Journal of Ambient Intelligence and Humanized Computing (2021) 12:9163–9180
https://doi.org/10.1007/s12652-020-02617-4

ORIGINAL RESEARCH

Big data analytics as a tool for fighting pandemics: a systematic review of literature

Alana Corsi[1] · **Fabiane Florencio de Souza**[1] · **Regina Negri Pagani**[1] · **João Luiz Kovaleski**[1]

Received: 10 June 2020 / Accepted: 10 October 2020 / Published online: 29 October 2020

Abstract
Infectious and contagious diseases represent a major challenge for health systems worldwide, either in private or public sectors. More recently, with the increase in cases related to these problems, combined with the recent global pandemic of COVID-19, the need to study strategies to treat these health disturbs is even more latent. Big Data, as well as Big Data Analytics techniques, have been addressed in this context with the possibility of predicting, mapping, tracking, monitoring, and raising awareness about these epidemics and pandemics. Thus, the purpose of this study is to identify how BDA can help in cases of pandemics and epidemics. To achieve this purpose, a systematic review of literature was carried out using the methodology Methodi Ordinatio. The rigorous search resulted in a portfolio of 45 articles, retrived from scientific databases. For the collection and analysis of data, the softwares NVivo 12 and VOSviewer were used. The content analysis sought to identify how Big Data and Big Data Analytics can help fighting epidemics and pandemics. The types and sources of data used in cases of previous epidemics and pandemics were identified, as well as techniques for treating these data. The results showed that the main sources of data come from social media and Internet search engines. The most common techniques for analyzing these data involve the use of statistics, such as correlation and regression, combined with other techniques. Results shows that there is a fruitful field of study to be explored by both areas, Big Data and Health.

Keywords Big data · Big data analytics · Pandemics · Epidemics · Systematic review of literature · COVID-19

1 Introduction

The term Big Data has become widely used to describe the exponential growth of data in recent years (Yu et al. 2017; Wasim et al. 2019). Due to its popularity, its meaning is quite diverse. It can be understood as a huge volume of data, with high speed, diversity of topics, and variety, coming from different sources, such as cell phones, social networks, sensors, and other devices, which need specific techniques and tools for their processing (Manogaran and Lopez 2018).

In this context, on a broader approach, Big Data Science is the study that encompasses different aspects of Big Data, including data storage capacity, and analysis speed, for instance. It can be employed on the study of different areas, such as, health data, genomes, environment, social media, and commercial activities (Rana and Mugavero 2019). All of these studies have in common the need for data analysis, and commom, or simpler analyzes do not have the same configurations or capacity that Big Data Analytcs can provide to users. Thus, Big Data Analytics (BDA) has become a modern practice for analyzing data, and identified as a specific analysis for this new scenario (Saggi and Jain 2018; Eken 2020).

BDA consists of a set of advanced analytical techniques, borrowed from related fields, such as statistics, data mining and business analysis, making it possible to discover knowledge from big volumes of data (Saggi and Jain 2018; Ajayi et al. 2020), thus enabling the extraction of valuable

✉ Alana Corsi
alanacorsi@alunos.utfpr.edu.br

Fabiane Florencio de Souza
fabianesouza@alunos.utfpr.edu.br

Regina Negri Pagani
reginapagani@utfpr.edu.br

João Luiz Kovaleski
kovaleski@utfpr.edu.br

[1] Federal University of Technology-Paraná (UTFPR)
Câmpus Ponta Grossa, Av. Monteiro Lobato, s/n-Km 04,
Ponta Grossa, PR 84016-210, Brazil

information generated by different devices (Babar and Arif 2019) connected with the Internet.

Among the different areas that benefit from BDA, health is one that can be further and proficuosly explored. The reason is that the correlation of a large amount of data, for instance, can make possible and more reliable to control, monitor and study patients, diseases and epidemiology (Huang et al. 2015; Yang et al. 2015; Kraemer et al. 2018). Specifically, in the field of epidemiology, which studies contamination, determinants, and control of health problems, such as outbreaks, endemics, epidemics and pandemics (Last 2001), BDA can be used as a useful tool in studies and analysis on the data generated during the the periodsof their occurrence.

Taking into account that in times of pandemics, data are generated in large quantities from different sources, and presenting several characteristics that can be difficult to correlate without a proper analysis, the objective of this study is to identify how BDA can help in cases of pandemics and epidemics. With this study it is expected to answer the following Research Questions:

RQ1 *What types of data and data sources are used to assist in the treatment of Epidemics/Pandemics?*

RQ2 *How Big Data Analytics/Analysis can assist in cases of Epidemics/Pandemics?*

Therefore, in order to achieve the objective of this study and to answer the research questions, a systematic review of literature was carried out using the methodology Methodi Ordinatio. The data collect from the robust portfolio of articles built is the basis of the discussions proposed.

This paper is composed of 5 sections, including this introductory one. Section 2 approaches the concepts on Big Data and Big Data Analytics (Sect. 2.1); Pandemics and Epidemics (Sect. 2.2); and the use of Big Data and Big Data Analytics in the context of pandemics and epidemics (Sect. 2.3), bringing contemporary theoretical contributions. Secion 3 describes the Research portfolio construction (Sect. 3.1), and data collection and analysis procedures (Sect. 3.2). Section 4 presents the results and disscussions sheding light on the literature and research findings, presenting the bibliometric (Sect. 4.1) and content analyzes, answering the RQs (Sect. 4.2). Finally, Sect. 5 concludes this study, listing the contributions and limitation.

2 Theoretical background

2.1 Big data and big data analytics

Big Data Science attempts to study the Big Data universe, focusing on the speed, storage and analysis of this data (Rana and Mugavero 2019). Therefore, it is interesting to know that the concept of Big Data goes far beyond a large volume of data, "massive data" and "whacking data". It has other characteristics, known as the "Vs" of Big Data (Hu et al. 2014). Ali and Abdullah (2019) point 6Vs of Big Data, which are: value, volume, velocity, variety, veracity, and variability. Each of these characteristics is shown in Fig. 1.

Thus, a large amount of data from different sources is generated every day, at great speed, becoming a challenge for academics and professionals to transform this data into valuable information (Sivarajah et al. 2017; Kauffmann et al. 2019).

In this context, BDA is increasingly being adopted by many different organizations, in an effort to turn large volumes of raw data into valuable information (Sivarajah et al. 2017; Kauffmann et al. 2019). In order to accomplish this task, BDA must effectively mine massive data sets as close to real-time as possible, so that correlations and associations among different variables can be found and insights can be revealed(Hu et al. 2014; Meera and Sundar 2020).

Large companies in the technology area, such as Amazon, Facebook, Google, Netflix, and Uber, whose data to be managed and saved are gigantic, have adopted the integration of BDA in their daily operational activities, which resulted in a greater capacity for analyzing their data (Wamba et al. 2020).

Despite the technology sector being one of those who makes the most of BDA, it is important to point out that this strategic tool is also applicable to other areas, such as transports, education and health. The latter is a prosperous field for analysis, as it has a large amount of data that can be analyzed with the BDA (Hu et al. 2018), providing great benefits not only to health organizations, but to final consumers, that is, pacients and health citizens as well.

The next section approaches the scenario of epidemics and pandemics and, in the sequence, how these phenomena are closely related to the Big Data context.

2.2 Epidemics and pandemics

According to the Centers for Disease Control and Prevention (CDC 2012), Epidemiology is a scientific discipline with solid methods of investigation, which is guided by data in a systematic and impartial approach to collection, analysis, and interpretation. In summary, this area of study makes use of valid groups of comparison to assess whether what has been observed—such as the number of disease cases in a specific area, during a specific period—differs from what might be expected. It is also based on methods from other scientific fields, including biostatistics and computer science, with biological, economic, social, and behavioral sciences.

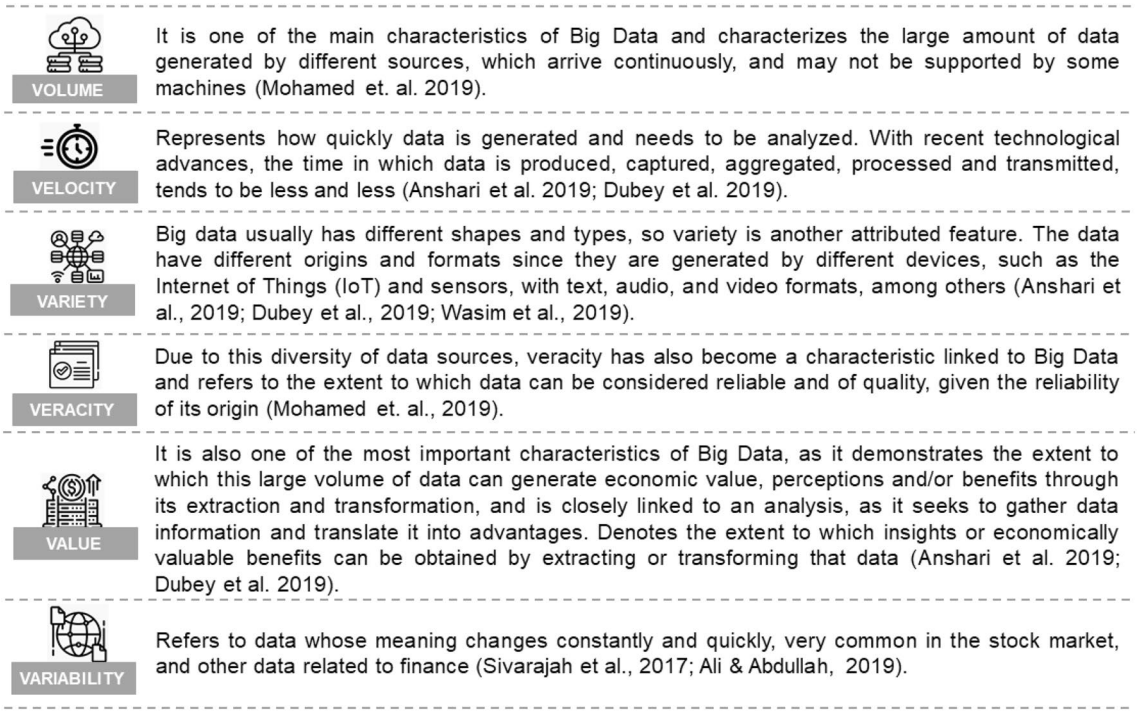


Fig. 1 Six V's of Big Data applied to health data

According to the CDC (2012), some terms used in epidemiological investigations are:

Outbreaks The occurrence of more cases of a disease than would normally be expected, in a specific location or group of people, during a given period. According to Association for Professionals in Infection Control and Epidemiology (APIC 2019), some outbreaks are expected per year, such as influenza, for instance.
Epidemic It has a similar meaning as in the outbreak, but connotes a more serious occurrence, such as the Severe Acute Respiratory Syndrome (SARS) epidemic, that occurred between 2003 and 2004 (WHO 2003). An outbreak may suggest something geographically limited, while an epidemic predicts a situation that could spread a little wider.
Pandemic It is a much broader widespread epidemic, often global, that affects an expressively large number of people. A pandemic does not necessarily represents a more serious disease than an epidemic, but a greater number of affected beings. The most recent global pandemics are: Swine Flu Pandemic (H1N1 Influenza), which started in 2009, and the New Coronavirus Pandemic (COVID-19), which started in 2019, in China (APIC 2019).

These classifications are used worldwide to determine the scale of infectious diseases.

According to reports from the World Health Organization (WHO), throughout history, several outbreaks, epidemics and also pandemics, have occurred in the world. Figure 2 shows the most recent ones.

As mentioned previously, the world is currently experiencing the pandemic of the recently discovered new coronavirus, called COVID-19 or SARS-CoV-2. The first report of COVID-19 occurred in December 2019, in China, as pneumonia of unknown cause. It was then declared as a pandemic in January 2020, by WHO. Since its emergence, 24,822,800 cases have been confirmed worldwide and 838,360 confirmed deaths, according to the August 30, 2020 report proposed by WHO (2019). Global actions and strategies are being developed and encouraged by public and private agencies to slow down and combat the COVID-19 pandemic. Among these initiatives we can mention platforms for real-time monitoring, such as Johns Hopkins Platform and Google Trends. Other strategies are the monitoring from official organizations, such as WHO and CDC, which justifies the development of this present study.

2.3 Big Data and Big Data Analytics in the epidemic and pandemic scenario

As in other fields, the volume of data produced every day is also increasing exponentially in the area of modern health (Ergüzen and Ünver 2018). Allied to this scenario, and with the development of smart devices and Cloud Computing, a

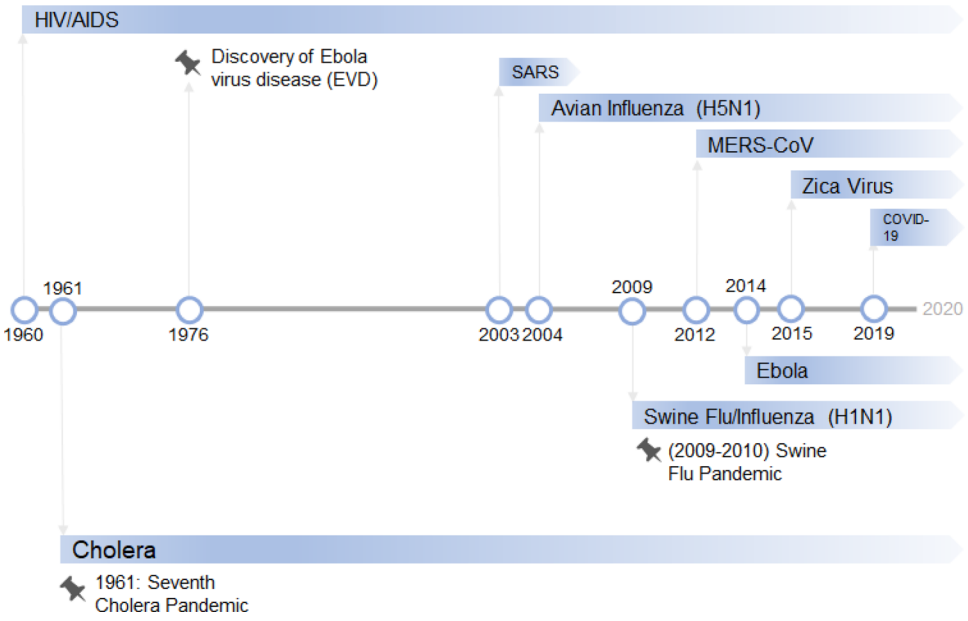


Fig. 2 Outbreaks, Epidemics and Pandemics Timeline. Source: Adapted from WHO (2019), CDC (2012), and Bloom and Cadarette (2019)

huge amount data coming from public health occurences can be collected from various sources and analyzed in different ways. The great social and academic impact of such developments has caused a great worldwide repercussion on Big Data applied to health (Huang et al. 2015).

The use of BDA in the health sector has shown promising results in its attempts to manage data that are expanding in the area (Štufi et al. 2020). An example of this application was the study carried out by the Massachusetts Institute of Technology (MIT) on BDA in intensive care units (Davoudi et al. 2015). The study reported that data analysis could positively predict critical information, such as length of hospital stay, number of patients requiring surgical intervention, and which patients could be at risk for sepsis or iatrogenic diseases. For all these patients, data analysis can save lives or prevent other complications that might happen (Štufi et al. 2020).

In another study (Chung et al. 2018), the authors propose a modeling of health risk assessment using a deep neural network, as a way of BDA, based on medical big data, concluding that the model can have a quite positive impact on human quality of life.

These detection methods, which can anticipate knowledge about an epidemic or pandemic, can be achieved by monitoring internet searches for health-related terms (Beam and Kohane 2018). The use of these detection methods allowed Google to develop an algorithm capable of estimating flu activity in different regions of the United States. This approach allowed the company to use the most searched terms in its base to detect the emergence of a possible epidemic, which has proven effective in areas where the population has regular access to the internet (Štufi et al. 2020).

In addition to internet research, the increasing availability of Big Data sources for diagnostic and pre-diagnostic in public and private health services allows the advancement of new generation methods for detecting and predicting pandemics (Spreco et al. 2017; Agbehadji et al. 2020). Allied to public and private health data, the billions users of cell phones, social media platforms, and other technologies generate an increasing volume of data that can be potentially used to solve challenges in the health field (Kraemer et al. 2018).

In face of the recent scenario of pandemics, it is clear that Big Data science can greatly help in finding solutions and strategies to help diminish the problems.

In the sequence, the materials and methods used to build this study are described.

3 Materials and methods

The methodological procedures were divided into two parts: (3.1) construction of the research portfolio, which will be the source of data collection and analysis; and (3.2) data collection and analysis procedures, as shown in Fig. 3, and described in the sequence.

3.1 Construction of the research portfolio

In order to identify how Big Data and Big Data analysis can assist in the treatment of Epidemics and Pandemics, more specifically contextualizing COVID-19, this study carried out a systematic review of literature, using the multi-criteria methodology Methodi Ordinatio (Pagani et al. 2015, 2017; Campos et al. 2018). The nine steps of Methodi Ordinatio were followed, as described:

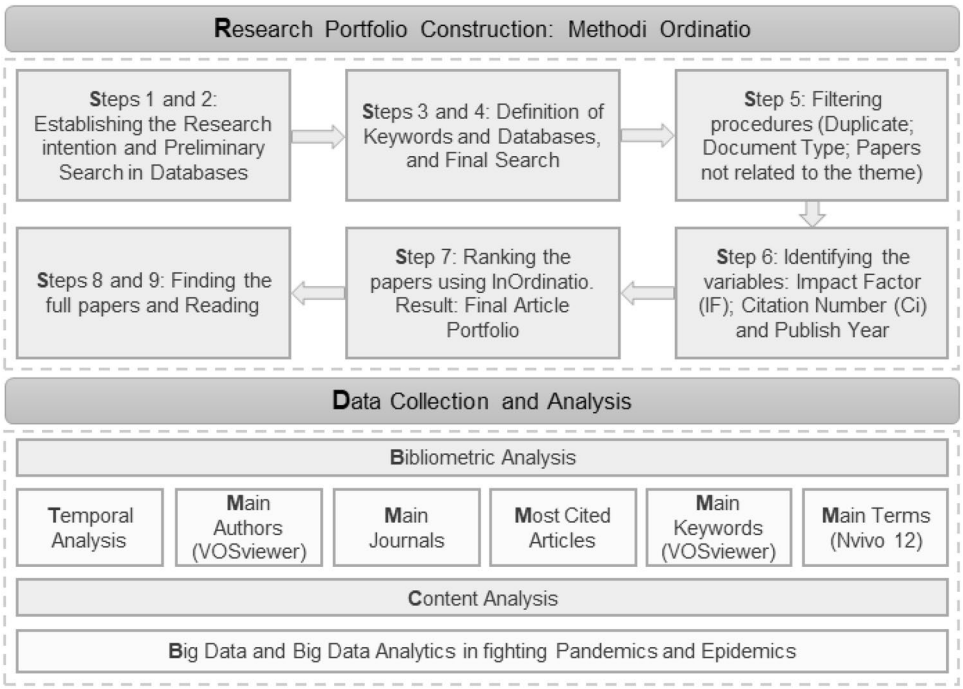


Fig. 3 Methodological procedures

Steps 1 and 2 Establishing the intention of the research and exploratory searches in the databases: since the objective of this work is to analyze the role of Big Data and Big Data Analytics in the treatment of pandemics and epidemics, the keywords used were: “Big Data”; “Big Data Analytic *”; "Epidemic *"; "Pandemic *"; "coronavirus" and "COVID-19". From this preliminary search, Web of Science and Scopus were selected for providing a large number of articles.

Steps 3 and 4 Definition of keywords combination and databases, and final search: the preliminary tests confirmed the keywords combinations, and the final searches were carried out. The search sintaxe and the results of the final search are shown in Table 1.

Step 5 Filtering procedure: the filtering procedures were performed in order to eliminate duplicate articles; conference articles, books, book chapters; and articles with themes outside the scope of this research, eliminated by reading the title, abstract, and keywords. The results obtained in the filtering procedures are shown in Table 2.

Step 6 Identifying the Impact Factor (IF), Year of publication and Number of Citations (Ci): The metrics of the papers were collected from CAPES and Scopus. The number of the articles’ citations was obtained from Google Scholar.

Step 7 After collecting the variables, the InOrdinatio Eq. (1) was applied, resulting in a portfolio of ordered scientific articles.

$$\text{InOrdinatio} = (\text{IF}/1000) + \alpha * [10 - (\text{ResearchYear} - \text{PublishYear})] + (\text{Ci}) \quad (1)$$

The elements of the equation are: IF (impact factor); α (alfa value, ranging from 1 to 10, to be defined by the

Table 1 Search sintaxe and results

Keywords combinations	Databases configurations: No time restriction; Search in: Title, abstract and keyword; Document type: Article and Review; Use of Boolean operator	
	Web of Science	Scopus
("Big Data" OR "Big Data Analytic*") AND ("Pandemic*" OR "Epidemic*")	115	121
("Big Data" OR "Big Data Analytic*") AND ("Coronavirus" OR "COVID-19″)	2	3
Total by Database	117	124

Table 2 Filtering procedures

Filtering Procedures	Number of articles excluded
Initial number of articles	241
Duplicate papers deleted	93
Deletion by document type	4
Deletion of articles outside the theme	99
Total articles deleted	196
The resulting number of articles in the portfolio	45

researcher according to the importance of the newness of the theme. For this study, the value of α was defined as 10 since the object of this study is published in very recent papers; ResearchYear (year in which the research was developed); PublishYear (year in which the paper was published); and Ci (number of times the paper has been cited).

Thus, the final portfolio, ordered by scientific relevance, was composed of 45 articles, as shown in Table 3 ("Appendix").

Steps 8 and 9 Finding full papers, articles systematic reading and analysis: with the final portfolio organized, the full papers were collected and archived so that systematic reading and analyzes could be carried out.

The next subsections details the procedures for collecting and analyzing data.

3.2 Data collection and analysis procedures

With the portfolio of articles organized and full texts retrieved, both bibliometric and content analysis were thoroughly carried out using the softwares VOSviewer and Nvivo 12. The aspects investigated in the bibliometric section were: number of publications per year, main authors of the portfolio; main sources of publications; most cited articles; main keywords of the articles; and main terms mentioned in the whole portfolio. In-depth content analysis sought to identificate correlations between Big Data and pandemics or epidemics, attempting to identify how Big Data Analytics can help managing pandemic situations, having as background the scenario of the current pandemic COVID-19. From this analysis, research questions 1 and 2 were answered.

4 Results and discussions

This section provides with the results of the analysis. The first subsection (4.1) presents the bibliometrics analysis, and the second one (4.2) presents the results and research correlations build from the in-depth content analysis.

4.1 Bibliometric analysis

The temporal analysis of the publications can be observed in Fig. 4. 2018 was the year with the largest number of publications, with about 27% of the articles in the portfolio. Curiously, number of publications in 2019 was not that expressive.

Nevertheless, it is possible to observe that the topic is of scientific increasingly interest since the last three years are responsible for approximately 50% of the articles in the portfolio. Since the year 2020 is marked by the global pandemic of COVID-19, there is a demand for strategies to control it, justifying the present research.

To find the authors of the portfolio, the density map functionality of the VOSviewer software was used, and results are shown in Fig. 5.

We can observe that the main author of the portfolio is Bragazzi, N. L., with three articles in the portfolio, one as the main author and the others as a co-author, adding 69 citations in their publications regarding Big Data and Pandemic/Epidemic. The second author in the portfolio is Martini, M., co-author of two articles written with Bragazzi, N. L. Finally, the third author in the portfolio is Gianfredi, V., with two articles, both as the main author and both written with Bragazzi, N. L.

The journals that have published a larger number of articles with the keywords combinations for this portfolio are shown in Fig. 6.

The journal with the largest number of articles in the portfolio is the Journal of Infectious Diseases, with four articles, followed by the Scientific Reports journal, with three articles. Both journals with Impact Factor above 4000. Of the remaining journals, 16% present two articles each, and more than 50% present only one article each. From the analysis, it is noticed that most of the articles are published in journals with main focus in the areas of health, biology and medicine, with interest in studies that address pathogenesis, diagnosis, and treatment of infectious diseases. It is also observed some journals focused on Big Data, Big Data Analytics and computational methods: the Journal of Big Data, Big Data Research, Computer Methods and Programs in Biomedicine, and Studies in Computational Intelligence. Nevertheless, the number of journals encompassing both main themes, health area and Big Data, it is still in a reduced number.

We conclude this point addressing the importance to produce more studies encompassing these themes, considering that the Big Data Science has a lot to contribute with the health and life ciences.

Figure 7 presents the most cited articles in the porfolio.

The article, "A review of data mining using big data in health informatics", published in 2014 in the Journal of Big Data, is the most cited article in the portfolio. Although not the newest, it is the article with the highest InOrdinatio, revealing its importante to the connection between the two themes. The second article with the

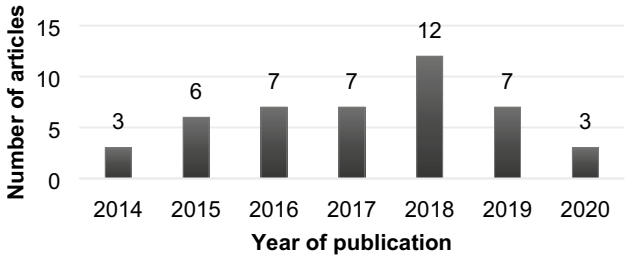


Fig. 4 Temporal analysis

Fig. 5 Main authors

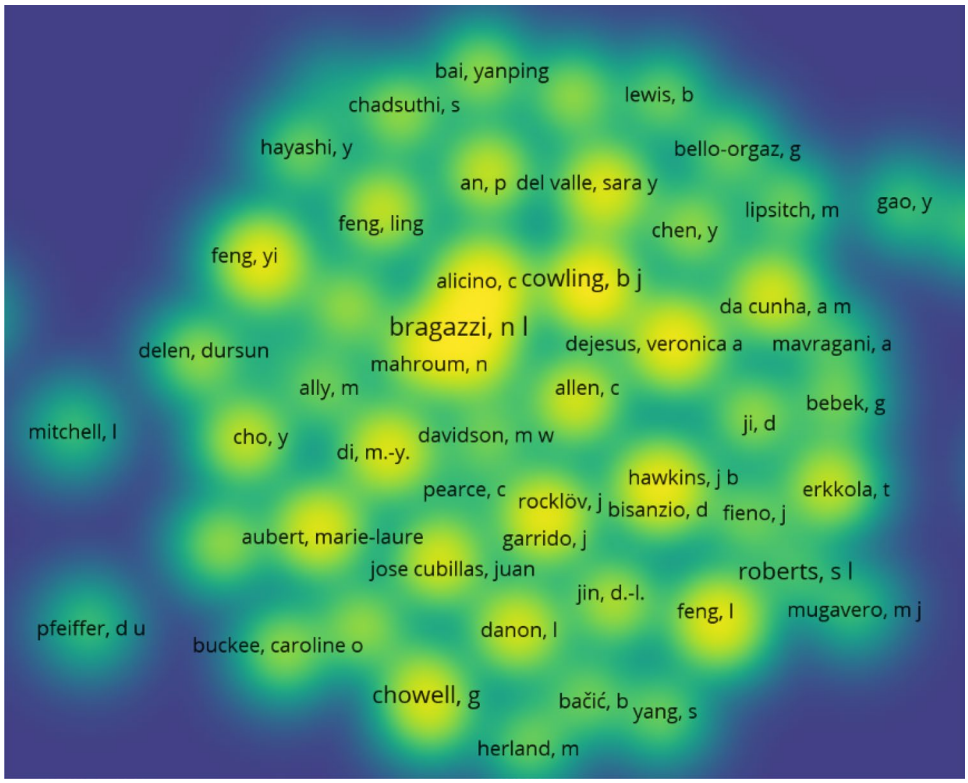
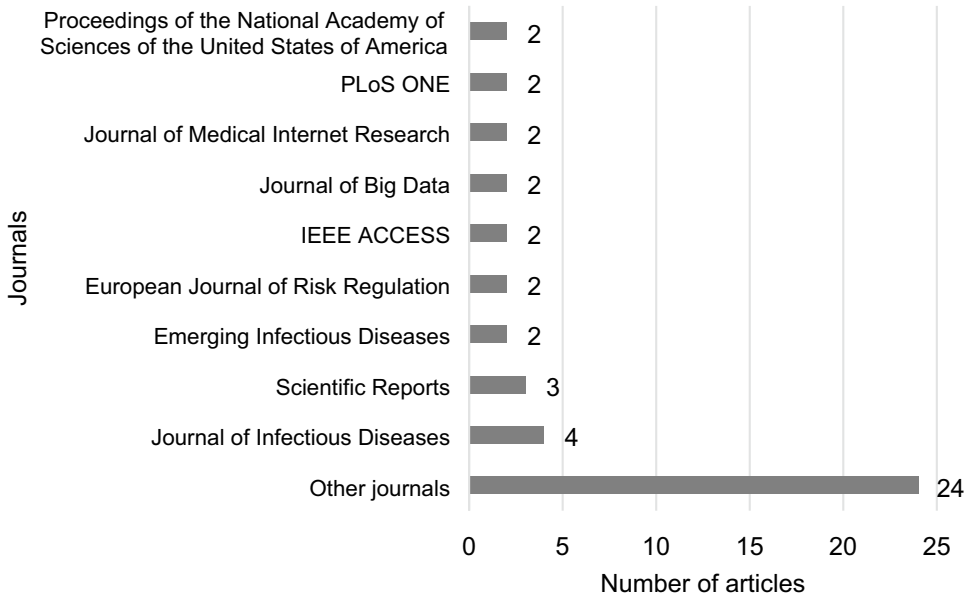


Fig. 6 Main journals



highest number of citations, and second in the InOrdinatio ranking, "Accurate estimation of influenza epidemics using Google search data via ARGO", was published in 2015 in the Proceedings of the National Academy of Sciences of the United States of America, journal with the second largest impact factor in the portfolio, 9580. As noted, the first five most cited articles are also the top five in the InOrdinatio ranking. Although not the most recent articles in the portfolio, they were published in high-impact journals and achieved wide dissemination, seen through the high number of citations. All of them show the connections between the two main streams of this research.

Furthermore, the first three articles with the greatest impact in the portfolio, according to the InOrdinatio ranking, present the main focus on computational methods, use and treatment of data and Big Data, to help and combat

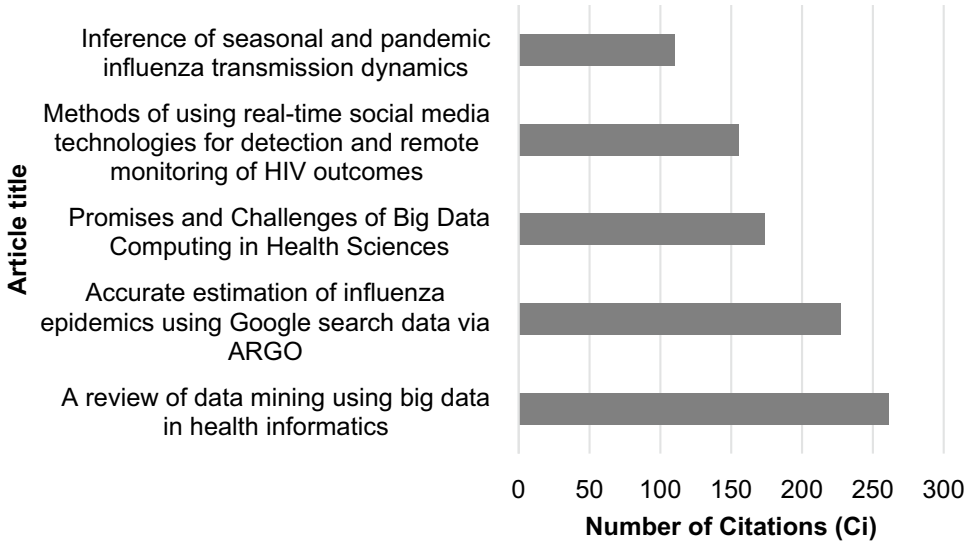


Fig. 7 Most cited articles

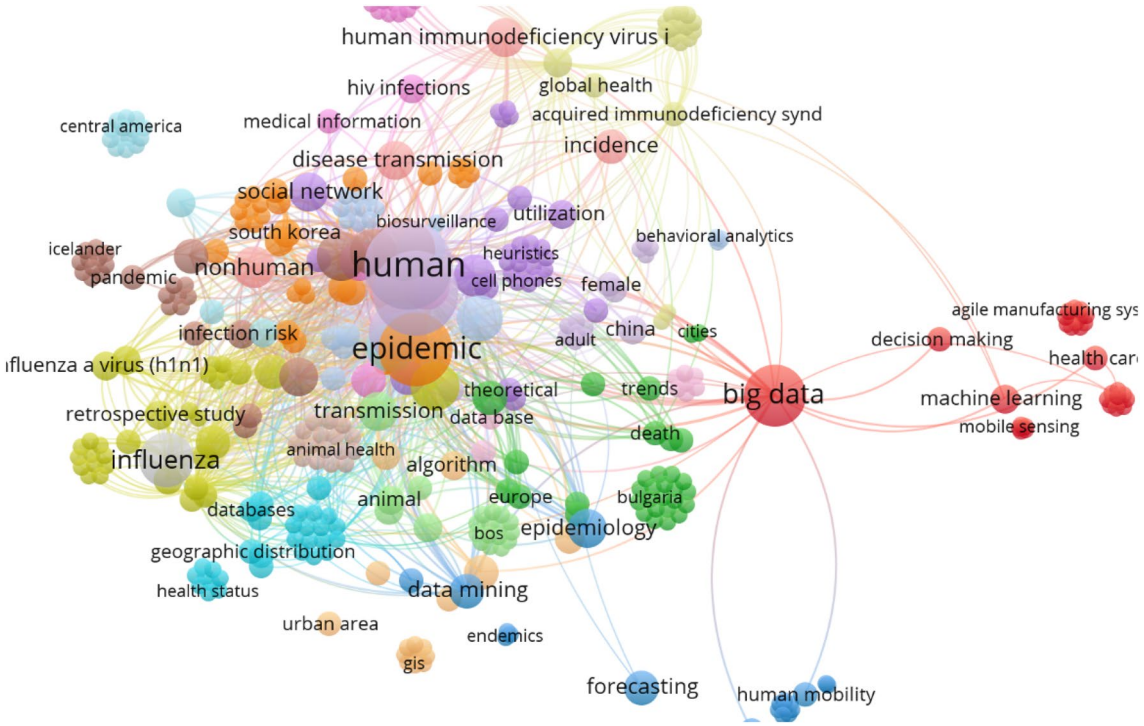


Fig. 8 Main keywords

pandemics/epidemics, being published in a journal with these focus. This demonstrates that although the topic of pandemics and epidemics is more recurrent in areas related to health, other areas can greatly contribute developing strategies to combat them.

Figure 8 shows the analysis performed with VOSviewer concerning the keywords of the articles.

From the words network, it is observed that the main keywords in the portfolio are "Human" and "Humans", present in 26 and 24 articles, respectively, revealing that the human and social factors of epidemics and pandemics are the most addressed themes in the articles. The words "Epidemic" and "Big Data" were also recurrent, present in 18 and 13 articles, respectively, revealing that the main terms searched are central in the final portfolio. These results evidence that the final portfolio is aligned with the research objective, confirming the effectiveness of the methodology used.

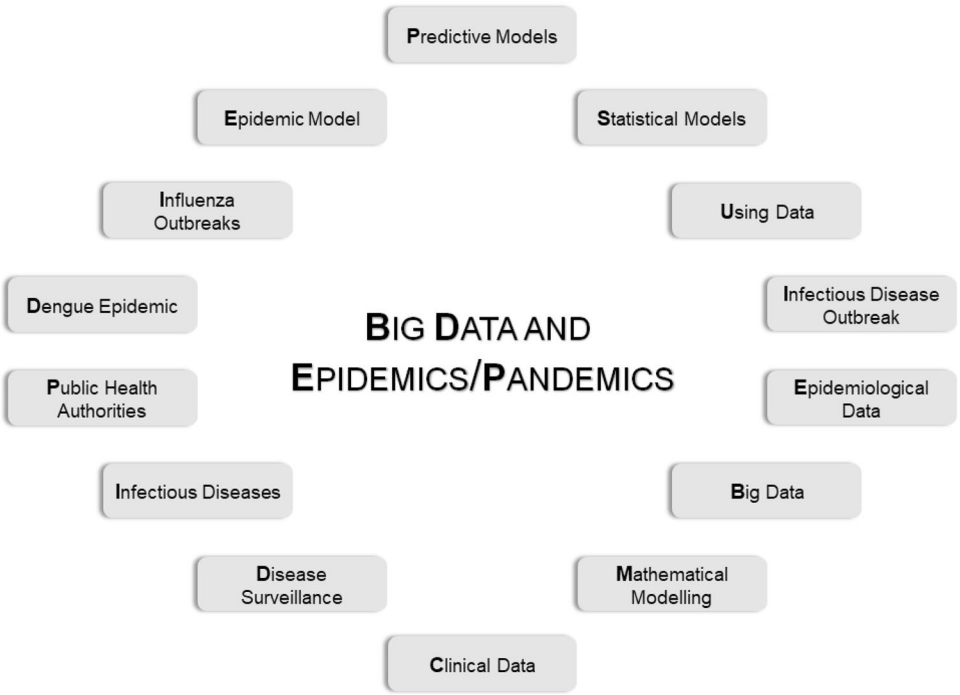


Fig. 9 Main themes addressed on the final portfolio

The words Internet and Social Media were also highlighted in the portfolio, present in 12 and 8 articles, respectively, evidencing the importance of these resources for population in times of pandemics and epidemics. Finally, aspects related to health and diseases were mentioned as keywords, among them, the most frequent were “Influenza”, “Health Survey” and “disease surveillance”, present in 8 articles each.

Based on Fig. 8, we can infer that the portfolio mainly addresses the issue of Big Data sources as coming from social media and internet, the issue of infectious outbreaks, and how these two themes affect the human/social factor.

The last aspect in the bibliometric analysis sought to identify the main topics addressed throughout the articles’ texts. For this task, the automatic coding functionality of the NVivo 12 software was used, which, based on a word frequency count, highlights the main terms covered, and demonstrates the number of sections that address those words. The results obtained are shown in Fig. 9.

It is observed that the main themes developed in the articles are on the use and availability of data, and mathematical and statistical models to treat these data related to epidemics and pandemics. Among the main terms coded, "Predictive Models" and "Statistical Models" were the most frequent terms, mapped in 58 and 56 sections, respectively. Other terms related to computational methods and Big Data were also highly mentioned, such as “Using Data” from search engine query (46 sections); “Epidemiological Data” and “Clinical Data”, with specific data on outbreaks, (38 and 31 sections each); and “Big Data” (34 sections).

The term Big Data was mentioned in conjunction with several other terms, such as “analytics” (44 sections); “sources” (14 sections); “bulks”, “sets”, “technologies”, “analysis”, “exchange”, “extraction” and “transparency” (11 sections each), an evidence that the Big Data concept is explored in order to extract and analyze its data as a tool to assist in the development of strategies for fighting pandemics/epidemics.

Other main terms were related to Epidemics and Pandemics, like “Infectious disease outbreak”, were identified in 43 sections. The main outbreaks, with the largest number of sections mapped, are Influenza and Dengue. Influenza was codified in 107 sections, along with: “epidemic”, "detection", "monitoring", "predicting", "curve", "identifying", "outbreak", “emerging”, "2013", “report” and “seasonal”. Dengue was codified in 30 sections, along with the terms "epidemic" and "fever epidemic". Influenza epidemics, Severe Acute Respiratory Syndrome (SARS) and AIDS were also mapped as frequent infectious outbreaks.

Based on the results obtained from the bibliometric analysis, it can be concluded that the portfolio is aligned with the main objective of this work, that is to identify how Big Data and Big Data Analytics can help fight epidemics and pandemics. The main keywords and terms, mapped in the articles addresses strategies and tools that can be exploited to handle and analyze big data, like the use of data coming from social media, for instance Twitter and Personal Blogs. With the aid of mathematical and statistical modeling, they serve to study pandemics, epidemics, and their behavior, facilitating the development of combat strategies, acting as

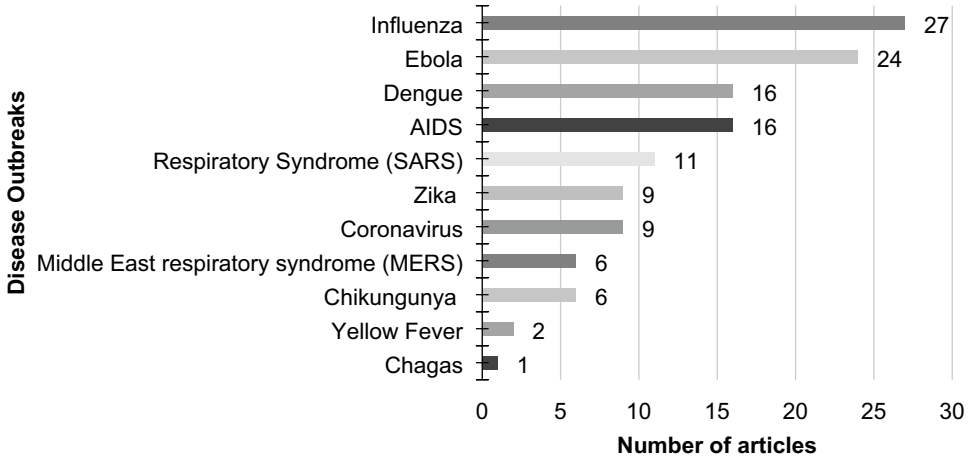


Fig. 10 Disease outbreaks

a computational tool to achieve benefic results for society, answering the main objective of this work.

4.2 Content analysis: big data and big data analytics to fight epidemics and pandemics

The systematic reading and content analysis revealed the types of epidemics and pandemics mentioned in the portfolio. The word search employed NVivo 12 software in order to quantify the number of articles that mentioned each disease. The results obtained are in Fig. 10.

It is observed that Influenza is most frequent outbreak, mentioned in 55% of the articles in the portfolio. The second most frequent was Ebola, present in 49% of the articles, followed by Dengue and AIDS, both present in 33% of the articles. Respiratory Syndrome (SARS) outbreaks were mentioned in 22%; the Middle East respiratory syndrome (MERS) was mentioned in 18% and Coronavirus in 12% of the articles.

After identifying the main epidemics/pandemics mentioned in the literature, the types of data and sources mentioned were used to propose techniques and methodologies that help fightingepidemics/pandemics or infectious outbreaks, answering RQ1: *What types of data and data sources are used to assist in the treatment of Epidemics/Pandemics?*

After the systematic reading, that evidenced the types of data and sources mentioned in the portfolio of articles, the word search function in the NVivo 12 software was used to quantify the number of articles that mentioned data and data sources. Figure 11 presents the results.

It can be observed that data from searches on search engines and social media are the main sources of data, mentioned in 22 and 14 articles, respectively. The research carried out in search engines revealed the users' concerns regarding the reality of their health status and diseases and infectious outbreaks (Nan and Gao 2018). Herland et al. (2014) claim that data from official sources, as in the reports of the Center of Disease Control and Prevention (CDC), can present a delay in the information. On the other hand, the search engines can be useful for detecting epidemics more quickly and accurately. This finding consists in a line of research to be explored, which could assist authorities in decision making and in reacting more quickly in case of future epidemics.

The most frequent data source from search engines is Google, mentioned in 35 articles, more than 70% of the portfolio. The other two mechanisms most frequently mentioned were Google Trends and Google Flu Trends, mentioned in 20 and 18 articles, representing 41% and 37%, respectively. In addition to Google, the search engine Baidu was mentioned in 11 articles, 22% of the portfolio. The Korean search engine Daum was also used as a source of data collection in two articles.

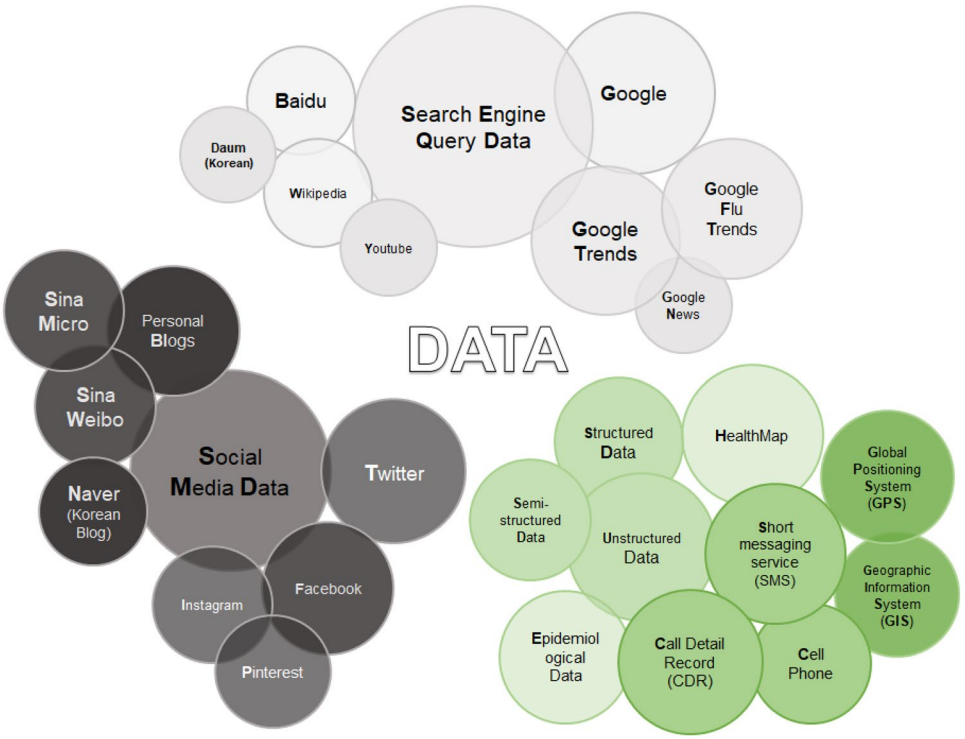
In 2008, Google developed the tool Google Flu Trends, which purpose was monitoring researchs related to influenza symptoms in the United States and predicting the outbreak of medical consultations for these symptoms. The tool proved to be useful in its first estimates. Nevertheless, between the years 2012 and 2013, the forecasts were incorrect, indicating twice as many medical consultations as the CDC (Carneiro and Mylonakis 2009; Huang et al. 2015; Drake and Lake 2015). Therefore, in 2015, the tool was discontinued and no longer publishes predictions, but maintains data from previous years for research purposes.[1]

Baidu is a search engine used more specifically by the Chinese, and it can provide insights into outbreaks of various diseases, including hepatitis, tuberculosis, venereal diseases, and influenza (Huang et al. 2015).

Another source of data collection widely addressed in the portfolio was social media. According to Bello-Orgaz et al. (2015), social media, such as Facebook, Twitter and Blogs,

[1] https://www.google.org/flutrends/about/.

Fig. 11 Data and data sources



are means of extracting the opinion of society in real-time, in addition to allowing the capture of geographic location information. This set of information available in the media, can be an important aid for public health once correctly used. Mitchell and Ross (2016) claim social media as tools to be explored in changing people's behavior during disease outbreaks. However, it is difficult to extract data from them, given the heterogeneous characteristics, unstructured data, and dynamic change (Bello-Orgaz et al. 2015).

Twitter was the most mentioned media, in 24 articles, almost 50% of the portfolio. The second most frequent medias were Blogs and Facebook, covered in 10 and 9 articles, respectively. Herland et al. (2014) present Twitter as a means of tracking influenza epidemics in real-time, with little error and with greater agility than reports from the CDC. Alessa and Faezipour (2018) add that the advantages of Twitter are the possibility of carrying out minute-by-minute analysis, compared to search engine records available. Mitchell and Ross (2016) claim Twitter to be an effective means of raising public awareness and may reduce the rate of transmission of influenza through publicity awareness.

As discussed, there are different sources of data that can be used to predict and contain disease outbreaks. For Moran et al. (2016), although data from official sources and records are accurate, and takes the necessary care with data privacy issues, it is a slow, bureaucratic, and expensive process. The delay between observation and the generation of the report on the outbreak prevents decision making in the early stages. On the other hand, data from the Internet, such as researchs on Google or Twitter, are free of cost or less expensive, and are available in real-time, representing a potential to be explored. Also, Woo et al (2016) add that influenza surveillance using social media data, such as blogs,Twitter and data from search engines, such as the Daum website, have the potential to detect influenza outbreaks, exhibiting congruence with traditional surveillance data.

Nevertheless, the data generated by all these means can be misleading and the result of their information is insufficient for the correct decision-making during critical health situations. Therefore, modern technologies cannot completely replace traditional epidemiological monitoring networks, like the CDC, but can play a complementary role if adequately used (Xue et al. 2018).

In addition to sources from search engines and social media, data from Global Position Systems (GPS) and Geographic Information System (GIS) were mentioned. Also, data from HealthMap, which presents infectious disease information and a monitoring system, was mentioned (Alessa and Faezipour 2018).

Finally, data characteristics were discussed, classifying them into unstructured, structured, and semi-structured data, identified in 10, 5, and 2 articles, respectively. Siriyasatien et al. (2018) defined the three types of data:

Unstructured Data Data stored as text, images, pictures, recordings, and videos. The processing of this type of data may require complex processing systems, such

as text mining. However, with the advent of Big Data technology, which uses the unstructured data storage method, unstructured data processing methods have gained greater visibility.

Structured Data Data stored in a structured format which facilitates its access. For instance, data stored in a relational database, organizing data into tables. It has advantages such as: consistent data; non-redundant data; data integrity; standardization; and data security. This type of data can be easily and efficiently manipulated, and accessed using the SQL language.

Semi-structured Data Relatively new form of storage, where data is stored together with tags or markers to indicate the type of data and other characteristics, such as XML and HTML. This format allows various types of data, such as text and image, that can be stored together.

Finally, some challenges concerning data and data sources were encountered and will be now discussed. According to Bello-Orgaz et al. (2015) and Alessa and Faezipour (2018), the unstructured configuration of data made available on social media, with different formats and dynamic variation, makes its collection and analysis more difficult. Besides, this type of data can present other difficulties, such as lack of accuracy; the size of the data, that makes it difficult to process; the language, that can be complex or use codes; the heterogeneity of information; sampling bias, since the social media population may not represent society; consistency, or lack of consistency, of the data set; lack of precise location of users; difficulty in defining a target population for the analysis; and presence of spam (Alessa and Faezipour 2018).

O'Shea (2017) addressed the issues of privacy and security as a challenge to data coming from social media, arguing that there is a need for regulations from these online sources, to ensure good governance and user privacy. Also, ethical issues must be taken into account, like maintaining the privacy and confidentiality of people who have the disease, essential to not cause discomfort. For this reason, the advancement of an ethical framework for assessing methods of making patient data available has become fundamental (Craven and Page 2015; Zhang et al. 2017).

According to Leclerc-Madlala et al. (2018), on the other hand, it becomes a difficulty when the data is subject to ownership and with restrict access, which prevents its wide use. This lack of data sharing between organizations, or even countries, can create a delay in the treatment of epidemics/pandemics.

Initially, the lack of data made forecasting outbreaks a challenge. However, with the advent of Big Data, the excess of data made available in real-time has also become a difficulty in global disease surveillance (Roberts 2019).

The HealthMap, which covers several sources of data to monitor outbreaks of infectious diseases worldwide, has the limitation of being configured for the English language, losing initial notifications of diseases reported in other languages (Erikson 2018). Thus, only later the outbreak can be detected which can become a larger problem than it could really be if detected in its early stages.

After mapping the types of infectious disease outbreaks addressed, and the types of data and sources mentioned, the techniques for treatment and analysis of a large number of data, available from different sources, and the effects that these techniques cause in epidemics/pandemics were identified, answering RQ2: *How Big Data Analytics/Analysis can assist in cases of Epidemics/Pandemics?*

The techniques mentioned, in two or more articles, were mapped and organized in a Treemap, in Fig. 12. The larger the rectangle, the greater the incidence of the term in the articles.

The data processing techniques found in the literature were related to the challenges also cited in the literature. Thus, the techniques mapped were divided according to their purpose, showing how such technique can help in situations of epidemics and pandemics: (1) Techniques for Forecasting and Monitoring Epidemics and Pandemics, and (2) Techniques for Awareness about the disease during the Epidemic or Pandemic.

1. *Techniques for Forecasting and Monitoring Epidemics and Pandemics* (Elkin et al. 2017; Spreco et al. 2017; Bouzillé et al. 2018).

Predicting future outbreaks and understanding how they are spreading from place to place can improve patient care and prevent the outbreak from becoming a pandemic. Big Data mining recently proved its ability to track patterns and trends around the world (Elkin et al. 2017). This form of Big Data, which brings additional challenges, such as Text Mining and noise manipulation, can lead to discoveries in the medical field (Herland et al. 2014).

Mathematical models, which predict the spread of epidemics, must overcome the challenges of associating incomplete and inaccurate data, estimating the probability of multiple possible scenarios, incorporating changes in human and/or pathogen behavior and environmental factors (Moran et al. 2016).

Statistical models and analysis are widely used in the processing of these data. For this reason, many studies on health and data mention regression models as a way to detect influenza outbreaks. Such models can be suitable and used to identify outbreaks of other diseases Bello-Orgaz et al. (2015).

In this scenario, ARGO (AutoRegression Google) is a rigorous statistical-based regression model that uses public

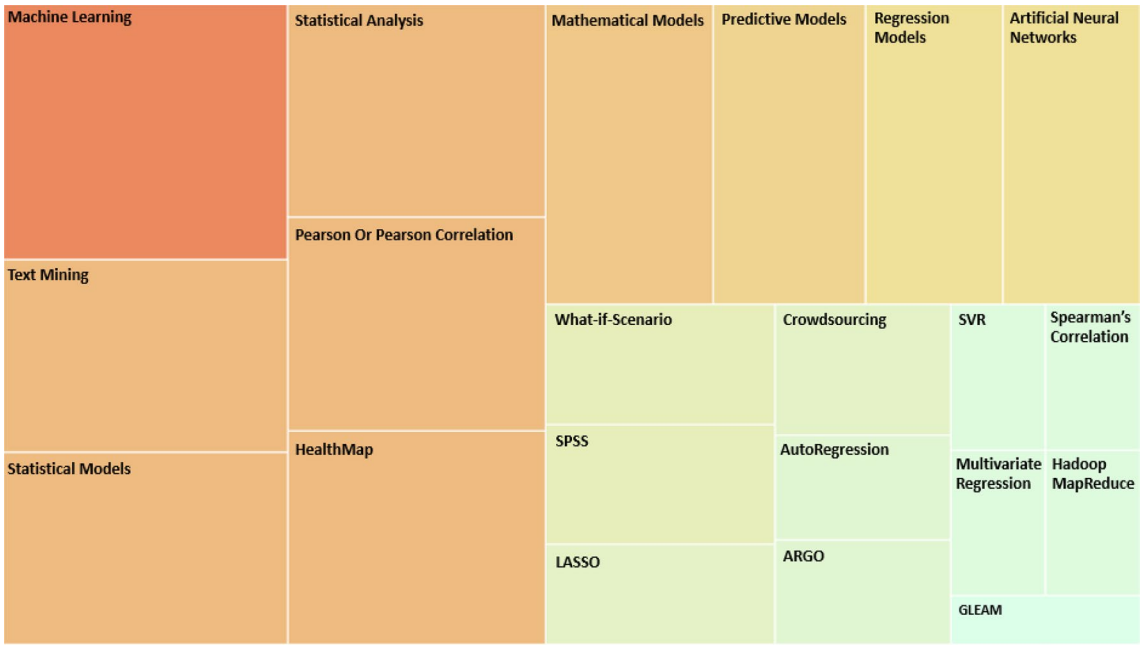


Fig. 12 Techniques used for data processing

data available online to predict influenza outbreaks around the world. ARGO not only incorporates seasonality in influenza epidemics but also captures changes in people's online research behavior over time. Also, it is flexible, robust, and scalable, making it a powerful tool that can be used for real-time tracking (Yang et al. 2015). These models generally integrate Auto Regression and Multivariate Regression, to achieve better results in tracking epidemics and pandemics in various locations (Zhang et al. 2019).

Still, in the area of statistical regression, the Support Vector Machine Regression Algorithm (SVR) is cited as useful in predicting disease outbreaks (Ritterman et al. 2015). In the work of Woo et al. (2016), the union between the SVR and Least Absolute Shrinkage and Selection Operator (Lasso), resulted in a model for predicting influenza epidemics. Already Bragazzi et al. (2017), used a model of Multivariate Regression widespread using the SVR to correlate the data on the Zika Virus.

The models of Correlation Statistics are also cited with the Pearson Correlation and Spearman's Correlation. Bragazzi et al. (2017), used the Spearman's Correlation in their work to correlate all data related to the Zika virus, which were extracted from various sources (Twitter, Google Notícias, Wikipedia, YouTube and outros). Also, Spearman was used to correlate the number of weekly Zika cases and search query volumes from Google Trends, Google News, and YouTube. Already Bouzillé et al. (2018), used Pearson Correlation and managed to correlate various data, such as patient reports, sex, age groups, among others, managing to associate these characteristics with serious epidemics.

To assist in statistical analysis, the *software* SPSS (Statistical Package for the Social Sciences), validated by different authors (Bragazzi et al. 2017; Siriyasatien et al. 2018; Gianfredi et al. 2018; Li et al. 2019; Zhang et al. 2020) is largely used.

According to Nan and Gao (2018), the use of Artificial Neural Networks together with the Machine Learning technique can predict incidents and deaths from AIDS epidemics. Data on AIDS-related research trends are collected from the largest Chinese Internet search engine Baidu, and are subsequently analyzed with the integration of the two techniques, taking into account three criteria: the absolute mean percentage error, the root mean square percentage error, and the agreement index, used to test the performance of the method prediction.

Moran et al. (2016) points out that the What-If-Scenario approach is beginning to be innovatively used to analyze the probability of an outbreak, but it is still far from being well understood and useful (Moran et al. 2016).

While an epidemic or pandemic is occurring, its data must be monitored as close to real-time as possible. In this regard, the Big Data generated by hospitals with disease data, patient monitoring, diagnostics, and treatments, can be used in disease monitoring, constituting a surveillance tool capable of providing relevant information in a short time. However, frequent evaluations of real efficacy must occur to validate the use of predictive models based on Big

Data of hospitals to correctly predict epidemics (Bouzillé et al. 2018).

Mobile devices and social media platforms also generate data that can be used to monitor infected patients, providing data on their location, for example (Kraemer et al. 2018). In this way, people who have had contact with the infected person can be notified through text messages, so that the necessary measures for their safety are taken, such as carrying out tests that indicate the contamination or not. Such initiative was carried out by the government of South Korea, during the COVID-19 pandemic, which started in 2020 (Harvard Business Review 2020).

However, this large amount of data generated by different devices requires data processing capable of handling this problem. Zadeh et al. (2019), used the IBM InfoSphere BigInsights as a platform for Big Data Analytics to derive underlying relationships in data sources and understand how that information can help achieve public health goals about epidemics. The BigInsights is the distribution of IBM the Hadoop, that combines open-source query language called JAQL with the usual components of Apache Hadoop, like the MapReduce, which is a Java-based technology for storing and batch processing large amounts of data (Štufi et al. 2020).

Another interesting point in the use of large amounts of data is the possibility of carrying out simulations with them, as though the public software GLEAM. This system simulates the spread of infectious diseases, from people to people, around the world. This simulation engine uses the structure Global Epidemic and Mobility (GLEaM), a computational stochastic method that integrates high resolution demographic and mobility data worldwide to simulate the spread of the disease on a global scale (Broeck et al. 2011; Liu et al. 2016).

With the recent COVID-19 pandemic, the authors Štufi et al. (2020), presented a system, which provides a visualization of the regional geographical map of the Czech Republic in 10 s and can exchange data with others systems medical assistance. In addition to data integration, geographic mapping functionality provides near real-time pandemic/epidemic tracking, outbreak propagation monitoring, and visualization of risk data. This system is an example that can be applied to other territories, thus integrating more information, becoming an essential tool, since it is a global pandemic.

2. *Awareness about the disease during the epidemic or pandemic* (Leclerc-Madlala et al., 2018).

The mining of data available on social networks and internet search platforms can be used by the BDA to identify the main doubts of the population. After this identification, mass media can be used as a tool to disseminate good information to guide people's behavior during the disease outbreak. These public awareness campaigns can increase the sharing of information on social media, bringing information to a larger number of people, which can increase, for instance, the vaccination rate of these (Mitchell and Ross 2016).

In this scenario, the Digital Humanitarians (Meier 2015) aim to apply Big Data Analytics to humanitarian aid through a participatory surveillance system, a form of crowdsourcing that creates real-time communication on social media (Erikson 2018). Participatory surveillance systems allow people to report via Internet, as in crowdsourcing. These systems encourage the regular and voluntary submission of health-related information by the general public, using computers, tablets, or smartphones similar to technologies (Wójcik et al. 2014; O'Shea 2017). By doing it, data can be shared and serve both to monitor outbreaks and to direct questions about the affected population.

Before the Big Data era, the results of health science were not easily disseminated, and the time to raise public awareness about the emergence of new diseases could be very long. Now, with Big data and the emergence of Big Data 'Analytics, this reality can be changed, as scientific data can be sent directly to mobile devices, through messages delivered almost instantly, for example. In this way, scientific reports can be widely disseminated and the impact of health science can be extended (Huang et al. 2015).

From the results obtained, it is observed that the data and techniques used in previous outbreaks, epidemics, and pandemics mentioned in the extant literature, can be applied in the current context of the COVID-19 pandemic, as they are generic technologies and applicable to different scenarios. Thus, the good practices reported can serve as a guide in the current pandemic scenario.

5 Conclusions

Contagious and infectious diseases represent a major challenge for health systems, both public and private, around the world. The increase in cases of these diseases, and the recent pandemic of COVID-19, raised the call for the use of new techniques and technologies capable of detecting, tracking, mapping and managing behavioral patterns in these diseases.

As it is a recent pandemic, little has been found in the literature on the use of data in the control and monitoring of the new COVID-19 pandemic. Thus, techniques and technologies used in other epidemics and pandemics have been mapped to relate them to the COVID-19 pandemic, seeking to highlight possible strategies that may help in this context.

Therefore, in order to identify how Big Data and Big Data Analytics can help in the fight against epidemics and pandemics, we mapped the types of data and sources used for analysis and creation of techniques that support the fight against epidemics and pandemics. Equally, the techniques used to treat these data were also mapped, thus showing their correlation with the current COVID-19 pandemic. To

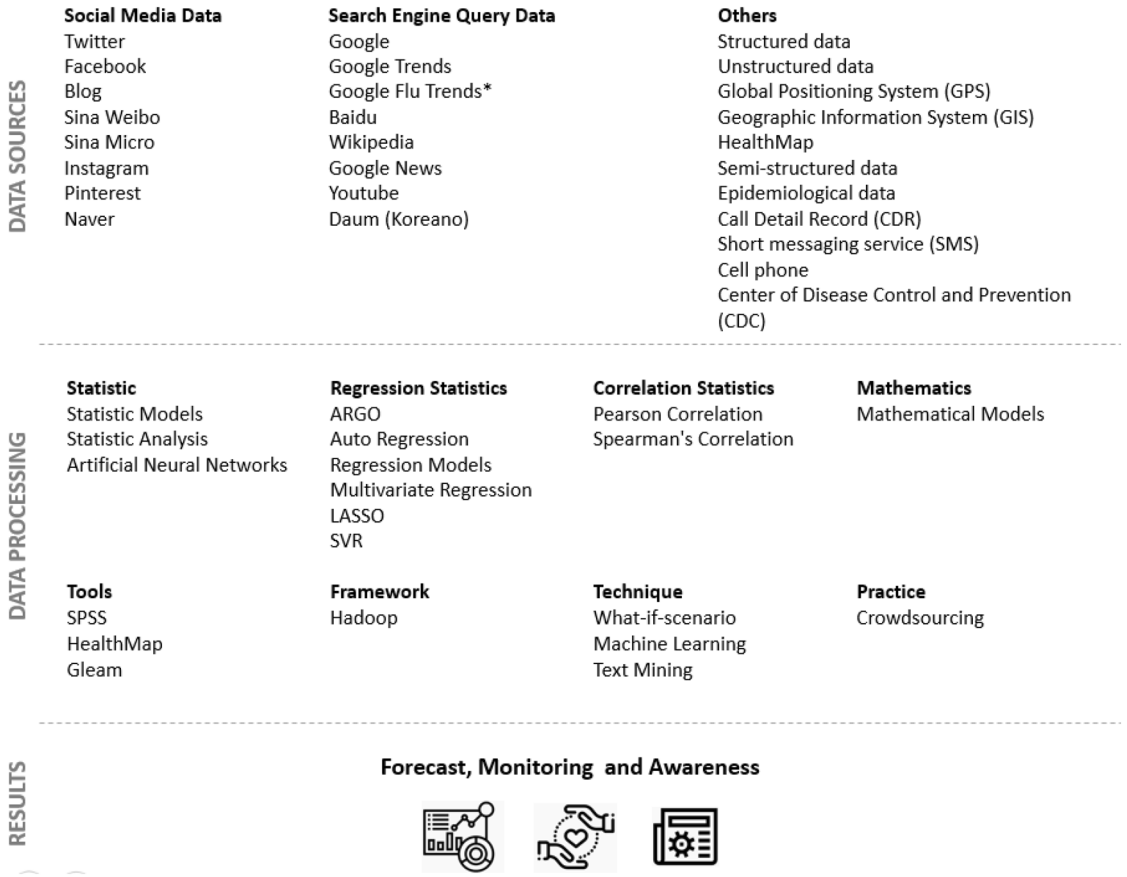


Fig. 13 Main results

accomplish the task proposed by research, a Systematic Literature Review was carried out, through Methodi Ordinatio, building a portfolio of articles with scientific relevance, which was a source of data collection and analysis. The results obtained are summarized in Fig. 13.

From this mapping and analysis, the research questions were answered, and the objective of the work achieved.

From the analysis and results obtained, it can be concluded that the types of outbreaks of infectious diseases most commonly addressed until now are Influenza and Ebola. The most recurrent types of data, on the other hand, come from social media and search engines, being data characterized as unstructured, carrying several difficulties in its treatment.

Finally, it was found that the main techniques mentioned can be divided by purpose, presenting techniques for forecasting and monitoring epidemics and pandemics, and techniques for raising awareness about diseases. Among the most discussed techniques are Machine Learning, Text Mining, and statistical models and analysis. From these results, it is possible to evidence strategies for combating the current pandemic of the new coronavirus (COVID-19).

The changes Big Data can bring to the healthcare sector promise to be much bigger than many, governments, companies and organizations, can realize. With the emergence of a series of smart devices capable of collecting, storing, analyzing, and sharing user data in a cloud, will make many data available to millions of people. This situation will change, almost completely, the way the health science outcome is achieved (Huang et al. 2015).

The limitations of this study is the fact that the review used only articles from journals, excluding studies published in other sources, such as books and conference papers.

We have realized that the Health Area is a fruitfl field for Big Data Science, in which not both can broadly benefit from future studies, but the whole society as well. In future researches, studies that analyze data from public and private hospitals, or even from social media, can be developed using BDA technologies, indicated in Fig. 13. From these studies, results, dashboards for monitoring information, and even insights on data from patients who contracted the disease, future outbreks could be early identified and prevent diseases to spread.

Acknowledgements The present study was carried out with support from the Higher Education Personnel Improvement Coordination—Brazil (CAPES)—Financing Code 001.

Funding Not applicable.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

Appendix

See Table 3.

Table 3 Final portfolio

Title	Inordinatio
A review of data mining using big data in health informatics	301,00
Accurate estimation of influenza epidemics using Google search data via ARGO	277,01
Promises and Challenges of Big Data Computing in Health Sciences	224,00
Methods of using real-time social media technologies for detection and remote monitoring of HIV outcomes	195,00
Inference of seasonal and pandemic influenza transmission dynamics	160,01
Supersize me: How whole-genome sequencing and big data are transforming epidemiology	146,01
A review of influenza detection and prediction through social networking sites	122,00
Connecting Mobility to Infectious Diseases: The Promise and Limits of Mobile Phone Data	118,01
Epidemic Forecasting is Messier Than Weather Forecasting: The Role of Human Behavior and Internet Data Streams in Epidemic Forecast	115,01
Using Networks to Combine 'Big Data' and Traditional Surveillance to Improve Influenza Predictions	115,00
Global reaction to the recent outbreaks of Zika virus: Insights from a Big Data analysis	106,00
Using big data to predict pertussis infections in Jinan city, China: a time series analysis	104,00
Estimating influenza outbreaks using both search engine query data and social media data in South Korea	103,00
Using Baidu Search Engine to Monitor AIDS Epidemics Inform for Targeted intervention of HIV/AIDS in China	102,00
Social Media for Nowcasting Flu Activity: Spatio-Temporal Big Data Analysis	102,00
Monitoring public interest toward pertussis outbreaks: an extensive Google Trends–based analysis	102,00
Predicting seasonal influenza epidemics using cross-hemisphere influenza surveillance data and local internet query data	101,00
Flying, phones and flu: Anonymized call records suggest that Keflavik International Airport introduced pandemic H1N1 into Iceland in 2009	101,00
Using big data to monitor the introduction and spread of Chikungunya, Europe, 2017	100,01
Digital disease detection: A systematic review of event-based internet biosurveillance systems	100,00
Big data analytics and processing platform in Czech Republic healthcare	100,00
Inferences about spatiotemporal variation in dengue virus transmission are sensitive to assumptions about human mobility: a case study using geolocated tweets from Lahore, Pakistan	98,00
Leveraging hospital big data to monitor flu epidemics	96,00
Reality mining: A prediction algorithm for disease dynamics based on mobile big data	95,00
Big data, algorithmic governmentality and the regulation of pandemic risk	95,00
Social Big Data Analysis of Information Spread and Perceived Infection Risk during the 2015 Middle East Respiratory Syndrome Outbreak in South Korea	93,00
Influenza Activity Surveillance Based on Multiple Regression Model and Artificial Neural Network	92,00
Forecasting AIDS prevalence in the United States using online search traffic data	92,00
A machine learning method to monitor China's AIDS epidemics with data from Baidu trends	91,00
Harnessing big data for communicable tropical and subtropical disorders: Implications from a systematic review of the literature	91,00
Signals, Signs and Syndromes: Tracing [Digital] Transformations in European Health Security	91,00
How Big Data Science Can Improve Linkage and Retention in Care	90,01
The 'end of AIDS' project: Mobilising evidence, bureaucracy, and big data for a final biomedical triumph over AIDS	87,00
Cell Phones ≠ Self and Other Problems with Big Data Detection and Containment during Epidemics	87,00
Dengue Epidemics Prediction: A Survey of the State-of-the-Art Based on Data Science Processes	86,00
Evaluation of nowcasting for detecting and predicting local influenza epidemics, Sweden, 2009–2014	85,01
A data-driven model for influenza transmission incorporating media effects	81,00
Integrated detection and prediction of influenza activity for real-time surveillance: Algorithm design	80,00
Social Media Monitoring of Discrimination and HIV Testing in Brazil, 2014–2015	80,00
EpiDMS: Data management and analytics for decision-making from epidemic spread simulation ensembles	77,01
Elucidating transmission patterns from internet reports: Ebola and middle east respiratory syndrome as case studies	74,01
Network based model of social media big data predicts contagious disease diffusion	73,00
Sources of spatial animal and human health data: Casting the net wide to deal more effectively with increasingly complex disease problems	71,00
Lyme disease: The promise of big data, companion diagnostics and precision medicine	70,00
A survey of social web mining applications for disease outbreak detection	56,00

References

APIC (Association for Professionals in Infection Control and Epidemiology) (2019) Outbreaks, epidemics and pandemics—what you need to know. https://apic.org/monthly_alerts/outbreaks-epidemics-and-pandemics-what-you-need-to-know/

Agbehadji I, Awuzie B, Ngowi A, Millham R (2020) Review of big data analytics, artificial intelligence and nature-inspired computing models towards accurate detection of COVID-19 pandemic cases and contact tracing. Int J Environ Res Public Health 17(15):5330–5342

Ajayi A, Oyedele L, Akinade O, Bilal M, Owolabi H, Akanbi L, Delgado JMD (2020) Optimised big data analytics for health and safety hazards prediction in power infrastructure operations. Saf Sci 125:1–12. https://doi.org/10.1016/j.ssci.2020.104656

Ali AH, Abdullah MZ (2019) A survey on vertical and horizontal scaling platforms for big data analytics. Int J Integr Eng 11(6):138–150

Babar M, Arif F (2019) Real-time data processing scheme using big data analytics in internet of things based smart transportation environment. J Ambient Intell Hum Comput 10:4167–4177. https://doi.org/10.1007/s12652-018-0820-5

Beam AL, Kohane IS (2018) Big data and machine learning in health care. JAMA 319(13):1–12. https://doi.org/10.1001/jama.2017.18391

Bello-Orgaz G, Hernandez-Castro J, Camacho DA (2015) Survey of social web mining applications for disease outbreak detection. Intell Distrib Comput 8:345–356. https://doi.org/10.1007/978-3-319-10422-5_36

Bloom DE, Cadarette D (2019) Infectious disease threats in the 21st century: strengthening the global response. Front Immunol 10:549. https://doi.org/10.3389/fimmu.2019.00549

Bouzillé G, Poirier C, Campillo-Gimenez B, Aubert ML, Chabot M, Chazard E, Lavenu A, Cuggia M (2018) Leveraging hospital big data to monitor flu epidemics. Comput Methods Progr Biomed 154:153–160. https://doi.org/10.1016/j.cmpb.2017.11.012

Bragazzi NL, Alicino C, Trucchi C, Paganino C, Barberis I, Martini M, Sticchi L, Trinka E, Brigo F, Ansaldi F (2017) Global reaction to the recent outbreaks of Zika virus: insights from a big data analysis Insights from a big data analysis. PLoS ONE 12(9):1–15. https://doi.org/10.1371/journal.pone.0185263

Campos EAR, Pagani RN, Resende LM, Pontes J (2018) Construction and qualitative assessment of a bibliographic portfolio using the methodology Methodi Ordinatio. Scientometrics 116(2):815–842

Carneiro HA, Mylonakis E (2009) Google Trends: a web: based tool for real: time surveillance of disease outbreaks: a web-based tool for real-time surveillance of disease outbreaks. Clin Infect Dis 49(10):1557–1564. https://doi.org/10.1086/630200

CDC (Centers for Disease Control and Prevention) (2012) Principles of epidemiology in public health practice, third edition an introduction to applied epidemiology and biostatistics. https://www.cdc.gov/csels/dsepd/ss1978/index.html. Accessed 20 Apr 2020

Chung K, Yoo H, Choe D (2018) Ambient context-based modeling for health risk assessment using deep neural network. J Ambient Intell Hum Comput 11(4):1387–1395

Craven M, Page CD (2015) Big data in healthcare: opportunities and challenges: opportunities and Challenges. Big Data 3(4):209–210. https://doi.org/10.1089/big.2015.29001.mcr

Davoudi S, Dooling JA, Glondys B, Jones TD, Kadlec L, Overgaard SM, Ruben K, Wendicke A (2015) Data quality management model (Updated). J AHIMA 86:62–65

Eken S (2020) An exploratory teaching program in big data analysis for undergraduate students. J Ambient Intell Hum Comput. https://doi.org/10.1007/s12652-020-02447-4

Elkin LS, Topal K, Bebek G (2017) Network based model of social media big data predicts contagious disease diffusion. Inf Disc Deliv 45(3):110–120. https://doi.org/10.1108/idd-05-2017-0046

Ergüzen A, Ünver M (2018) Developing a file system structure to solve healthy big data storage and archiving problems using a distributed file system. Appl Sci 8(6):2–20. https://doi.org/10.3390/app8060913

Erikson SL (2018) Cell phones ≠ self and other problems with big data detection and containment during epidemics. Med Anthropol Q 32(3):315–339. https://doi.org/10.1111/maq.12440

Gianfredi V, Nl B, Mahamid M, Bisharat B, Mahroum N, Amital H, Adawi M (2018) Monitoring public interest toward pertussis outbreaks: an extensive google trends-based analysis: an extensive Google Trends–based analysis. Public Health 165:9–15. https://doi.org/10.1016/j.puhe.2018.09.001

Harvard Business Review (2020) How digital contact tracing slowed covid-19 in east asia. https://hbr.org/2020/04/how-digital-contact-tracing-slowed-covid-19-in-east-asia. Accessed 20 Apr 20

Herland M, Khoshgoftaar TM, Wald R (2014) A review of data mining using big data in health informatics. J Big Data 1(1):2–12. https://doi.org/10.1186/2196-1115-1-2

Hu H, Wen W, Chua T-S, Li X (2014) Toward scalable systems for big data analytics: a technology tutorial. IEEE Access 2:652–687. https://doi.org/10.1109/ACCESS.2014.2332453

Hu F, Liu W, Tsai S-B, Gao J, Bin N, Chen Q (2018) An empirical study on visualizing the intellectual structure and hotspots of big data research from a sustainable perspective. Sustainability 10(3):1–19. https://doi.org/10.3390/su10030667

Huang T, Lan L, Fang X, An P, Min J, Wang F (2015) Promises and challenges of big data computing in health sciences. Big Data Res 2(1):2–11. https://doi.org/10.1016/j.bdr.2015.02.002

Kauffmann E, Peral J, Gil D, Ferrández A, Sellers R, Mora H (2019) A framework for big data analytics in commercial social networks: a case study on sentiment analysis and fake review detection for marketing decision-making. Ind Market Manag 2019:1–13

Kraemer MUG, Bisanzio D, Reiner RC, Zakar R, Hawkins JB, Freifeld CC, Smith DL, Hay SI, Brownstein JS, Perkins TA (2018) Inferences about spatiotemporal variation in dengue virus transmission are sensitive to assumptions about human mobility: a case study using geolocated tweets from lahore, pakistan: a case study using geolocated tweets from Lahore, Pakistan. EPJ Data Sci 7(1):1–17. https://doi.org/10.1140/epjds/s13688-018-0144-x

Lake P, Drake R (2015) Information systems management in the big data era. Springer, Berlin

Last JM (2001) Dictionary of epidemiology. Oxford University Press, New York, p 61

Leclerc-Madlala S, Broomhall L, Fieno J (2018) The 'end of AIDS' project: mobilising evidence, bureaucracy, and big data for a final biomedical triumph over aids: Mobilising evidence, bureaucracy, and big data for a final biomedical triumph over AIDS. Glob Public Health 13(8):972–981. https://doi.org/10.1080/17441692.2017.1409246

Li K, Liu M, Feng Y, Ning C, Ou W, Sun J, Wei W, Liang H, Shao Y (2019) Using Baidu search engine to monitor AIDS epidemics inform for targeted intervention of HIV/AIDS in China. Sci Rep 9(1):1–12. https://doi.org/10.1038/s41598-018-35685-w

Liu S, Poccia S, Candan KS, Chowell G, Sapino ML (2016) EpiDMS: data management and analytics for decision-making from epidemic spread simulation ensembles: data management and analytics for decision-making from epidemic spread simulation ensembles. J Infect Dis 214(4):427–432. https://doi.org/10.1093/infdis/jiw305

Manogaran G, Lopez D (2018) Spatial cumulative sum algorithm with big data analytics for climate change detection. Comput Electr Eng 65:207–221

Meera S, Sundar C (2020) A hybrid metaheuristic approach for efficient feature selection methods in big data. J Ambient Intell Hum Comput 2020:1–9. https://doi.org/10.1007/s12652-019-01656-w

Meier P (2015) Digital humanitarians: how big data is changing the face of humanitarian response. CRC Press, Boca Raton

Mitchell L, Ross JV (2016) A data-driven model for influenza transmission incorporating media effects. R Soc Open Sci 3(10):1–10. https://doi.org/10.1098/rsos.160481

Moran KR, Fairchild G, Generous N, Hickmann K, Osthus D, Priedhorsky R, Hyman J, Valle SY (2016) Epidemic forecasting is messier than weather forecasting: the role of human behavior and internet data streams in epidemic forecast: the role of human behavior and internet data streams in epidemic forecast. J Infect Dis 214(4):404–408. https://doi.org/10.1093/infdis/jiw375

O'shea J (2017) Digital disease detection: a systematic review of event-based internet biosurveillance systems: a systematic review of event-based internet biosurveillance systems. Int J Med Inf 101:15–22. https://doi.org/10.1016/j.ijmedinf.2017.01.019

Pagani RN, Kovaleski JL, Resende LM (2015) Methodi Ordinatio: a proposed methodology to select and rank relevant scientific papers encompassing the impact factor, number of citation, and year of publication. Scientometrics 105(3):2109–2135

Pagani RN, Kovaleski JL, Resende LM (2017) Tics na composição da methodi ordinatio: construção de portfólio bibliográfico sobre modelos de Transferência de Tecnologia. Ciência Inf 46:2

Rana AI, Mugavero MJ (2019) How big data science can improve linkage and retention in care. Infect Dis Clin N Am 33(3):807–815. https://doi.org/10.1016/j.idc.2019.05.009

Saggi MK, Jain S (2018) A survey towards an integration of big data analytics to big insights for value-creation. Inf Process Manage 54(5):758–790

Siriyasatien P, Chadsuthi S, Jampachaisri K, Kesorn K (2018) Dengue epidemics prediction: a survey of the state-of-the-art based on data science processes: a survey of the state-of-the-art based on data science processes. IEEE Access 6:53757–53795. https://doi.org/10.1109/access.2018.2871241

Sivarajah U, Kamal MM, Irani Z, Weerakkody V (2017) Critical analysis of big data challenges and analytical methods. J Business Res 70(1):263–286

Spreco A, Eriksson O, Dahlström Ö, Cowling BJ, Timpka T (2017) Integrated detection and prediction of influenza activity for real-time surveillance: algorithm design: algorithm design. J Med Internet Res 19(6):1–22. https://doi.org/10.2196/jmir.7101

van Broeck WD, Gioannini C, Gonçalves B, Quaggiotto M, Colizza V, Vespignani A (2011) The GLEaMviz computational tool, a publicly available software to explore realistic epidemic spreading scenarios at the global scale. BMC Infect Dis 11(1):1–12. https://doi.org/10.1186/1471-2334-11-37

Wamba SF, Gunasekaran A, Akter S, Dubey R (2020) The performance effects of big data analytics and supply chain ambidexterity: the moderating effect of environmental dynamism. Int J Prod Econ 222(1):1–14

Wasim ASK, Varalakshmi M, Sudeepthi J (2019) Big data analytics—current status, challenges and connection of unbounded data processing platforms. Int J Innovat Technol Explor Eng 8(92):698–700. https://doi.org/10.35940/ijitee.I1144.0789S219

WHO (World Health Organization) (2003) Severe acute respiratory syndrome (SARS). https://www.who.int/csr/sars/en/ea56r29.pdf?ua=1

WHO (World Health Organization) (2019) Emergencies: disease outbreaks. https://www.who.int/emergencies/diseases/en/

Wójcik OP, Brownstein JS, Chunara R, Johansson M (2014) Public health for the people: participatory infectious disease surveillance in the digital age: participatory infectious disease surveillance in the digital age. Emerg Themes Epidemiol 11(1):1–12. https://doi.org/10.1186/1742-7622-11-7

Woo H, Cho Y, Shim E, Lee J, Lee C, Kim SH (2016) Estimating influenza outbreaks using both search engine query data and social media data in South Korea. J Med Internet Res 18(7):1–12. https://doi.org/10.2196/jmir.4955

Xue H, Bai Y, Hu H, Liang H (2018) Influenza activity surveillance based on multiple regression model and artificial neural network. IEEE Access 6:563–575. https://doi.org/10.1109/access.2017.2771798

Yang W, Lipsitch M, Shaman J (2015) Inference of seasonal and pandemic influenza transmission dynamics. Proc Natl Acad Sci 112(9):2723–2728. https://doi.org/10.1073/pnas.1415012112

Yu S, Liu M, Dou W, Liu X, Zhou S (2017) Networking for big data: a survey. IEEE Commun Surv Tutor 19(1):531–549. https://doi.org/10.1109/COMST.2016.2610963

Zadeh AH, Zolbanin HM, Sharda R, Delen D (2019) Social media for nowcasting flu activity: spatio-temporal big data analysis: spatio-temporal big data analysis. Inf Syst Front 21(4):743–760. https://doi.org/10.1007/s10796-018-9893-0

Zhang X, Zheng Y, Wang D, Zhou F (2017) Solid-liquid triboelectrification in smart U-tube for multifunctional sensors. Nano Energy 40:95–106. https://doi.org/10.1016/j.nanoen.2017.08.010

Zhang Y, Yakob L, Bonsall MB, Hu W (2019) Predicting seasonal influenza epidemics using cross-hemisphere influenza surveillance data and local internet query data. Sci Rep 9(1):1–9. https://doi.org/10.1038/s41598-019-39871-2

Zhang Y, Bambrick H, Mengersen K, Tong S, Feng L, Zhang L, Liu G, Xu A, Hu W (2020) Using big data to predict pertussis infections in Jinan city, China: a time series analysis: a time series analysis. Int J Biometeorol 64(1):95–104. https://doi.org/10.1007/s00484-019-01796-w