Computers & Industrial Engineering 138 (2019) 106120

Contents lists available at ScienceDirect

Computers & Industrial Engineering

journal homepage: www.elsevier.com/locate/caie

Data mining and machine learning techniques applied to public health problems: A bibliometric analysis from 2009 to 2018

Bruno Samways dos Santos[a,*], Maria Teresinha Arns Steiner[a], Amanda Trojan Fenerich[a], Rafael Henrique Palma Lima[b]

[a] *Industrial and Systems Engineering Graduate Program, Pontifícia Universidade Católica do Paraná (PUCPR), Curitiba, PR, Brazil*
[b] *Department of Industrial Engineering, Universidade Tecnológica Federal do Paraná (UTFPR), Londrina, PR, Brazil*

ARTICLE INFO

Keywords:
Data mining
Machine learning
Public health
Systematic review
Bibliometric analysis

ABSTRACT

The objective of this paper is to present a bibliometric analysis of the applications of Data Mining (DM) and Machine Learning (ML) techniques in the context of public health from 2009 to 2018. A systematic review of the literature was conducted considering three major scientific databases: Scopus, Web of Science and Science Direct. This enabled an analysis of the number of papers by journal, the countries where the applications were carried out, which databases are more commonly used, the most studied topics in public health, and the techniques, programming languages and software applications most frequently used by researchers. Our results showed a slight increase in the number of papers published in 2014 and a significative increase since 2017, focusing mostly on infectious, parasitic and communicable diseases, chronic diseases and risk factors for chronic diseases. The Journal of Medical Internet Research and PLoS ONE published the highest number of papers. Support Vector Machines (SVM) were the most common technique, while R and WEKA were the most common programming language and software application, respectively. The U.S. was the most common country where the studies were conducted. In addition, Twitter was the most frequently used source of data by researchers. Hence, this paper provides an overview of the literature on DM and ML in the field of public health and serves as a starting point for beginner and experienced researchers interested in this topic.

1. Introduction

Public health may be defined as the art and science of preventing diseases, promoting health and prolonging society's life. It is common sense to associate the problems of the dengue virus, malaria and Ebola with public health. However, public health involves a far wider range of problems, such as climate change, the use of tobacco and its derivative products, domestic violence, racism and vaccines (Association, 2019). The consequences of industrialization have impacted biodiversity, ecosystems and the climate on Earth (Lang & Rayner, 2015), causing health problems for the population regarding air quality (Zhu, Wang, Zhang, & Sun, 2012), water pollution (Bichler, Neumaier, & Hofmann, 2014) and other issues.

The discussion over the importance of public health has been highlighted in news reports as well as in academia. Recent examples are the Ebola virus outbreak that hit western Africa between 2014 and 2016 (World Health Organization - WHO. (2018a) (2018a), 2018a), the Zika virus infections, feared since 1947 and with recent cases in Brazil, especially in 2015 (Organization, 2017), and the increasing cases of measles due to gaps in vaccination coverage (World Health Organization - WHO. (2018c) (2018c), 2018c). Society is aware of many of these cases of epidemics and outbreaks, as the authorities have widely and forcefully publicized them in order to alert the population with regard to the dangers and the care that has to be taken in a given area or situation. These situations include warnings against traveling to countries with many cases of yellow fever, a region with an outbreak of the Zika virus or a city with high levels of air pollution. These cases are frequently published in newspapers and on social media because of their widespread impact, justifying the emphasis placed on these problems. Nevertheless, some important studies are not frequently publicized by the media, including acute malnutrition and anemia in children in a refugee camp in Bangladesh (Leidman et al., 2018), excessive pollution caused by dust affecting children in South Korea (Kim, 2018), traffic accidents in Marrakesh, Morocco (Ait-Mlouk, Gharnati, & Agouti, 2017), concerns over the flu virus in Hong Kong (Xu et al., 2017) and warnings in China with regard to an outbreak of Ebola West in Africa (Liu et al., 2016). These cases are very important scientific articles concerning the advances made in detecting and preventing

[*] Corresponding author.
E-mail address: brunosantos@utfpr.edu.br (B.S. dos Santos).

https://doi.org/10.1016/j.cie.2019.106120
Received 30 April 2019; Received in revised form 19 September 2019; Accepted 7 October 2019
Available online 11 October 2019
0360-8352/

situations of public health on both a small and a large scale.

A common issue in healthcare management has to do with public health policies that typically define a set of actions intended to improve some fundamental indicators of public utilities (Anisetti et al., 2018). In addition to the policies, the financial issues of public health have become a delicate and widely debated subject. In the context of the European continent, this concern is now being fiercely debated due to the policy of financial restrictions adopted by governments, resulting in cuts in funding in countries such as Bulgaria, the United Kingdom, Estonia and Lithuania (Rechel, 2019). This issue means that there is a significant risk to the quality of treatment in the public health systems in question. It should be highlighted that the outcomes regarding healthcare are dependent on investment inside and outside the health system, as stated by the Organization for Economic Cooperation and Development (OECD, 2017). The OECD showed that in 2017 only a small fraction of healthcare expenditure (2.8%) is earmarked for prevention activities (an issue prioritized by public health), also highlighting that many of these activities involve healthcare monitoring programs that are not very effective (Gmeinder, Morgan, & Mueller, 2017). An aggravating condition was identified by Muennig (2015) and by Partington, Papakroni, and Menzies (2014), who commented on the difficulty of financing research on public health and emphasized the need to optimize the cost-benefit of projects of this nature.

In the 1930s, public healthcare in Brazil began to be viewed more critically. At the time, actions were taken to create agencies to prevent and monitor diseases. Historically, the country had faced serious administrative difficulties concerning preventive healthcare due to limited scientific, technological and industrial knowledge. However, the feature with the greatest impact was the slow formation of awareness of the rights of citizenship (Fundação Nacional da Saúde - FUNASA, 2017). A common way of addressing the problems experienced by Brazilian states is to set up specialized committees to make decisions on public health that are similar to those that exist in several American states, Europe and Australia. The Lancet Public Health (The Lancet Public Health, 2017) explains that a long-term committee was formed to take care of the British public health system, known as the National Health System (NHS), because the government acknowledged that a culture of short-term thinking dominated the system. The primary focus of the NHS is to devise policies for two major problems, mental health and obesity, which were formerly neglected by health authorities.

Mental health encompasses different types of diseases such as dementia, depression, epilepsy, headaches, schizophrenia and autism (World Health Organization - WHO. (2014) (2014), 2014). This dimension is an example in which research on medicine and psychology could be aided by statistics and data mining (DM) techniques. Due to technological innovations and the mining of social media data, it has been argued that Natural Language Processing (NLP) and Machine Learning (ML) can assist researchers in these fields and ultimately impact the health of individuals and the general population (Conway & O'Connor, 2016). Obesity is another example where NLP and ML techniques can help researchers to categorize patients and devise treatments, as well as preventive measures. Obesity has been linked to several cases of mortality, cardiovascular diseases and metabolic disorders (Ortega Hinojosa et al., 2014). This issue is even more serious when studying children's behavior (Kang, Wang, Zhang, & Zhou, 2017), since the number of overweight children rose to 41 million in 2016 (World Health Organization - WHO. (2018b) (2018b), 2018b).

Technological innovations have created new possibilities for electronically recording vast amounts of data in public and private databases. A term that has been widely used in the literature is "big data", which has to do with the discovery of information from a vast quantity of data, based on the premise that most activities generate data at a low cost and can be used as a basis for decision making (Torrecilla & Romo, 2018). Institutions from different fields have used big data to understand better the issues related to healthcare and their corresponding variables, always taking due care with the implications of this type of interaction. In this respect, they have cited privacy, data security and the consent of those involved in the studies (Salerno, Knoppers, Lee, Hlaing, & Goodman, 2017). In the Brazilian context, opportunities arose to use big data from multiple sources of information on healthcare, such as the databases of the Health Ministry, the Information System on Born-Alive Children (SINASC), the Information System on Mortality (SIM), cards of the Brazilian National Health Service (SUS) and partnerships between national and international institutions, characterizing the development of multicenter research (Chiavegatto Filho, 2015).

The vast amount of data has been analyzed by researchers to extract specialized and non-trivial knowledge. This can be done by a process known as Knowledge Discovery in Databases (KDD), the main stage of which is the use of DM and ML techniques to extract knowledge from databases, i.e. identify patterns and make predictions based on the stored data (Fayyad, Piatetsky-Shapiro, & Smyth, 1996; Lara, Lizcano, Martínez, & Pazos, 2014). Because DM techniques can efficiently analyze large databases related to public health issues, it is important to verify how the scientific literature has applied such approaches in the context of public health to discover new knowledge. An analytical approach to surveys helps to understand the evolution and trends in specific study areas, including their related tools and methods. This also applies to the wide field of public health.

The objective of this paper is to present a bibliometric analysis based on a systematic review of the literature on the applications of DM techniques on a broader scope of problems related to public health. More specifically, this paper aims to: (i) analyze the number of papers published from 2009 to 2018 (10 years) due to the increasing number of publications and dissemination of ML in public health; (ii) identify the journals with the greatest number of papers; (iii) determine which techniques, programming languages and software tools are most widely used in the field of DM applied to public health; (iv) identify which countries and databases were targeted by these studies; (v) analyze which public health classes were tackled by these papers and (vi) identify which papers were most frequently cited in the literature.

The remainder of this paper is organized as follows. Section 2 presents the theoretical background of DM and ML applied to public health problems. Section 3 discusses the methodological procedures, including the research terms and the flowchart used in the systematic literature review. Section 4 presents and discusses the results. Section 5 concludes the paper and suggests directions for future research in this field.

2. Theoretical background

This section presents the DM approaches and concepts and correlated works on the semantic and/or bibliometric review in the context of public health.

2.1. Data mining and machine learning applications in public health

DM and ML have been extensively used to research health problems. Several studies have been carried out using a wide variety of tools. In the field of obesity, Ortega Hinojosa et al. (2014) used data from the U.S. census. Kang et al. (2017) and Ghosh and Guha (2013) analyzed textual data from Twitter. In addition, Zhang et al. (2009) studied the topic using a specialized database from the United Kingdom (UK), whereas Lazarou, Karaolis, Matalas, and Panagiotakos (2012) applied pattern identification techniques using a database from Cyprus.

Examples of DM and ML usage can be found in other health-related research. Huang, Lewis, and Britton (2014) and Myslín, Zhu, Chapman, and Conway (2013) studied smoking cessation in the UK and applied sentiment analysis using data from Twitter. Authors such as Deb and Liew (2016), Kumar and Toshniwal (2016), Pakgohar, Tabrizi, Khalili, and Esmaeili (2011) and Griselda, Juan, and Joaquín (2012) worked with data on traffic accidents from different parts of the world. Banks et al. (2005), Dunn, Leask, Zhou, Mandl, and Coiera (2015), Massey et al. (2016) and Du, Xu, Song, and Tao (2017) conducted studies on vaccines and their reactions. It should be noted that many public health problems can be researched using databases available around the world.

2.2. Systematic literature review and related work

A systematic review of the literature consists of a search for papers and scientific publications related to one or more topics using a predefined protocol including the search terms (Kitchenham & Charters, 2007), the scientific databases to be searched, as well as the selection and assessment criteria. The protocol should be defined in such a manner as to allow other researchers to reproduce it (Ahmed, Ahmad, Ahmad, & Zakaria, 2018; Kitchenham & Charters, 2007; Kitchenham, 2004). Tranfield, Denyer, and Smart (2003) contend that a systematic review includes quantitative and qualitative analyses of the literature to broaden the debate on the topics in question. Bibliometric measures are useful to represent the bibliographic materials collected during the review (Cancino, Merigó, Coronado, Dessouky, & Dessouky, 2017). Petticrew and Roberts (2008) argue that systematic reviews can overcome limitations of older studies by aggregating samples from different databases, which enables more relevant and up-to-date results to be found.

The last decade has seen an increase in the number of scientific studies that use systematic reviews as their core method. This has also been the case in the fields of public health applications of DM and ML, with many recent contributions to this topic. For example, Carroll et al. (2014) conducted a systematic review considering the period between 1980 and 2013 to identify visualization tools used by public health professionals with an emphasis on social network analysis and geographic information systems. Dallora, Eivazzadeh, Mendes, Berglund, and Anderberg (2016) investigated the main objectives and variables of the ML and micro-simulation studies applied to the prognosis of dementia. They verified that using neuroimaging to predict Alzheimer characteristics was the most frequent objective of the 37 studies found. Bellinger, Mohomed Jabbar, Zaïane, and Osornio-Vargas (2017) conducted a systematic review on the application of DM in the epidemiology of air pollution using three scientific databases covering the period up to October 2017 and analyzed 47 relevant papers.

Kadi, Idri, and Fernandez-Aleman (2017) conducted a systematic review on the field of cardiology to identify the main techniques used by researchers in the period between 2000 and 2015 and found 149 papers in accordance with their research protocol. Kavakiotis et al. (2017) conducted a broad systematic review covering the application of DM techniques in studies on diabetes. The authors found that 85% of the papers used supervised learning, with Support Vector Machines (SVM) being the most common technique. As for the 15% of studies that used unsupervised learning, most of them relied on association rules. In the context of bipolar disorder, Librenza-Garcia et al. (2017) did a systematic review using three scientific databases to understand the application of DM on this topic considering all papers published up to January 2017. O'Shea (2017) reviewed three databases to identify internet-based bio-vigilance systems that handle large amounts of data. The authors found 99 papers that reported 50 event-based biovigilance systems, which underlines the importance of official and unofficial sources of data for the biovigilance of disease outbreaks.

Ahmadi, Gholamzadeh, Shahmoradi, Nilashi, and Rashvand (2018) demonstrated the contribution of fuzzy logic methods in the diagnosis of diseases. For this purpose, eight databases were selected, limited to the works found from January 2005 to June 2017, identifying 46 articles that met the inclusion criteria. Alonso et al. (2018) systematically reviewed the DM techniques and algorithms in the context of mental illnesses from 2008 to 2018 in the Google Scholar, IEEE Xplore, PubMed, ScienceDirect, Scopus and Web of Science databases. As for the predictive model for risk of readmission to hospitals, Artetxe, Beristain, and Graña (2018) conducted a systematic review of the predictive methods applied in this field in databases such as PubMed and Google Scholar, identifying 77 studies published up to September of 2017 that satisfied the inclusion criteria. Egan et al. (2018) carried out a systematic review using six scientific databases to study dementia focusing on intervention programs using the internet to train caregivers of patients with this disability. Islam et al. (2018) performed a systematic review and a later meta-analysis on the performance of ML models to predict sepsis, with seven studies satisfying the criteria for the research protocol they defined. For this purpose, documents from January 2000 to March 2008 were analyzed from the PubMed, EMBASE, Google Scholar and Scopus databases.

Lee et al. (2018) investigated how ML algorithms aid the selection of treatment and personalization of therapies in people suffering from depression through a systematic review and meta-analysis using the Ovid MEDLINE/PubMed, Cochrane Controlled Trials Register, ClinicalTrials.gov and Google Scholar databases from the earliest publications up to 8 February 2018. A meta-analysis was conducted by Nindrea, Aryandono, Lazuardi, and Dwiprahasto (2018) to gauge the accuracy of diagnosis of ML algorithms regarding the risk of breast cancer. The period under study was from January 2000 to May 2018 in databases such as PubMed, ProQuest and EBSCO. The study found in 11 articles that SVM had better results compared with the other methods that were identified, such as Artificial Neural Networks (ANN), Decision Tree (DT), Naïve Bayes (NB) and KNN. Rybarczyk and Zalakeviciute (2018) used a systematic review protocol to identify recent studies on ML applied to air pollution. The study involved articles published in the SCOPUS database limited to the period ranging from 2010 to 2018, identifying 46 articles relevant to the scope of the study. William, Ware, Basaza-Ejiri, and Obungoloch (2018) reviewed the state of the art of recent publications focused on the automated detection of cervical cancer through pap-smear images. Studies were identified in four scientific bases (Google Scholar, Scopus, IEEE and ScienceDirect) over the last 15 years (2004–2018), enabling the authors to determine that SVM and K-Nearest Neighbour (KNN) proved to be excellent classifiers of two classes of problems (normal and abnormal). Yassin, Omran, El Houby, and Allam (2018) used four scientific databases and found 154 papers that led to the identification of ML techniques and other assessment indicators used to detect breast cancer.

Burke, Ammerman, and Jacobucci (2019) conducted a systematic literature review on the application of ML techniques to predict thoughts and behaviors regarding self-inflicted wounds (suicidal and non-suicidal) from five databases up to February 2018. Finally, Dwivedi, Imtiaz, and Rodriguez-Villegas (2019) critically analyzed all existing approaches for the automatic identification and classification of heart sounds based on 117 peer-reviewed articles found for the period ranging from 1963 to 2018. The bases used for this purpose were IEEE Xplore, Scopus, PubMed, Web of Science, ScienceDirect, Google Scholar, EMBASE and ACM Digital Library.

The works that used a systematic review in a field of public health are summarized in chronological order in Table 1.

The systematic reviews indicate that many authors study the application of DM and ML to specific public health problems, such as dementia, diabetes or air pollution. However, there are no systematic reviews that address the application of DM and ML considering multiple public health problems. Therefore, the main contribution of this paper is that it conducts a systematic review and bibliometric analysis of the literature on the application of DM and ML on the broader context of public health problems covering the past 10 years (from 2009 until 2018) of peer-reviewed research.

3. Methodology

From the premise that there is no unique rule for conducting a systematic literature review, this paper used a methodology similar to those reported by Ngai, Xiu, and Chau (2009) and Hasan et al. (2017) to define the search criteria and the systematic review flowchart. The

Table 1
Summary of works from a systematic review in public health.

Authors (Year)	Focus of the review	N. of databases
Carroll et al. (2014)	Visualization tools used by public health professionals with an emphasis on social network analysis and geographic information systems	5
Dallora et al. (2016)	Objectives and variables of the ML and micro-simulation studies applied to the prognosis of dementia	3
Bellinger et al. (2017)	DM in the epidemiology of air pollution	3
Kadi et al. (2017)	Main DM techniques applied to cardiology	8
Kavakiotis et al. (2017)	Application of DM techniques in studies on diabetes	2
Librenza-Garcia et al. (2017)	Application of DM on bipolar disorder	3
O'Shea (2017)	Internet-based bio-vigilance systems that handle large amounts of data	3
Ahmadi et al. (2018)	Fuzzy logic methods in the diagnosis of diseases	8
Alonso et al. (2018)	DM techniques and algorithms in the context of mental illnesses	6
Artetxe et al. (2018)	Predictive models for the risk of readmission in hospitals	2
Egan et al. (2018)	Dementia focusing on intervention programs using the internet to train caregivers	6
Islam et al. (2018)	Performance of ML models for predicting sepsis	4
Lee et al. (2018)	ML algorithms aid the selection of treatment and personalization of therapies for people with depression	4
Nindrea et al. (2018)	Gauging the accuracy of diagnosis of ML algorithms on the risk of breast cancer	3
Rybarczyk and Zalakeviciute (2018)	ML applied to air pollution	3
William et al. (2018)	Automated detection of cervical cancer detected by pap-smear images	4
Yassin et al. (2018)	ML techniques and other assessment indicators used to detect breast cancer	4
Burke et al. (2019)	Application of ML techniques to predict thoughts and behaviors regarding self-inflicted wounds (suicidal and non-suicidal)	5
Dwivedi et al. (2019)	Existing approaches for the automatic identification and classification of heart sounds	8

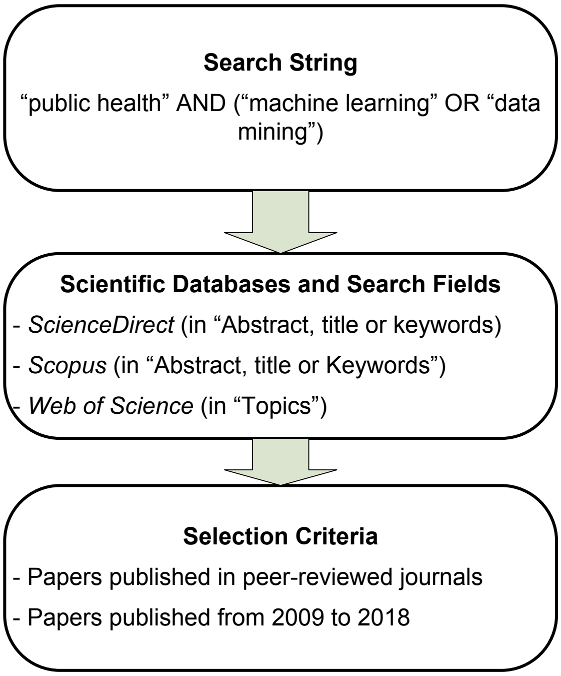
Search String
"public health" AND ("machine learning" OR "data mining")
Scientific Databases and Search Fields
- ScienceDirect (in "Abstract, title or keywords)
- Scopus (in "Abstract, title or Keywords")
- Web of Science (in "Topics")
Selection Criteria
- Papers published in peer-reviewed journals
- Papers published from 2009 to 2018

Fig. 1. Initial search parameters.

research question that guided this research was stated as follows: "What is the present scenario on the utilization of DM and ML techniques for problems related to public health?" This led to the definition of the search scope, search string and selection criteria described in Fig. 1.

After conducting the search using the parameters in Fig. 1, 1057 papers were found. Following this initial stage, all the references were organized using Mendeley Desktop software, which automatically identifies most of the duplicate references. The papers were then separated into two distinct groups according to the effort required to analyze them. The papers that only required the reading of the title and abstract were placed in Group A. The criteria for excluding papers from further analysis in this group were:

- Papers not published in English;
- Papers that did not have any relationship with the research question, such as debates on general issues concerning public health problems.

The papers placed in Group B required a thorough reading of their content. The criteria for excluding papers from further analysis in this group were:

- Papers in which DM and/or ML techniques were not mentioned;
- Papers whose focus was more closely related to developing a particular software tool or system, not addressing the application of DM and/or ML techniques to study public health problems in any sense;
- Papers whose full texts were not available for download;
- Papers in which the applications were too specific, such as DM and/or ML for genetic analysis or the veterinary field, for example.

The flowchart in Fig. 2 summarizes the procedures used in the systematic review of the literature reported in this paper.

As shown in Fig. 2, following all the stages of the systematic review, 250 papers were deemed suitable for further analysis and were considered during the bibliometric phase.

4. Results and discussion

Considering all 250 papers selected for the bibliometric analysis, it was possible to separate them by year of publication, which resulted in the time series shown in Fig. 3.

Fig. 3 indicates a significant increase in the number of papers that discuss applications of DM and ML to public health problems. In 2014, there was a slight rise in the number of publications compared with previous years. This small rise can mainly be explained by the growth of electronic databases and the scanning of documents, encouraging researchers to extract useful information aided by computational approaches. However, the number of articles published in 2017 and 2018 should be highlighted, showing that the field of data science has also been consolidated in the field of public health, mainly due to the dissemination of techniques, courses and access to electronic databases. Fig. 4 shows the number of papers by journal.

An interesting finding from Fig. 4 is that the journal titles often combine terms related to "informatics" and "medical", such as the *Journal of Medical Internet Research*, *Journal of Biomedical Informatics* and *BMC Medical Informatics and Decision Making*, while two journals have the term "public health" (in this case, *International Journal of Environmental Research and Public Health* and *BMC Public Health*). The

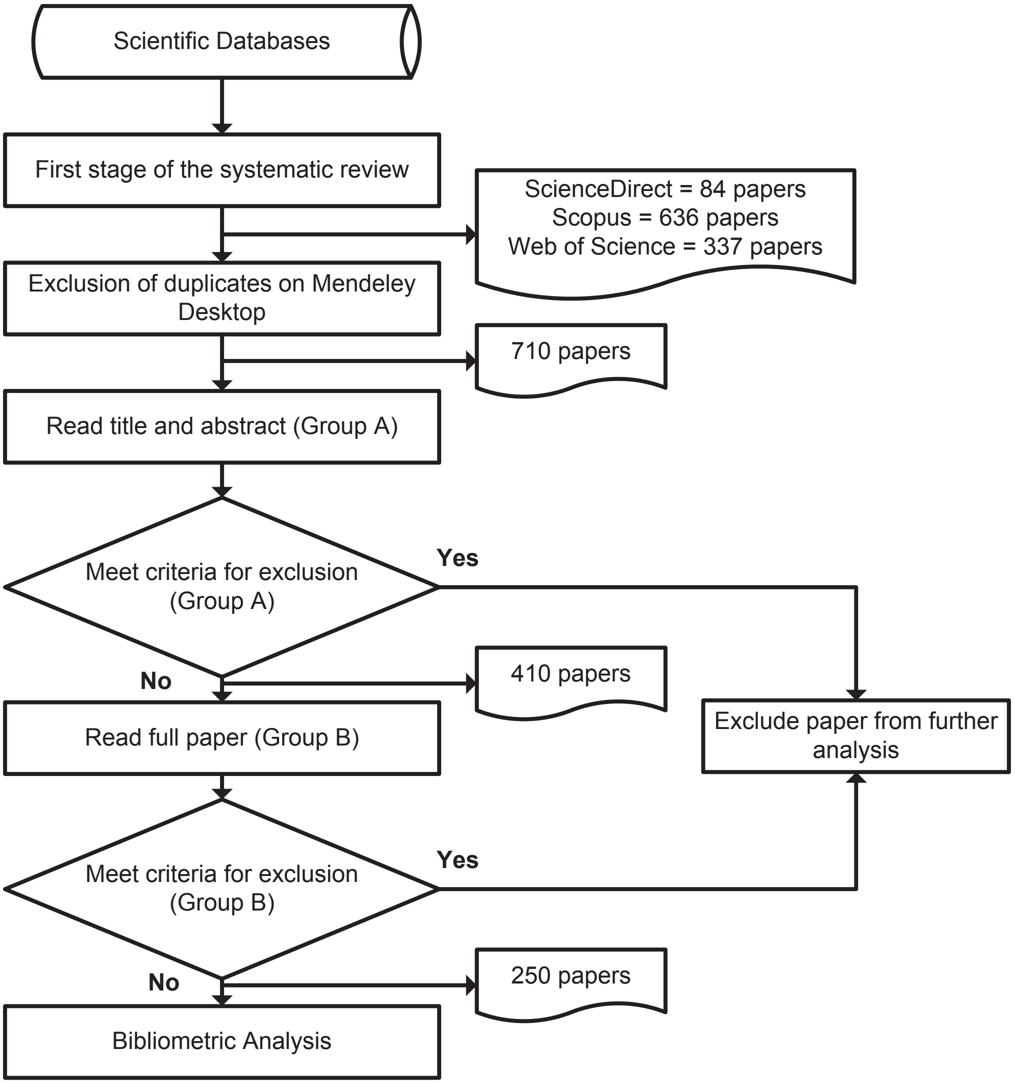
Scientific Databases
First stage of the systematic review
ScienceDirect = 84 papers
Scopus = 636 papers
Web of Science = 337 papers
Exclusion of duplicates on Mendeley Desktop
710 papers
Read title and abstract (Group A)
Meet criteria for exclusion (Group A)
Yes
No
410 papers
Read full paper (Group B)
Exclude paper from further analysis
Meet criteria for exclusion (Group B)
Yes
No
250 papers
Bibliometric Analysis

Fig. 2. Sytematic literature review flowchart.

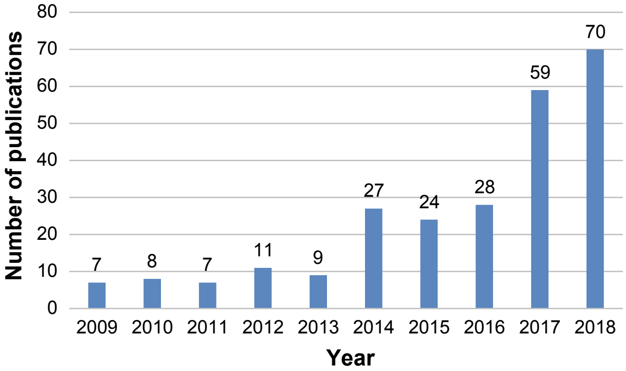

Fig. 3. Number of papers (250) published by year (2009–2018).

journal with the highest number of papers is the *Journal of Medical Internet Research*, with 16 studies published. This open-access journal has an impact factor of 4.671 (2017) and has published scientific research for more than 20 years. The second-best ranking journal was PLoS ONE with 15 papers and an impact factor of 2.766 (2015), followed by the *International Journal of Environmental Research and Public Health* and *BMC Medical Informatics and Decision Making* with 7 papers each and impact factors of 2.145 (2017) and 2.134 (2017), respectively. It should be highlighted that many journals ended up not entering this ranking due to their scope focusing on innovative DM or ML techniques, in addition to applications in distinct areas of public health.

Fig. 5 shows which DM and ML techniques were most used in the papers considered during the bibliometric phase. SVM was the most common technique, appearing in 65 papers, followed by DT, with 57 papers, Random Forest (RF), with 48 papers, Logistic Regression (LR), with 43 papers, NB, with 33, ANN, with 27, and finally K-nearest neighbor (KNN), with 21 papers. All the aforementioned techniques are commonly used for supervised learning, whereas K-Means Clustering (KMC) and Least Absolute Shrinkage and Selection Operator (LASSO), combining for 17 papers and Association Rule Mining (ARM), with 15 articles, are used for unsupervised learning. It is worth noting that many papers used text mining and sentiment analysis to classify the concerns of individuals towards specific public health problems, such as virus outbreaks and diabetes. This may explain why so many classification techniques have been employed by researchers.

According to Fig. 6, the R language, Waikato Environment for Knowledge Analysis (WEKA) software, the Python language and MATLAB® software are the most popular among the studies analyzed during the bibliometric phase. Some of these tools can be downloaded for free, and include several toolboxes or packages that implement DM and ML techniques and make it easier for researchers to apply them.

Even though MATLAB® is a paid software tool, it is still very popular

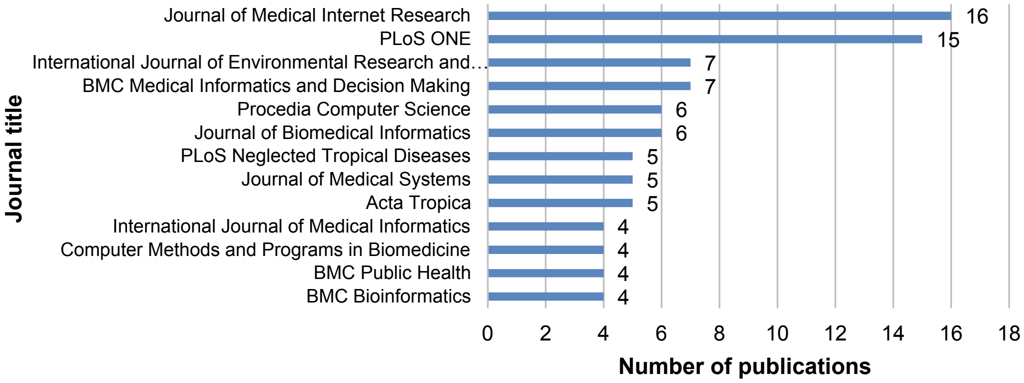

Fig. 4. Number of papers (250) researched by journal (2009–2018).

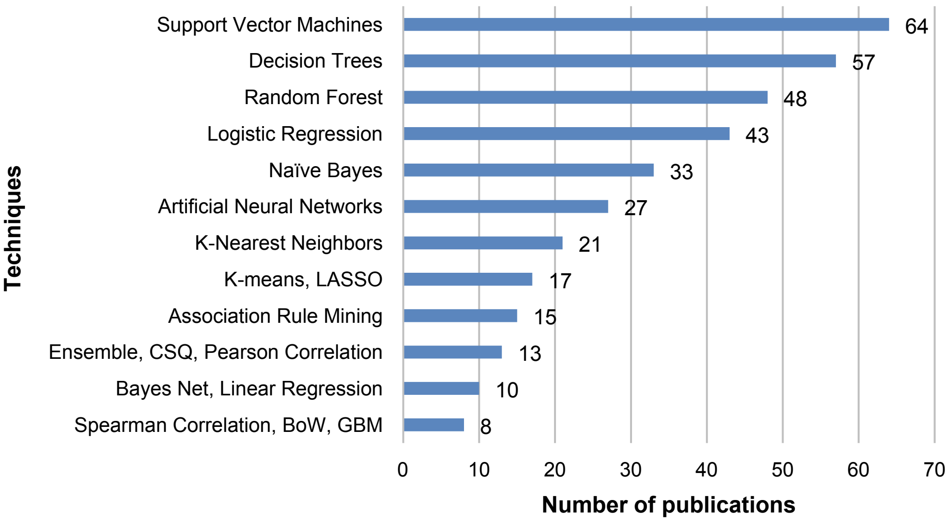

Fig. 5. Most common techniques used by researchers in the 250 documents. Legend: BoW (Bag-of-Words); CSQ (Chi-Squared Test); GBM (Gradient Boosted Machine); LASSO (Least Absolute Shrinkage and Selection Operator).

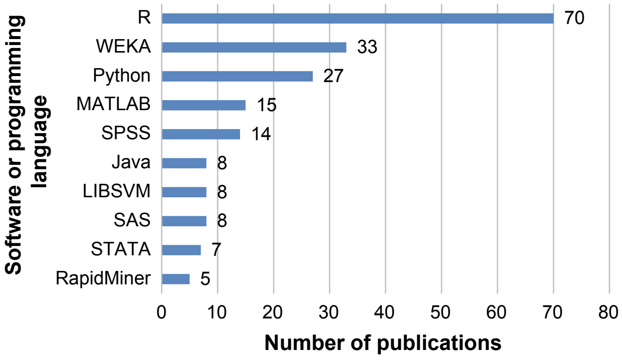

Fig. 6. Most common software tools and programming languages.

Table 2
Regions and databases most targeted by researchers.

Region	Frequency	Database	Frequency
U.S.	60	Twitter	36
China	17	Google Trends	8
Taiwan	14	UCI	7
South Korea	11	Google News, Weibo, VAERS, Bing, BRFSS, Baidu	3
United Kingdom	9		
Iran	7		
Brazil, India	5		
Spain, Turkey	4		

UCI (University of California at Irvine); VAERS (Vaccine Adverse Event Reporting System); BRFSS (Behavioral Risk Factor Surveillance System).

among researchers due to its powerful toolboxes. However, many have opted for free alternatives, such as the R and Python languages. As for WEKA, which is also a free alternative, one of its main advantages is its friendly and self-explanatory user interface, which attracts many researchers that are not interested in programming or algorithm development, only the easy-to-use concept and results achieved.

Table 2 shows the number of papers by the regions targeted by the DM and ML applications and the databases used, when available. It is important to highlight that this analysis is not about the country where researchers work, but rather the countries where the reported application took place. The greatest number of papers (60) report applications in the U.S., which is four times higher than the 2nd and 3rd ranking countries (China and Taiwan), with 17 and 14 papers, respectively. South Korea was studied in 11 cases, followed by the United Kingdom (9), Iran (7) Brazil and India (5), Turkey and Spain (4).

As for the data sources and databases used in the papers, it should be noted that a significant number of studies used Twitter as the main data source and later applied techniques for DM and ML for classification. Google Trends was another common source of data, while the Vaccine Adverse Event Reporting System (VAERS) database was a common data source for studies on adverse responses to vaccines and other medicines. The University of California at Irvine (UCI) has a public repository of databases commonly used to test DM techniques for classification, association and clustering, using some database related

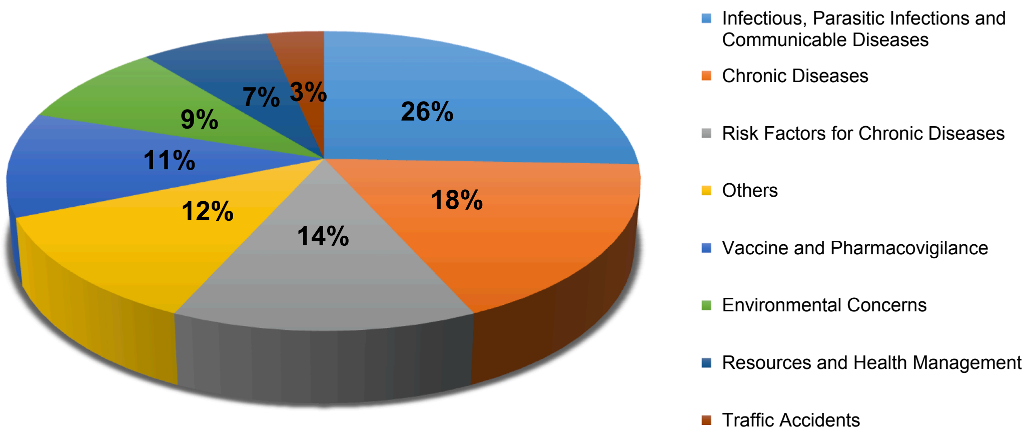

Fig. 7. Fields of public health addressed by the 250 papers.

to a public health problem. Fig. 7 shows the main topics addressed by the papers considered during the bibliometric analysis.

Fig. 7 illustrates that research on public health revolves around three main topics: infectious, parasitic, and communicable diseases; chronic diseases, and; risk factors for chronic diseases. These topics were divided into eight different classes, many of them in accordance with the classification suggested by the Center for Disease Control and Prevention (CDC). It is possible for an article to be assigned to more than one class. Fig. 8 shows the number of papers published according to the area explored in the past 10 years (2009–2018). The classes and some examples of public health issues are shown in Table 3.

Analyzing Figs. 7 and 8 and also Table 3, the interest in infectious diseases may be explained by two main factors:

i. The growth in the number of studies related to the likelihood of outbreaks of diseases such as Ebola, dengue, malaria and the influenza virus;
ii. The increasing popularity of social media tools, such as Weibo and Twitter, which enable individuals to post information that can be used later by researchers.

Analyzing the articles on infectious diseases, there are many opportunities to explore social media to chart diseases of this nature, as there has been a clear manifestation on the part of society with regard to these issues on these networks, showing that people feel comfortable about revealing their feelings on Twitter, Facebook and Weibo, for instance. Studies using space-time approaches have proved to be effective when it comes to detecting signs that might indicate the prevalence of some kind of infectious disease, predicting in which locations and at what time a certain disease could become an epidemic. With this type of analysis, public health authorities could define effective prevention strategies and promote healthcare to meet the demands of these situations.

Finally, the most frequently cited documents, in order of their total number of citations (TC), are shown in Table 4.

Table 4 shows that six articles were cited at least 100 times, demonstrating the importance of applications of DM and/or ML in the field of public health. Summarizing the first three works (with over 200 citations), the study by Nikfarjam, Sarker, O'Connor, Ginn, and Gonzalez (2015) sought to mine adverse drug reactions on social media using a new system of information extraction called ADRMine (based on Conditional Random Fields – CRF). This application used data collected from Twitter and DailyStrength (the latter dedicated to support groups) and classified reactions based on methods such as SVM, MetaMap and Lexicon-based, all compared with ADRMine. The results showed that ADRMine had a better performance in both databases. Myslín et al. (2013) analyzed feelings related to tobacco consumption using posts on Twitter to gauge users' perceptions with regard to tobacco and emerging products, focusing particularly on electronic cigarettes and hookahs. 7362 tweets were collected with the help of the Twitter Application Programming Interface (API) program, and the related tweets were classified (or not) with regard to the use of tobacco using the KNN, SVM and NB methods with 10-fold cross-validation, using the program developed in C language known as Rainbow Toolkit. The results showed that positive feelings were highly prevalent, which is correlated with social image, personal experience and the products, such as the hookah and electronic cigarettes. Furthermore, the SVM technique succeeded in classifying more accurately posts related to

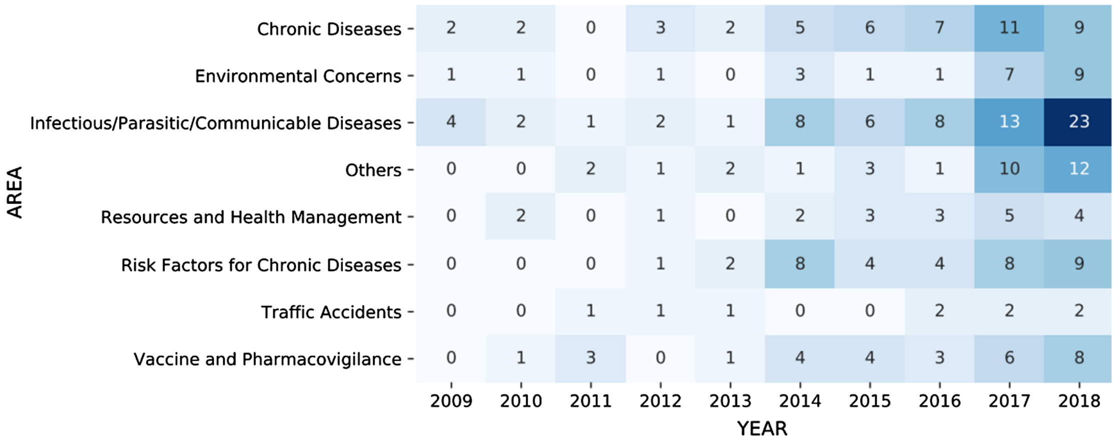

Fig. 8. Number of papers published according to the area explored in the past 10 years (2009–2018).

Table 3
Classes and examples in public health areas.

Class	Examples
Infectious, Parasitic and Communicable Diseases	Arboviruses: Dengue, Chikungunya and Zika viruses; malaria; hepatitis B; Sexually Transmitted Infections (HIV, syphilis, gonorrhea); tuberculosis; flu
Chronic Diseases	Alzheimer’s diseases; breast, cervical and skin cancer; diabetes; heart disease; obesity; high blood pressure; strokes
Risk Factors for Chronic Diseases	Tobacco use; secondhand smoke; poor nutrition; lack of physical activity; excessive alcohol use
Others	Preterm birth; verbal autopsy analysis; elderly people concerns; self-care; suicide
Vaccine and Pharmacovigilance	Pharmacological risk factors; vaccine adverse reactions
Environmental Concerns	Landscape analysis; air pollution; water quality
Resources and Health Management	Clinical notes classification; health care social media; health outcomes
Traffic Accidents	Traffic accidents with general vehicles
Others	Preterm birth; verbal autopsy analysis; elderly people concerns; self-care; suicide

tobacco from those that were not related. Finally, J. Huang, Kornfield, Szczypka, and Emery (2014) collected and analyzed tweets related to electronic cigarettes and possible commercial relationships. For this purpose, Twitter Firehouse was used to collect data from May 2012 to June 2012, classifying the tweets as “commercial” or “non-commercial” with the NB technique, achieving a better performance in the classification of commercial tweets compared with non-commercial tweets.

Table 4 also makes it clear that there is a community of researchers that opt to conduct studies using social media or online news, and these have been references for other works, showing that this kind of approach may be due to the easy access to information on people from any part of the world. For this finding, the studies of Nikfarjam et al. (2015), Myslín et al. (2013), Huang et al. (2014), Dunn et al. (2015), Odlum and Yoon (2015), Paul and Dredze (2014), Zhang et al. (2009) and Ghosh and Guha (2013) may be cited.

5. Concluding remarks

The present study analyzed 250 articles related to the DM and ML techniques published from 2009 to 2018 (10 years) through a systematic review based on the structures of Ngai et al. (2009) and Hasan et al. (2017). Based on the proposed research methodology, the field of public health was found to have been significantly investigated using DM and ML techniques, showing a slight growth starting in 2014 and greater interest in the last two years, surpassing the mark of 50 articles (59 and 70 in 2017 and 2018, respectively; Fig. 3). As expected, journals that combine themes related to health and computer science published the most articles, with particular interest in applications that use data obtained from social media, showing that data from sources of this nature have great potential for exploitation, especially in public health. Many authors used classical techniques for the task of classification, with emphasis on SVM, DT, RF, LR, NB, ANN and KNN, with programming languages such as R, Python and MATLAB®. However, the free software WEKA proved to be the one favored by researchers who preferred a friendly interface and direct analysis. Most researchers are American, and chose Twitter as the base of application and preferred to address public health issues such as infectious, parasitic and communicable diseases, followed by chronic diseases and related risk factors.

The results obtained in this article allow researchers to use the references and trends identified to guide relevant future research in this field, exploring gaps regarding the type of problem or the DM and ML techniques cited here. Thus, people linked to public health, academics and data analysts can use this study to find new approaches to improve public health, making an impact at the local and worldwide level, using the most commonly investigated current problems as a reference.

Nevertheless, it is important to point out some limitations identified in the development of this work, examples of which are listed below:

- The key words used in the systematic review were chosen based on

Table 4
The 15 out of 250 most cited documents.

R	TC	Title	Author/s (Year)
1	249	Pharmacovigilance from social media: Mining adverse drug reaction mentions using sequence labeling with word embedding cluster features	Nikfarjam et al. (2015)
2	231	Using Twitter to examine smoking behavior and perceptions of emerging tobacco products	Myslín et al. (2013)
3	220	A cross-sectional examination of marketing of electronic cigarettes on Twitter	Huang et al. (2014)
4	152	Intelligible support vector machines for diagnosis of diabetes mellitus	Barakat, Bradley, and Barakat (2010)
5	129	The definition of insulin resistance using HOMA-IR for Americans of Mexican descent using machine learning	Qu, Li, Rentfro, Fisher-Hoch, and McCormick (2011)
6	124	Detecting drug interactions from adverse-event reports: Interaction between paroxetine and pravastatin increases blood glucose levels	Tatonetti et al. (2011)
7	87	Associations between exposure to and expression of negative opinions about human papillomavirus vaccines on social media: An observational study	Dunn et al. (2015)
8	82	– What can we learn about the Ebola outbreak from tweets? – Prediction of adverse drug reactions using decision tree modelling	Odlum and Yoon (2015), Hammann, Gutmann, Vogt, Helma, and Drewe (2010)
9	80	Discovering health topics in social media using topic models	Paul and Dredze (2014)
10	69	Data mining to generate adverse drug events detection rules	Chazard, Ficheur, Bernonville, Luyckx, and Beuscart (2011)
11	66	Automatic online news monitoring and classification for syndromic surveillance	Zhang, Dang, Chen, Thurmond, and Larson (2009)
12	64	What are we 'tweeting' about obesity? Mapping tweets with topic modeling and Geographic Information System	Ghosh and Guha (2013)
13	53	Spatial analysis of plague in California: Niche modeling predictions of the current distribution and potential response to climate change	Holt, Salkeld, Fritz, Tucker, and Gong (2009)
14	43	Mortality risk score prediction in an elderly population using machine learning	Rose (2013)
15	41	Feasibility of Obtaining Measures of Lifestyle From a Smartphone App The MyHeart Counts Cardiovascular Health Study	McConnell et al. (2017)

R (Rank); TC (Total citations).

preliminary tests using different combinations of search terms. Some examples of previously tested terms were "data mining technique*", "big data", "public health problems" and "public health issues". These terms did not have a satisfactory return that answered the research question of the review. On the other hand, the definitive terms used in the study (Fig. 1), may have failed to identify relevant articles within the scope of the research;
- Within the scope of the systematic review, no filter was used with regard to the quality of the journals, such as having a minimum number of citations, forming a highly diversified portfolio;
- The number of databases that were chosen was established by the authors' knowledge and access. Therefore, some other scientific databases that were not used might have included relevant articles that fell within the scope of the study.

Based on the findings of this study, suggestions for future research can be made:

- Add more databases to the present study in an attempt to find more relevant and current articles;
- Conduct an in-depth analysis of the variation of techniques over the years, as it has been noted that some of the most current articles used deep learning approaches (as can be seen, for example, in the studies of Marinelarena-Dondena, Ferretti, Maragoudakis, Sapino, and Luis Errecalde (2017), Song and Kim (2017), Choi, Lee, Yoon, Won, and Kim (2018), Kwon, Lee, Lee, Lee, and Park (2018), Miao and Miao (2018) and Subramani, Wang, Vu, and Li (2018)) and also an ensemble (works such as Marucci-Wellman, Corns, and Lehto (2017), Rao and Makkithaya (2017), Sun, Ren, and Ye (2017), Zhan et al. (2017), Bentley, Baker, Simons, Simpson, and Blakely (2018) and Herrera et al. (2018)). In this context, the hypothesis could be raised that there is a tendency to increase the number of more sophisticated approaches that could result in a better performance of different DM tasks.

Acknowledgements

This study was partially funded by PUC-PR and by the Coordination for the Improvement of Higher Education Personnel – Brazil (CAPES; 3rd author) and by the National Council for Scientific and Technological Development – Brazil (CNPq; 2nd author).

References

Ahmadi, H., Gholamzadeh, M., Shahmoradi, L., Nilashi, M., & Rashvand, P. (2018). Diseases diagnosis using fuzzy logic methods: A systematic and meta-analysis review. *Computer Methods and Programs in Biomedicine, 161*, 145–172. https://doi.org/10.1016/j.cmpb.2018.04.013.

Ahmed, Y. A., Ahmad, M. N., Ahmad, N., & Zakaria, N. H. (2018). Social media for knowledge-sharing: A systematic literature review. *Telematics and Informatics, 37*, 72–112. https://doi.org/10.1016/j.tele.2018.01.015.

Ait-Mlouk, A., Gharnati, F., & Agouti, T. (2017). An improved approach for association rule mining using a multi-criteria decision support system: A case study in road safety. *European Transport Research Review, 9*(3), https://doi.org/10.1007/s12544-017-0257-5.

Alonso, S. G., de la Torre-Díez, I., Hamrioui, S., López-Coronado, M., Barreno, D. C., Nozaleda, L. M., & Franco, M. (2018). Data mining algorithms and techniques in mental health: A systematic review. *Journal of Medical Systems, 42*(9), https://doi.org/10.1007/s10916-018-1018-2.

America Public Health Association - APHA. (2019). Topics & Issues. Retrieved March 12, 2019, from https://www.apha.org/topics-and-issues.

Anisetti, M., Ardagna, C., Bellandi, V., Cremonini, M., Frati, F., & Damiani, E. (2018). Privacy-aware Big Data Analytics as a service for public health policies in smart cities. *Sustainable Cities and Society, 39*, 68–77. https://doi.org/10.1016/j.scs.2017.12.019.

Artetxe, A., Beristain, A., & Graña, M. (2018). Predictive models for hospital readmission risk: A systematic review of methods. *Computer Methods and Programs in Biomedicine, 164*, 49–64. https://doi.org/10.1016/j.cmpb.2018.06.006.

Banks, D., Woo, E. J., Burwen, D. R., Perucci, P., Braun, M. M., & Ball, R. (2005). Comparing data mining methods on the VAERS database. *Pharmacoepidemiology and Drug Safety, 14*(9), 601–609. https://doi.org/10.1002/pds.1107.

Barakat, N., Bradley, A. P., & Barakat, M. N. H. (2010). Intelligible support vector machines for diagnosis of diabetes mellitus. *IEEE Transactions on Information Technology in Biomedicine, 14*(4), 1114–1120. https://doi.org/10.1109/TITB.2009.2039485.

Bellinger, C., Mohomed Jabbar, M. S., Zaïane, O., & Osornio-Vargas, A. (2017). A systematic review of data mining and machine learning for air pollution epidemiology. *BMC Public Health, 17*(1), https://doi.org/10.1186/s12889-017-4914-3.

Bentley, R., Baker, E., Simons, K., Simpson, J. A., & Blakely, T. (2018). The impact of social housing on mental health: Longitudinal analyses using marginal structural models and machine learning-generated weights. *International Journal of Epidemiology, 47*(5), 1414–1422. https://doi.org/10.1093/ije/dyy116.

Bichler, A., Neumaier, A., & Hofmann, T. (2014). A tree-based statistical classification algorithm (CHAID) for identifying variables responsible for the occurrence of faecal indicator bacteria during waterworks operations. *Journal of Hydrology, 519*(Part), 909–917. https://doi.org/10.1016/j.jhydrol.2014.08.013.

Burke, T. A., Ammerman, B. A., & Jacobucci, R. (2019). The use of machine learning in the study of suicidal and non-suicidal self-injurious thoughts and behaviors: A systematic review. *Journal of Affective Disorders, 245*(August 2018), 869–884. https://doi.org/10.1016/j.jad.2018.11.073.

Cancino, C., Merigó, J. M., Coronado, F., Dessouky, Y., & Dessouky, M. (2017). Forty years of computers & industrial engineering: A bibliometric analysis. *Computers and Industrial Engineering, 113*, 614–629. https://doi.org/10.1016/j.cie.2017.08.033.

Carroll, L. N., Au, A. P., Detwiler, L. T., Fu, T., Painter, I. S., & Abernethy, N. F. (2014). Visualization and analytics tools for infectious disease epidemiology: A systematic review. *Journal of Biomedical Informatics, 51*, 287–298. https://doi.org/10.1016/j.jbi.2014.04.006.

Chazard, E., Ficheur, G., Bernonville, S., Luyckx, M., & Beuscart, R. (2011). Data mining to generate adverse drug events detection rules. *IEEE Transactions on Information Technology in Biomedicine, 15*(6), 823–830. https://doi.org/10.1109/TITB.2011.2165727.

Chiavegatto Filho, A. D. P. (2015). Uso de big data em saúde no Brasil: Perspectivas para um futuro próximo. *Epidemiologia e Serviços de Saúde, 24*(2), 325–332. https://doi.org/10.5123/s1679-49742015000200015.

Choi, S. B., Lee, W., Yoon, J.-H., Won, J.-U., & Kim, D. W. (2018). Ten-year prediction of suicide death using Cox regression and machine learning in a nationwide retrospective cohort study in South Korea. *Journal of Affective Disorders, 231*, 8–14. https://doi.org/10.1016/j.jad.2018.01.019.

Conway, M., & O'Connor, D. (2016). Social media, big data, and mental health: Current advances and ethical implications. *Current Opinion in Psychology*. https://doi.org/10.1016/j.copsyc.2016.01.004.

Dallora, A. L., Eivazzadeh, S., Mendes, E., Berglund, J., & Anderberg, P. (2016). Prognosis of dementia employing machine learning and microsimulation techniques: A systematic literature review. *Procedia Computer Science, 100*, 480–488. https://doi.org/10.1016/j.procs.2016.09.185.

Deb, R., & Liew, A. W.-C. (2016). Missing value imputation for the analysis of incomplete traffic accident data. *Information Sciences, 339*, 274–289. https://doi.org/10.1016/j.ins.2016.01.018.

Du, J., Xu, J., Song, H.-Y., & Tao, C. (2017). Leveraging machine learning-based approaches to assess human papillomavirus vaccination sentiment trends with Twitter data. *BMC Medical Informatics and Decision Making, 17*. https://doi.org/10.1186/s12911-017-0469-6.

Dunn, A. G., Leask, J., Zhou, X., Mandl, K. D., & Coiera, E. (2015). Associations between exposure to and expression of negative opinions about human papillomavirus vaccines on social media: An observational study. *Journal of Medical Internet Research, 17*(6), e144. https://doi.org/10.2196/jmir.4343.

Dwivedi, A. K., Imtiaz, S. A., & Rodriguez-Villegas, E. (2019). Algorithms for automatic analysis and classification of heart sounds – A systematic review. *IEEE Access, 7*, 8316–8345. https://doi.org/10.1109/ACCESS.2018.2889437.

Egan, K. J., Pinto-Bruno, Á. C., Bighelli, I., Berg-Weger, M., van Straten, A., Albanese, E., & Pot, A.-M. (2018). Online training and support programs designed to improve mental health and reduce burden among caregivers of people with dementia: A systematic review. *Journal of the American Medical Directors Association, 19*(3), 200–206.e1. https://doi.org/10.1016/j.jamda.2017.10.023.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI Magazine, 17*(3), 37. https://doi.org/10.1609/aimag.v17i3.1230.

FUNASA (2017). Retrieved February 26, 2019, from *Cronologia Histórica da Saúde Pública*.

Ghosh, D., & Guha, R. (2013). What are we "tweeting" about obesity? Mapping tweets with topic modeling and Geographic Information System. *Cartography and Geographic Information Science, 40*(2), 90–102. https://doi.org/10.1080/15230406.2013.776210.

Gmeinder, M., Morgan, D., & Mueller, M. (2017). How much do OECD countries spend on prevention? *OECD Health Working Papers, 101*. https://doi.org/10.1787/f19e803c-en.

Griselda, L., Juan, D. O., & Joaquín, A. (2012). Using decision trees to extract decision rules from police reports on road accidents. *Procedia – Social and Behavioral Sciences, 53*, 106–114. https://doi.org/10.1016/j.sbspro.2012.09.864.

Hammann, F., Gutmann, H., Vogt, N., Helma, C., & Drewe, J. (2010). Prediction of adverse drug reactions using decision tree modeling. *Clinical Pharmacology and Therapeutics, 88*(1), 52–59. https://doi.org/10.1038/clpt.2009.248.

Hasan, H., Muhammed, T., Yu, J., Taguchi, K., Samargandi, O. A., Howard, A. F., ... Goddard, K. (2017). Assessing the methodological quality of systematic reviews in radiation oncology: A systematic review. *Cancer Epidemiology, 50*(February), 141–149. https://doi.org/10.1016/j.canep.2017.08.013.

Herrera, R., Berger, U., Von Ehrenstein, O. S., Díaz, I., Huber, S., Muñoz, D. M., & Radon, K. (2018). Estimating the causal impact of proximity to gold and copper mines on respiratory diseases in Chilean children: An application of targeted maximum likelihood estimation. *International Journal of Environmental Research and Public Health, 15*(1), https://doi.org/10.3390/ijerph15010039.

Holt, A. C., Salkeld, D. J., Fritz, C. L., Tucker, J. R., & Gong, P. (2009). Spatial analysis of plague in California: Niche modeling predictions of the current distribution and potential response to climate change. *International Journal of Health Geographics, 8*(1), https://doi.org/10.1186/1476-072X-8-38.

Huang, J., Kornfield, R., Szczypka, G., & Emery, S. L. (2014). A cross-sectional examination of marketing of electronic cigarettes on Twitter. *Tobacco Control, 23*, iii26–iii30. https://doi.org/10.1136/tobaccocontrol-2014-051551.

Huang, Y., Lewis, S., & Britton, J. (2014). Use of varenicline for smoking cessation treatment in UK primary care: An association rule mining analysis. *BMC Public Health, 14*(1), https://doi.org/10.1186/1471-2458-14-1024.

Islam, M. M., Li, Y.-C., Wu, C.-C., Walther, B. A., Nasrin, T., & Yang, H.-C. (2018). Prediction of sepsis patients using machine learning approach: A meta-analysis. *Computer Methods and Programs in Biomedicine, 170*, 1–9. https://doi.org/10.1016/j.cmpb.2018.12.027.

Kadi, I., Idri, A., & Fernandez-Aleman, J. L. (2017). Knowledge discovery in cardiology: A systematic literature review. *International Journal of Medical Informatics, 97*, 12–32. https://doi.org/10.1016/j.ijmedinf.2016.09.005.

Kang, Y., Wang, Y., Zhang, D., & Zhou, L. (2017). The public's opinions on a new school meals policy for childhood obesity prevention in the U.S.: A social media analytics approach. *International Journal of Medical Informatics, 103*, 83–88. https://doi.org/10.1016/j.ijmedinf.2017.04.013.

Kavakiotis, I., Tsave, O., Salifoglou, A., Maglaveras, N., Vlahavas, I., & Chouvarda, I. (2017). Machine learning and data mining methods in diabetes research. *Computational and Structural Biotechnology Journal, 15*, 104–116.

Kim, P. W. (2018). Operating an environmentally sustainable city using fine dust level big data measured at individual elementary schools. *Sustainable Cities and Society, 37*, 1–6. https://doi.org/10.1016/j.scs.2017.10.019.

Kitchenham, B. (2004). Procedures for performing systematic reviews. *Procedures for Performing Systematic Reviews*.

Kitchenham, B., & Charters, S. (2007). Guidelines for performing systematic literature reviews in software engineering. *Guidelines for Performing Systematic Literature Reviews in Software Engineering*.

Kumar, S., & Toshniwal, D. (2016). Analysis of hourly road accident counts using hierarchical clustering and cophenetic correlation coefficient (CPCC). *Journal of Big Data, 3*(1), https://doi.org/10.1186/s40537-016-0046-3.

Kwon, J.-M., Lee, Y., Lee, Y., Lee, S., & Park, J. (2018). An algorithm based on deep learning for predicting in-hospital cardiac arrest. *Journal of the American Heart Association, 7*(13), https://doi.org/10.1161/JAHA.118.008678.

Lang, T., & Rayner, G. (2015). Beyond the Golden Era of public health: Charting a path from sanitarianism to ecological public health. *Public Health, 129*(10), 1369–1382. https://doi.org/10.1016/j.puhe.2015.07.042.

Lara, J. A., Lizcano, D., Martínez, M. A., & Pazos, J. (2014). Data preparation for KDD through automatic reasoning based on description logic. *Information Systems, 44*, 54–72. https://doi.org/10.1016/j.is.2014.03.002.

Lazarou, C., Karaolis, M., Matalas, A.-L., & Panagiotakos, D. B. (2012). Dietary patterns analysis using data mining method. An application to data from the CYKIDS study. *Computer Methods and Programs in Biomedicine, 108*(2), 706–714. https://doi.org/10.1016/j.cmpb.2011.12.011.

Lee, Y., Ragguett, R. M., Mansur, R. B., Boutilier, J. J., Rosenblat, J. D., Trevizol, A., ... McIntyre, R. S. (2018). Applications of machine learning algorithms to predict therapeutic outcomes in depression: A meta-analysis and systematic review. *Journal of Affective Disorders, 241*(August), 519–532. https://doi.org/10.1016/j.jad.2018.08.073.

Leidman, E., Humphreys, A., Cramer, C. G., Mil, L. T.-V., Wilkinson, C., Narayan, A., & Bilukha, O. (2018). Acute Malnutrition and Anemia Among Rohingya Children in Kutupalong Camp, Bangladesh. *JAMA, 319*(14), 1505–1506. https://doi.org/10.1001/jama.2018.2405.

Librenza-Garcia, D., Kotzian, B. J., Yang, J., Mwangi, B., Cao, B., Pereira Lima, L. N., ... Passos, I. C. (2017). The impact of machine learning techniques in the study of bipolar disorder: A systematic review. *Neuroscience and Biobehavioral Reviews, 80*, 538–554. https://doi.org/10.1016/j.neubiorev.2017.07.004.

Liu, K., Li, L., Jiang, T., Chen, B., Jiang, Z., Wang, Z., ... Gu, H. (2016). Chinese public attention to the outbreak of ebola in West Africa: Evidence from the online big data platform. *International Journal of Environmental Research and Public Health, 13*(8), https://doi.org/10.3390/ijerph13080780.

Marinelarena-Dondena, L., Ferretti, E., Maragoudakis, M., Sapino, M., & Luis Errecalde, M. (2017). Predicting Depression: A comparative study of machine learning approaches based on language usage. *Cuadernos de Neuropsicologia-Panamerican Journal of Neuropsychology, 11*(3), 42–54. https://doi.org/10.7714/CNPS/11.3.201.

Marucci-Wellman, H. R., Corns, H. L., & Lehto, M. R. (2017). Classifying injury narratives of large administrative databases for surveillance—A practical approach combining machine learning ensembles and human review. *Accident Analysis and Prevention, 98*, 359–371. https://doi.org/10.1016/j.aap.2016.10.014.

Massey, P. M., Leader, A., Yom-Tov, E., Budenz, A., Fisher, K., & Klassen, A. C. (2016). Applying multiple data collection tools to quantify human papillomavirus vaccine communication on Twitter. *Journal of Medical Internet Research, 18*(12), https://doi.org/10.2196/jmir.6670.

McConnell, M. V., Shcherbina, A., Pavlovic, A., Homburger, J. R., Goldfeder, R. L., Waggot, D., ... Ashley, E. A. (2017). Feasibility of obtaining measures of lifestyle from a smartphone app the MyHeart counts cardiovascular health study. *JAMA Cardiology, 2*(1), 67–76. https://doi.org/10.1001/jamacardio.2016.4395.

Miao, K. H., & Miao, J. H. (2018). Coronary heart disease diagnosis using deep neural networks. *International Journal of Advanced Computer Science and Applications, 9*(10), 1–8. https://doi.org/10.14569/IJACSA.2018.091001.

Muennig, P. A. (2015). How automation can help alleviate the budget crunch in public health research. *American Journal of Public Health, 105*(9), e19–e22. https://doi.org/10.2105/AJPH.2015.302782.

Myslín, M., Zhu, S.-H., Chapman, W., & Conway, M. (2013). Using twitter to examine smoking behavior and perceptions of emerging tobacco products. *Journal of Medical Internet Research, 15*(8), https://doi.org/10.2196/jmir.2534.

Ngai, E. W. T., Xiu, L., & Chau, D. C. K. (2009). Application of data mining techniques in customer relationship management: A literature review and classification. *Expert Systems with Applications, 36*(2 PART 2), 2592–2602. https://doi.org/10.1016/j.eswa.2008.02.021.

Nikfarjam, A., Sarker, A., O'Connor, k., Ginn, R., & Gonzalez, G. (2015). Pharmacovigilance from social media: Mining adverse drug reaction mentions using sequence labeling with word embedding cluster features. *Journal of the American Medical Informatics Association, 22*(3), 671–681. https://doi.org/10.1093/jamia/ocu041.

Nindrea, R. D., Aryandono, T., Lazuardi, L., & Dwiprahasto, I. (2018). Diagnostic accuracy of different machine learning algorithms for breast cancer risk calculation: A meta-analysis. *Asian Pacific Journal of Cancer Prevention: APJCP, 19*, 1747–1752. https://doi.org/10.22034/APJCP.2018.19.7.1747.

O'Shea, J. (2017). Digital disease detection: A systematic review of event-based internet biosurveillance systems. *International Journal of Medical Informatics, 101*, 15–22. https://doi.org/10.1016/j.ijmedinf.2017.01.019.

Odlum, M., & Yoon, S. (2015). What can we learn about the Ebola outbreak from tweets? *American Journal of Infection Control, 43*(6), 563–571. https://doi.org/10.1016/j.ajic.2015.02.023.

OECD (2017). *Health at a glance 2017: OECD indicators.* Paris: OECD Publishing.

Ortega Hinojosa, A. M., Davies, M. M., Jarjour, S., Burnett, R. T., Mann, J. K., Hughes, E., ... Jerrett, M. (2014). Developing small-area predictions for smoking and obesity prevalence in the United States for use in Environmental Public Health Tracking. *Environmental Research, 134*, 435–452. https://doi.org/10.1016/j.envres.2014.07.029.

Pakgohar, A., Tabrizi, R. S., Khalili, M., & Esmaeili, A. (2011). The role of human factor in incidence and severity of road crashes based on the CART and LR regression: A data mining approach. *Procedia Computer Science, 3*, 764–769. https://doi.org/10.1016/j.procs.2010.12.126.

Pan American Health Organization - PAHO. (2017). Zika - Epidemiological report Brazil. Washington, D.C.

Partington, S. N., Papakroni, V., & Menzies, T. (2014). Optimizing data collection for public health decisions: A data mining approach. *BMC Public Health, 14*(1), https://doi.org/10.1186/1471-2458-14-593.

Paul, M. J., & Dredze, M. (2014). Discovering health topics in social media using topic models. *PLoS One, 9*(8), https://doi.org/10.1371/journal.pone.0103408.

Petticrew, M., & Roberts, H. (2008). Systematic reviews in the social sciences: A practical guide. *Systematic Reviews in the Social Sciences: A Practical Guide*. https://doi.org/10.1002/9780470754887.

Qu, H.-Q., Li, Q., Rentfro, A. R., Fisher-Hoch, S. P., & McCormick, J. B. (2011). The definition of insulin resistance using HOMA-IR for americans of mexican descent using machine learning. *PLoS One, 6*(6), https://doi.org/10.1371/journal.pone.0021041.

Rao, R. R., & Makkithaya, K. (2017). Learning from a class imbalanced public health dataset: A cost-based comparison of classifier performance. *International Journal of Electrical and Computer Engineering, 7*(4), 2215–2222. https://doi.org/10.11591/ijece.v7i4.pp2215-2222.

Rechel, B. (2019). Funding for public health in Europe in decline? *Health Policy, 123*(1), 21–26. https://doi.org/10.1016/j.healthpol.2018.11.014.

Rose, S. (2013). Mortality risk score prediction in an elderly population using machine learning. *American Journal of Epidemiology, 177*(5), 443–452. https://doi.org/10.1093/aje/kws241.

Rybarczyk, Y., & Zalakeviciute, R. (2018). Machine learning approaches for outdoor air quality modelling: A systematic review. *Applied Sciences, 8*(12), 2570. https://doi.org/10.3390/app8122570.

Salerno, J., Knoppers, B. M., Lee, L. M., Hlaing, W. M., & Goodman, K. W. (2017). Ethics, big data and computing in epidemiology and public health. *Annals of Epidemiology, 27*(5), 297–301. https://doi.org/10.1016/j.annepidem.2017.05.002.

Song, S.-H., & Kim, D. K. (2017). Development of a stress classification model using deep belief networks for stress monitoring. *Healthcare Informatics Research, 23*(4), 285–292. https://doi.org/10.4258/hir.2017.23.4.285.

Subramani, S., Wang, H., Vu, H. Q., & Li, G. (2018). Domestic violence crisis identification from facebook posts based on deep learning. *IEEE Access, 6*, 54075–54085. https://doi.org/10.1109/ACCESS.2018.2871446.

Sun, X., Ren, F., & Ye, J. (2017). Trends detection of flu based on ensemble models with emotional factors from social networks. *IEEJ Transactions on Electrical and Electronic Engineering, 12*(3), 388–396. https://doi.org/10.1002/tee.22389.

Tatonetti, N. P., Denny, J. C., Murphy, S. N., Fernald, G. H., Krishnan, G., Castro, V., ... Altman, R. B. (2011). Detecting drug interactions from adverse-event reports: Interaction between paroxetine and pravastatin increases blood glucose levels. *Clinical Pharmacology and Therapeutics, 90*(1), 133–142. https://doi.org/10.1038/clpt.2011.83.

The Lancet Public Health (2017). Prospects for public health in a sustainable NHS. *The Lancet Public Health, 2*(5), e202.

Torrecilla, J. L., & Romo, J. (2018). Data learning from big data. *Statistics and Probability Letters, 136*, 15–19. https://doi.org/10.1016/j.spl.2018.02.038.

Tranfield, D., Denyer, D., & Smart, P. (2003). Towards a methodology for developing evidence-informed management knowledge by means of systematic review. *British of Journal Management, 14*, 207–222.

William, W., Ware, A., Basaza-Ejiri, A. H., & Obungoloch, J. (2018). A review of image analysis and machine learning techniques for automated cervical cancer screening from pap-smear images. *Computer Methods and Programs in Biomedicine, 164*, 15–22.

https://doi.org/10.1016/j.cmpb.2018.05.034.

World Health Organization - WHO. (2014). Mental health: A state of well-being. Retrieved February 22, 2019, from https://www.who.int/features/factfiles/mental_health/en/.

World Health Organization - WHO. (2018a). Ebola virus disease – Fact sheet. Retrieved August 20, 2019, from http://www.who.int/news-room/fact-sheets/detail/ebola-virus-disease.

World Health Organization - WHO. (2018b). Global strategy on diet, physical activity and health: Childhood overweight and obesity. Retrieved February 22, 2019, from https://www.who.int/dietphysicalactivity/childhood/en/.

World Health Organization - WHO. (2018c). Measles cases spike globally due to gaps in vaccination coverage. Retrieved March 12, 2019, from https://www.who.int/news-room/detail/29-11-2018-measles-cases-spike-globally-due-to-gaps-in-vaccination-coverage.

Xu, Q., Gel, Y. R., Ramirez, L. L. R., Nezafati, K., Zhang, Q., & Tsui, K.-L. (2017). Forecasting influenza in Hong Kong with Google search queries and statistical model fusion. *PLoS One, 12*(5), https://doi.org/10.1371/journal.pone.0176690.

Yassin, N. I. R., Omran, S., El Houby, E. M. F., & Allam, H. (2018). Machine learning techniques for breast cancer computer aided diagnosis using different image modalities: A systematic review. *Computer Methods and Programs in Biomedicine, 156*, 25–45. https://doi.org/10.1016/j.cmpb.2017.12.012.

Zhan, Y., Luo, Y., Deng, X., Chen, H., Grieneisen, M. L., Shen, X., ... Zhang, M. (2017). Spatiotemporal prediction of continuous daily PM2.5 concentrations across China using a spatially explicit machine learning algorithm. *Atmospheric Environment, 155*, 129–139. https://doi.org/10.1016/j.atmosenv.2017.02.023.

Zhang, Y., Dang, Y., Chen, H., Thurmond, M., & Larson, C. (2009). Automatic online news monitoring and classification for syndromic surveillance. *Decision Support Systems, 47*(4), 508–517. https://doi.org/10.1016/j.dss.2009.04.016.

Zhang, S., Tjortjis, C., Zeng, X., Qiao, H., Buchan, I., & Keane, J. (2009). Comparing data mining methods with logistic regression in childhood obesity prediction. *Information Systems Frontiers, 11*(4), 449–460. https://doi.org/10.1007/s10796-009-9157-0.

Zhu, W., Wang, J., Zhang, W., & Sun, D. (2012). Short-term effects of air pollution on lower respiratory diseases and forecasting by the group method of data handling. *Atmospheric Environment, 51*, 29–38. https://doi.org/10.1016/j.atmosenv.2012.01.051.