Review Article

Healthc Inform Res. 2018 October;24(4):253-262.
https://doi.org/10.4258/hir.2018.24.4.253
pISSN 2093-3681 • eISSN 2093-369X

HIR
Healthcare Informatics Research

Digital Epidemiology: Use of Digital Data Collected for Non-epidemiological Purposes in Epidemiological Studies

Hyeoun-Ae Park, PhD, FAAN, FACMI, RN, Hyesil Jung, MSN, RN, Jeongah On, RN, Seul Ki Park, RN, Hannah Kang, MSN, RN
College of Nursing, Seoul National University, Seoul, Korea

Objectives: We reviewed digital epidemiological studies to characterize how researchers are using digital data by topic domain, study purpose, data source, and analytic method. **Methods:** We reviewed research articles published within the last decade that used digital data to answer epidemiological research questions. Data were abstracted from these articles using a data collection tool that we developed. Finally, we summarized the characteristics of the digital epidemiological studies. **Results:** We identified six main topic domains: infectious diseases (58.7%), non-communicable diseases (29.4%), mental health and substance use (8.3%), general population behavior (4.6%), environmental, dietary, and lifestyle (4.6%), and vital status (0.9%). We identified four categories for the study purpose: description (22.9%), exploration (34.9%), explanation (27.5%), and prediction and control (14.7%). We identified eight categories for the data sources: web search query (52.3%), social media posts (31.2%), web portal posts (11.9%), webpage access logs (7.3%), images (7.3%), mobile phone network data (1.8%), global positioning system data (1.8%), and others (2.8%). Of these, 50.5% used correlation analyses, 41.3% regression analyses, 25.6% machine learning, and 19.3% descriptive analyses. **Conclusions:** Digital data collected for non-epidemiological purposes are being used to study health phenomena in a variety of topic domains. Digital epidemiology requires access to large datasets and advanced analytics. Ensuring open access is clearly at odds with the desire to have as little personal data as possible in these large datasets to protect privacy. Establishment of data cooperatives with restricted access may be a solution to this dilemma.

Keywords: Public Health Surveillance, Epidemiology, Epidemiological Monitoring, Social Media, Internet

Submitted: September 26, 2018
Revised: October 24, 2018
Accepted: October 24, 2018

Corresponding Author
Hyesil Jung, MSN, RN
College of Nursing, Seoul National University, 103 Daehak-ro, Jongno-gu, Seoul 03080, Korea. Tel: +82-2-740-8844, E-mail: soledad7@snu.ac.kr (https://orcid.org/0000-0002-8346-9343)

I. Introduction

The explosion of internet and mobile phone usage has led to a new type of epidemiology: digital epidemiology. Digital epidemiology has increased in the last decade due to the increasing availability of big data, and advancements in computing power and data analytics methods. The goal of epidemiology is to understand the distribution (who, when, where), and determinants of health and disease conditions in a defined population. Surveillance and descriptive studies can be performed to assess the distribution and analytical studies to identify determinants. The findings of epidemiological studies can be used to control disease and health problems.

Digital epidemiology can be broadly defined as epidemiology that uses digital methods from data collection to data analysis [1]. The goal of digital epidemiology is identical to that of traditional epidemiology. Then, what is the difference between the two? A narrower definition of digital epidemiology is that which uses digital data that was not generated with the primary purpose of epidemiological studies [2]. Examples of such data include search queries, social media posts, webpage access logs, mobile phone network data, data generated by sensors, and data collected at call centers.

A large proportion of sick people search for relevant health information using internet search engines, and many share their experience with the rest of us on social media. Descriptions of health problems, time-stamped and geo-tagged, are available. Thus, it is possible for us to study the health of a population in real time using such digital traces. Researchers have already started to use digital data to support public health surveillance and infectious disease monitoring or to understand public attitudes, perceptions, and behaviors towards health issues.

Google Flu Trends (GFT) is an early example of digital epidemiology, using search queries for the purpose of tracking influenza-like illnesses (ILIs) [3]. In 2009, researchers from Google and the US Centers for Disease Control and Prevention (CDC) published a method to estimate flu activity by region using search queries. For many years, Google Trends (GT) has served as a prime data source for digital epidemiology. Nuti et al. [4] were able to identify 70 articles in a systematic review which used GT in health care research.

However, search queries of GT frequently overestimate the incidence of illness. According to research carried out by a team at Northeastern University and Harvard University, GFT forecasted twice as many influenza cases as actually occurred in the United States during the 2012–2013 flu season [5]. Furthermore, the estimates cannot be reproduced easily because Google data is not publicly available. Twitter became an alternative data source because anyone with an internet connection can retrieve Twitter data.

A group of researchers used data from Twitter to track level of disease activity and concern about the influenza H1N1 pandemic in 2011 [6]. Twitter has also been used to assess health sentiments such as those about vaccination [7] and to monitor drug safety [8]. Wikipedia, another publicly accessible data source, has also been used for digital epidemiology. Researchers at Boston Children's Hospital introduced a method of estimating the level of ILIs, in near-real-time, in the United States by monitoring the rate of influenza-related Wikipedia article views on a daily basis [9].

Mobile phone data has been used to examine the movement of humans and its influence on infectious disease dynamics. For example, Wesolowski et al. [10] analyzed mobile phone call records as indicators of the travel patterns of 15 million mobile phone owners in Kenya over the course of 1 year. They combined the travel patterns with a detailed malaria risk map to estimate the movements of malaria parasites that could be caused by human movement. This information enabled detailed analyses of parasite sources and sinks among hundreds of local settlements. Bengtsson et al. [11] used the position data of subscriber identity module (SIM) cards from the largest mobile phone company in Haiti to estimate the magnitude and trends of population movements following the 2010 Haiti earthquake and cholera outbreak.

Increasing numbers of epidemiological studies are using digital data generated for a purpose other than epidemiology. There are also more freely accessible tools that allow users access to digital data, which may provide deep insight into health-related phenomena and population behavior. However, there is limited knowledge of the uses and limitations of digital data for epidemiological studies. Therefore, we reviewed epidemiological studies that used digital data and classified them by topic domain, purpose, data source, and analytic method, and evaluated their limitations for use in research.

II. Methods

1. Study Selection

We included all research articles published within the last decade that used digital data to answer epidemiological

Table 1. PubMed search criteria

digital epidemiology[Title/Abstract] OR ((epidemiology[Title/Abstract] OR disease surveillance[Title/Abstract]) AND (image[Title/Abstract] OR social media[Title/Abstract] OR social network[Title/Abstract] OR twitter[Title/Abstract] OR google[Title/Abstract] OR social data[Title/Abstract] OR digital data[Title/Abstract] OR real time data[Title/Abstract] OR mobile data[Title/Abstract] OR online encyclopedia[Title/Abstract])) AND (hasabstract[text] AND English[lang]) AND ("2008/09/24"[PDat] : "2018/09/21"[PDat])

research questions within the domain of healthcare. We included only studies of human epidemiology written in the English language. We excluded studies that primarily focused on plant and animal epidemiology. We also excluded review articles and articles that did not make substantial use of digital data.

2. Search Strategy

We identified relevant articles written in English by searching PubMed using the search terms on digital epidemiology published from September 24, 2008 to September 21, 2018. The key words used in the PubMed search are presented in Table 1.

This search identified 853 potential articles for inclusion. A multistage review process was used to select articles for review. Four authors independently reviewed the titles and abstracts of the retrieved publications. We excluded 753 articles that met at least one of the exclusion criteria. These articles were review, commentary, or perspective articles. They did not use substantial digital data, and they focused on plant or animal epidemiology. In total, 100 articles met our inclusion criteria and were subjected to full-text review by the authors. This resulted in the exclusion of seven survey or intervention studies that utilized social media (Figure 1). A review of the references of the retrieved studies resulted in the identification of 16 further studies. Thus, 109 studies were included for analysis in this review study.

3. Article Classification

To characterize how researchers use various types of digital data for epidemiological studies, we examined a review article for epidemiological studies [12] and a systematic review of studies with GT data [4] and extracted variables and their categories. The variables included year of publication, purpose of research [13], topic domain, geographical region, time period, study design, outcome measures, and use of empirical data. The topic domain variable is composed of six topics: four of these were based on the classification used by Nuti et al. [4] and two were added by the authors.

Since the data sources and analytic methods of digital epidemiology can be different from those of traditional epidemiology, we included data sources and analytic methods as variables. We created a general descriptive classification of these variables. First, we examined the full text of the articles, and extracted the data sources and analytical methods. Next, we identified and categorized data sources and analytic methods. Then, we classified the articles using these categories.

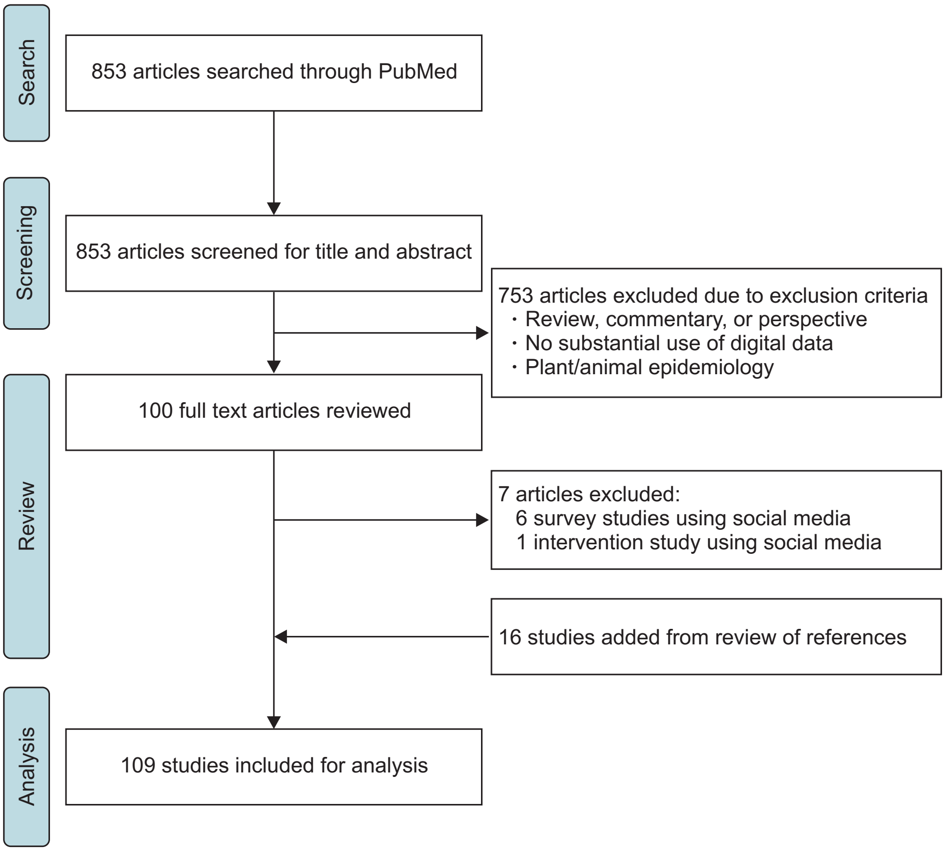
Search
853 articles searched through PubMed
Screening
853 articles screened for title and abstract
753 articles excluded due to exclusion criteria
· Review, commentary, or perspective
· No substantial use of digital data
· Plant/animal epidemiology
Review
100 full text articles reviewed
7 articles excluded:
6 survey studies using social media
1 intervention study using social media
16 studies added from review of references
Analysis
109 studies included for analysis

Figure 1. Flowchart of literature selection.

We classified data sources into the following eight categories: web search query (e.g., Google, Baidu), social media post (e.g., Twitter, blog), web portal post (e.g., HealthMap, proMED), webpage access log (e.g., page views of the National Travel Health Network and Centre and Wikipedia views), image (e.g., Pinterest), mobile phone network data, global positioning system (GPS), and others (drone/balloon, meteorological data, and call center data). We classified analytical methods into the following four categories: descriptive analyses (e.g., relative search volume and number of page views), correlation analyses (e.g., Pearson correlation coefficient, Spearman rank correlation coefficient, and

Table 2. Definitions of variables

Variable	Definition	Categories of variable
Year of publication	The year in which the article was published	From 2008 to 2018
Purpose of research	The primary purpose of the study as derived from the introduction section	Description Exploration Explanation Prediction and control
Topic domain	The phenomenon or subject studied	Infectious disease Non-communicable disease Mental health and substance use General population behavior Environmental, dietary, and lifestyle Vital status
Geographical region of study	The region included in the study	City or State Single country Multiple countries
Time period	The time period chosen for the study	-
Data source	The types of digital data studied	Web search query Social media's post Web portal's post Image Webpage access log Mobile phone network data GPS Others
Study design	The formulation/type of study	Cross-sectional study (Time series/Single point in time) Case-control study (Time series/Single point in time) Cohort study (Time series/Single point in time) Intervention study (Time series/Single point in time)
Outcome measures	Different measures of health outcomes used in the study	Prevalence Incidence Risk Relative risk
Analytic methods	The statistical methods used in the study	Descriptive statistics Correlation analyses Regression analyses Machine learning
Use of empirical data	Use of empirical data in the study	Yes No
Primary findings	The main findings of the study	Free text

pairwise cross-correlation), regression analyses (e.g., linear regression, logistic regression, jointpoint regression, and negative binomial regression), and machine learning (e.g., support vector machine, decision tree, and artificial neural networks).

4. Variable Abstraction

Data was abstracted from these studies using a data collection tool we developed (Table 2) and disagreements were resolved by consensus. This tool pertains to the purpose of research, methodology (study design, type of analysis) and findings of the studies. We identified the purpose of research by examining the introduction section of the articles. We evaluated the geographic region, data collection period, additional data sources, design, outcome measures, analytic methods, and use of empirical data by examining the methods section of the articles. We extracted primary findings by examining the results section of the articles.

We did not attempt to pool the results due to the heterogeneity of the study outcomes; we instead provide summary statistics of the studies.

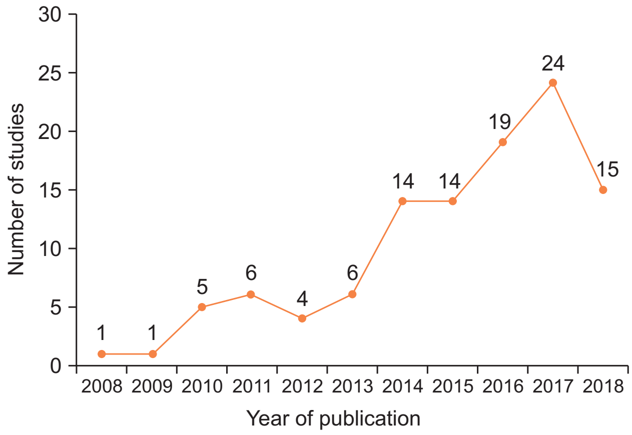

Figure 2. Distribution of articles included in review by year of publication.

III. Results

1. Study Sample

As shown in Figure 2, the number of publications on digital epidemiology has increased over the past decade. Approximately 80% of articles were published from 2014 to 2018.

Of these studies, 56.9% were conducted in a single country (Figure 3). Of the 62 studies conducted in a single country, more than half were performed in the United States followed in order by the UK, France, and Italy.

2. Characteristics of the Studies

The characteristics of the 109 articles included in this review are summarized in Table 3.

1) Topic domain

We classified the articles by their primary topic. We identified the following six main topic domains: infectious diseases (58.7%), non-communicable diseases (29.4%), mental health and substance use (8.3%), general population behavior (4.6%), environmental, dietary, and lifestyle (4.6%), and vital status (0.9%). The infectious diseases studied included influenza; ILIs; dengue fever; Ebola virus; human immunodeficiency virus (HIV); malaria; Zika virus, tuberculosis; meningococcal disease; cellulitis; chickenpox; chikungunya; hand, foot, and mouth disease; Mayaro virus; West Nile virus disease; Lyme disease; Middle East Respiratory Syndrome (MERS); respiratory virus; norovirus; and Lassa fever. The non-communicable diseases studied included cancer, allergic rhinitis, ragweed pollen allergy, antiphospholipid syndrome, multiple sclerosis, dental caries, type 2 diabetes, asthma, stroke, silicosis, tick paralysis, status epilepticus, interstitial cystitis, chronic cerebrospinal venous insufficiency migraine headache, and Willis-Ekbom disease. The general population behaviors studied included suicide, cancer screening, and population movement. The environmental, dietary, and lifestyle category included foodborne illness, vitamin D, and

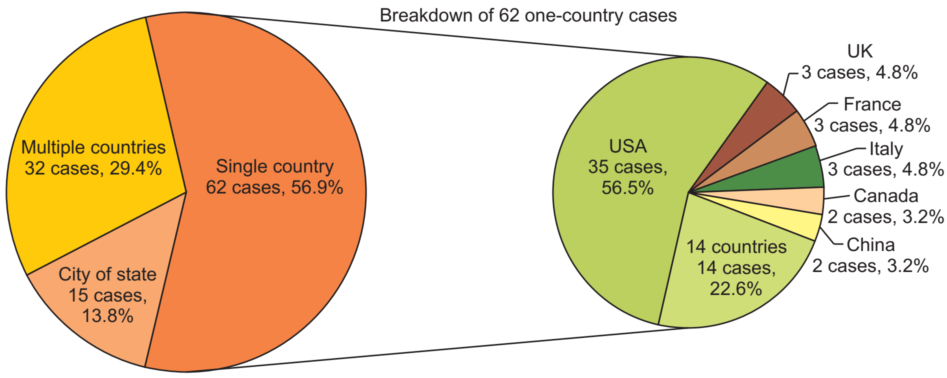

Figure 3. Geographic locations.

Table 3. Characteristics of studies reviewed (n = 109)

Category	n (%)
Topic domain[a]	
Infectious disease	64 (58.7)
Non-communicable disease	32 (29.4)
Mental health and substance use	9 (8.3)
General population behavior	5 (4.6)
Environmental, dietary, and lifestyle	5 (4.6)
Vital status	1 (0.9)
Data sources[a]	
Web search query	57 (52.3)
Social media post	34 (31.2)
Web portal post	13 (11.9)
Image	8 (7.3)
Webpage access log	8 (7.3)
Mobile phone network data	2 (1.8)
GPS	2 (1.8)
Others (drone/balloon, metrological, and call center data)	3 (2.7)
Purpose of research	
Description	25 (22.9)
Exploration	38 (34.9)
Explanation	30 (27.5)
Prediction and control	16 (14.7)
Study design	
Cross-sectional study	
Time series	77 (70.6)
Single point in time	25 (22.9)
Case-control study (Single point in time)	4 (3.7)
Cohort study	
Time series	1 (0.9)
Single point in time	1 (0.9)
Intervention study (Single point in time)	1 (0.9)
Use of external data	
Yes	83 (76.1)
No	26 (23.9)
Outcome measures	
Incidence	37 (48.7)
Prevalence	36 (47.4)
Risk	2 (2.6)
Relative risk	1 (1.3)
Analytic methods[a]	
Descriptive statistics	21 (19.3)
Correlation analyses	55 (50.5)
Regression analyses	45 (41.3)
Machine learning	29 (26.6)

[a] Multiple responses.

air quality.

2) Data sources

More than half of the studies (52.3%) used web search queries via search engines such as Google, Baidu, Yandex, Daum and Parsijoo, and 31.2% used social media posts. Most of the social media data was from Twitter and blogs. Online obituaries from funeral home web pages were used to retrieve the vital status of patients with cancer [14] and to examine the cause of death in a cancer-related epidemiological discovery study [15]. Data from balloons and/or drones equipped with infrared cameras and sensors were used as crowdsourced data (such as body temperature) to detect emerging infectious diseases [16].

3) Purpose of research

We identified four study purpose categories: description (22.9%), exploration (34.9%), explanation (27.5%), and prediction and control (14.7%). Descriptive studies focus on presenting phenomena with incidence, size, and/or other measurable attributes. An example of a descriptive study is a study by Moon et al. [17], which investigated internet search rates for vitamin D using GT data and reported seasonal variation in public interest in vitamin D. An explorative study investigates the full nature of the phenomenon, the manner in which it manifests, and other related factors through correlation analyses. An example of such a study is Chary et al. [18], which demonstrated a strong correlation between the geographic variation in social media posts mentioning prescription opioid misuse and government estimates of misuse of prescription opioids. An explanatory study focuses on understanding the underlying causes of a phenomenon or the systematic relationships among variables [13]. An example is Towers et al. [19], which examined whether news coverage was a significant factor in the temporal patterns of Ebola-related Twitter data using a linear regression model. A predictive and controlled study attempts to predict a phenomenon based on research findings. An example is Ram et al. [20], which analyzed Twitter, Google search interests, and air quality index data using machine learning techniques to predict the number of asthma-related emergency department visits in a specific area.

4) Study design

Overall, 70.6% of studies were time series cross-sectional studies, 22.9% were cross-sectional studies at a single point in time, 3.7% were case-control studies, 1.8% were cohort studies, and 0.9% were intervention studies. Most time series cross-sectional studies analyzed digital data to compare phenomena across time periods. An example is Zhang et al. [21], which examined seasonal variation in the volume of Google search queries for cellulitis from 2004 to 2016. An example of cross-sectional study at single point in time is a study that analyzed the types of discourse about Zika virus on Twitter for 2 months [22]. An example of a case-control study is a study by Tourassi et al. [15], which classified study subjects into a case group (females for whom cancer is the stated or inferred cause of death) or a control group (females for whom there was no mention of cancer) using online obituary announcements in order to examine the association between parity and cancer risk. An example of an intervention study is Edoh's study [16] which conducted an experiment using large range/distance temperature sensors and drones in order to collect infectious diseases-related data from study participants.

5) Use of external data

Most articles (76.1%) used ground truth data to measure the relationship between digital data and the gold standard or to develop and validate models. Examples of ground truth data include reports published by governments or the World Health Organization, census statistics, data obtained from scientific studies, and news data. For example, Phillips et al. [23] used cancer incidence reported by the CDC to characterize the relationship between cancer incidence and online Google search volumes in the United States for six common types of cancer.

6) Outcome measures

Of the studies with outcome measures, 48.7% estimated the incidence of infectious diseases or other problems. For example, McGough et al. [24] evaluated the feasibility of using Zika-related Google search queries, Twitter, and news reports collected by HealthMap to dynamically track and predict the incidence of Zika virus up to 3 weeks ahead of the release of official reports. In addition, 47.4% assessed disease prevalence. For example, McIver and Brownstein [9] estimated the prevalence of ILIs in the United States in near-real-time by monitoring the rates of Wikipedia article views.

7) Analytic methods

Overall, 50.5% of the studies used correlation analyses, 41.3% used regression analyses, 25.6% used machine learning techniques, and 19.3% used descriptive analyses. Linear regression analyses were the most frequently used type of regression analysis. Generous et al. [25] used a linear regression model to examine the potential of Wikipedia access logs as an emerging data source for global disease surveillance and forecasting. Machine learning techniques have prominently been used since 2014, and support vector machine has been the most frequently used. Adrover et al. [26] assessed whether adverse effects of HIV drug treatment and associated sentiments can be determined using Twitter. They utilized boosted decision trees, support vector machines, and artificial neural networks as machine-learning classifiers.

IV. Discussion

Researchers are increasingly utilizing digital data in a wide variety of areas in multiple ways. Examples include use of online obituaries from funeral home websites for near real-time surveillance of mortality [14], a study of the relationship between restaurant table availabilities and an increase in disease incidence, specifically ILIs [27], use of meteorological data to study the spatiotemporal clustering of dengue cases and climate [28], and use of a female-oriented social media site, Pinterest, for skin cancer education [29].

We reviewed 109 epidemiological studies that investigated the distribution, and determinants of health and conditions using digital data. The number of such publications increased over time from 1 article in 2008 to 24 articles in 2017 and 15 articles in the first half of 2018. This trend suggests increasing awareness and leveraging of digital data for epidemiological studies.

The geographic regions in which studies were conducted varied from a single city to multiple countries. More than half of studies (56.9%) used digital data collected in a single country. The United States topped the list of countries with 35 studies, followed by the UK, France, and Italy with 3 studies each. Use of digital methods in collecting and analyzing data can be challenging in the resource-poor countries. Examples of epidemiological studies in a city or state include a study of a community outbreak of meningococcal disease using a regional online newspaper in Sardinia, Italy [30], and a study of the main drivers of the temporal and spatiotemporal dynamics of the 2014 chikungunya outbreak using Twitter data in Martinique, France [31]. Examples of epidemiological studies in more than one country include evaluation of GFT data in low-to middle-income Latin America [32], and use of HealthMap to categorize and quantify MERS alerts [33].

The studies included in this review reflect a wide variety of topics. Infectious disease was the most frequently studied topic (64 studies), followed by non-communicable diseases (32 studies). Of the 64 studies on infectious disease, 22 focused on influenza or ILIs. This finding is slightly different from the review by Nuti et al. [4] of GT articles, in which infectious disease was the second most popular topic domain following general population behavior (including all health-related behaviors).

Most studies have used internet-based data sources such as web search queries (57 studies) followed by social media posts (34 studies), web portal posts (13 studies), and webpage access logs (8 studies). Other data sources have included mobile phone networks, GPS, drones/balloons, and call centers. Epidemiological studies using digital data in underdeveloped countries tend to use data from devices because of the typically poor internet connections in such regions. For example, long-range/-distance temperature sensors and drones were used to collect infectious disease-related data from study participants in Africa [16]. Mobile phone data was used to study the travel patterns and malaria risk for the population of Kenya [10], and to examine the magnitude and trends of population movements following the earthquake and cholera outbreak in Haiti [11].

Regarding the purposes of the studies, almost equal numbers of studies used digital data for descriptive, explorative, explanatory, and predictive and controlled studies. However, explorative studies were the most frequent type (38 studies) followed by explanatory studies (30 studies).

Regarding study design, time series cross-sectional studies were the most frequent (77 studies, 70.6%) followed by single point in time cross-sectional studies (25, 22.9%). In traditional epidemiological studies, the most frequently used method is cross-sectional regression. However, the use of digital data collected across a time period enables the modeling of effects across time and space.

A majority of the studies (76.1%) used external datasets as outcome variables for model development or outcome validation. Regarding outcome measures, incidence and prevalence were the most common measures used in digital epidemiological studies.

A majority of the studies used correlation analyses to examine the relationships among variables (55 studies) followed by various regression analyses (45 studies), such as linear regression, jointpoint regression, and Least Absolute Shrinkage and Selection Operator regression.

We examined how digital data collected for non-epidemiological purposes is being used for epidemiologic purposes. Digital epidemiological studies require large datasets and advanced analytics such as machine learning. Most machine learning algorithms are openly available due to the strong open source software movement. Thus, it is important to ensure that as much data as possible are openly accessible. As Salathe [2] elaborated so well in his review article, this is clearly at odds with the desire to have as little personal data as possible publicly accessible to protect individual privacy. There is no straightforward solution to this conflict of interest between open access to large data sets and privacy protection, but Salathe [2] proposed data cooperatives with restricted access as a solution.

Our study had certain limitations. First, given the diversity of topics and uses, there were inherent challenges in the classification of articles. However, four authors independently reviewed each article and category of abstraction, and disagreements were resolved by consensus. Second, there are no prior standards to follow for evaluating literature from novel digital data sources such as search engines, social media services, and mobile phones. Finally, there was the possibility of incomplete retrieval of articles on digital epidemiology using our search strategy. We reviewed the references sections of articles to capture as many articles as possible and performed an extensive search of the PubMed database. Notably, we focused on evaluating the use of digital data for epidemiological studies and refrained from making any comments on the conclusions drawn by researchers. Further studies should evaluate the interpretation and validity of use of digital data for epidemiological studies.

Conflict of Interest

No potential conflict of interest relevant to this article was reported.

Acknowledgments

This study was supported by a grant from the National R&D Program for Cancer Control, Ministry of Health and Welfare, Republic of Korea (No. 1720250).

References

1. Eckhoff PA, Tatem AJ. Digital methods in epidemiology can transform disease control. Int Health 2015;7(2):77-8.
2. Salathe M. Digital epidemiology: what is it, and where is it going? Life Sci Soc Policy 2018;14(1):1.
3. Ginsberg J, Mohebbi MH, Patel RS, Brammer L, Smolinski MS, Brilliant L. Detecting influenza epidemics using search engine query data. Nature 2009;457(7232):1012-4.
4. Nuti SV, Wayda B, Ranasinghe I, Wang S, Dreyer RP, Chen SI, et al. The use of google trends in health care research: a systematic review. PLoS One 2014;9(10):e109583.
5. Lazer D, Kennedy R, King G, Vespignani A. Big data: the parable of Google Flu: traps in big data analysis. Science 2014;343(6176):1203-5.
6. Signorini A, Segre AM, Polgreen PM. The use of Twitter to track levels of disease activity and public concern in the U.S. during the influenza A H1N1 pandemic. PLoS One 2011;6(5):e19467.
7. Salathe M, Khandelwal S. Assessing vaccination sentiments with online social media: implications for infectious disease dynamics and control. PLoS Comput Biol 2011;7(10):e1002199.
8. Freifeld CC, Brownstein JS, Menone CM, Bao W, Filice R, Kass-Hout T, et al. Digital drug safety surveillance: monitoring pharmaceutical products in twitter. Drug Saf 2014;37(5):343-50.
9. McIver DJ, Brownstein JS. Wikipedia usage estimates prevalence of influenza-like illness in the United States in near real-time. PLoS Comput Biol 2014;10(4):e1003581.
10. Wesolowski A, Eagle N, Tatem AJ, Smith DL, Noor AM, Snow RW, et al. Quantifying the impact of human mobility on malaria. Science 2012;338(6104):267-70.
11. Bengtsson L, Lu X, Thorson A, Garfield R, von Schreeb J. Improved response to disasters and outbreaks by tracking population movements with mobile phone network data: a post-earthquake geospatial study in Haiti. PLoS Med 2011;8(8):e1001083.
12. Zaccai JH. How to assess epidemiological studies. Postgrad Med J 2004;80(941):140-7.
13. Polit DF, Beck CT. Nursing research: generating and assessing evidence for nursing practice. 10th ed. Philadelphia (PA): Lippincott Williams & Wilkins; 2017.
14. Sylvestre E, Bouzille G, Breton M, Cuggia M, Campillo-Gimenez B. Retrieving the vital status of patients with cancer using online obituaries. Stud Health Technol Inform 2018;247:571-5.
15. Tourassi G, Yoon HJ, Xu S, Han X. The utility of web mining for epidemiological research: studying the association between parity and cancer risk. J Am Med Inform Assoc 2016;23(3):588-95.
16. Edoh T. Risk prevention of spreading emerging infectious diseases using a hybridcrowdsensing paradigm, optical sensors, and smartphone. J Med Syst 2018;42(5):91.
17. Moon RJ, Curtis EM, Davies JH, Cooper C, Harvey NC. Seasonal variation in Internet searches for vitamin D. Arch Osteoporos 2017;12(1):28.
18. Chary M, Genes N, Giraud-Carrier C, Hanson C, Nelson LS, Manini AF. Epidemiology from Tweets: estimating misuse of prescription opioids in the USA from social media. J Med Toxicol 2017;13(4):278-86.
19. Towers S, Afzal S, Bernal G, Bliss N, Brown S, Espinoza B, et al. Mass media and the contagion of fear: the case of Ebola in America. PLoS One 2015;10(6):e0129179.
20. Ram S, Zhang W, Williams M, Pengetnze Y. Predicting asthma-related emergency department visits using big data. IEEE J Biomed Health Inform 2015;19(4):1216-23.
21. Zhang X, Dang S, Ji F, Shi J, Li Y, Li M, et al. Seasonality of cellulitis: evidence from Google Trends. Infect Drug Resist 2018;11:689-93.
22. Miller M, Banerjee T, Muppalla R, Romine W, Sheth A. What are people tweeting about Zika? An exploratory study concerning its symptoms, treatment, transmission, and prevention. JMIR Public Health Surveill 2017;3(2):e38.
23. Phillips CA, Barz Leahy A, 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;20(1):e6.
24. 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;11(1):e0005295.
25. Generous N, Fairchild G, Deshpande A, Del Valle SY, Priedhorsky R. Global disease monitoring and forecasting with Wikipedia. PLoS Comput Biol 2014;10(11):e1003892.
26. Adrover C, Bodnar T, Huang Z, Telenti A, Salathe M. Identifying adverse effects of HIV drug treatment and associated sentiments using Twitter. JMIR Public Health Surveill 2015;1(2):e7.

27. Nsoesie EO, Buckeridge DL, Brownstein JS. Guess who's not coming to dinner? Evaluating online restaurant reservations for disease surveillance. J Med Internet Res 2014;16(1):e22.
28. Valson JS, Soman B. Spatiotemporal clustering of dengue cases in Thiruvananthapuram district, Kerala. Indian J Public Health 2017;61(2):74-80.
29. Park SE, Tang L, Bie B, Zhi D. All pins are not created equal: communicating skin cancer visually on Pinterest. Transl Behav Med 2018 Apr 17 [Epub]. https://doi.org/10.1093/tbm/iby044.
30. Dettori M, Arru B, Azara A, Piana A, Mariotti G, Camerada MV, et al. In the digital era, is community outrage a feasible proxy indicator of emotional epidemiology? The case of meningococcal disease in Sardinia, Italy. Int J Environ Res Public Health 2018;15(7):E1512.
31. Roche B, Gaillard B, Leger L, Pelagie-Moutenda R, Sochacki T, Cazelles B, et al. An ecological and digital epidemiology analysis on the role of human behavior on the 2014 Chikungunya outbreak in Martinique. Sci Rep 2017;7(1):5967.
32. 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;64(1):34-41.
33. Hossain N, Househ M. Using HealthMap to analyse Middle East Respiratory Syndrome (MERS) data. Stud Health Technol Inform 2016;226:213-6.