# Role of Participatory Health Informatics in Detecting and Managing Pandemics: Literature Review

Elia Gabarron[1*], Octavio Rivera-Romero [2*], Talya Miron-Shatz[3,4], Rebecca Grainger[5], Kerstin Denecke[6]

[1] Norwegian Centre for E-health Research, University Hospital of North Norway, Tromsø, Norway
[2] Department of Electronic Technology, Universidad de Sevilla, Spain
[3] Faculty of Business Administration, Ono Academic College, Israel
[4] Winton Centre for Risk and Evidence Communication, Cambridge University, England
[5] Department of Medicine, University of Otago, Wellington, New Zealand
[6] Institute for Medical Informatics, Bern University of Applied Sciences, Bern, Switzerland

## Summary

**Objectives**: Using participatory health informatics (PHI) to detect disease outbreaks or learn about pandemics has gained interest in recent years. However, the role of PHI in understanding and managing pandemics, citizens' role in this context, and which methods are relevant for collecting and processing data are still unclear, as is which types of data are relevant. This paper aims to clarify these issues and explore the role of PHI in managing and detecting pandemics.

**Methods**: Through a literature review we identified studies that explore the role of PHI in detecting and managing pandemics. Studies from five databases were screened: PubMed, CINAHL (Cumulative Index to Nursing and Allied Health Literature), IEEE Xplore, ACM (Association for Computing Machinery) Digital Library, and Cochrane Library. Data from studies fulfilling the eligibility criteria were extracted and synthesized narratively.

**Results**: Out of 417 citations retrieved, 53 studies were included in this review. Most research focused on influenza-like illnesses or COVID-19 with at least three papers on other epidemics (Ebola, Zika or measles). The geographic scope ranged from global to concentrating on specific countries. Multiple processing and analysis methods were reported, although often missing relevant information. The majority of outcomes are reported for two application areas: crisis communication and detection of disease outbreaks.

**Conclusions**: For most diseases, the small number of studies prevented reaching firm conclusions about the utility of PHI in detecting and monitoring these disease outbreaks. For others, e.g., COVID-19, social media and online search patterns corresponded to disease patterns, and detected disease outbreak earlier than conventional public health methods, thereby suggesting that PHI can contribute to disease and pandemic monitoring.

## Keywords

Epidemics, public health surveillance, social media

Yearb Med Inform 2021:
http://dx.doi.org/10.1055/s-0041-1726486



## 1 Introduction

Detecting the spread of infection in an epidemic allows health systems and governments to implement timely public health interventions. The term 'epidemic intelligence' refers to "all the activities related to early identification of potential health threats, their verification, assessment and investigation in order to recommend public health measures to control them" [1]. Traditional epidemic intelligence systems mainly use clinical epidemiological data, such as reports from hospitals and healthcare providers, which often lead to time delays. In the last 20 years, the use of journalistic and other unofficial online information sources for epidemic intelligence has gained interest. News aggregators such as HealthMap [2] or MediSys [3] collect online news and information from websites or blogs to provide an overview of the worldwide disease situation for the purpose of real-time surveillance. Data are also collected from search engines or messages on social media, such as Twitter. These data, which are now being utilized as a form of participatory health informatics (PHI), may provide a complementary source of information to traditional sources such as health system, thereby helping to detect and predict the volume and spread of infection in epidemics [4].

PHI is a multidisciplinary field that uses information technology as provided through the web, smartphones, or wearables to increase participation of individuals in their care process and to enable them in practicing self-care and shared decision-making [5]. PHI deals with the resources, devices, and methods required to support active participation and engagement of the stakeholders, such as social media [5]. Goals to be achieved through PHI include improving and maintaining health and well-being; improving the healthcare system and health outcomes; sharing experiences; achieving life goals; and gaining self-education [6]. Beyond eliciting epidemic intelligence, participatory health is used to engage or inform citizens of disease outbreaks or governmental activities related to an outbreak.

[*] Contributed equally and share first authorship

The COVID-19 pandemic has highlighted the potential role of PHI in pandemics. For example, during the COVID-19 crisis Chinese central government agencies used social media to promote citizen engagement [4, 7]. Other studies during the COVID-19 pandemic have used online data to assess citizens' risk perceptions or attitudes and opinions related to the pandemic [8, 9].

Given these research developments, we aim to examine which methods and features of PHI are considered, and which roles PHI plays in assessing, managing and controlling pandemics. Furthermore, we aimed to identify and summarize the research about what roles citizens play in this process. Specifically, we aim to use a literature search and synthesis to address the following research questions:

- Which epidemics have been studied by means of a PHI approach to disease surveillance?
- Which tools of PHI and which methods are used to analyze citizens' contributions?
- Does citizen input correspond with epidemic data?
- What are the barriers to the use of social media for pandemic detection and management?

# 2 Methods

We undertook a literature review to identify studies that help in answering the listed questions above. The Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) criteria guided the conduct and reporting of the review [10].

## 2.1 Search Strategy

The full search was carried out on June 10th, 2020 (see Appendix 1). The search covered PubMed; ACM Digital Library; IEEE Xplore; CINAHL and Cochrane library using the following keywords:

- **Keywords related to epidemics or pandemics**: Epidemics (MeSH) OR Pandemics (MeSH) OR Disease Outbreaks (MeSH) OR Chikungunya OR Cholera OR Crimean-Congo haemorrhagic fever OR Ebola virus disease OR Hendra virus infection OR Influenza OR Lassa fever OR Marburg virus disease OR Meningitis OR MERS-CoV OR Monkeypox OR Nipah virus infection OR coronavirus OR 2019-nCoV OR Covid OR Plague OR Rift Valley fever OR SARS OR Smallpox OR Tularaemia OR Yellow fever OR Zika virus disease.
- **Keywords related to participatory health**: Social media OR social network site OR online social network OR online community OR Facebook OR Twitter OR YouTube OR Instagram OR WhatsApp OR mHealth OR mobile health OR e-health OR ehealth OR mobile applications OR apps.
- **Keywords related to treatment / interventions**: Management OR Detection OR surveillance OR Infoveillance OR infodemic.

## 2.2 Eligibility

We uploaded all search references to Rayyan (https://rayyan.qcri.org) and removed duplicates. To assess the eligibility of the articles, in a first step, all titles and abstracts were divided among two reviewers (EG, OR) where each reviewer looked at half of the papers. After title and abstract screening, in a second step, full text of all potentially eligible articles was obtained, and articles were reviewed to confirm their eligibility by two reviewers independently (EG, OR). Conflicts were discussed with a third reviewer (KD) until consensus was reached.

## 2.3 Inclusion and Exclusion Criteria

Articles were included if they: a) focused on epidemics, disease outbreaks or any of the 20 pandemic diseases recognized by WHO (https://www.who.int/emergencies/diseases/en/); b) addressed the role or features of social media, mobile health, or other PHI; c) were primary studies reporting results; and d) were in the English language. Articles were excluded if they: did not deal with epidemics, outbreak diseases or pandemics; did not deal with PHI; were not primary studies or did not report results (i.e., study protocols, opinion, frameworks or review papers); or were published in languages other than English.

The selected articles were divided among three authors (EG, KD, OR) for data extraction. We extracted: 1) Disease/epidemic, settings and country; 2) Objective, data source, type of information provided, user group, epidemic and considered region; 3) Data preprocessing, analysis techniques and features; 4) Outcome and reasons that limit the outcome. Data were abstracted into a Microsoft Excel spreadsheet standardized for this review, piloted and refined with 10 preliminary papers. The selected articles were included in the narrative synthesis.

# 3 Results

## 3.1 Sample

The database search retrieved 461 records, with 417 records remaining after duplicate removal; 53 papers met the inclusion criteria after full text review and were included in the qualitative synthesis (Figure 1). A summary of the included studies can be found in Appendix 2.

## 3.2 Which Epidemics Have Been Studied?

The 53 papers considered various epidemics; influenza-like illnesses (27/53; 51%) and COVID-19 (11/53; 19%) were the most frequently studied epidemics (Appendix 3). Almost all studies analyzed retrospectively collected social media data where contributors were unaware of the usage of their content for the purpose of epidemic surveillance. In all these papers, the data sets were created based on predefined keywords such as example tweets collected using keywords like flu, influenza etc. (e.g., [11,12]). Only one study was conducted prospectively [13] with data actively generated by users.

About a quarter (14/53; 26%) of the papers had a global scope, with 10/53 (19%) focused on the USA and Canada, and 7/53 (13%) on China. Australia and Malaysia

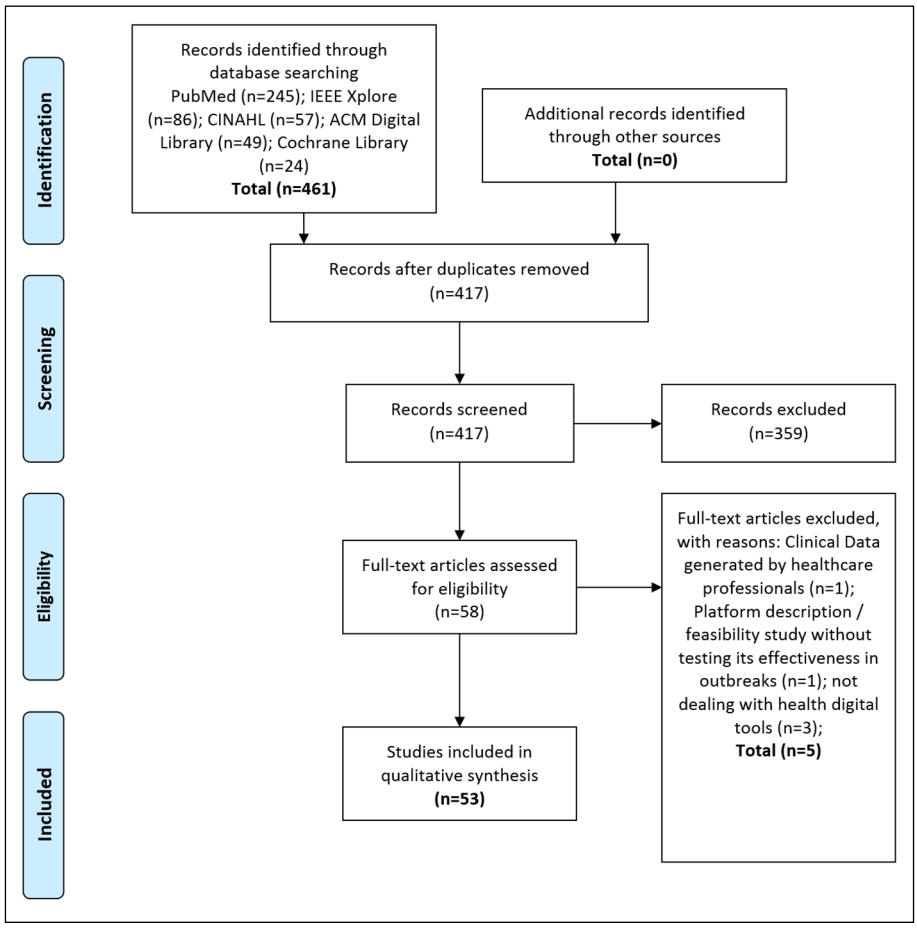


**Fig. 1** PRISMA Flowchart of the paper selection procedure.

were considered a region in three papers (6%). Japan, Madagascar, the Netherlands, and Italy were the focus of two papers each (4%). Greece or the UK were target locations of one paper (2%). Just over one in ten papers (6/53, 11%) did not specify a region.

The objectives of the 53 studies were classified into six main categories: analyzing contents related to epidemics (i.e., reactions, opinion, attitudes, quality of the information, sentiment analysis, distribution patterns, etc.); detecting disease; disease monitoring or disease surveillance; comparing number of posts with official disease numbers; predicting future epidemics; and tracing contacts (Appendix 4). The three most common objectives were analyzing content related to the epidemic (20/53; 38%), detecting disease (12/53; 23%), and monitoring or surveillance (10/53; 19%).

## 3.3 Which Tools of PHI Were Examined, and Which Methods Were Used?

Social media was the data source in the majority of the included studies (49/53; 92%). Among those, Twitter was the most commonly used channel (34/53; 64%), followed by Sina Weibo (10/53, 19%) (Table 1). Other data sources were official disease outbreak data (15/53; 28%), Google Trends (6/53; 11%), and Internet search engines (6/53; 11%).

More than two thirds of the included studies provided information about the content of social media posts (37/53; 70%). The next most common types of provided information were the reporting of spatiotemporal data from social media, found in 14 papers (26%); and Internet search queries in 12 papers (23%) (Appendix 5).

A preprocessing stage that prepares data to be analyzed is required when automated analyses techniques are used. Preprocessing techniques clean data and transform them to the predictable and analyzable format required by analysis algorithms. Examples of common related techniques in natural language processing are lowercasing, stemming, lemmatization, stop word removal, and normalization. Preprocessing is crucial when data sources present a dynamic and specific language like those used in social networks. Reporting the data preprocessing techniques used in automated analysis is relevant for research reproducibility. Thirteen studies (13/53; 25%) did not require a data preprocessing stage for two main reasons [11, 13, 18, 25, 26, 28, 30, 39, 41, 58, 59, 61, 62]: they were based on quantitative data such as web search index, or authors followed a manual approach to analyze the collected data. Although a preprocessing stage was required in the remaining included studies, four of them (4/53; 7.5%) did not report any information regarding this stage [14, 17, 29, 60].

A total of 36 studies reported data preprocessing. Among those, data preprocessing and filtering was reported in 19 studies (19/40; 48%). Several approaches to filtering data were reported. Nineteen of the studies reporting preprocessing information included a data preparation stage (19/40; 48%). Information regarding data cleaning was the most frequently reported. However, many of these studies reported vaguely about data preparation that did not include details on specific techniques and tools used. Data aggregation was implemented in 14 studies (14/40; 35%) in which data from several sources such as Google Trends data, Twitter, Web search indexes, or official Centre for Disease control (CDC) data were combined on a weekly or daily basis. See Appendix 6 for further details.

Regarding the analysis techniques, the most common analyses in the included studies were the correlation analysis (23/53; 43%), followed by spatiotemporal analysis using various techniques (22/53; 42%), and classification problem (14/53; 26%) (Table 2).

When it comes to features used in the analysis, spatiotemporal features were the most common data used in the analysis of

**Table 1** Data source used by the studies included in the review (n=53)

| Data source | References using this data source | Number (%) of papers citing this data source |
|---|---|---|
| Social media | [7,11,12,14-58] | 49 (92%) |
| Twitter | [11,12,15,16,20,23,25–30,32–38,40,42–52,55,57,59] | 34 (64%) |
| Sina Weibo | [7,17–19,22,24,31,39,53,54] | 10 (19%) |
| Baidu | [17,18,31,39] | 4 (7.5%) |
| Social media (in general) | [17,21,56,59] | 4 (7.5%) |
| Coosto (social media monitoring tool) | [32,41] | 1 (2%) |
| Facebook | [32] | 1 (2%) |
| WeChat | [14] | 1 (2%) |
| Wikipedia | [28] | 1 (2%) |
| YouTube | [58] | 1 (2%) |
| Official disease outbreak data (CDC, WHO, laboratory,… etc) | [12,17,18,20,21,24,29,31,38,39,41,46,59–61] | 15 (28%) |
| Google Trends | [11,12,20,39,41,55] | 6 (11%) |
| Internet search / search engines | [17,23,26,28,61,62] | 6 (11%) |
| News, online news | [18,20,26,56,59] | 5 (9%) |
| Surveys (any type) | [13,21] | 2 (4%) |
| Spinn3r (Web and social media indexing service) | [60] | 1 (2%) |

* Note: some studies included more than one data source

the included studies. Thirty-four studies used time stamp or time series in the analysis (34/53; 64%). Most of the studies used the location information to filter data out in the collection stage, but only 16 of them compared results from different geolocation (areas, cities, or countries) (16/53; 30%).

Twenty-three of the included studies analyzed features extracted from post contents. Seven of those studies analyzed words or keywords (7/53; 13%). Six studies used word frequencies or the Term Frequency-Inverse Document Frequency (TF-IDF) values (6/53; 11%). Five studies used "bag of words" in their analysis (5/53; 9%). Other features such as linguistic characteristics were used in a single study. Further details about features used in the analyses are reported in Appendix 7.

## 3.4 Does Citizen Input Correspond With Epidemic Data?

Table 3 summarizes the reported outcomes by disease. Several papers reported correlations between social media data related to COVID-19, influenza-like illnesses, measles, MERS-CoV, MRSA and the plague and official case numbers [11, 30, 41, 53, 61, 62]. Results regarding the timeliness of the detected events, i.e., whether PHI can help in detecting outbreaks ahead of official statistics were rare and contradictory. For conjunctivitis and COVID-19, two papers (3.8%) reported that social media data can detect outbreaks as early as, or earlier than, the official reporting mechanisms [53, 62]. For Ebola, one paper concluded it is unlikely that online surveillance provides an alert more than a week before the official announcement [47].

Some papers mentioned a positive effect of using social media for crisis communication. Posting of latest news and information on how the government is handling the pandemic can affect citizens' engagement [7, 59], thereby demonstrating that posting can be valuable for citizen education [18]. Other papers question the usefulness or effectiveness of social media for communication on an epidemic [29, 32, 52] since there are discrepancies between the interest or concerns of the population in general and the provided information by public health authorities. Misinformation and false alarms were mentioned in two papers [48, 58] (Table 3).

## 3.5 Barriers to the Use of Social Media for Pandemic Detection and Management

In the included papers, we identified four groups of issues that impact the outcome of PHI for detecting or retrieving knowledge on pandemics. These are usage of app data, data collection, behavior of individuals, and analysis and interpretation of social media data.

When apps are used for disease surveillance, privacy issues and concerns about personal confidentiality can hamper the data usage. Authorizing social media apps to use personal data including personal information, activity status data, and spatiotemporal data is still not acceptable [14]. Another barrier relates to data collection. Specifically, not all data generated on social media is available for analysis. Several research papers used the Twitter API which only allows a collection of a subset of the data posted on Twitter, which may have resulted in leaving relevant tweets unconsidered [12, 33, 34]. Only one paper considered popular tweets, i.e., those that were re-tweeted many times [52], which means that these tweets are not representative of all tweets.

Another bias may arise from censorship in countries like China, thus limiting the completeness of the data [22]. Data collection from Google Trends only provides relative and not absolute values, thus hindering the possibility of further refining and processing them [61]. Finally, data related to

**Table 2** Analysis techniques used in included papers (n=53)

| Analysis | Algorithm or technique (References using this technique) | Number of papers citing this technique |
|---|---|---|
| **Correlation analysis** | Spearman ([11,18,31,33,39,54])<br>Pearson ([19,20,37,38,43,44,48,49,55,60])<br>unspecified ([12,22,25,30,50,51,61]) | 23 (43%) |
| **Spatiotemporal analysis** | | |
| Time series analysis | FDR [23], DBNM [24], KLD [47], JI [51], GCt [53], DFt [53], ARIMAX [55], ESA [57]<br>unspecified ([13,14,30,35,38,43,49,50,61,62]) | 17 (32%) |
| Seasonal analysis | SARIMA [17], SDA [17], LiR [17], BCP [28], SH-ESD [42], STL [54]<br>unspecified [48] | 5 (9%) |
| **Classification** | | |
| Relevance classification | SVM ([15,16,27,35,36,38,44,49,53])<br>LR ([35,46,55])<br>NB ([36,42])<br>LiR ([46])<br>ME ([48])<br>RF [53], DT ([53]), ET ([53]), KNN ([53]), MP ([53]) | 13 (25%) |
| Emotions classification | NB [47] | 1 (2%) |
| **Detection** | | |
| Topic Identification | LDA ([11,27,44,54])<br>BTM ([33])<br>RF ([54])<br>Km ([57])<br>unspecified ([40,51]) | 8 (15%) |
| Symptom Recognition | LDA ([21])<br>LR ([21]) | 1 (2%) |
| Language detection | Unspecified ([32]) | 1 (2%) |
| **Prediction/Estimation** | LiR ([12,21,22,37,38,46])<br>MRM ([30,46,61]) | 8 (15%) |
| **Sentiment analysis** | NB ([34])<br>ME ([34])<br>DLMC ([34])<br>SSA ([47])<br>unspecified ([7,47,50]) | 5 (9%) |
| **Other analysis** | | |
| Qualitative analysis | Thematic analysis ([59])<br>Classification ([31]) | 2 (4%) |
| Link analysis, Influence analysis, and/or Communities identification | GNA ([60]) | 1 (2%) |

* Note: some studies reported more than one type of information

FDR: False Discovery Rate; DBNM: Dynamic Bayesian Network Model; KLD: Kullback-Leibler Divergence; JI: Jaccard Index; GCt: Granger Causality Test; DFt: Dickey-Fuller tests; ARIMAX: AutoRegressive Integrated Moving Average model with eXogenous covariates; ESA: Exponential Smoothing Algorithm; SARIMA: Seasonal AutoRegressive Integrated Moving Average; SDA: Seasonal Decomposition Analysis; LiR: Linear Regression; BCP: Bayesian Change Point; SH-ESD: Seasonal-Hybrid Extreme Studentized Deviate; STL: Seasonal-Trend decomposition based on Loess; SVM: Supported Vector Machine; LR: Logistic Regression; NB: Naïve Bayes; ME: Maximum Entropy; RF: Random Forest; DT: Decision Tree; ET: Extra Tree; KNN: K nearest neighbors; MP: Multilayer Perceptron; LDA: Latent Dirichlet Allocation; BTM; Biterm Topic Model; Km: K-means; MRM: Multivariable Regression Model; DLMC: Dynamic Language Model Classifier; SSA: Stanford Sentiment Analyzer; GNA: Girvan-Newman Algorithm.

disease surveillance, such as geolocation or other user data, might be unavailable or imprecise (e.g., when using search logs from Google or access rates of Wikipedia pages) [28, 62].

The third group of problems is related to the behavior of individuals, or citizens´ input. When a disease (e.g., influenza) becomes a hot topic, people do not post about it [15]. In contrast, when celebrities are concerned about a disease, there are more people posting about it [62] which may generate false concern. Public awareness of disease surveillance methods using social media could influence behavior and consequently lead to false reporting [16].

The fourth group of issues concerns the analysis and interpretation of data, i.e., data processing for the purpose of disease surveillance. First of all, when considering free text as a data source, misspellings, abbreviations, and use of slang hamper processing [27]. Second, social media data is dynamic: new words can appear (e.g., new slang referring to a disease). This requires retraining classifiers so that new vocabulary and new anomalies in social signals can be learned [16, 27].

# 4 Discussion

## 4.1 Summary of Findings

This literature review and synthesis confirms that PHI has been used to address a wide variety of public health issues relating to pandemics. Most literature has focused on influenza or influenza-like illness or, in 2020, COVID-19. The vast majority of studies have used data from social media posts or web search patterns with a wide variety of data analysis techniques. For most diseases, the small number of studies identified means that firm conclusions about the utility of PHI in detecting and monitoring these disease outbreaks cannot be reached. In comparison, the extensive literature on influenza and COVID-19 (in spite of the fact that the literature search ended in June 2020) provides valuable insights into the potential for PHI to provide additional, more timely or efficient pandemic monitoring.

**Table 3** Reported outcomes related to PHI per epidemic in the reviewed papers

| Epidemic | Outcome |
|---|---|
| Conjunctivitis | PHI (Google search data) enable earlier outbreak detection. [62] |
| COVID-19 | PHI (number of tweets) correlates positively with daily case numbers [19,22,39,54]<br>PHI (Reports of symptoms and diagnosis of COVID-19 on social media) enabled to predict daily case counts up to 14 days ahead of official statistics [53]<br>Value for communication and information provision: Media richness negatively predicts citizen engagement through government social media, but dialogic loop facilitates engagement. Information relating to the latest news about the crisis and the government's handling of the event positively affects citizen engagement through government social media [7] |
| Dengue | Social media user trust in information shared by health professionals and others in their online social networks [40] |
| Ebola | PHI did not enable to earlier outbreak detection [26].<br>Analysis of emotions in social media microblogging data (Twitter) may be utilized as a source of evidence for disease outbreak detection and monitoring [47]. |
| Influenza, Seasonal flu, Influenza-like illnesses, Avian influenza | Outbreak detection:<br>▪ Partially positive correlation between local influenza-like illness percentages and tweet rates [16]<br>▪ Positive correlation between the number of tweets, search volume or frequent daily discussions and daily case numbers [20,24,25,27,31,35,36,38,43,44,49,50,56,60].<br>▪ PHI (Twitter) enabled earlier detection of outbreaks [42]. Differences in the degree of sensitivity exist between social media: A high sensitivity of 92% was found for Google and a low sensitivity of 50% was calculated for Twitter. Wikipedia had the lowest sensitivity of 33% [28].<br>▪ PHI led to false alarms: Twitter flu surveillance erroneously indicated a typical flu season during 2011-2012. |
| Measles | Value for communication and information provision: Social media provides insight into the opinions regarding the pandemic that are at a certain moment salient among the public [59]<br>There is a positive correlation between the weekly number of social media messages and the weekly number of online news articles [59] |
| MERS-CoV | High correlations between social media data and the number of confirmed MERS cases. High correlations between social media data and the number of quarantined cases [11] |
| Methicillin-resistant Staphylococcus aureus (MRSA) | PHI enabled rapid identification of potential MRSA outbreaks [41] |
| Plague (bubonic plague, pneumonic plague) | Statistically significant positive correlations were found between Google trends search data and confirmed, suspected, and probable cases [30,61] |
| Zika | Value for communication and information provision: Social media is unlikely to be useful or effective for communication on an epidemic; There are discrepancies between what the general public was most interested in, or concerned about, and what public health authorities provided [29,32,52] |

## 4.2 Epidemics and PHI

Although most of the articles included in our review focused on influenza-like illnesses and COVID-19, other epidemics have also been considered in PHI research (i.e., Ebola, Zika, Dengue, etc.). PHI research on previous pandemics has probably facilitated fast developments in relation to the COVID-19 emergency. Likewise, PHI research on COVID-19 may be applicable for detecting and managing future epidemics.

PHI is an imperfect source of data with its own biases such that its content and frequency of posting are not equally distributed across the population. For example, a recent secondary analysis of survey data revealed that women were more likely to post on COVID-19 than men; that black, Latino, and other non-white males were more likely to post on the topic than whites, and that people age 65 and above were more likely to post than younger people [68]. However, existing analysis tools take this into account, and use the frequency of posting as a variable in and of itself. Research has yet to study the association between differential frequency of posting, and outbreak detection.

It should also be noted that PHI is varied and, as such, some types of PHI are better at answering specific questions than others. For example, search data on the CDC website was better and faster at detecting influenza trends on a national, but not state level [69]. There is a need to clarify which questions are best suited to be answered by which source of PHI. The same holds true for analysis techniques. As individuals generate increasingly more PHI, and its use for detecting and managing pandemics persists, newer, more refined tools and analyses are required to assess how PHI best assists in promoting health and, along with that, what its characteristics are.

Finally, recent work analyzing Tweets to capture public sentiment about COVID-19, identified five dominant themes: health care environment, emotional support, business economy, social change, and psychological stress [70]. These are not captured in electronic health records, and yet they provide invaluable insights into population needs and concern which should receive public health attention. Since some papers claim that early alerts cannot be achieved from social media, more information needs to be collected on various diseases to understand to what degree patterns generalize.

## 4.3 PHI Tools, Methods and Citizens´ Input

Analysis of social media and web searches shows that posting and search frequencies have consistent positive correlations with official disease incidence numbers in the cases of influenza, COVID-19 and for the

related MERS Co-V. Most recent studies in COVID-19 suggest that analysis of these data may even predict an increase in case numbers ahead of official health system generated data [53]. Previous literature synthesis has also concluded that online social networks generate data that can track pandemic development [63] and other disease outbreaks [64]. As these data are created by citizens in their daily lives and do not incur additional data collection costs, they therefore represent an attractive additional source of information to complement traditional disease surveillance data. The timeliness of PHI provides other significant benefits, such as noting a certain level of awareness of and concern with COVID-19, which epidemiological records do not convey. Further studies validating these observations are a high priority, particularly as the COVID-19 pandemic unfolds.

Since citizen behavior and input may change as the pandemic evolves, these correlations between infection incidence and secondary data sources may not remain stable. On the other hand, this might reflect psychological responses to the pandemic which are related to, but not the same as, the actual prevalence of COVID-19. Furthermore, lack of correlation between disease incidence and media trends might result from adaptation and familiarity, leading, for example, to fewer information searches (e.g., on Wikipedia).

The requirements for data cleaning and analytic methods, which may need to rapidly evolve, may present additional barriers to this approach to disease surveillance being widely adopted in multiple geographic settings. The availability of standardized surveillance approaches and efficient development of effective algorithms have previously been identified as barriers to use of social media in surveillance of illicit drug use [65]. Furthermore, the uncertainties about how representative data are and if sufficient population coverage is reached remain unresolved.

Our synthesis of the outcomes of using PHI in pandemics suggests that analysis of social media posting is useful in assessing disease-related informational needs, such as reducing vaccination hesitancy [71].

The analysis of social media posts can also be useful for assessing the effectiveness of government or health authority communication with their populations. Again, issues of how representative social media users are of the wider population remain unresolved. At present, each analysis requires a bespoke approach to data collection and analysis. It would seem likely that this use of social media and web browsing data will complement traditional research approaches with the advantages of more immediacy of data.

## 4.4 Barriers of Using Social Media During Pandemics

We found several barriers of using social media for detecting or retrieving knowledge on pandemics: data privacy and concerns about personal confidentiality, data collection (technical limitation like the Twitter API data sample, lack of data completeness, censorship, or potential inaccuracies), behavior trends, and complexity of data analysis and interpretation.

Data privacy and personal confidentiality are two of the most relevant issues of using social media for participatory health purposes [72, 73]. Included studies reported privacy and confidentiality barriers showing that there are still unresolved ethical, legal, and technological questions. Therefore, new models of responsible and transparent data collection and treatment addressing these questions are needed, especially in public health emergencies [74]. Limitations in data collection from social media sources is an issue that is commonly reported in the scientific literature [75, 76]. The social influence that an individual's posts on social media may have on others' behaviors is also reported as a relevant aspect to be considered in digital surveillance systems using social media [77]. Simple methods are commonly used to collect data from social media sources, resulting in a dataset including noise (data not related to specific pandemic). Then, a filtering stage is required to select efficiently the data sample. Both manual and automated filtering are commonly used to classify collected data. Most automated methods are based on artificial intelligence. Due to social media data characteristics, several processing stages are required to prepare data to be used by analysis models. However, artificial intelligence supporting participatory health is still in its infancy [78]. Although the most common application of artificial intelligence in participatory health is the secondary analysis of social media data [78], there are several challenges that must be addressed [78]. Additionally, a combination of epidemiologic expertise, analytical expertise, and advanced computational skills are required to interpretate data for pandemic surveillance [77].

## 4.5 Limitations

One limitation of our work is that the data collection ended on June 10th, 2020, while the COVID-19 pandemic continued to evolve. Thus, for example, we did not include a study published in August 2020 on how media coverage influences Google search trends, so that they cannot be assumed to only reflect people's health [66]. Likewise, we did not include a study on natural language processing that revealed changes in large mental health groups (e.g., SuicideWatch and Depression) on Reddit during the COVID-19 pandemic [67]. This novel, illuminating work was published after our inclusion date. Regardless, we believe it was important to end the search at that date so that COVID-19 was still included in the review and so that the results of the review are released when COVID-19 is still of relevance, so they can be utilized by health officials and researchers.

In order to move the field of PHI forward for detecting and managing future pandemics we recommend:

- ✓ Finding the best way to deal with the current barriers to fuller impact of PHI data (i.e., privacy issues, commercial practices, governmental practices, etc.)
- ✓ Clarifying which questions are best suited to be answered by which source of PHI
- ✓ Creating more refined tools and analyses is required to assess how PHI best assists in promoting health during pandemics]

# 5 Conclusions and Recommendations for Future Research

This paper explored the role of PHI in managing and detecting pandemics. We conclude that PHI provides an unmediated, authentic, and readily available source of information that can be highly useful in the detection and management of pandemics. Our findings highlight the ways in which social media can be used as a form of participatory health, to manage and detect pandemics. They also illuminate the barriers to fuller impact of such data: some of these barriers stem from privacy issues, some from commercial practices such as providing relative but not absolute rankings of trends, while others are rooted in governmental practices such as censorship. There is a series of questions that future studies could aim to answer: What are the issues that hamper citizens' contribution and the value of their contribution, and what can facilitate their contribution? To what extent are citizens invited to contribute to outbreak detection and crisis communication using PHI? Given that citizen input is instrumental in early detection of diseases and is crucial in detection of mental distress resulting from diseases, governments should strive to invite such input in a standardized, anonymized manner.

## References


1. Kaiser R, Coulombier D, Baldari M, Morgan D, Paquet C. What is epidemic intelligence, and how is it being improved in Europe? Euro Surveill 2006 Feb 2;11(2):E060202.4
2. Freifeld CC, Mandl KD, Reis BY, Brownstein JS. HealthMap: global infectious disease monitoring through automated classification and visualization of Internet media reports. J Am Med Inform Assoc 2008 Mar-Apr;15(2):150-7.
3. Rortais A, Belyaeva J, Gemo M, Van der Goot E, Linge JP. MedISys: An early-warning system for the detection of (re-) emerging food-and feed-borne hazards. Food Research International 2010 Jun 1;43(5):1553-6.
4. Samaras L, García-Barriocanal E, Sicilia MA. Syndromic surveillance using web data: a systematic review. Innovation in Health Informatics 2020:39-77.
5. Castillo-Sánchez G, Marques G, Dorronzoro E, Rivera-Romero O, Franco-Martín M, De la Torre-Díez I. Suicide Risk Assessment Using Machine Learning and Social Networks: a Scoping Review. J Med Syst 2020 Nov 9;44(12):205.
6. Syed-Abdul S, Gabarron E, Lau AY, Househ M. Chapter 1 - An introduction to participatory health through social media. In: Participatory Health Through Social Media. Academic Press; 2016 Jan 1. p. 1-9.
7. Chen Q, Min C, Zhang W, Wang G, Ma X, Evans R. Unpacking the black box: How to promote citizen engagement through government social media during the COVID-19 crisis. Comput Human Behav 2020 Sep;110:106380.
8. Lohiniva AL, Sane J, Sibenberg K, Puumalainen T, Salminen M. Understanding coronavirus disease (COVID-19) risk perceptions among the public to enhance risk communication efforts: a practical approach for outbreaks, Finland, February 2020. Euro Surveill 2020 Apr;25(13):2000317.
9. World Health Organization. Coronavirus (COVID-19) events as they happen. Geneva: World Health Organization; 2020:1-0. Available from: https://www.who.int/emergencies/diseases/novel-coronavirus-2019/interactive-timeline/
10. Liberati A, Altman DG, Tetzlaff J, Mulrow C, Gøtzsche PC, Ioannidis JP, et al. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: explanation and elaboration. J Clin Epidemiol 2009 Oct;62(10):e1-34.
11. Shin SY, Seo DW, An J, Kwak H, Kim SH, Gwack J, et al. High correlation of Middle East respiratory syndrome spread with Google search and Twitter trends in Korea. Sci Rep 2016 Sep 6;6:32920.
12. Comito C, Forestiero A, Pizzuti C. Improving influenza forecasting with web-based social data. In: 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM) 2018 Aug 28; IEEE; 2018. p. 963-70.
13. Moberley S, Carlson S, Durrheim D, Dalton C. Flutracking: Weekly online community-based surveillance of influenza-like illness in Australia. 2017 Annual Report. Commun Dis Intell 2019;43.
14. Wang S, Ding S, Xiong L. A New System for Surveillance and Digital Contact Tracing for COVID-19: Spatiotemporal Reporting Over Network and GPS. JMIR Mhealth Uhealth 2020 Jun 10;8(6):e19457.
15. Wakamiya S, Kawai Y, Aramaki E. After the boom no one tweets: microblog-based influenza detection incorporating indirect information. In: Proceedings of the Sixth International Conference on Emerging Databases: Technologies, Applications, and Theory 2016 Oct 17. p. 17-25.
16. Allen C, Tsou MH, Aslam A, Nagel A, Gawron JM. Applying GIS and Machine Learning Methods to Twitter Data for Multiscale Surveillance of Influenza. PLoS One 2016 Jul 25;11(7):e0157734.
17. Chen Y, Zhang Y, Xu Z, Wang X, Lu J, Hu W. Avian Influenza A (H7N9) and related Internet search query data in China. Sci Rep 2019 Jul 18;9(1):10434.
18. Liu K, Li L, Jiang T, Chen B, Jiang Z, Wang Z, et al. Chinese Public Attention to the Outbreak of Ebola in West Africa: Evidence from the Online Big Data Platform. Int J Environ Res Public Health 2016 Aug 4;13(8):780.
19. Zhao Y, Cheng S, Yu X, Xu H. Chinese Public's Attention to the COVID-19 Epidemic on Social Media: Observational Descriptive Study. J Med Internet Res 2020 May 4;22(5):e18825.
20. Samaras L, García-Barriocanal E, Sicilia MA. Comparing Social media and Google to detect and predict severe epidemics. Sci Rep 2020 Mar 16;10(1):4747.
21. Daughton AR, Chunara R, Paul MJ. Comparison of Social Media, Syndromic Surveillance, and Microbiologic Acute Respiratory Infection Data: Observational Study. JMIR Public Health Surveill 2020 Apr 24;6(2):e14986.
22. Li J, Xu Q, Cuomo R, Purushothaman V, Mackey T. Data Mining and Content Analysis of the Chinese Social Media Platform Weibo During the Early COVID-19 Outbreak: Retrospective Observational Infoveillance Study. JMIR Public Health Surveill 2020 Apr 21;6(2):e18700.
23. 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.
24. Huang J, Zhao H, Zhang J. Detecting flu transmission by social sensor in China. In: 2013 IEEE International Conference on Green Computing and Communications and IEEE Internet of Things and IEEE Cyber, Physical and Social Computing 2013 Aug 20. IEEE; 2013. p. 1242-7
25. Lee B, Yoon J, Kim S, Hwang BY. Detecting social signals of flu symptoms. In: 8th International Conference on Collaborative Computing: Networking, Applications and Worksharing (CollaborateCom) 2012 Oct 14. IEEE; 2012. p. 544-5.
26. 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; May 18. p. 115-9.
27. 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;19:e315
28. 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.
29. Khatua A, Khatua A. Immediate and long-term effects of 2016 Zika Outbreak: A Twitter-based study. In: 2016 IEEE 18th International Conference on e-Health Networking, Applications and Services (Healthcom) 2016 Sep 14. IEEE; 2016. p. 1-6.
30. Al-Mohrej A, Agha S. Are Saudi medical students aware of middle east respiratory syndrome coronavirus during an outbreak? J Infect Public Health 2017 Jul-Aug;10(4):388-95.
31. 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.

32. Barata G, Shores K, Alperin JP. Local chatter or international buzz? Language differences on posts about Zika research on Twitter and Facebook. PLoS One 2018 Jan 5;13(1):e0190482.
33. Mackey T, Purushothaman V, Li J, Shah N, Nali M, Bardier C, et al. Machine Learning to Detect Self-Reporting of Symptoms, Testing Access, and Recovery Associated With COVID-19 on Twitter: Retrospective Big Data Infoveillance Study. JMIR Public Health Surveill 2020 Jun 8;6(2):e19509.
34. Byrd K, Mansurov A, Baysal O. Mining Twitter Data for Influenza Detection and Surveillance. In: 2016 IEEE/ACM International Workshop on Software Engineering in Healthcare Systems (SEHS) May 14. IEEE; 2016. p. 43-9.
35. Broniatowski DA, Paul MJ, Dredze M. National and local influenza surveillance through Twitter: an analysis of the 2012-2013 influenza epidemic. PLoS One 2013 Dec 9;8(12):e83672.
36. Collier N, Son NT, Nguyen NM. OMG U got flu? Analysis of shared health messages for bio-surveillance. J Biomed Semantics 2011 Oct 6;2 Suppl 5(Suppl 5):S9.
37. Chew C, Eysenbach G. Pandemics in the age of Twitter: content analysis of Tweets during the 2009 H1N1 outbreak. PLoS One 2010 Nov 29;5(11):e14118.
38. 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 Apr 3; p. 1327-34.
39. Garazzino S, Montagnani C, Donà D, Meini A, Felici E, Vergine G, et al. Italian SITIP-SIP SARS-CoV-2 paediatric infection study group. Multicentre Italian study of SARS-CoV-2 infection in children and adolescents, preliminary data as at 10 April 2020. Euro Surveill 2020 May;25(18):2000600.
40. He Z, Zhang CJ, Huang J, Zhai J, Zhou S, Chiu JW, Sheng et al. A New Era of Epidemiology: Digital Epidemiology for Investigating the COVID-19 Outbreak in China. J Med Internet Res 2020 Sep 17;22(9):e21685.
41. Yousefinaghani S, Dara R, Poljak Z, Bernardo TM, Sharif S. The Assessment of Twitter's Potential for Outbreak Detection: Avian Influenza Case Study. Sci Rep 2019 Dec 3;9(1):18147.
42. Nagel AC, Tsou MH, 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.
43. Aslam AA, Tsou MH, 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.
44. Abd-Alrazaq A, Alhuwail D, Househ M, Hamdi M, Shah Z. Top Concerns of Tweeters During the COVID-19 Pandemic: Infoveillance Study. J Med Internet Res 2020 Apr 21;22(4):e19016.
45. Culotta A. Towards detecting influenza epidemics by analyzing Twitter messages. In Proceedings of the 1st Workshop on Social Media Analytics (SOMA'10), 2010 Jul 25. p. 115-22.
46. Ofoghi B, Mann M, Verspoor K. Ealry discovery of salient health threats: a social media emotion classification technique. Pac Symp Biocomput 2016;21:504-15.
47. Mowery J. Twitter Influenza Surveillance: Quantifying Seasonal Misdiagnosis Patterns and their Impact on Surveillance Estimates. Online J Public Health Inform 2016 Dec 28;8(3):e198.
48. 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.
49. Comito C, Forestiero A, Pizzuti C. Twitter-based Influenza Surveillance: An Analysis of the 2016-2017 and 2017-2018 Seasons in Italy. In: Proceedings of the 22nd International Database Engineering & Applications Symposium 2018 Jun 18. p. 175-82.
50. Kanhabua N, Nejdl W. Understanding the diversity of tweets in the time of outbreaks. In: Proceedings of the 22nd International Conference on World Wide Web 2013 May 13; p. 1335-42.
51. Gui X, Wang Y, Kou Y, Reynolds TL, Chen Y, Mei Q, et al. Understanding the Patterns of Health Information Dissemination on Social Media during the Zika Outbreak. AMIA Annu Symp Proc 2018 Apr 16;2017:820-9.
52. Shen C, Chen A, Luo C, Zhang J, Feng B, Liao W. Using Reports of Symptoms and Diagnoses on Social Media to Predict COVID-19 Case Counts in Mainland China: Observational Infoveillance Study. J Med Internet Res 2020 May 28;22(5):e19421.
53. Han X, Wang J, Zhang M, Wang X. Using Social Media to Mine and Analyze Public Opinion Related to COVID-19 in China. Int J Environ Res Public Health 2020 Apr 17;17(8):2788.
54. 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 Jan-Jun;1(1):e5.
55. Corley CD, Cook DJ, Mikler AR, Singh KP. Adv Exp Med Biol 2010;680:559-64.
56. Odlum M, Yoon S. What can we learn about the Ebola outbreak from tweets? Am J Infect Control 2015 Jun;43(6):563-71.
57. Li HO, Bailey A, Huynh D, Chan J. YouTube as a source of information on COVID-19: a pandemic of misinformation? BMJ Glob Health 2020 May;5(5):e002604.
58. 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.
59. Corley CD, Cook DJ, Mikler AR, Singh KP. Text and structural data mining of influenza mentions in web and social media. Int J Environ Res Public Health 2010 Feb;7(2):596-615.
60. Bragazzi NL, Mahroum N. Google Trends Predicts Present and Future Plague Cases During the Plague Outbreak in Madagascar: Infodemiological Study. JMIR Public Health Surveill 2019 Mar 8;5(1):e13142.
61. Deiner MS, McLeod SD, Wong J, Chodosh J, Lietman TM, Porco TC. Google Searches and Detection of Conjunctivitis Epidemics Worldwide. Ophthalmology 2019 Sep;126(9):1219-1229.
62. Al-Garadi MA, Khan MS, Varathan KD, Mujtaba G, Al-Kabsi AM. Using online social networks to track a pandemic: A systematic review. J Biomed Inform 2016 Aug;62:1-11.
63. 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.
64. Kazemi DM, Borsari B, Levine MJ, Dooley B. Systematic review of surveillance by social media platforms for illicit drug use. J Public Health (Oxf) 2017 Dec 1;39(4):763-76.
65. Sousa-Pinto B, Anto A, Czarlewski W, Anto JM, Fonseca JA, Bousquet J. Assessment of the Impact of Media Coverage on COVID-19–Related Google Trends Data: Infodemiology Study. J Med Internet Res 2020 Aug 10;22(8):e19611.
66. Low DM, Rumker L, Talkar T, Torous J, Cecchi G, Ghosh SS. Natural Language Processing Reveals Vulnerable Mental Health Support Groups and Heightened Health Anxiety on Reddit During COVID-19: Observational Study. J Med Internet Res 2020 Oct 12;22(10):e22635.
67. Campos-Castillo C, Laestadius LI. Racial and Ethnic Digital Divides in Posting COVID-19 Content on Social Media Among US Adults: Secondary Survey Analysis. J Med Internet Res 2020 Jul 3;22(7):e20472.
68. Caldwell WK, Fairchild G, Del Valle SY. Surveilling Influenza Incidence With Centers for Disease Control and Prevention Web Traffic Data: Demonstration Using a Novel Dataset. J Med Internet Res 2020 Jul 3;22(7):e14337.
69. Hung M, Lauren E, Hon ES, Birmingham WC, Xu J, Su S, et al. Social Network Analysis of COVID-19 Sentiments: Application of Artificial Intelligence. J Med Internet Res 2020 Aug 18;22(8):e22590.
70. Wilson SL, Wiysonge C. Social media and vaccine hesitancy. BMJ Global Health 2020 Oct 1;5(10):e004206.
71. Rivera-Romero O, Konstantinidis S, Denecke K, Gabarrón E, Petersen C, Househ M, et al. Ethical Considerations for Participatory Health through Social Media: Healthcare Workforce and Policy Maker Perspectives: Contribution of the IMIA Participatory Health and Social Media Working Group. Yearb Med Inform 2020 Aug;29(1):71-6.
72. Golinelli D, Boetto E, Carullo G, Nuzzolese AG, Landini MP, Fantini MP. Adoption of Digital Technologies in Health Care During the COVID-19 Pandemic: Systematic Review of Early Scientific Literature. J Med Internet Res 2020 Nov 6;22(11):e22280.
73. Almeida BD, Doneda D, Ichihara MY, Barral-Netto M, Matta GC, Rabello ET, et al. Personal data usage and privacy considerations in the COVID-19 global pandemic. Ciência & Saúde Coletiva 2020

Jun 5;25:2487-92.
74. Castillo-Sánchez G, Marques G, Dorronzoro E, Rivera-Romero O, Franco-Martín M, De la Torre-Díez I. Suicide Risk Assessment Using Machine Learning and Social Networks: a Scoping Review. J Med Syst 2020 Nov 9;44(12):205.
75. Al-Garadi MA, Khan MS, Dewi K, Varathan GM, Al-Kabsi AM. Using online social networks to track a pandemic: A systematic review. J Biomed Inform 2016 Aug;62:1-11.
76. Salathé M, Bengtsson L, Bodnar TJ, Brewer DD, Brownstein JS, Buckee C, et al. Digital epidemiology. PLoS Comput Biol 2012;8(7):e1002616.
77. Denecke K, Gabarron E, Grainger R, Konstantinidis ST, Lau A, Rivera-Romero O, et al. Artificial Intelligence for Participatory Health: Applications, Impact, and Future Implications. Yearb Med Inform 2019 Aug;28(1):165-73.

**Correspondence to:**
Kerstin Denecke
Bern University of Applied Sciences
Institute for Medical Informatics
Quellgasse 1
2502 Biel / Switzerland
E-mail: kerstin.denecke@bfh.ch