel DR, Karmacharya P, Pathak A, Aryal MR, Mahmood M, et al. YouTube as a source of information on Ebola virus disease. N Am J Med Sci 2015 Jul;7(7):306-309 [FREE Full text] [doi: 10.4103/1947-2714.161244] [Medline: 26258077]
104. Pollett S, Wood N, Boscardin WJ, Bengtsson H, Schwarcz S, Harriman K, et al. Validating the use of Google Trends to enhance pertussis surveillance in California. PLoS Curr 2015 Oct 19;7 [FREE Full text] [doi: 10.1371/currents.outbreaks.7119696b3e7523faa4543faac87c56c2] [Medline: 26543674]
105. Ram S, Zhang W, Williams M, Pengetnze Y. Predicting asthma-related emergency department visits using big data. IEEE J Biomed Health Inform 2015 Jul;19(4):1216-1223. [doi: 10.1109/JBHI.2015.2404829] [Medline: 25706935]
106. Santillana M, Nguyen AT, Dredze M, Paul MJ, Nsoesie EO, Brownstein JS. Combining search, social media, and traditional data sources to improve influenza surveillance. PLoS Comput Biol 2015 Oct;11(10):e1004513 [FREE Full text] [doi: 10.1371/journal.pcbi.1004513] [Medline: 26513245]
107. Sueki H. The association of suicide-related Twitter use with suicidal behaviour: a cross-sectional study of young internet users in Japan. J Affect Disord 2015 Jan 1;170:155-160. [doi: 10.1016/j.jad.2014.08.047] [Medline: 25240843]
108. Thangarajan N, Green N, Gupta A, Little S, Weibel N. Analyzing Social Media to Characterize Local HIV At-Risk Populations. In: Proceedings of the conference on Wireless Health. 2015 Presented at: WH'15; October 14 - 16, 2015; Bethesda, Maryland p. 1-8. [doi: 10.1145/2811780.2811923]
109. Yom-tov E. Ebola data from the Internet: An Opportunity for Syndromic Surveillance or a News Event? In: Proceedings of the 5th International Conference on Digital Health 2015. 2015 Presented at: DH'15; May 18 - 20, 2015; Florence, Italy p. 115-119. [doi: 10.1145/2750511.2750512]
110. Wang H, Chen D, Yu H, Chen Y. Forecasting the incidence of dementia and dementia-related outpatient visits with Google Trends: evidence from Taiwan. J Med Internet Res 2015 Nov 19;17(11):e264 [FREE Full text] [doi: 10.2196/jmir.4516] [Medline: 26586281]
111. Yang S, Santillana M, Kou SC. Accurate estimation of influenza epidemics using Google search data via ARGO. Proc Natl Acad Sci U S A 2015 Nov 24;112(47):14473-14478 [FREE Full text] [doi: 10.1073/pnas.1515373112] [Medline: 26553980]
112. Yin Z, Fabbri D, Rosenbloom ST, Malin B. A scalable framework to detect personal health mentions on Twitter. J Med Internet Res 2015 Jun 5;17(6):e138 [FREE Full text] [doi: 10.2196/jmir.4305] [Medline: 26048075]
113. Yom-Tov E, Borsa D, Hayward AC, McKendry RA, Cox IJ. Automatic identification of web-based risk markers for health events. J Med Internet Res 2015 Jan 27;17(1):e29 [FREE Full text] [doi: 10.2196/jmir.4082] [Medline: 25626480]
114. Yuan Q, Nsoesie EO, Lv B, Peng G, Chunara R, Brownstein JS. Monitoring influenza epidemics in china with search query from baidu. PLoS One 2013;8(5):e64323 [FREE Full text] [doi: 10.1371/journal.pone.0064323] [Medline: 23750192]
115. Zaldumbide J, Sinnott RO. Identification and Validation of Real-Time Health Events through Social Media. In: Proceedings of the 2015 IEEE International Conference on Data Science and Data Intensive Systems. 2015 Presented at: DSDIS'15; December 11-13, 2015; Sydney, Australia. [doi: 10.1109/dsdis.2015.27]
116. Zhang EX, Yang Y, di Shang R, Simons JJ, Quek BK, Yin XF, et al. Leveraging social networking sites for disease surveillance and public sensing: the case of the 2013 avian influenza A(H7N9) outbreak in China. Western Pac Surveill Response J 2015;6(2):66-72 [FREE Full text] [doi: 10.5365/WPSAR.2015.6.1.013] [Medline: 26306219]
117. Zuccon G, Khanna S, Nguyen A, Boyle J, Hamlet M, Cameron M. Automatic detection of tweets reporting cases of influenza like illnesses in Australia. Health Inf Sci Syst 2015;3(Suppl 1 HISA Big Data in Biomedicine and Healthcare 2013 Con):S4 [FREE Full text] [doi: 10.1186/2047-2501-3-S1-S4] [Medline: 25870759]

118. Allen C, Tsou M, Aslam A, Nagel A, Gawron J. Applying GIS and machine learning methods to Twitter data for multiscale surveillance of influenza. PLoS One 2016;11(7):e0157734 [FREE Full text] [doi: 10.1371/journal.pone.0157734] [Medline: 27455108]
119. Bakker KM, Martinez-Bakker ME, Helm B, Stevenson TJ. Digital epidemiology reveals global childhood disease seasonality and the effects of immunization. Proc Natl Acad Sci U S A 2016 Jun 14;113(24):6689-6694 [FREE Full text] [doi: 10.1073/pnas.1523941113] [Medline: 27247405]
120. Byrd K, Mansurov A, Baysal O. Mining Twitter Data for Influenza Detection and Surveillance. In: Proceedings of the International Workshop on Software Engineering in Healthcare Systems. 2016 Presented at: SEHS'16; May 14 - 22, 2016; Austin, Texas p. 43-49. [doi: 10.1145/2897683.2897693]
121. Chen L, Hossain KS, Butler P, Ramakrishnan N, Prakash BA. Syndromic surveillance of Flu on Twitter using weakly supervised temporal topic models. Data Min Knowl Discov 2016;30(3):681-710. [doi: 10.1007/s10618-015-0434-x]
122. Deiner MS, Lietman TM, McLeod SD, Chodosh J, Porco TC. Surveillance tools emerging from search engines and social media data for determining eye disease patterns. JAMA Ophthalmol 2016 Sep 1;134(9):1024-1030 [FREE Full text] [doi: 10.1001/jamaophthalmol.2016.2267] [Medline: 27416554]
123. Gosh S, Chakraborty P, Cohn E, Brownstein JS, Ramakrishnan N. Characterizing Diseases from Unstructured Text: A Vocabulary Driven Word2vec Approach. In: Proceedings of the 25th ACM International on Conference on Information and Knowledge Management. 2016 Presented at: CIKM'16; October 24 - 28, 2016; Indianapolis, Indiana p. 1129-1138. [doi: 10.1145/2983323.2983362]
124. Ji X, Chun SA, Geller J. Knowledge-based tweet classification for disease sentiment monitoring. In: Sentiment Analysis and Ontology Engineering. Cham: Springer; 2016:425-454.
125. Klembczyk JJ, Jalalpour M, Levin S, Washington RE, Pines JM, Rothman RE, et al. Google Flu Trends spatial variability validated against emergency department influenza-related visits. J Med Internet Res 2016 Jun 28;18(6):e175 [FREE Full text] [doi: 10.2196/jmir.5585] [Medline: 27354313]
126. Ling R, Lee J. Disease monitoring and health campaign evaluation using Google search activities for HIV and AIDS, stroke, colorectal cancer, and marijuana use in Canada: a retrospective observational study. JMIR Public Health Surveill 2016 Oct 12;2(2):e156 [FREE Full text] [doi: 10.2196/publichealth.6504] [Medline: 27733330]
127. Majumder MS, Santillana M, Mekaru SR, McGinnis DP, Khan K, Brownstein JS. Utilizing nontraditional data sources for near real-time estimation of transmission dynamics during the 2015-2016 Colombian Zika virus disease outbreak. JMIR Public Health Surveill 2016 Jun 1;2(1):e30 [FREE Full text] [doi: 10.2196/publichealth.5814] [Medline: 27251981]
128. Martin LJ, Lee BE, Yasui Y. Google Flu Trends in Canada: a comparison of digital disease surveillance data with physician consultations and respiratory virus surveillance data, 2010-2014. Epidemiol Infect 2016 Jan;144(2):325-332. [doi: 10.1017/S0950268815001478] [Medline: 26135239]
129. Moss R, Zarebski A, Dawson P, McCaw JM. Forecasting influenza outbreak dynamics in Melbourne from internet search query surveillance data. Influenza Other Respir Viruses 2016 Jul;10(4):314-323 [FREE Full text] [doi: 10.1111/irv.12376] [Medline: 26859411]
130. Pollett S, Boscardin WJ, Azziz-Baumgartner E, Tinoco YO, Soto G, Romero C, et al. Evaluating Google Flu Trends in Latin America: important lessons for the next phase of digital disease detection. Clin Infect Dis 2017 Jan 1;64(1):34-41 [FREE Full text] [doi: 10.1093/cid/ciw657] [Medline: 27678084]
131. Priest C, Knopf A, Groves D, Carpenter JS, Furrey C, Krishnan A, et al. Finding the patient's voice using Big Data: analysis of users' health-related concerns in the ChaCha Question-and-Answer Service (2009-2012). J Med Internet Res 2016 Mar 9;18(3):e44 [FREE Full text] [doi: 10.2196/jmir.5033] [Medline: 26960745]
132. Radzikowski J, Stefanidis A, Jacobsen KH, Croitoru A, Crooks A, Delamater PL. The measles vaccination narrative in Twitter: a quantitative analysis. JMIR Public Health Surveill 2016;2(1):e1 [FREE Full text] [doi: 10.2196/publichealth.5059] [Medline: 27227144]
133. Sharpe JD, Hopkins RS, Cook RL, Striley CW. Evaluating Google, Twitter, and Wikipedia as tools for influenza surveillance using Bayesian change point analysis: a comparative analysis. JMIR Public Health Surveill 2016 Oct 20;2(2):e161 [FREE Full text] [doi: 10.2196/publichealth.5901] [Medline: 27765731]
134. Shin SY, Kim T, Seo DW, Sohn CH, Kim S, Ryoo SM, et al. Correlation between National Influenza Surveillance Data and Search Queries from Mobile Devices and Desktops in South Korea. PLoS One 2016;11(7):e0158539 [FREE Full text] [doi: 10.1371/journal.pone.0158539] [Medline: 27391028]
135. Sidana S, Amer-Yahia S, Clausel M, Rebai M, Mai S, Amini M. Health Monitoring on Social Media over Time. IEEE Trans Knowl Data Eng 2018 Aug;30(8):1467-1480. [doi: 10.1109/TKDE.2018.2795606]
136. Tourassi G, Yoon HJ, Xu S. A novel web informatics approach for automated surveillance of cancer mortality trends. J Biomed Inform 2016 Jun;61:110-118 [FREE Full text] [doi: 10.1016/j.jbi.2016.03.027] [Medline: 27044930]
137. Woo H, Cho Y, Shim E, Lee J, Lee C, Kim SH. Estimating influenza outbreaks using both search engine query data and social media data in South Korea. J Med Internet Res 2016 Jul 4;18(7):e177 [FREE Full text] [doi: 10.2196/jmir.4955] [Medline: 27377323]

138. Zhang W, Ram S, Burkart M, Pengetnze Y. Extracting Signals from Social Media for Chronic Disease Surveillance. In: Proceedings of the 6th International Conference on Digital Health Conference. 2016 Presented at: DH'16; April 11 - 13, 2016; Montreal, Canada p. 79-83. [doi: 10.1145/2896338.2897728]
139. Zou B, Gorton R, Cox IJ, Lampos V. On Infectious Intestinal Disease Surveillance using Social Media Content. In: Proceedings of the 6th International Conference on Digital Health Conference. 2016 Presented at: DH'16; April 11 - 13, 2016; Montreal, Canada p. 157-161. [doi: 10.1145/2896338.2896372]
140. Albinati J, Meira W, Pappa GL, Teixeira M, Marques-toledo C. Enhancement of Epidemiological Models for Dengue Fever Based on Twitter Data. In: Proceedings of the 2017 International Conference on Digital Health. 2017 Presented at: DH'17; July 2 - 5, 2017; London, United Kingdom p. 109-118. [doi: 10.1145/3079452.3079464]
141. Chiu APY, Lin Q, He D. News trends and web search query of HIV/AIDS in Hong Kong. PLoS One 2017;12(9):e0185004 [FREE Full text] [doi: 10.1371/journal.pone.0185004] [Medline: 28922376]
142. Dai X, Bikdash M, Meyer B. From Social Media to Public Health Surveillance: Word Embedding Based Clustering Method for Twitter Classification. In: Proceedings of the IEEE SoutheastCon 2017. 2017 Presented at: SoutheastCon'17; March 30 - April 2, 2017; Charlotte, NC, USA. [doi: 10.1109/secon.2017.7925400]
143. Du Z, Xu L, Zhang W, Zhang D, Yu S, Hao Y. Predicting the hand, foot, and mouth disease incidence using search engine query data and climate variables: an ecological study in Guangdong, China. BMJ Open 2017 Oct 6;7(10):e016263 [FREE Full text] [doi: 10.1136/bmjopen-2017-016263] [Medline: 28988169]
144. Fairchild G, del Valle SY, de Silva L, Segre A. Eliciting disease data from Wikipedia articles. Proc Int AAAI Conf Weblogs Soc Media 2015 May;2015:26-33 [FREE Full text] [doi: 10.5210/ojphi.v8i1.6526] [Medline: 28721308]
145. Fung IC, Zeng J, Chan C, Liang H, Yin J, Liu Z, et al. Twitter and Middle East respiratory syndrome, South Korea, 2015: A multi-lingual study. Infect Dis Health 2018 Mar;23(1):10-16. [doi: 10.1016/j.idh.2017.08.005] [Medline: 30479298]
146. Ghosh S, Chakraborty P, Nsoesie EO, Cohn E, Mekaru SR, Brownstein JS, et al. Temporal topic modeling to assess associations between news trends and infectious disease outbreaks. Sci Rep 2017 Jan 19;7:40841 [FREE Full text] [doi: 10.1038/srep40841] [Medline: 28102319]
147. Guo P, Zhang J, Wang L, Yang S, Luo G, Deng C, et al. Monitoring seasonal influenza epidemics by using internet search data with an ensemble penalized regression model. Sci Rep 2017 Apr 19;7:46469 [FREE Full text] [doi: 10.1038/srep46469] [Medline: 28422149]
148. Haghighi PD, Kang Y, Buchbinder R, Burstein F, Whittle S. Investigating subjective experience and the influence of weather among individuals with fibromyalgia: a content analysis of Twitter. JMIR Public Health Surveill 2017 Jan 19;3(1):e4 [FREE Full text] [doi: 10.2196/publichealth.6344] [Medline: 28104577]
149. Hartley DM, Giannini CM, Wilson S, Frieder O, Margolis PA, Kotagal UR, et al. Coughing, sneezing, and aching online: Twitter and the volume of influenza-like illness in a pediatric hospital. PLoS One 2017;12(7):e0182008 [FREE Full text] [doi: 10.1371/journal.pone.0182008] [Medline: 28753678]
150. Kandula S, Hsu D, Shaman J. Subregional nowcasts of seasonal influenza using search trends. J Med Internet Res 2017 Nov 6;19(11):e370 [FREE Full text] [doi: 10.2196/jmir.7486] [Medline: 29109069]
151. Lampos V, Zou B, Cox I. Enhancing Feature Selection Using Word Embeddings: The Case of Flu Surveillance. In: Proceedings of the 26th International Conference on World Wide Web. 2017 Presented at: WWW'17; April 3 - 7, 2017; Perth, Australia p. 695-704. [doi: 10.1145/3038912.3052622]
152. Lee K, Agrawal A, Choudhary A. Forecasting Influenza Levels using Real-Time Social Media Streams. In: 2017 IEEE International Conference on Healthcare Informatics. 2017 Presented at: ICHI'17; August 23-26, 2017; Park City, UT, USA. [doi: 10.1109/ichi.2017.68]
153. Li Z, Liu T, Zhu G, Lin H, Zhang Y, He J, et al. Dengue Baidu Search Index data can improve the prediction of local dengue epidemic: a case study in Guangzhou, China. PLoS Negl Trop Dis 2017 Mar;11(3):e0005354 [FREE Full text] [doi: 10.1371/journal.pntd.0005354] [Medline: 28263988]
154. Liu K, Huang S, Miao Z, Chen B, Jiang T, Cai G, et al. Identifying potential norovirus epidemics in China via internet surveillance. J Med Internet Res 2017 Aug 8;19(8):e282 [FREE Full text] [doi: 10.2196/jmir.7855] [Medline: 28790023]
155. McGough SF, Brownstein JS, Hawkins JB, Santillana M. Forecasting Zika incidence in the 2016 Latin America outbreak combining traditional disease surveillance with search, social media, and news report data. PLoS Negl Trop Dis 2017 Jan;11(1):e0005295 [FREE Full text] [doi: 10.1371/journal.pntd.0005295] [Medline: 28085877]
156. Morsy S, Dang TN, Kamel MG, Zayan AH, Makram OM, Elhady M, et al. Prediction of Zika-confirmed cases in Brazil and Colombia using Google Trends. Epidemiol Infect 2018 Oct;146(13):1625-1627. [doi: 10.1017/S0950268818002078] [Medline: 30056812]
157. Mowery D, Smith H, Cheney T, Stoddard G, Coppersmith G, Bryan C, et al. Understanding depressive symptoms and psychosocial stressors on Twitter: a corpus-based study. J Med Internet Res 2017 Feb 28;19(2):e48 [FREE Full text] [doi: 10.2196/jmir.6895] [Medline: 28246066]
158. Noll-Hussong M. Whiplash syndrome reloaded: digital echoes of whiplash syndrome in the European internet search engine context. JMIR Public Health Surveill 2017 Mar 27;3(1):e15 [FREE Full text] [doi: 10.2196/publichealth.7054] [Medline: 28347974]

159. Priedhorsky R, Osthus D, Daughton A, Moran K, Generous N, Fairchild G, et al. Measuring Global Disease with Wikipedia: Success, Failure, and a Research Agenda. In: Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing. 2017 Presented at: CSCW'17; February 25 - March 1, 2017; Portland, Oregon, USA p. 1812-1834 URL: http://europepmc.org/abstract/MED/28782059 [doi: 10.1145/2998181.2998183]
160. Reichert JR, Kristensen KL, Mukkamala RR, Vatrapu R. A Supervised Machine Learning Study of Online Discussion Forums about Type-2 Diabetes. In: 2017 IEEE 19th International Conference on e-Health Networking, Applications and Services. 2017 Presented at: Healthcom'17; October 12-15, 2017; Dalian, China. [doi: 10.1109/healthcom.2017.8210815]
161. Roccetti M, Marfia G, Salomoni P, Prandi C, Zagari RM, Kengni FL, et al. Attitudes of Crohn's disease patients: infodemiology case study and sentiment analysis of Facebook and Twitter posts. JMIR Public Health Surveill 2017 Aug 9;3(3):e51 [FREE Full text] [doi: 10.2196/publichealth.7004] [Medline: 28793981]
162. Rudra K, Imran M, Sharma A, Ganguly N. Classifying Information from Microblogs during Epidemics. In: Proceedings of the 2017 International Conference on Digital Health. 2017 Presented at: DH'17; July 2 - 5, 2017; London, United Kingdom p. 104-108. [doi: 10.1145/3079452.3079491]
163. Samaras L, García-Barriocanal E, Sicilia M. Syndromic surveillance models using web data: the case of influenza in Greece and Italy using Google Trends. JMIR Public Health Surveill 2017 Nov 20;3(4):e90 [FREE Full text] [doi: 10.2196/publichealth.8015] [Medline: 29158208]
164. Strauss RA, Castro JS, Reintjes R, Torres JR. Google dengue trends: an indicator of epidemic behavior. The Venezuelan Case. Int J Med Inform 2017 Aug;104:26-30. [doi: 10.1016/j.ijmedinf.2017.05.003] [Medline: 28599813]
165. Teng Y, Bi D, Xie G, Jin Y, Huang Y, Lin B, et al. Dynamic Forecasting of Zika Epidemics Using Google Trends. PLoS One 2017;12(1):e0165085 [FREE Full text] [doi: 10.1371/journal.pone.0165085] [Medline: 28060809]
166. Tkachenko N, Chotvijit S, Gupta N, Bradley E, Gilks C, Guo W, et al. Google Trends can improve surveillance of Type 2 diabetes. Sci Rep 2017 Jul 10;7(1):4993 [FREE Full text] [doi: 10.1038/s41598-017-05091-9] [Medline: 28694479]
167. Marques-Toledo CD, Degener CM, Vinhal L, Coelho G, Meira W, Codeço CT, et al. Dengue prediction by the web: tweets are a useful tool for estimating and forecasting Dengue at country and city level. PLoS Negl Trop Dis 2017 Jul;11(7):e0005729 [FREE Full text] [doi: 10.1371/journal.pntd.0005729] [Medline: 28719659]
168. Xu Q, Gel YR, Ramirez LL, Nezafati K, Zhang Q, Tsui K. Forecasting influenza in Hong Kong with Google search queries and statistical model fusion. PLoS One 2017;12(5):e0176690 [FREE Full text] [doi: 10.1371/journal.pone.0176690] [Medline: 28464015]
169. Zhang K, Arablouei R, Jurdak R. Predicting Prevalence of Influenza-Like Illness From Geo-Tagged Tweets. In: Proceedings of the 26th International Conference on World Wide Web Companion. 2017 Presented at: WWW'17; April 3 - 7, 2017; Perth, Australia p. 1327-1334. [doi: 10.1145/3041021.3051150]
170. Zhang Q, Perra N, Perrotta D, Tizzoni M, Paolotti D, Vespignani A. Forecasting Seasonal Influenza Fusing Digital Indicators and a Mechanistic Disease Model. In: Proceedings of the 26th International Conference on World Wide Web. 2017 Presented at: WWW'17; April 3 - 7, 2017; Perth, Australia p. 311-319. [doi: 10.1145/3038912.3052678]
171. Zhang Y, Milinovich G, Xu Z, Bambrick H, Mengersen K, Tong S, et al. Monitoring pertussis infections using internet search queries. Sci Rep 2017 Sep 5;7(1):10437 [FREE Full text] [doi: 10.1038/s41598-017-11195-z] [Medline: 28874880]
172. Barros JM, Duggan J, Rebholz-Schuhmann D. Disease mentions in airport and hospital geolocations expose dominance of news events for disease concerns. J Biomed Semantics 2018 Jun 12;9(1):18 [FREE Full text] [doi: 10.1186/s13326-018-0186-9] [Medline: 29895320]
173. Brigo F, Lattanzi S, Bragazzi N, Nardone R, Moccia M, Lavorgna L. Why do people search Wikipedia for information on multiple sclerosis? Mult Scler Relat Disord 2018 Feb;20:210-214. [doi: 10.1016/j.msard.2018.02.001] [Medline: 29428464]
174. Chen TH, Chen YC, Chen JL, Chang FC. Flu Trend Prediction Based on Massive Data Analysis. In: Proceedings of the 2018 IEEE 3rd International Conference on Cloud Computing and Big Data Analysis. 2018 Presented at: ICCCBDA'18; April 20-22, 2018; Chengdu, China. [doi: 10.1109/icccbda.2018.8386532]
175. Chen B, Shao J, Liu K, Cai G, Jiang Z, Huang Y, et al. Does eating chicken feet with pickled peppers cause avian influenza? Observational case study on Chinese social media during the avian influenza A (H7N9) outbreak. JMIR Public Health Surveill 2018 Mar 29;4(1):e32 [FREE Full text] [doi: 10.2196/publichealth.8198] [Medline: 29599109]
176. Gianfredi V, Bragazzi NL, Mahamid M, Bisharat B, Mahroum N, Amital H, et al. Monitoring public interest toward pertussis outbreaks: an extensive Google Trends-based analysis. Public Health 2018 Dec;165:9-15. [doi: 10.1016/j.puhe.2018.09.001] [Medline: 30342281]
177. Ho HT, Carvajal TM, Bautista JR, Capistrano JD, Viacrusis KM, Hernandez LF, et al. Using Google Trends to examine the spatio-temporal incidence and behavioral patterns of dengue disease: a case study in Metropolitan Manila, Philippines. Trop Med Infect Dis 2018 Nov 11;3(4):pii: E118 [FREE Full text] [doi: 10.3390/tropicalmed3040118] [Medline: 30423898]
178. Kagashe I, Yan Z, Suheryani I. Enhancing seasonal influenza surveillance: topic analysis of widely used medicinal drugs using Twitter data. J Med Internet Res 2017 Sep 12;19(9):e315 [FREE Full text] [doi: 10.2196/jmir.7393] [Medline: 28899847]
179. Karisani P, Agichtein E. Did You Really Just Have a Heart Attack?: Towards Robust Detection of Personal Health Mentions in Social Media. In: Proceedings of the 2018 World Wide Web Conference. 2018 Presented at: WWW'18; April 23 - 27, 2018; Lyon, France p. 137-146. [doi: 10.1145/3178876.3186055]

180. Livelo ED, Cheng C. Intelligent Dengue Infoveillance Using Gated Recurrent Neural Learning and Cross-Label Frequencies. In: Proceedings of the 2018 IEEE International Conference on Agents. 2018 Presented at: ICA'18; July 28-31, 2018; Singapore. [doi: 10.1109/agents.2018.8459963]
181. Mavragani A, Ochoa G. Infoveillance of infectious diseases in USA: STDs, tuberculosis, and hepatitis. J Big Data 2018;5:30. [doi: 10.1186/s40537-018-0140-9]
182. Nan Y, Gao Y. A machine learning method to monitor China's AIDS epidemics with data from Baidu trends. PLoS One 2018;13(7):e0199697 [FREE Full text] [doi: 10.1371/journal.pone.0199697] [Medline: 29995920]
183. Phillips CA, Leahy AB, Li Y, Schapira MM, Bailey LC, Merchant RM. Relationship between state-level Google online search volume and cancer incidence in the United States: retrospective study. J Med Internet Res 2018 Jan 8;20(1):e6 [FREE Full text] [doi: 10.2196/jmir.8870] [Medline: 29311051]
184. Shah MP, Lopman BA, Tate JE, Harris J, Esparza-Aguilar M, Sanchez-Uribe E, et al. Use of internet search data to monitor rotavirus vaccine impact in the United States, United Kingdom, and Mexico. J Pediatric Infect Dis Soc 2018 Feb 19;7(1):56-63 [FREE Full text] [doi: 10.1093/jpids/pix004] [Medline: 28369477]
185. Verma M, Kishore K, Kumar M, Sondh AR, Aggarwal G, Kathirvel S. Google search trends predicting disease outbreaks: an analysis from India. Healthc Inform Res 2018 Oct;24(4):300-308 [FREE Full text] [doi: 10.4258/hir.2018.24.4.300] [Medline: 30443418]
186. Wakamiya S, Kawai Y, Aramaki E. Twitter-based influenza detection after flu peak via tweets with indirect information: text mining study. JMIR Public Health Surveill 2018 Sep 25;4(3):e65 [FREE Full text] [doi: 10.2196/publichealth.8627] [Medline: 30274968]
187. Young SD, Torrone EA, Urata J, Aral SO. Using search engine data as a tool to predict syphilis. Epidemiology 2018 Jul;29(4):574-578 [FREE Full text] [doi: 10.1097/EDE.0000000000000836] [Medline: 29864105]
188. Zou B, Lampos V, Cox I. Multi-Task Learning Improves Disease Models from Web Search. In: Proceedings of the 2018 World Wide Web Conference. 2018 Presented at: WWW'18; April 23 - 27, 2018; Lyon, France p. 87-96. [doi: 10.1145/3178876.3186050]
189. Flu Near You. URL: https://flunearyou.org [accessed 2019-09-16]
190. Statista. 2018. Worldwide Desktop Market Share of Leading Search Engines From January 2010 to July 2019 URL: https://www.statista.com/statistics/216573/worldwide-market-share-of-search-engines/ [accessed 2018-08-28]
191. Statcounter. Desktop, Mobile & Console Search Engine Market Share Portugal, 2009-2019 URL: https://gs.statcounter.com/search-engine-market-share/desktop-mobile-console/portugal/#yearly-2009-2019-bar [accessed 2019-09-04]
192. Chen S, Zhang H, Lin M, Lv S. Comparision of Microblogging Service Between Sina Weibo and Twitter. In: Proceedings of 2011 International Conference on Computer Science and Network Technology. 2011 Presented at: ICCSNT'11; December 24-26, 2011; Harbin, China. [doi: 10.1109/iccsnt.2011.6182424]
193. Eysenbach G. Infodemiology and infoveillance tracking online health information and cyberbehavior for public health. Am J Prev Med 2011 May;40(5 Suppl 2):S154-S158. [doi: 10.1016/j.amepre.2011.02.006] [Medline: 21521589]
194. Kass-Hout TA, Alhinnawi H. Social media in public health. Br Med Bull 2013;108:5-24. [doi: 10.1093/bmb/ldt028] [Medline: 24103335]
195. Diaz-Aviles E, Stewart A, Velasco E, Denecke K, Nejdl W. arXiv e-Print archive. 2012. Epidemic Intelligence for the Crowd, by the Crowd (Full Version) URL: http://arxiv.org/abs/1203.1378 [accessed 2020-01-23]
196. Kroll M, Phalkey RK, Kraas F. Challenges to the surveillance of non-communicable diseases--a review of selected approaches. BMC Public Health 2015 Dec 16;15:1243 [FREE Full text] [doi: 10.1186/s12889-015-2570-z] [Medline: 26672992]
197. Pew Research Center. 2013 Feb 12. The Internet and Health URL: http://www.pewinternet.org/2013/02/12/the-internet-and-health/ [accessed 2018-08-28]
198. Omnicore. 2018. Twitter by the Numbers: Stats, Demographics & Fun Facts URL: https://www.omnicoreagency.com/twitter-statistics/ [accessed 2018-08-28]
199. Zickuhr K, Rainie L. Pew Research Center. 2011 Jan 13. Wikipedia, Past and Present: A Snapshot of Current Wikipedia Users URL: http://www.pewinternet.org/2011/01/13/wikipedia-past-and-present/ [accessed 2020-01-23]
200. Butler D. When Google got flu wrong. Nature 2013 Feb 14;494(7436):155-156. [doi: 10.1038/494155a] [Medline: 23407515]
201. Lazer D, Kennedy R, King G, Vespignani A. Big data. The parable of Google Flu: traps in big data analysis. Science 2014 Mar 14;343(6176):1203-1205. [doi: 10.1126/science.1248506] [Medline: 24626916]

Abbreviations

IBS: internet-based source
ILI: influenza-like illness
NCD: noncommunicable disease
P2P: peer-to-peer
RQ: research question
SFI: Science Foundation Ireland

Edited by G Eysenbach; submitted 24.02.19; peer-reviewed by V Gianfredi, D Paolotti, D Gunasekeran, N Bragazzi; comments to author 18.05.19; revised version received 18.09.19; accepted 26.11.19; published 13.03.20

Please cite as:
Barros JM, Duggan J, Rebholz-Schuhmann D
The Application of Internet-Based Sources for Public Health Surveillance (Infoveillance): Systematic Review
J Med Internet Res 2020;22(3):e13680
URL: http://www.jmir.org/2020/3/e13680/
doi: 10.2196/13680
PMID: