Journal of the Operational Research Society

ISSN: 0160-5682 (Print) 1476-9360 (Online) Journal homepage: https://www.tandfonline.com/loi/tjor20

A review of the literature on big data analytics in healthcare

Panagiota Galetsi & Korina Katsaliaki

To cite this article: Panagiota Galetsi & Korina Katsaliaki (2019): A review of the literature on big data analytics in healthcare, Journal of the Operational Research Society, DOI: 10.1080/01605682.2019.1630328

To link to this article: https://doi.org/10.1080/01605682.2019.1630328

Published online: 05 Jul 2019.

JOURNAL OF THE OPERATIONAL RESEARCH SOCIETY
https://doi.org/10.1080/01605682.2019.1630328

ORIGINAL ARTICLE

A review of the literature on big data analytics in healthcare

Panagiota Galetsi and Korina Katsaliaki

School of Economics, Business & Legal Studies, International Hellenic University, Thessaloniki, Greece

ABSTRACT

Big data analytics (BDA) is of paramount importance in healthcare aspects such as patient diagnostics, fast epidemic recognition, and improvement of patient management. The objective of this profiling study is (a) to provide an overview of the BDA publication dynamics in the healthcare domain and (b) to discuss this scientific field through related examples. A sampling literature review has been conducted. A total of 804 papers have been identified and content analysis has been performed to mine knowledge in the domain for the years 2000–2016. The findings show that co-authors' backgrounds are from the subject areas of medicine and computer sciences. Most articles are experimental in nature and use modeling and machine learning techniques to exploit clinical data, for health monitoring and prediction purposes. Many articles are relevant to the medical specialties of neurology/neurosurgery/neuropsychiatry, medical oncology, and cardiology. Well-cited papers investigate the identification and management of high-risk/cost patients, the use of big data, Hadoop and cloud computing in genomics, and the development of mobile applications for disease management. Important is also the research about improving disease prediction by investigating patients' medical results using advanced analysis (such as segmentation and predictive modelling, machine learning, visualisation, etc.).

ARTICLE HISTORY
Received 20 April 2018
Accepted 28 May 2019

KEYWORDS
Health analytics; big data analytics; techniques; capabilities; bibliometrics; sources of data

Introduction

Big data analytics (BDA), in Healthcare or Health Analytics, is a method of analysis of the wide amount of electronic data related to patient healthcare and well-being which is very diverse and difficult to measure by traditional software or hardware. To illustrate data volume magnitude, the health data explosion from 500 petabytes in 2012 will reach 25.000 petabytes in 2020 (Feldman, Martin, & Skotnes, 2012). These data include clinical data, patient data, machine generated data, emergency care data, news feeds, articles in medical journals, web pages, social media posts, and blogs. All these data may come from internal (e.g., electronic health records, clinical decision support systems, etc.) and external sources (government sources, laboratories, pharmacies, insurance companies, etc.), in multiple formats (flat files, relational tables, text, etc.) and from multiple locations (different healthcare providers' sites) (Raghupathi & Raghupathi, 2013).

The term BDA includes two perspectives: big data and analytics (Wang, Gunasekaran, Ngai, & Papadopoulos, 2016). Big data in healthcare involves all three characteristics, the so-called 3Vs: volume--due to the incredible size of data, velocity—due to the rapid and real-time accumulation, variety—due to the differentiated formats (structured, unstructured, and semi-structured) (Raghupathi & Raghupathi, 2014). For the most recently added 4th V of big data, veracity—concerning the trustworthiness of data, improvements are made, as we move from handwritten notes to electronic and machine generated data. Therefore, the term "big data" does not characterise only the volume but it also highlights the analytical workloads associated with some combination of data variety and velocity, as well as volume (Ferguson, 2012).

The term "analytics" pulls together Management Information Systems (MIS), Operational Research (OR), and statistics. It describes the combination of Business Intelligence reporting and descriptive analysis, advanced statistical methods in data mining and forecasting, and other OR methods such as optimisation and simulation (Gorman & Klimberg, 2014). OR has been benefited from big data processing and analytics for advanced problem solving and better decision-making (Turaga, 2018). The exploitation of big data via advanced modeling techniques creates great opportunities (Hazen, Skipper, Boone, & Hill, 2018). Analytics are categorised as descriptive, predictive and prescriptive. This taxonomy refers to the nature of the analysis techniques and the information gained. Descriptive analytics aim to identify problems and trends in existing processes and functions, predictive involve the use of mathematical algorithms to discover predictive patterns

CONTACT Korina Katsaliaki k.katsaliaki@ihu.edu.gr School of Economics, Business & Legal Studies, International Hellenic University, Thessaloniki 57001, Greece

and prescriptive determine decisions based on certain objectives for improving performance (Wang et al., 2016).

The effort of the interested parties (clinicians, patients, healthcare organisations, researchers, etc.) to address issues in order to harness and maximise the potential of BDA in healthcare is noteworthy. It is argued that only a small part of the available amount of data is currently captured, stored and organised so that it can be processed by computers and analysed for useful information. Therefore, healthcare organisations need more efficient ways to manage them (Raghupathi & Raghupathi, 2014). From the beginning of the Twenty-first century, healthcare organisations face challenges, such as reduced fee schemes, demands for faster turnaround times, diminished numbers of qualified technologists, etc. (Horowitz et al., 2005). To meet these challenges, hospitals and healthcare systems rely more and more on automation and management of those data that come from clinical and operational information systems such as Electronic Health Records (EHR) and Laboratory Information Management Systems (LIMS) (Ward, Marsolo & Froehle, 2014). It is suggested that the management of healthcare data could be beneficial with regards to fraud detection and prevention, production of effective drugs and devices for patients' well-being and improvement of public health surveillance and speed of service (Raghupathi & Raghupathi, 2014).

Searching in the international literature for systematic reviews in BDA in healthcare, we noticed that in the existing literature there is a lack of holistic bibliometric approach towards the characteristics of these publications. We identified a number of very informative papers which profile research in the field of BDA (Chen, Chiang & Storey, 2012; Peng, Shi, Fantinato & Chen, 2017, etc.), however only a few focus on the use of big data in the health sector, such as the study of Wamba, Anand, and Carter (2013), which is a review of 215 papers about "RFID-enabled healthcare applications" and the study of West, Borland, and Hammond (2014) that examines 18 articles on the issue of "innovative information visualisation of electronic health records." These studies are of limited spectrum in terms of the number of papers analysed and/or the discussed content. On the other hand, there are reviews and profiling papers in OR healthcare which study the use of techniques to solve complex healthcare problems (Jun, Jacobson & Swisher, 1999; Brandeau Sainfort & Pierskalla, 2004; Katsaliaki & Mustafee, 2011). Although the published material in the field of BDA in health is increasing in recent years (Wills, 2014; Ivan & Velicanu, 2015;Thouin, Hoffman & Ford, 2008) and despite the fact that BDA systems provide organisational benefits (Cosic, Shanks, & Maynard, 2012), to the best of our knowledge, there is no published work that provides both a wide bibliometric and content analysis of this material by categorising at the same time the applications, tools and methodologies of BDA in healthcare for understanding how and why these benefits are achieved over time. The high number of publications on the medical field makes systematic reviews valuable to researchers in order to keep pace with the recent developments.

The scope of this study is twofold. Firstly, it aims to record the production of articles between 2000 and 2016 and provide an overview of the publication activity. Secondly, it offers unique information through targeted examples in order to explain the use of 'Big Data Analytics in Healthcare." More specifically, the present study attempts to summarise the "state of the art in the subject field" and based on this mined knowledge indicate the medical specialties connected with this research, the big data/OR techniques which are applied for data analysis together with the types of data used and determine overall the nature of analytics along with the capabilities of big data in healthcare. We hope that this study would be a beneficial contribution to researchers and the sector itself, as, for the time being no other study provides such a wide overview of this complex but very promising field.

Materials and methods

For this purpose, we developed a concept centric literature review using the systematic framework of the following schema: (1) Input, (2) Processing, and (3) Output (Figure 1).

In preparing this article, we conducted a search in two well-known and extensive electronic databases: Web of Science® and Scopus. We only reviewed papers published between 2000 and 2016. We used the year 2000 as the starting point since the term "analytics" was first introduced in the late 2000 (Chen et al., 2012), the term "business intelligence," which we consider similar (both terms investigate the capabilities of analytical tools in the business processes, Chae & Olson, 2013) has been established after 2000 (Chuah & Wong, 2011) and the term "big data" started appearing in many well-cited publications even later (Davenport, 2006; Akter & Wamba, 2016). Therefore, for our keyword search we used the combination of the terms (a) "business intelligence" (b) "analytics," and (c) "big data", which are the keywords used in many reviews as "unified terms" (Chen et al. 2012; O'Connell, 2012; Nie & Li, 2011; Duan & Xiong, 2015) and added the term "health*" and its derivatives

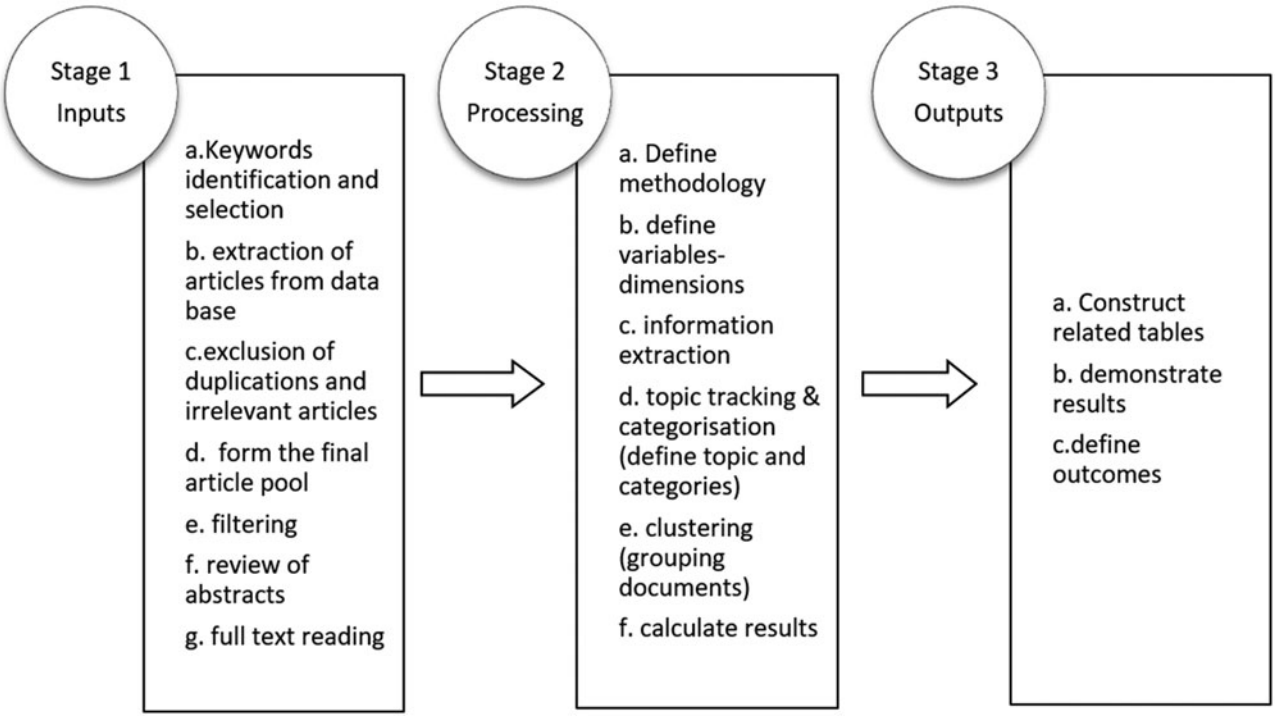

Figure 1. Systematic framework.

(symbol"*"), for example, "healthcare," "health sector," health records," 'health datasets," etc. We also used the combination of the terms (a), (b), and (c) with the term "medical," for example, "medical records," "medical data" etc. (Iqbal et al., 2016) and the term "clinical". These three terms (health, medical, and clinical) are mostly used in scientific papers to describe the nature of the data in the healthcare domain and the purpose of analysis and decision-making (Huang et al., 2018, Schnitzer & Blais, 2018, Kim, Lee, Kim & Kim, 2018). To avoid bias, we preferred not to use in our search more specific terms to describe: analytics, such as "machine learning," or health, such as "cancer." Thus, we queried the database with the following combination of keywords which could be identified in the title, abstract and/or keywords of any published item in order to download the maximum possible number of papers:

- "ANALYTICS" AND "HEALTH*" OR "BUSINESS INTELLIGENCE" AND "HEALTH*" OR "BIG DATA" AND "HEALTH*"
- "ANALYTICS" AND "MEDICAL" OR "BUSINESS INTELLIGENCE" AND "MEDICAL" OR "BIG DATA" AND "MEDICAL"
- ANALYTICS" AND "CLINICAL" OR "BUSINESS INTELLIGENCE" AND "CLINICAL" OR "BIG DATA" AND "CLINICAL"

Only articles and reviews written in English were included in the search, for capturing the full information about a specific study and in particular the results which are usually better presented in a full published article. As indicated in Figure 2, 6817 records were retrieved from the initial keyword search in the two databases. After duplicates exclusion, we ended up with 3241 papers. From the first screening, based on the content of the title and the abstract, we excluded 1364 out of the 3241 papers as they were not deemed relevant either to the health sector or the field of BDA. Nonetheless, we decided to include papers that tested, even with the use of quite a small sample, the capabilities of a proposed new technique or technology for collecting, storing or harnessing potentially big healthcare data. In line with this, we have also included papers that refer to biomarkers analysis and although a small sample of patients may be involved, they use a large number and a variety of biological parameters for testing, which overall lead to the creation of big datasets. Finally, in our dataset we have also included overviews and case study papers which are descriptive of the existence, benefits, and use of big data and their tools and technologies in the health sector. The broadness of our keywords and the variety of the subjects related to the health domain concluded to a dataset that incorporates papers from the areas of information technology, medical, biology, pharmacology and other disciplines.

After having completed the text screening of the 1877 articles, a further number of 1073 papers were excluded for the same reasons, leaving 804 articles in our final dataset which were submitted to content analysis using text analysis software (Zhang, Sun, & Xie, 2015; Mittelstadt & Floridi, 2016). Both authors assessed all abstracts and full-text independently and

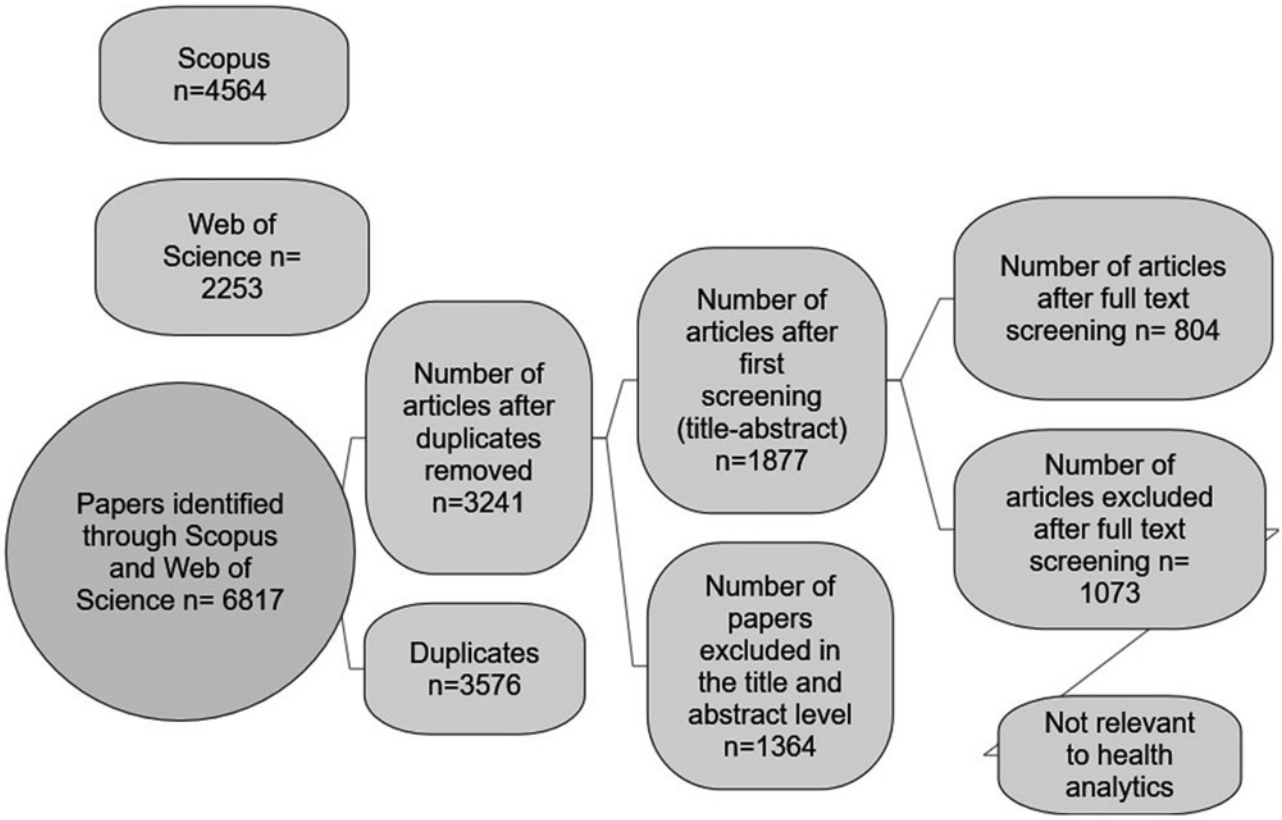

Figure 2. Dataset identification process.

Table 1. Classification Framework.

a/a	Classification Dimensions	Reference Source	Tables	Questions
1	Medical specialties	U.S. Government site of medicare	Table 7	What medical specialties have benefited the most?
2	Research approach	Wamba et al., 2015	Table 8	What type of research has been conducted?
3	Nature of analytics	Wang et al., 2016	Table 9	What level of analysis have they reached?
4	Types of data	Groves et al., 2013	Table 10	What kind of data has been used for the analysis?
5	Big data techniques	Wang et al., 2016; Waller & Fawcett, 2013	Table 11	What techniques have been used?
6	Big data capabilities	Groves et al., 2013	Table 12	What is the purpose of the analysis?

the results were compared. In cases of divergence, the authors discussed the paper's inclusion or exclusion.

The descriptive statistics of this dataset are presented in a number of tables and graphs with regards to the sources of publication, authors, affiliations, citations, and others. We further proceeded to a co-citation analysis of references that have been cited in the 804 papers of our study in order to capture the high impact publication activity of the broader field. We used the VosViewer co-citation analysis and visualisation software (Perianes-Rodriguez, Waltman, & Van Eck, 2016) to analyse the data retrieved from the ISI WoS and Scopus databases. Altogether, 26,998 references were cited in the 804 articles (33.5 references on average per article). From these references, we identified the articles and authors that were co-cited the most.

The stage II of our research indicates information extraction, topic tracking and categorisation. We applied text mining to derive information from the unstructured data of these papers (Dinov, 2016). To achieve this, the authors together built a text classification presented in Table 1. Every category of this classification contains several subcategories. The particular categories and their sub-categories, which act as the guide to the dataset content analysis, were inspired by a number of prominent review and overview papers relevant to BDA in general (Wamba, Akter, Edwards, Chopin, & Gnanzou, 2015 for the category of "research approach," Wang et al., 2016 for the category of nature of analytics, Groves, Kayyali, Knott, & Van Kuiken, 2013 for the categories of types of data & capabilities related to health, and finally, Chen & Zhang, 2014 and Waller & Fawcett, 2013 for the category of BDA techniques). Through the analysis of the selected categories, we made an attempt to answer questions like: What is the type and the frequency of the big data which have been used in the healthcare domain?; what big data are mostly used in health analytics techniques?; what is the level of analysis that has been reached (nature of analytics) and what type of research has been conducted (research approach)?;

what capabilities have been acquired from the application of BDA in the health sector? Also, a new category (medical specialties) was added to identify to whom this research is relevant. The selection of this classification attempts to map the knowledge in the field and explain the elements of BDA in healthcare through examples.

Continuing with the methodology, content analysis and text mining were performed by one of the authors through the use of NVivo10 software. After reading the full-text of each paper, the relevant section which signifies and explains its link to a sub-dimension was recognised/highlighted by the author and it was coded with the use of the software. From the NVivo menu, all relevant papers to a particular sub-dimension can be retrieved to bring up the highlighted information all at once. In many occasions a paper may fall in more than one subcategories of a certain variable.

Results

Part 1: Bibliometric analysis and descriptive results

Our review is not an exhaustive one, since the dataset derived from a sampling process. However, the presentation of descriptive statistics of this dataset could shed some light in the research that has been conducted in this area so far. We demonstrate the descriptive statistics of the publications in our dataset in several tables.

a. Years of publication, country of origin, source of publication, subject areas and authors' multidisciplinarity

Figure 3 and Tables 2–4 present the publication movement per year, per country, per source of publication and subject areas.

In Figure 3, it is notable that till 2008 there has been no or little publication related to BDA in healthcare. From 2009 until 2013 a bigger publication activity has begun and after 2014 we observe an explosion of articles. This phenomenal growth has also been mentioned in other reviews (Wang et al., 2016; Baro, Degoul, Beuscart & Chazard, 2015; Andreu-Perez, Poon, Merrifield, Wong, & Yang, 2015, etc.). Of course, the terms of our keywords search (business intelligence/analytics/big data) appeared in the literature after 2000 (Chen et al., 2012) and therefore a time-lag to the wider adoption of the terms from the academic community was anticipated.

Table 2. Top 10 most popular countries of authors' origin.

Num Publications	466	67	65	50	48	41	21	21	17	17
Country	US	CN	UK	AU	CA	DE	IN	KR	ES	IT

Table 3. Top 10 most popular journals in the dataset with publications in Health Analytics.

Journals	N	IF/2016
Journal of the American Medical Informatics Association	31	3.698
Journal of Biomedical Informatics	24	2.753
PLoS ONE	20	2.806
Big Data	19	1.239
Journal of Medical Systems	16	2.456
BMC Bioinformatics	14	2.448
Healthcare financial management: journal of the Healthcare Financial Management Association	11	0.000
Health Affairs	10	4.980
IEEE Journal of Biomedical and Health Informatics	10	3.451
Journal of Medical Internet Research	10	5.175
Indian Journal of Science and Technology	10	2.108

Table 4. Most popular subject areas.

SUBJECT AREAS	N
MEDICINE	293
TELECOMMUNICATIONS/COMPUTER SCIENCE	262
MEDICAL INFORMATICS	124
HEALTH CARE SCIENCES SERVICES	117
ENGINEERING	81
MATHEMATICAL COMPUTATIONAL BIOLOGY	62
GENETICS/BIOCHEMISTRY MOLECULAR BIOLOGY	53
BUSINESS ECONOMICS/MANAGEMENT	50
PHARMACOLOGY/PHARMACY	42
BIOTECHNOLOGY APPLIED MICROBIOLOGY/IMMUNOLOGY	40
SCIENCE/TECHNOLOGY/OTHER TOPICS	36
INFORMATION SCIENCE LIBRARY SCIENCE	33
CHEMISTRY	22
RESEARCH EXPERIMENTAL MEDICINE	20
PSYCHOLOGY/PSYCHIATRY	20
NURSING	19
SOCIAL SCIENCES/OTHER TOPICS	10
ENERGY/ENVIRONMENTAL SCIENCES/ECOLOGY	6
INSTRUMENTS/INSTRUMENTATION	5
EDUCATION/EDUCATIONAL RESEARCH	3
MEDICAL ETHICS	3

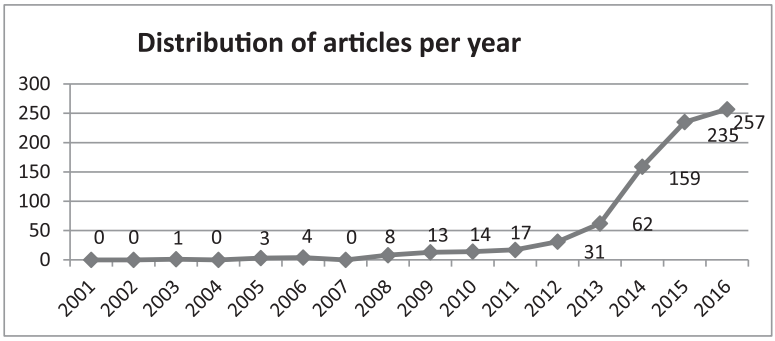

Figure 3. Distribution of articles per year.

The country with the greatest number of publications in our dataset (counting the number of authors affiliated with that country) is the USA with 466 published articles, followed by China (67) and the UK (65), as indicated in Table 2.

Overall, the 804 papers are spread in 460 journals. Table 3 shows the ten most popular journals that have published the higher number of articles from our dataset. The "Journal of the American Medical Informatics Association (JAMIA)" is the journal with the biggest number of publications (31), covering articles in the areas of clinical care, implementation science, imaging, education, consumer & public health and policy and holds an Impact Factor (IF) of 3.698, as of 2016. The next journal is the "Journal of Biomedical Informatics" with 24 articles and IF 2.753, which includes studies in the area of biomedical informatics and a special issue on a field related to health analytics "Methods of Clinical Research Informatics" in 2014. The "PLoS ONE," which is the world's first multidisciplinary Open Access journal, comes next with 20 publications. "Big Data" follows with 19 publications, and even though it has only been launched in March 2013, it has already 1.239 IF. In the review of Wang et al. (2016), about the trends of Big Data in social science, the "Big Data" journal was a popular publishing outlet as well. In our analysis, the next popular journal is the "Journal of Medical Systems" (16) with IF 2.456, an established journal in its field which has been published since 1977. The majority of the first ten journals are oriented mostly towards health and medical matters and less towards informatics or engineering. The first OR journal that appears in the list is the "Health Care Management Science" with five articles, followed by "Interfaces" with three articles and a few others each with one paper. In the future, the growth of BDA in healthcare may lead to the launch of more specialised journals of the field. Currently no OR journal has made it to the top 10.

Table 4 presents a description of the broad medical research/subject areas (as of Web of Science and Scopus), covered in the investigated papers. In the general area of "Medicine" all medical specialties, such as oncology, pathology, cardiology etc. are included. "Medicine" has gathered the majority of papers (293) followed by "Computer Science" (262) and Medical Informatics (124).

The study continues with the examination of authors' multi-disciplinarity. Due to the nature of the scientific field that involves different subject areas and scientists from different affiliations, we investigated the level of multi-disciplinarity of the co-authors in the articles of our dataset. Hence, we identified the authors' affiliated departments from the authors' list of each paper. We counted the specialty/discipline of every researcher that contributed to each article. Only in 11% (91) of the articles the authors come from three or more research areas, in 37.5% (302) of the articles authors come from two disciplines, with the 72% (216 papers) of these coming from medicine and informatics, and in 411 (51%) articles all authors come from only one discipline, including in this number the 51 single-authored papers. This discipline could be medical, informatics, business, engineering or other. In our dataset, there are 138 papers with two authors, and 64 articles with more than 10 authors. The remaining 551 articles are written by three to nin authors. The most populous article (Allen et al., 2016) includes 101 authors who form a panel of experts on the evaluation of the predicting performance of Alzheimer's disease by a computational crowdsourced project.

b. Citation and co-citation analysis based on bibliographic data

Table 5 presents the citation report of the 10 most cited articles of our dataset by also providing a short summary of their research. In the summary, an emphasis is given to the nature of the studies which is related to big data. This practice was also followed to all the examples provided in this study. The citation numbers have been retrieved from Google Scholar database in November 2018. The most cited paper out of a total number of 804 papers is Bates, Saria, Ohno-Machado, Shah, and Escobar (2014) with 356 total citations and 36.50 average citations per year. It currently holds the highest scores in both categories, and the number of total citations considering the young age of the paper is noticeable. The article presents six cases of high-risk patients as examples of opportunities to reduce costs through the use of big data. It discusses the insights emerging from clinical analytics (types of data, algorithms, registries, assessment scores, monitoring devices, and so forth) for the healthcare organisations, which can lead to better decision-making and the implementation of changes that will improve care while at the same time reduce costs. The second most highly cited article is O'Driscoll, Daugelaite, and Sleator (2013) with 319 citations and 21.33 average citations per year. It holds the second place in both score categories. This study provides an overview of big data technologies describing the example of the Apache Hadoop software and its current usage within the bioinformatics. The most co-cited paper (the most commonly referenced paper among the 804 articles in our dataset) is by

Table 5. Top 10 most cited articles (as of Nov 2018).

a/a	Articles	Summary	Total citations	Average cit. per year
1	Bates et al. (2014)	Presents six examples of high-cost patients and ways to reduce risk and costs through the use of big data. It discusses the types of data needed and the infrastructure (e.g. wear devices that monitor in real-time physiological parameters and remotely send data to clinicians; develop machine learning algorithms to learn from previous experience and optimise patient allocation to therapies, etc.).	356	36.60
2	O'Driscoll et al. (2013)	An overview of cloud computing and big data technologies handling large biology datasets, such as sequencing million human genomes to understand biological pathways and the genomic variation of a tumor.	319	21.33
3	Quinn et al. (2008)	Describes the evaluation of a smartphone diabetes management software which analyses, through its statistical model, users' logged data, trends, and behaviour and then, through its therapy optimisation tools, provides users real-time advice on diabetes for better managing their disease.	306	14.18
4	Chawla and Davis (2013)	An overview of the role of Big Data analytics and computation in personalised healthcare and biomedical discovery. Creates a personalised disease risk profile for an individual patient by leveraging the big data resident in electronic medical records, patients' experiences and histories, along with the biological information of diseases and their interactions. Delivers a personalised plan to an individual by leveraging similarities across a large group of patient pool, in real-time.	221	12.83
5	Andreu-Perez et al. (2015)	An overview of the progress in biomedical and health informatics through big data. It explains the benefits for medical, sensor and imaging informatics, and translational bioinformatics from piecing together different personalised information from unstructured and structured data, such as clinical diagnosis, imaging, continuous physiological sensing and genomics, proteomics, metabolomics.	220	28
6	Hilario, Kalousis, Pellegrini, and Müller (2006)	A study on data analytics that takes mass spectra data of biological specimens, like DNA microarray data and discovers patterns between different pathological states applying classification algorithms and reporting predictive performance. A mass spectrum contains thousands of different mass/charge ratios. The reduction of the size of the high dimensionality variable set through classification is crucial to biomarker discovery.	199	8.23
7	Zhang et al. (2014)	An approach for data anonymisation techniques of large-scale electronic health records that masks sensitive information specialising the level of information in a top-down manner until a minimum privacy requirement is compromised, making it possible to capture, manage, and process them within a tolerable elapsed time.	192	17.60
8	Castellanos, Di Martino, Craddock, Mehta, and Milham (2013)	Focuses on predictive modeling approaches for diagnosis through resting state fMRI, a method of functional magnetic resonance imaging that is used in brain mapping, a tool for brain-based biomarker identification for neurological and psychiatric illness. The convergence of dimensional approaches and large dataset of images processing and sharing is propitious for improving predictions.	190	21.17
9	Krumholz (2014)	Explores the ways in which big data, such as biological, clinical, behavioural, and outcomes data can be analysed through advanced methods to predict, discover, and compare effectiveness to tackle the complexity of patients, populations, and health-related organisations in a similar way that it is done in other businesses.	189	19.60
10	Alyass et al. (2015)	A review that discusses recent advances in high-throughput large omics (genomics, epigenomic, metagenomics, metabolomics, nutriomics, etc) technologies which have led to more precise modeling of complex diseases accelerating the global transition to personalised medicine. It also touches upon ethics and equity issues.	188	20.00

Murdoch and Detsky (2013), "The inevitable application of big data to health care," published in JAMA. It is included in the reference list of 24 out of 804 articles. It is closely followed by the papers of Breiman (2001) and Dean and Ghemawat (2008). Table 6 presents the 11 most co-cited papers. The majority of the papers are related to health with the exception of two: Breiman (2001), which is about a machine learning method often used in health applications and Dean and Ghemawat (2008) which discusses MapReduce; both techniques are often used in the health care sector. There is also the report from McKinsey (Manyika et al., 2011), which is about big data opportunities in general (with a special reference to the health sector in the US), and it is the only one not published in an academic journal. No overlaps are observed between the most cited (Table 5) and the most co-cited (Table 6) top 10 papers, except for two: Chawla and Davis (2013) and Bates et al. (2014) with the latter coming last in the list of the most co-cited papers as it shares the 10th position with Murphy et al. (2010). Overall, the great majority of the most cited and co-cited articles consists of overview and review papers.

Table 6. Top 10 most co-cited articles.

a/a	Citation	Summary	Frequency
1	Murdoch and Detsky (2013)	A viewpoint that discusses the applications and opportunities of big data (deriving from electronic health records) in health care, using an economic framework, to improve quality and efficiency of health care delivery.	24
2	Breiman (2001)	Gives insight into the random forests' capabilities for classification and prediction.	23
3	Dean and Ghemawat (2008)	Describes the function and success of MapReduce programming.	19
4	Raghupathi and Raghupathi (2014)	Describes the opportunities in healthcare from the analysis of big data, such as electronic health records, financial and operational data, clinical data, genomic data, real-time data from health monitoring devises.	19
5	Chawla and Davis (2013)	Look at Table 3 entry 4	16
6	Lazer et al. (2014)	Presents the Google Flu Trends, a flu tracking system from social media posts, as a case study to provide critical lessons for the future of big data analysis.	16
7	Jensen, Jensen, and Brunak (2012)	A review focusing on the potential knowledge discovery of genotype–phenotype relationship from integrating EHR data with genetic data and ethical, legal and technical reasons currently hindering the systematic deposition of these data in EHRs and their mining.	15
8	Manyika et al. (2011)	A research of McKinsey Global Institute about big data analytics in healthcare and other 4 domains focusing on the economic impact of the technology. For healthcare it provides examples of health insurance organisations deploying electronic health records, health monitoring data from devises, R&D data and mostly financial and pricing data.	15
9	Ginsberg et al. (2009)	Presents a method of analysing Google search queries to track influenza in a population. This approach may make it possible to use search queries to detect influenza epidemics in areas with a large population of web search users.	14
10	Murphy et al. (2010)	Presents a software, i2b2, that uses large datasets of patient medical record data, such as diagnoses, medications, and laboratory values and provides clinical investigators with the ability to identify sets of patients with special health characteristics while preserving patient privacy.	13
11	Bates et al. (2014)	Look at Table 3 entry 1	13

Part II: Content analysis results

In this section, we demonstrate the results and some indicative examples of papers based on the classification framework presented in the methodology section. The selection of the particular examples is based on their popularity (high-cited papers in their categories) and the fact that they provide a clear and comprehensive case for each sub-category as per authors' judgment.

Medical Specialties: The first categorisation concerns the allocation of papers to the medical specialties. Our groupings are based on the official U.S. government site for medicine.[1] Table 7 shows the relevant specialties and the number of identified papers with an indicative reference. In many cases articles are allocated in more than one subcategory. For example, an article about breast cancer was categorised in "Gynecology" and "Medical Oncology." The ten most popular specialties are presented according to the number of the allocated articles. The most popular subcategory is the one that includes articles referring to all kind of medical specialties or to "no specific specialty" (49%). The next popular subcategory with 77 articles (10%) combines three medical specialties (Neurology/Neuropsychiatry/Psychiatry) since many articles refer to these categories together. No similar research has been identified for comparison of results, however given that health analytics is a fast-growing field of research we expect to see in the near future a great number of studies on specific medical specialties. The specialties with the biggest number of studies, like neurology, oncology, and cardiology, are on the spotlight of the World Health Organization (WHO). Based on World Health Statistics 2016, mental disorders affect one in ten people on the planet and almost 40% of premature deaths are caused by cardiovascular diseases, cancer, diabetes, and chronic respiratory diseases. Therefore, the healthcare community is expected to give great importance to the evolution of these specialties with the help of the new capabilities offered by analytics.

Research approach: We also distributed the selected publications according to their research approach, as of Wamba et al. (2015) and Chen and Zhang (2014) with some additions. For the

diversification of the articles according to their research approach, some best examples of each approach are presented here. The results of the categorisation indicate that most papers are experimental studies, followed by review studies. In general, papers under the categorisation of "experiment" cover both a theoretical contribution (an advancement of an algorithm, experimentation with program running time, etc.) and a part where this theoretical advancement is tested or evaluated in relation to the under discussion application. In a field with an expected growth in the following years (according to the literature above), it is anticipated that the academia will provide at least this volume of experimental studies. Many articles have been categorised in more than one subcategories (e.g. the research of Barrett, Mondick, Narayan, Vijayakumar, & Vijayakumar, 2008 and Blakely

Table 7. Distribution of articles to Medical Specialties.

Specialties	N	%	Indicative Reference
No specific specialty	391	48.9	Althebyan et al., 2016
Neurology/ neuropsychiatry/ psychiatry	77	9.6	Maccione et al., 2015
Medical oncology	56	7.0	Miriovsky, Shulman, & Abernethy, 2012
Cardiology	54	6.7	Bardhan et al., 2015
Infectious disease	23	2.9	Carroll et al., 2014
Endocrinology	22	2.7	De Silva, Burstein, Jelinek, & Stranieri, 2015
Emergency medicine	22	2.7	Baum, 2010
Pediatric medicine	21	2.6	Basole et al., 2015
Radiology	19	2.4	Cook & Nagy, 2014
Pathology	19	2.4	Angelelli et al., 2014
Pulmonary disease	13	1.6	Kenner, 2016

Table 8. Research Approach and indicative research examples.

Research approach	N	%	Examples	Context
Experiment Incorporates papers that provide experimental results of new models	318	39.6	Barrett, Mondick, Narayan, Vijayakumar, & Vijayakumar, 2008	The study proposes mixed effect models and Bayesian forecasting algorithms to develop drug–specific dashboards for better decision-making and education of patient caregivers on clinical pharmacology principals which lead to fewer medication errors, reduced toxicity, reduced length of hospital stay, etc. Data visualisation tools summarise patient profiles from hospital electronic medical records of pediatric populations, such as lab values, vital signs, and associated biomarker and interface those data by a web-based decision support system.
Review & Overview Includes literature review papers that present a summary of the research methods and outcomes in a specific field	178	22.1	Gligorijević, Malod-Dognin, & Pržulj, 2016	The study reviewed recent big data integrative methods for disease sub typing, biomarkers discovery, and drug repurposing, and listed the tools that are available to domain scientists while highlighting key issues in the context of personalised medicine.
Data analysis Papers that contain methods and results from analysed data	146	18.1	Bello-Orgaz, Hernandez-Castro & Camacho, 2017	The study analysed large scale text related to vaccine opinions retrieved from Twitter for measuring the potential influence of these opinions based on the variation in the vaccination coverage rates. This method can be used to detect and locate communities against vaccination that could generate future disease outbreaks in different parts of the world.
Conceptual Studies that provide conceptual frameworks and general discussions on the investigated scientific fields	140	17.4	Kuiler, 2014	This study presents a conceptual framework for data analytics. An IT-supported ontology-based approach for health data to address the semantic challenges presented by big data sets and discusses architectural considerations. Future research will focus on developing the specifications for the lexicon, ontology, and other architectural artifacts to support software development.
Case study Qualitative research based on a case and designed to suit the research question	86	10.6	Chute, Beck, Fisk & Mohr, 2010	A case study about Mayo Clinic and its "semantically integrated warehouse of biomedical data." An information management initiative that integrates a huge amount of different medical data types.
Survey Studies that gathered and analysed questionnaires and/or participant opinions	22	2.7	Yildirim, Majnarić, Ekmekci & Holzinger, 2014	The authors dealt with the analysis of 1941 children clinical data, in a Health Center of Croatia, and interviewed their parents for more details of family history on antibiotics and other allergic and chronic diseases with the purpose of investigating reactions and allergy from antibiotics in children. Their analysis involved structure and unstructured data from a big population to present outcomes in biomedical research.

et al., 2015) as they include more than one approach in their methodology (Table 8).

Nature of Analytics: The following category allocates the articles according to their descriptive, predictive or prescriptive nature as of Wang et al., (2016). In many cases, we distributed the articles in more than one sub-categories because there is evidence of more than one dimension in some papers. Our most popular subcategory with 47% (377 papers) is that of "Predictive analytics." The second subcategory in our classification is "Prescriptive Analytics," with 33% of papers (263 out of 804) and the last one is "Descriptive analytics" (24% with 190 articles). In the paper of Raghupathi and Raghupathi (2013) descriptive analytics were found to be the most commonly used type due to their explanatory and easy approached nature. However, in healthcare, prediction is more valuable than explanation because the outcomes are measured in lives (Agarwal & Dhar, 2014). The majority of articles in our dataset are published after 2013, and therefore later than the publication of Raghupathi and Raghupathi (2013). Healthcare is a growing sector and as a result advanced technology and skills are needed for the application of models with predictive or prescriptive character. The industry may have a time-lag in the adoption of the more advanced nature of analytics, but the research must pave the way (Groves et al., 2013). In our article pool, the majority of papers (40%) included experiments of new models with the hope that these predictive/prescriptive models will become part of a software and will be adopted by analysts for use in the decision-making in healthcare organisations or systems (Table 9).

Types of data: In our research, we adopted the detailed description of the types of primary data related to healthcare from the study of McKinsey and Co (Groves et al., 2013). These include: A. Clinical data, B. Patient and sentiment data, C. Administration and cost activity data, and D. Pharmaceutical and R&D data. The review papers did not take part in this classification. In Table 10 together with the allocation of papers according to the type of the analysed data, we also provide in the second column a definition of these types of data. Overall, the adopted types are in line with the categorisation used by other researchers too (Gaitanou, Garoufallou & Balatsoukas, 2014). The most popular data that have been analysed in the articles of our dataset are "clinical data" with a 70% (562 articles out of 804) representation. Our results are consistent with the literature which has identified that significant research has been focused on EHRs implementation, but relatively few studies exploited other types of big data (Gaitanou et al., 2014).

Big Data Techniques: While investigating extensively the literature, we realised that the boundaries of BDA techniques are difficult to be completely distinguished (Royston, 2013). For better understanding the use of the different techniques, we present a definition in the second column and some indicative examples in the last. The listed BDA techniques have derived from the literature (Chen & Zhang, 2014; Waller & Fawcett, 2013) and although some may overlap with each other or consist a subcategory of another, they are as inclusive as possible. For example, in Table 11, "web-mining" is presented separately from "data-mining" although it can be

Table 9. Nature of Analytics and indicative research examples.

Nature of analytics	N	%	Examples	Context
Predictive Involve the use of mathematical algorithms to discover predictive patterns within data and project what will happen in the future.	377	46.9	Bardhan, Zheng & Kirksey 2015	Presented a novel model to predict readmission of patients with congestive heart failure. The model tracks patient demographic, clinical, and administrative data across 67 hospitals in North Texas over a 4-year period.
Prescriptive Involve the use of data and mathematical algorithms to determine decisions that involve objectives with the aim to improve performance.	263	32.7	Sir et al., 2015	Surveyed 2865 patients from the surgery unit and 3241 from the oncology unit and proposed nurse–patient assignment models to achieve a balanced assignment workload. Patient metrics used from QuadraMed AcuityPlus patient classification system (which accumulates hospital's patient indicators over 20 years) to classify patients on nurses' workload.
Descriptive Techniques such as online analytical processing (OLAP) that aim to identify problems and trends in existing processes and functions.	190	23.6	Basole et al., 2015	Presented a visual analytic tool that used clinical data from 5784 pediatric asthma emergency department patients and reported that asthma is the most common pediatric chronic disease and is the third leading cause of hospitalisation among children, affecting 9.3% of children in the US. Their results assist in the improvement of health care quality. The data were obtained from Population Discovery, Children's data warehouse. This included patient and provider information, administrative events, clinical observations, medications, laboratory tests, and charges in a relational database.

Table 10. Types of data and indicative research examples.

Types of data	N	%	Example	Context
Clinical data Patient data such as EHR and medical images	562	69.9	Forsberg et al., 2015	Collected biomarker and clinical information from 73 patients who sustained 116 life threatening combat wound by conflicts in Afghanistan and Iraq, and tried to determine if those data could be used to predict the likelihood of wound failure. The collected data included clinical information, serum, wound effluent, and tissue and their analysis model indicated that it would improve clinical outcomes and reduce unnecessary surgical procedures. The same approach was also tested and performed equally well with larger samples of patients (67,486 patients with traumatic extremity wounds).
Patient behaviour and sentiment data Data collected from wearable sensors and social sites	133	16.5	Boulos et al., 2010	Described the analysis of predictive tools that gather posts and queries from Social Web ("Web 2.0") tools such as blogs, micro-blogging and social networking sites to form coherent representations of real-time health events like flu out-breaks. Harvested data in the form of human feelings from a large number of blogs and social pages such as those hosted by MySpace.
Administrative & cost data Financial and operational data and patient profiling data and choices	59	7.3	Abbas, Bilal, Zhang & Khan, 2015	Used a vast number of individuals' administrative and clinical data to create a cloud-based solution (Software as a Service) that provides personalised recommendations about the health insurance plans according to the user specified criteria.
Pharmaceutical R&D data Drug therapeutic mechanisms, R& D data from target behaviour in the body, such as effects of toxicity etc.	38	4.7	Calabrese, Minkoff & Kristine, 2014	Described "Pharmachosychrony" as a new concept of analytical pharmacy solutions to improved care coordination and provided a high quality and patient-centric model of care. Among the data that this solution elaborates, pharmacy data are included for the effective and safe use of medication. Elaborated claims data from call centers, web portals, mobile technology, and decentralised clinical staff.

seen as a subcategory of the latter, acknowledging the fact that this mining field is represented by a quite large number of papers and has gathered momentum because of the high usage of internet data in the very last decades. The case with the allocation of papers to the "modeling" and "simulation" techniques is also similar. The criterion for allocating a paper to the "modeling" subcategory was whether the modeling technique mainly included mathematical formulations of variable relationships presented in a static form, and that for the allocation to the "simulation" subcategory was whether the data variability was addressed by running the model many times with different values taken from a distribution. Willing to address both approaches and present indicative research examples to explain them, we separated the techniques. Moreover, in the statistics subcategory, the majority of articles are allocated to another technique too, and overall many of the papers have multiple entries as the handling of big data requires a combination of techniques for their analysis. As seen in Table 11, "modeling" emerges as the most popular technique, as it is also the most general amongst the categories. It is followed by "machine learning" (which includes the design of algorithms), a fast-growing technique with lots of successful cases in the field of health, such as the classification of medical data and symptoms for disease diagnosis and prediction (Chen, Hao, Hwang, Wang & Wang, 2017; Khalaf et al., 2017).

Big Data Capabilities: The term "big data capabilities" refers to the different organisational competencies created by IT models that analyse vast amounts of complex and different types of data, processed in daily operations (Bharadwaj, 2000) and resulting to better decision-making. But, on which big data capabilities healthcare should focus, in order to achieve its goals? To assist managers in better decision-making, organisations must develop infrastructure with essential big data capabilities (Groves et al., 2013). Groves et al. (2013) acknowledged five important BDA capabilities, which have been adopted in this study. These are: (a) "monitoring," which includes articles that present monitoring efficiencies, and collect and analyse data (using analytical methods) describing "what is happening now," (b) "prediction/simulation," which includes articles that present methods that provide information about future outcomes (what will happen), (c) "data mining," which incorporates articles that involve methods enabling extraction and categorisation of knowledge (what happened), (d) "evaluation," which includes articles that demonstrate methods for testing the performance of BDA techniques or explain the outcomes of the application of BDA (why did it happen?), and (e) "reporting capability," which includes articles with methods for organising the collected data in an informative format.

Due to the plethora of capabilities described in the papers, many of them have been allocated to

Table 11. Big Data Techniques and indicative research examples.

Techniques	N	%	Examples	Context
Modeling Methods of analytical mathematical analysis with approximate relationships between variables (Waller & Fawcett, 2013)	344	42.8	Ajorlou, Shams & Yang, 2015	Developed a linear predictive Bayesian model indicating that risk adjustment for patient health conditions can improve the prediction power. Data from 82,000 patients from 888 facilities assembled for a total capture period of 1 year and assessed from the Veteran Health Administration.
Machine learning Artificial intelligence aimed to design algorithms that allow computers to evolve behaviours based on empirical data. (Chen & Zhang, 2014)	327	40.7	Dugan et al., 2015	Experimented with six different machine learning methods to identify the best one for predicting future obesity in children above 2 years old with 85% accuracy. Data collected from a pediatric clinical decision support system (CHICA) and used for the analysis. The data included 9 years of clinical information collected from 4 different community health centers.
Data mining A set of techniques to extract information from data (Chen & Zhang, 2014)	200	24.9	Delen, 2009	Used three popular data mining techniques (decision trees, artificial neural networks and support vector machines) to develop prediction models for prostate cancer survivability. The researchers obtained around 120,000 records from the Surveillance, Epidemiology, and End Results Program and formed 77 variables for statistical analysis. They concluded that data mining methods are capable of extracting patterns and relationships but are useless without medical experts' feedback.
Visualisation approaches The techniques used to create tables, images, diagrams and other intuitive display ways to understand data (Chen & Zhang, 2014)	153	19	Angelelli et al., 2014	Presented a visualisation tool "brain atlas" with cohort data analysis of 100+ participants. The tool, which was assessed by neuropsychological testing, genetic analysis and multimodal magnetic-resonance (MR) imaging, enables a first quick analysis of the identified hypotheses.
Statistics The methods of organising and interpreting data for exploiting causal relationships between different objectives (Chen & Zhang, 2014)	132	16.4	Demir, 2014	Proposed a method to compare predictive analytic capabilities of emergency readmissions. Using data from the emergency department from 963 patients with chronic obstructive pulmonary disease and asthma within 45 days after a patient has been discharged from hospital. This data set was divided into derivation and validation samples 1000 times. They actually proved that predictive logistic regression and regression trees could be a valuable decision support tool for clinicians for the prediction of readmissions.
Simulation Quantitative analysis of a system in a stochastic setting (Waller & Fawcett, 2013)	55	6.8	Liu & Wu, 2014	Developed an agent-based simulation model to study accountable care organisations. It identified the critical determinants for the payment model design that can motivate provider behaviour changes to achieve maximum financial and quality outcomes that considers payers, healthcare providers, and patients as agents under the shared saving payment model of care. It constructed a healthcare system analytics model that can help inform health policy and healthcare management decisions.
Web mining The process of information discovery from sources across the World Wide Web (Cooley, Mobasher, & Srivastava, 1997)	54	6.7	Chen & Kotecha, 2014	Developed an analytics platform, called "Cytobank," for community cytometry data analysis (to track cells and subsets in blood and tissue) using large computing resources for analysis on the Internet. These platforms can simultaneously measure up to 100 parameters.
Optimisation methods Methods that find the minimum or maximum of a function, subject to constraints and solve quantitative problems, improve the accuracy of forecasting and algorithms (Waller & Fawcett, 2013)	49	6.1	Katircioglu et al., 2014	IBM Research developed a scenario modeling and analysis tool, supply chain scenario modeler (SCSM), for McKesson (the largest healthcare services company) to optimise its pharmaceutical supply chain policies. SCSM optimises the distribution network, supply flow and inventory policies and quantifies the impacts of changes on financial, operational, and environmental metrics. They developed complex queries to generate all input needs and rigorously tested them. The resulting data model has over 200 tables with a combined size of tens of millions of records.
Text mining Techniques based on machine learning and data mining to find useful patterns in text data (Holzinger & Jurisica, 2014)	42	5.2	Holzinger & Jurisica, 2014	Presented an overview of some selected text mining methods, i.e. Latent Semantic Analysis, and Probabilistic Latent Semantic Analysis along with examples from the biomedical domain by extracting data from texts (unstructured patient data and, structured patient data e.g. biometrics or laboratory results), and biomedical images, which will benefit clinical decision support. It provided machine learning solutions for large and complex biomedical data analysis.
	22	2.7	Toerper et al., 2016	

(*continued*)

Table 11. Continued.

Techniques	N	%	Examples	Context
Forecasting Is about predicting the future, while also evaluates what could happen under different circumstances using predictive analytic methods (Waller & Fawcett, 2013)				Developed and evaluated a web-based forecast tool that predicts the daily bed need for admissions from the cardiac catheterisation laboratory. The forecast model was derived using a 13-month retrospective cohort of 7029 catheterisation patients and included predictor variables such as demographics, scheduled procedures, and clinical indicators mined from free-text notes.
Social Network Analysis A technique that views and analyses data from social networks	20	2.5	Abbas et al., 2016	Proposed a cloud-based framework for BDA in health that uses the Internet and social media. The framework offers users disease risk assessment and consultation service from health experts on Twitter with high accuracy results. It utilises collective data of people's health status from whole populations.

Table 12. Big Data Capabilities and indicative research examples.

Big data capabilities	N	%	Examples	Context
Monitoring What is happening now?	264	32.8	Althebyan et al., 2016	Proposed an e-healthcare monitoring system that targets a crowd of individuals in a wide geographical area that integrates emerging technologies such as mobile computing, wearable sensors, cloud computing etc. to offer remote monitoring of patients anytime and anywhere. The monitoring BDA capability provided through this system can enhance the decision support system in order to reduce risk of patient health decisions.
Prediction /simulation What will happen?	258	32	Abdelrahman, Zhang, Bray & Kawamoto, 2014	Proposed a new analytical approach to develop high-performing predictive models for congestive heart failure (CHF) readmission using an operational dataset with incomplete records and changing data over time. Data came from 2,787 CHF hospitalisations at University of Utah Health Care Center from January 2003 to June 2013.
Data mining What did it happen?	230	28.6	Chen et al., 2016	Developed a bootstrapping method for global module detection on features across breast cancer cohorts. They used electronic medical records' data from a Medical Center annotated with BioCarta signaling signatures and provided new insights into breast cancer, such as the association of patient's cultural background with preferences for surgical procedure. The modeling tool demonstrated unique ability to discover clinically meaningful and actionable knowledge across highly heterogeneous biomedical big data sets.
Evaluation Why did it happen?	105	13	Catlin et al., 2015	Proposed a web-based analytics system for conducting in house evaluations and comparisons of "infusion pump data" across hospital systems allowing users to select any number and combination of hospital data. Smart pump infusions are customisable libraries with dose limits and administration rates specific to medications and care areas and provide information in order to avoid medication errors like the delivery of wrong drugs or delivery to the wrong patient or assessing the wrong dose.
Reporting What happens on a regular basis	72	8.9	Curcin, Woodcock, Poots, Majeed, & Bell, 2014	Presented a software—Web Improvement Support in Healthcare (WISH)—which is a prototype tool that attempts to translate research into practice using local improvement projects. This approach facilitates electronic data collection and reporting in health settings and is tested on a Chronic Obstructive Pulmonary Disease improvement project run in Northwest London Hospitals. Data are gathered from a large class of tasks, particularly local ones that cannot be adequately measured by exclusively using routinely collected data residing in hospital's EHRs.

more than one subcategory. We also want to clarify here that the difference between "data mining capability" and "data mining technique" (Table 11) is that the latter applies algorithms and mathematical modeling to perform clustering, etc. while the term *capability* refers to the process of applying these methods with the purpose of uncovering hidden patterns in large datasets. There is also a link between the BDA capabilities as described by Groves et al. (2013) and the nature of analytics as described by Wang et al. (2016). For example, we can identify that the prediction/simulation capability is connected to the predictive and prescriptive nature of analytics respectively, and the monitoring

and reporting capabilities are associated with the descriptive nature of analytics. However, the capabilities focus on the IT functionalities and the nature of analytics focuses on the technique's goal.

Table 12 provides examples of articles for each subcategory for further clarifying each capability. The most popular subcategory is "monitoring" with 33% of articles (264 papers out of 804), which shows the importance of the use of analytics methods to assist managers to maintain a view of "what is happening now." The next dimension for the distributed articles is "prediction and simulation" with 32%. Having already demonstrated that in our systematic review predictive analytics is the most exploited type used in the examined articles, it can be justified that a good percentage of all the papers would provide information about predictive BDA capabilities. The next popular sub-dimension is "data mining" with 29% (230 articles), followed by the "evaluation" with 13% (105), reflecting the publishing activity on evaluating the applications of BDA in healthcare. In this subcategory, articles that produce methods to evaluate the performance of other applications are often encountered. Finally, the last capability in Table 12 is "reporting" with a percentage of 9% (72 articles). A closer look at the allocation of articles reveals that the majority of the papers are almost equally distributed in the three first BDA capabilities (a) monitoring (33%), (b) prediction/simulation (32%), and (c) "data mining" (29%). However, in the reality of the health sector, reporting, and monitoring activities are already in effect but predictive modeling and simulation techniques have not been used at scale yet (Groves et al., 2013).

Discussion and conclusions

The descriptive characteristics of the 804 articles which have been reviewed in this study reveal an explosion of publications in the field of health analytics the last years. Medicine and computer sciences are the most common subject areas and there is some multi-disciplinarity amongst authors' backgrounds in less than half of the examined papers. The "Journal of the American Medical Informatics Association" has published a good number of papers in the field. Currently the most cited paper is about predicting and managing high-risk and high-cost patients (Bates et al., 2014), published in Health Affairs and the most co-cited among the 804 articles is about the application of big data in health care published in Jama (Murdoch & Detsky, 2013). Both are overview papers. Most of the examined papers follow an "experimental" approach. Many of the papers in the article pool deal with the medical specialities of neurology, medical oncology and cardiology. Machine learning is a popular BDA technique that researchers use in these studies. Almost half of the papers are predictive in nature. One third of the studies develop monitoring capabilities and 70% of studies use clinical data.

The indicative examples provided for each subcategory aim to improve the comprehension of the categorisation and to further increase the general understanding of the type of research conducted in the field of health analytics.

Summarising our examples, indicative is the research in the field of population health management in terms of (a) disease surveillance by determining disease outbreaks (mostly using social media and web analytics) and ensuring speedy response and needs in new vaccines (Lazer, Kennedy, King, & Vespignani, 2014; Boulos, Sanfilippo, Corley, & Wheeler, 2010; Ginsberg et al., 2009), and mostly (b) (chronic) disease management by prediction of disease by patient profiling, in terms of symptoms, lab results, medical images, and patient history details, for individuals' accurate health diagnosis, by applying advanced analysis (such as segmentation and predictive modeling, machine learning, visualisation, etc.; Krumholz, 2014; Delen, 2009). Likewise, medical staff, through decision support systems, can identify the most fitting treatment and medication for each patient to avoid possible complications (Barrett, Humblet, Hiatt, & Adler 2013) and identify patients with high health risk profiles to offer proactive care options such as screening, brief interventions, etc. (Dugan, Mukhopadhyay, Carroll, & Downs, 2015, Bates et al., 2014) so as to avoid hospitalisation or readmissions (Bardah et al., 2015; Demir, 2014; Bardhan, Oh, Zheng, & Kirksey, 2015). On a similar thematic area, research is concentrated on offering customised e-healthcare solutions, mainly at home, but in hospitals as well, by constantly monitoring and analysing inbound clinical data from wearables and sensors and alerting health specialists about negative trends for conditions that need attention and possible hospitalisation (Althebyan, Yaseen, Jararweh, & Al-Ayyoub, 2016; Baum, 2010). Another very important type of research that emerges from the indicative examples is with regards to the use of big data technologies handling large biology datasets, such as sequencing human genomes to understand biological pathways and the genomic variation of, for example, tumor, which have led to personalised medicine, meaning offering different therapeutic schemes based on a patient's biomarkers (Alyass, Turcotte, & Meyre, 2015; Chawla & Davis, 2013; O'Driscoll, 213).

A different type of research is about the provision of cloud services and mobile software that

accumulates specific disease-based knowledge deriving from the collection of a vast amount of data from its targeted users and offers them "personalised" consultation for better disease management (Abbas, Ali, Khan, & Khan, 2016; Quinn et al., 2008). From the organisational perspective, there are studies that focus on the improvement of health services processes and cost-reduction by evaluating performance of resources, monitoring workload and human error, understanding clinical and other processes and identifying bottlenecks in care quality with the use of modeling and simulation (Sir, Dundar, Steege, & Pasupathy, 2015; Catlin et al., 2015). A similar type of research also focuses on more customised products/services, such as health insurance plan offerings based on the modeling and forecasting of a vast number of individuals' clinical and administrative data (Abbas, Bilal, Zhang, & Khan 2015).

In the reviewed papers, a number of issues for future directions emerged. Since the volumes of health data will grow globally in an intense rate, the demand for Information Technology (IT) infrastructure will consequently increase (Abbas, Bilal, Zhang & Khan, 2015). Another field for future research is the assurance of data privacy and cybersecurity which will enable healthcare organisations and researchers to safely manage/exploit the big health datasets for further value creation (Zhang, Yang, Liu, & Chen, 2014). Further research is also desirable towards enabling clinical decision-making in real-time, based on the patients' individual characteristics, where large groups of patient data can be pooled from across institutions so that each patient and their clinicians can find "patients like me" to help with real-time clinical decision-making (Broughman & Chen, 2016).

This study aims to verify the status of published research on BDA in Healthcare and create a descriptive classification. We believe that this research adds to the existing body of knowledge and provides a more thorough analysis of the field with the use of content analysis, through easily comprehensible information based on examples. It also attempts to continue the effort of other researchers (Waller & Fawcett, 2013) to explore the possibilities of big data and OR.

This research is multifaceted as it deals with different health issues (different diseases or quality of care), it examines them from a different perspective (for monitoring, reporting, prediction, etc.) and with the use of different types of data (clinical, administrative, pharmaceutical, etc.). Overall, considering the distribution of papers per medical specialty it is noticeable that BDA have a crucial role to play in the research of the most severe diseases that humanity faces nowadays (cancer, Alzheimer, diabetes, etc.) and reveal the importance of innovative technological solutions to unanswered medical questions. Researchers from different disciplines (medicine, information technology, operational researchers, business administrators, etc.) collaborate to gather and actually use the vast amount of data that cannot be managed from commonly implemented technology. New modeling and machine learning methods explore new capabilities and reveal hidden information. Since OR professionals are in the front line of offering improved decision-making via innovative modeling tools, this study provides them with the big picture of the BDA research that has been conducted in the health sector. The presented overview aims to answer questions like, when (chronologically), where (country and publishing journals) by whom (popular authors, subject areas) and what (medical specialties, research approach, nature of analytics) research is conducted in health BDA, with what means this is achieved (BDA data types, techniques) and through what competencies for health-related organisations and information analysts (BDA capabilities).

Future research under this agenda could investigate the benefits and the values created by BDA in healthcare and could focus on the new tools that are used for the analysis of the vast amount of data in the domain along with issues that restrict its extensive use. Therefore, future research could involve broadening the categories of our literature review with more technical content (tools and applications) or by identifying issues, benefits and detailed future perspectives of BDA techniques and capabilities. Lastly, the presentation of a more detailed co-citation analysis could shed more light to the broader body of literature.

However, our research comes with limitations. We realised that the boundaries between BDA and data analytics as well as the boundaries between techniques and other subcategories that we have created are not always easily discernible and therefore a small fragment of our article selection or the categorisation may be debatable based on the reader's point of view. Furthermore, we attempted to explore a sample of the related literature; by no means is this an exhaustive literature review of the field.

There seems to be a deficiency of studies relevant to population and public health compared to these related to medicine. Although both keywords "health" and "medicine" were used in our sampling method and our results demonstrate that papers are distributed to a variety of different disciplines (Table 4), it is true that the majority of the derived papers are more related to medical advancements and clinical decision support. This may imply that

more research has been conducted towards this direction. Although indicative examples are also well presented herein and especially under "data types analysis" (Table 10), future research could solely focus on examining the use of BDA in population and public health for shedding more light in its progress and potentiality.

Furthermore, our classification system is not exhaustive, and it could be expanded to include further categories and subcategories. In our case, we have only included the categories that could clarify certain questions in the area of BDA in healthcare and specifically with regard to what medical specialties have benefited the most, what kind of data are used for the analysis, what techniques are used, what is the purpose of the analysis (capabilities) and what is the level of analysis that has been reached (predictive, etc.).

Notes

1. https://www.medicare.gov/physiciancompare/staticpages/resources/specialtydefinitions.html?AspxAutoDetectCookieSupport = 1

Disclosure Statement

No potential conflict of interest was reported by the author(s).

References

Abbas, A., Ali, M., Khan, M. U. S., & Khan, S. U. (2016). Personalized healthcare cloud services for disease risk assessment and wellness management using social media. *Pervasive and Mobile Computing*, *28*, 81–99. doi:10.1016/j.pmcj.2015.10.014

Abbas, A., Bilal, K., Zhang, L., & Khan, S. U. (2015). A cloud based health insurance plan recommendation system: A user centered approach. *Future Generation Computer Systems*, *43*, 99–109.

AbdelRahman, S. E., Zhang, M., Bray, B. E., & Kawamoto, K. (2014). A three-step approach for the derivation and validation of high-performing predictive models using an operational dataset: Congestive heart failure readmission case study. *BMC Medical Informatics and Decision Making*, *14*(1), 41.

Agarwal, R., & Dhar, V. (2014). Editorial—big data, data science, and analytics: The opportunity and challenge for IS research. *Information Systems Research*, *25*(3), 443–448. doi:10.1287/isre.2014.0546

Ajorlou, S., Shams, I., & Yang, K. (2015). An analytics approach to designing patient centered medical homes. *Health Care Management Science*, *18*(1), 3–18. doi:10.1007/s10729-014-9287-x

Akter, S., & Wamba, S. F. (2016). Big data analytics in E-commerce: A systematic review and agenda for future research. *Electronic Markets*, *26*(2), 173–194. doi:10.1007/s12525-016-0219-0

Allen, G. I., Amoroso, N., Anghel, C., Balagurusamy, V., Bare, C. J., Beaton, D., … Caberlotto, L. (2016). Crowdsourced estimation of cognitive decline and resilience in Alzheimer's disease. *Alzheimer's & Dementia*, *12*(6), 645–653.

Althebyan, Q., Yaseen, Q., Jararweh, Y., & Al-Ayyoub, M. (2016). Cloud support for large scale e-healthcare systems. *Annals of Telecommunications*, *71*(9–10), 503–515. doi:10.1007/s12243-016-0496-9

Alyass, A., Turcotte, M., & Meyre, D. (2015). From big data analysis to personalized medicine for all: Challenges and opportunities. *BMC Medical Genomics*, *8*(1), 33doi:10.1186/s12920-015-0108-y

Andreu-Perez, J., Poon, C. C., Merrifield, R. D., Wong, S. T., & Yang, G. Z. (2015). Big data for health. *IEEE Journal of Biomedical and Health Informatics*, *19*(4), 1193–1208. doi:10.1109/JBHI.2015.2450362

Angelelli, P., Oeltze, S., Turkay, C., Haasz, J., Hodneland, E., Lundervold, A., … Hauser, H. (2014). Interactive visual analysis of heterogeneous cohort study data. *IEEE Computer Graphics and Applications, 34*, 70–82.

Bardhan, I., Oh, J. H., Zheng, Z., & Kirksey, K. (2015). Predictive analytics for readmission of patients with congestive heart failure. *Information Systems Research*, *26*(1), 19–39. doi:10.1287/isre.2014.0553

Baro, E., Degoul, S., Beuscart, R., & Chazard, E. (2015). Toward a literature-driven definition of big data in healthcare. *BioMed Research International*, *2015*, 1. doi:10.1155/2015/639021

Barrett, M. A., Humblet, O., Hiatt, R. A., & Adler, N. E. (2013). Big data and disease prevention: from quantified self to quantified communities. *Big data*, *1*(3), 168–175.

Barrett, J. S., Mondick, J. T., Narayan, M., Vijayakumar, K., & Vijayakumar, S. (2008). Integration of modeling and simulation into hospital-based decision support systems guiding pediatric pharmacotherapy. *BMC Medical Informatics and Decision Making*, *8*(1), 6.

Basole, R. C., Braunstein, M. L., Kumar, V., Park, H., Kahng, M., Chau, D. H. (P.)., … Thompson, M. (2015). Understanding variations in pediatric asthma care processes in the emergency department using visual analytics. *Journal of the American Medical Informatics Association*, *22*(2), 318–323.

Bates, D. W., Saria, S., Ohno-Machado, L., Shah, A., & Escobar, G. (2014). Big data in health care: Using analytics to identify and manage high-risk and high-cost patients. *Health Affairs*, *33*(7), 1123–1131. doi:10.1377/hlthaff.2014.0041

Baum, D. (2010). An intelligent patient focus. Cambridge Memorial Hospital is increasing efficiency and improving patient care with a new emergency room tracking board and business-intelligence system. *Health Management Technology*, *31*(4), 12–14.

Bello-Orgaz, G., Hernandez-Castro, J., & Camacho, D. (2017). Detecting discussion communities on vaccination in twitter. *Future Generation Computer Systems*, *66*, 125–136. doi:10.1016/j.future.2016.06.032

Bharadwaj, A. S. (2000). A resource-based perspective on information technology capability and firm performance: An empirical investigation. *MIS Quarterly*, *24*(1), 169–196. doi:10.2307/3250983

Blakely, T., Atkinson, J., Kvizhinadze, G., Nghiem, N., McLeod, H., Davies, A., & Wilson, N. (2015). Updated New Zealand health system cost estimates from health events by sex, age and proximity to death: further improvements in the age of 'big data'. *New Zealand Medical Journal*, *128*(1422), 13–23.

Boulos, M. N. K., Sanfilippo, A. P., Corley, C. D., & Wheeler, S. (2010). Social Web mining and exploitation

for serious applications: Technosocial Predictive Analytics and related technologies for public health, environmental and national security surveillance. *Computer Methods and Programs in Biomedicine*, *100*(1), 16–23. doi:10.1016/j.cmpb.2010.02.007

Brandeau, M. L., Sainfort, F., & Pierskalla, W. P. (Eds.). (2004). *Operations research and health care: A handbook of methods and applications* (Vol. 70). New York: Springer.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32. doi:10.1023/A:1010933404324

Broughman, J. R., & Chen, R. C. (2016). Using big data for quality assessment in oncology. *Journal of Comparative Effectiveness Research*, *5*(3), 309–319. doi:10.2217/cer-2015-0021

Calabrese, D., Minkoff, N. B., & Kristine, R. (2014). Pharmacosynchrony: Road map to transformation in pharmacy benefit management. *The American Journal of Pharmacy Benefits*, *6*, 76–80.

Carroll, L. N., Au, A. P., Detwiler, L. T., Fu, T. C., 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. doi:10.1016/j.jbi.2014.04.006

Castellanos, F. X., Di Martino, A., Craddock, R. C., Mehta, A. D., & Milham, M. P. (2013). Clinical applications of the functional connectome. *Neuroimage*, *80*, 527–540. doi:10.1016/j.neuroimage.2013.04.083

Catlin, A. C., Malloy, W. X., Arthur, K. J., Gaston, C., Young, J., Fernando, S., & Fernando, R. (2015). Comparative analytics of infusion pump data across multiple hospital systems. *American Journal of Health-System Pharmacy*, *72*(4), 317–324. doi:10.2146/ajhp140424

Chae, B., & Olson, D. L. (2013). Business analytics for supply chain: A dynamic-capabilities framework. *International Journal of Information Technology & Decision Making*, *12*(1), 9–26. doi:10.1142/S0219622013500016

Chawla, N. V., & Davis, D. A. (2013). Bringing big data to personalized healthcare: A patient-centered framework. *Journal of General Internal Medicine*, *28*(S3), 660–665. doi:10.1007/s11606-013-2455-8

Chen, C. P., & Zhang, C. Y. (2014). Data-intensive applications, challenges, techniques and technologies: A survey on Big Data. *Information Sciences*, *275*, 314–347. doi:10.1016/j.ins.2014.01.015

Chen, H., Chen, W., Liu, C., Zhang, L., Su, J., & Zhou, X. (2016). Relational network for knowledge discovery through heterogeneous biomedical and clinical features. *Scientific Reports*, *6*, 29915.

Chen, H., Chiang, R. H., & Storey, V. C. (2012). Business intelligence and analytics: From big data to big impact. *MIS Quarterly*, *36*, 1165–1188.

Chen, M., Hao, Y., Hwang, K., Wang, L., & Wang, L. (2017). Disease prediction by machine learning over big data from healthcare communities. *IEEE Access*, *5*, 8869–8879. doi:10.1109/ACCESS.2017.2694446

Chen, T. J., & Kotecha, N. (2014). Cytobank: Providing an analytics platform for community cytometry data analysis and collaboration. In *High-dimensional single cell analysis* (pp. 127–157). Berlin: Springer.

Chuah, M. H., & Wong, K. L. (2011). A review of business intelligence and its maturity models. *African Journal of Business Management*, *5*(9), 3424–3428.

Chute, C. G., Beck, S. A., Fisk, T. B., & Mohr, D. N. (2010). The Enterprise Data Trust at Mayo Clinic: A semantically integrated warehouse of biomedical data. *Journal of the American Medical Informatics Association*, *17*(2), 131–135. doi:10.1136/jamia.2009.002691

Cook, T. S., & Nagy, P. (2014). Business intelligence for the radiologist: Making your data work for you. *Journal of the American College of Radiology: JACR*, *11*(12 Pt B), 1238–1240. doi:10.1016/j.jacr.2014.09.008

Cooley, R., Mobasher, B., & Srivastava, J. (1997). Web Mining: Information and Pattern Discovery on the World Wide Web. In *ictai* (Vol. 97, pp. 558–567).

Cosic, R., Shanks, G., & Maynard, S. (2012). Towards a business analytics capability maturity model. In *ACIS 2012: Location, Location, Location: Proceedings of the 23rd Australasian Conference on Information Systems 2012* (pp. 1–11). ACIS.

Curcin, V., Woodcock, T., Poots, A. J., Majeed, A., & Bell, D. (2014). Model-driven approach to data collection and reporting for quality improvement. *Journal of Biomedical Informatics*, *52*, 151–162. doi:10.1016/j.jbi.2014.04.014

Davenport, T. H. (2006). Competing on analytics. *Harvard Business Review*, *84*(1), 98.

Dean, J., & Ghemawat, S. (2008). MapReduce: Simplified data processing on large clusters. *Communications of the ACM*, *51*(1, 107–113. doi:10.1145/1327452.1327492

Delen, D. (2009). Analysis of cancer data: A data mining approach. *Expert Systems*, *26*(1), 100–112. doi:10.1111/j.1468-0394.2008.00480.x

Demir, E. (2014). A decision support tool for predicting patients at risk of readmission: A comparison of classification trees, logistic regression, generalized additive models, and multivariate adaptive regression splines. *Decision Sciences*, *45*(5), 849–880. doi:10.1111/deci.12094

De Silva, D., Burstein, F., Jelinek, H. F., & Stranieri, A. (2015). Addressing the complexities of big data analytics in healthcare: The diabetes screening case. *Australasian Journal of Information Systems*, 19.

Dinov, I. D. (2016). Methodological challenges and analytic opportunities for modeling and interpreting Big Healthcare Data. *Gigascience*, *5*(1), 12.

Duan, L., & Xiong, Y. (2015). Big data analytics and business analytics. *Journal of Management Analytics*, *2*(1), 1–21. doi:10.1080/23270012.2015.1020891

Dugan, T. M., Mukhopadhyay, S., Carroll, A., & Downs, S. (2015). Machine learning techniques for prediction of early childhood obesity. *Applied Clinical Informatics*, *6*(3), 506–520. doi:10.4338/ACI-2015-03-RA-0036

Feldman, B., Martin, E. M., & Skotnes, T. (2012). Big data in healthcare hype and hope. *Dr. Bonnie*, *360*, 122–125.

Ferguson, M. (2012). Architecting a big data platform for analytics. A Whitepaper prepared for IBM, 30.

Forsberg, J. A., Potter, B. K., Wagner, M. B., Vickers, A., Dente, C. J., Kirk, A. D., & Elster, E. A. (2015). Lessons of war: Turning data into decisions. *EBioMedicine*, *2*(9), 1235–1242. doi:10.1016/j.ebiom.2015.07.022

Gaitanou, P., Garoufallou, E., & Balatsoukas, P. (2014). The effectiveness of big data in health care: A systematic review. In *Research conference on metadata and semantics research* (pp. 141–153). Cham: Springer.

Ginsberg, J., Mohebbi, M. H., Patel, R. S., Brammer, L., Smolinski, M. S., & Brilliant, L. (2009). Detecting influenza epidemics using search engine query data. *Nature*, *457*(7232), 1012. doi:10.1038/nature07634

Gligorijević, V., Malod-Dognin, N., & Pržulj, N. (2016). Integrative methods for analyzing big data in precision medicine. *Proteomics*, *16*(5), 741–758. doi:10.1002/pmic.201500396

Gorman, M. F., & Klimberg, R. K. (2014). Benchmarking academic programs in business analytics. *Interfaces*, *44*(3), 329–341. doi:10.1287/inte.2014.0739

Groves, P., Kayyali, B., Knott, D., & Van Kuiken, S. (2013). The 'big data' revolution in healthcare. *McKinsey Quarterly*, *2*(3).

Hazen, B. T., Skipper, J. B., Boone, C. A., & Hill, R. R. (2018). Back in business: Operations research in support of big data analytics for operations and supply chain management. *Annals of Operations Research*, *270*(1–2), 201–211. doi:10.1007/s10479-016-2226-0

Hilario, M., Kalousis, A., Pellegrini, C., & Müller, M. (2006). Processing and classification of protein mass spectra. *Mass Spectrometry Reviews*, *25*(3), 409–449. doi:10.1002/mas.20072

Holzinger, A., & Jurisica, I. (2014). Knowledge discovery and data mining in biomedical informatics: The future is in integrative, interactive machine learning solutions. In *Interactive knowledge discovery and data mining in biomedical informatics* (pp. 1–18). Berlin: Springer.

Horowitz, G. L., Zaman, Z., Blanckaert, N. J., Chan, D. W., Dubois, J. A., Golaz, O., … Marocchi, A. (2005). Modular analytics: A new approach to automation in the clinical laboratory. *Journal of Automated Methods & Management in Chemistry*, *1*, 8–25. doi:10.1155/JAMMC.2005.8

Huang, M., Han, H., Wang, H., Li, L., Zhang, Y., & Bhatti, U. A. (2018). A clinical decision support framework for heterogeneous data sources. *IEEE Journal of Biomedical and Health Informatics*, *22*(6), 1824–1833. doi:10.1109/JBHI.2018.2846626

Iqbal, U., Hsu, C.-K., Nguyen, P. A. (A.)., Clinciu, D. L., Lu, R., Syed-Abdul, S., … Li, Y.-C. (J.). (2016). Cancer-disease associations: A visualization and animation through medical big data. *Computer Methods and Programs in Biomedicine*, *127*, 44–51. doi:10.1016/j.cmpb.2016.01.009

Ivan, M., & Velicanu, M. (2015). Healthcare industry improvement with business intelligence. *Informatica Economica*, *19*(4/2015), 81. doi:10.12948/issn14531305/19.2.2015.08

Jensen, P. B., Jensen, L. J., & Brunak, S. (2012). Mining electronic health records: Towards better research applications and clinical care. *Nature Reviews Genetics*, *13*(6), 395. doi:10.1038/nrg3208

Jun, J. B., Jacobson, S. H., & Swisher, J. R. (1999). Application of discrete-event simulation in health care clinics: A survey. *Journal of the Operational Research Society*, *50*(2), 109–123. doi:10.2307/3010560

Katircioglu, K., Gooby, R., Helander, M., Drissi, Y., Chowdhary, P., Johnson, M., & Yonezawa, T. (2014). Supply chain scenario modeler: A holistic executive decision support solution. *Interfaces*, *44*(1), 85–104. doi:10.1287/inte.2013.0725

Katsaliaki, K., & Mustafee, N. (2011). Applications of simulation within the healthcare context. *Journal of the Operational Research Society*, *62*(8), 1431–1451. doi:10.1057/jors.2010.20

Kenner, A. (2016). Asthma on the move: How mobile apps remediate risk for disease management. *Health, Risk & Society*, *17*(7–8), 510–529. doi:10.1080/13698575.2015.1136408

Khalaf, M., Hussain, A. J., Keight, R., Al-Jumeily, D., Fergus, P., Keenan, R., & Tso, P. (2017). Machine learning approaches to the application of disease modifying therapy for sickle cell using classification models. *Neurocomputing*, *228*, 154–164. doi:10.1016/j.neucom.2016.10.043

Kim, H., Lee, C. H., Kim, S. H., & Kim, Y. D. (2018). Epidemiology of complex regional pain syndrome in Korea: An electronic population health data study. *PLoS One*, *13*(6), e0198147. doi:10.1371/journal.pone.0198147

Krumholz, H. M. (2014). Big data and new knowledge in medicine: The thinking, training, and tools needed for a learning health system. *Health Affairs*, *33*(7), 1163–1170. doi:10.1377/hlthaff.2014.0053

Kuiler, E. W. (2014). From big data to knowledge: An ontological approach to big data analytics. *Review of Policy Research*, *31*(4), 311–318. doi:10.1111/ropr.12077

Lazer, D., Kennedy, R., King, G., & Vespignani, A. (2014). Big data. The parable of Google Flu: traps in big data analysis. *Science (New York, N.Y.)*, *343*(6176), 1203–1205. doi:10.1126/science.1248506

Liu, P., & Wu, S. (2016). An agentbased simulation model to study accountable care organizations. *Health care management science*, *19*(1), 89–101.

Maccione, A., Gandolfo, M., Zordan, S., Amin, H., Di Marco, S., Nieus, T., … Berdondini, L. (2015). Microelectronics, bioinformatics and neurocomputation for massive neuronal recordings in brain circuits with large scale multielectrode array probes. *Brain Research Bulletin*, *119*, 118–126. doi:10.1016/j.brainresbull.2015.07.008

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., & Byers, A. H. (2011). *Big data: The next frontier for innovation, competition, and productivity*. Seattle: McKinsey & Company.

Miriovsky, B. J., Shulman, L. N., & Abernethy, A. P. (2012). Importance of health information technology, electronic health records, and continuously aggregating data to comparative effectiveness research and learning health care. *Journal of Clinical Oncology*, *30*(34), 4243–4248. doi:10.1200/JCO.2012.42.8011

Mittelstadt, B. D., & Floridi, L. (2016). The ethics of big data: Current and foreseeable issues in biomedical contexts. *Science and Engineering Ethics*, *22*(2), 303–341. doi:10.1007/s11948-015-9652-2

Murdoch, T. B., & Detsky, A. S. (2013). The inevitable application of big data to health care. *JAMA*, *309*(13), 1351–1352. doi:10.1001/jama.2013.393

Murphy, S. N., Weber, G., Mendis, M., Gainer, V., Chueh, H. C., Churchill, S., & Kohane, I. (2010). Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2). *Journal of the American Medical Informatics Association*, *17*(2), 124–130. doi:10.1136/jamia.2009.000893

Nie, P., & Li, B. (2011). A cluster-based data aggregation architecture in WSN for structural health monitoring. In *Wireless Communications and Mobile Computing Conference (IWCMC), 2011 7th International* (pp. 546–552). IEEE.

O'Connell, M. (2012). Big Data Analytics: Scaling Up and Out in the Event-Enabled Enterprise. Wall Street Technology Association, Ticker, (3).

O'Driscoll, A., Daugelaite, J., & Sleator, R. D. (2013). 'Big data', Hadoop and cloud computing in genomics. *Journal of Biomedical Informatics*, *46*(5), 774–781. doi:10.1016/j.jbi.2013.07.001

Peng, Y., Shi, J., Fantinato, M., & Chen, J. (2017). A study on the author collaboration network in big data. *Information Systems Frontiers*, *19*(6), 1329–1342. doi:10.1007/s10796-017-9771-1

Perianes-Rodriguez, A., Waltman, L., & Van Eck, N. J. (2016). Constructing bibliometric networks: A comparison between full and fractional counting. *Journal of Informetrics*, *10*(4), 1178–1195. doi:10.1016/j.joi.2016.10.006

Quinn, C. C., Clough, S. S., Minor, J. M., Lender, D., Okafor, M. C., & Gruber-Baldini, A. (2008). WellDoc™ mobile diabetes management randomized controlled trial: Change in clinical and behavioral outcomes and patient and physician satisfaction. *Diabetes Technology & Therapeutics*, *10*(3), 160–168. doi:10.1089/dia.2008.0283

Raghupathi, W., & Raghupathi, V. (2013). An overview of health analytics. *Journal of Health Medical Information*, *4*, 132.

Raghupathi, W., & Raghupathi, V. (2014). Big data analytics in healthcare: Promise and potential. *Health Information Science and Systems*, *2*(1), 3.

Royston, G. (2013). Operational Research for the Real World: Big questions from a small island. *Journal of the Operational Research Society*, *64*(6), 793–804. doi:10.1057/jors.2012.188

Schnitzer, M. E., & Blais, L. (2018). Methods for the assessment of selection bias in drug safety during pregnancy studies using electronic medical data. *Pharmacology Research & Perspectives*, *6*(5), e00426. doi:10.1002/prp2.426

Sir, M. Y., Dundar, B., Steege, L. M. B., & Pasupathy, K. S. (2015). Nurse–patient assignment models considering patient acuity metrics and nurses' perceived workload. *Journal of Biomedical Informatics*, *55*, 237–248. doi:10.1016/j.jbi.2015.04.005

Thouin, M. F., Hoffman, J. J., & Ford, E. W. (2008). The effect of information technology investment on firm-level performance in the health care industry. *Health Care Management Review*, *33*(1), 60–68. doi:10.1097/01.HMR.0000304491.03147.06

Toerper, M. F., Flanagan, E., Siddiqui, S., Appelbaum, J., Kasper, E. K., & Levin, S. (2016). Cardiac catheterization laboratory inpatient forecast tool: A prospective evaluation. *Journal of the American Medical Informatics Association*, *23*(e1), e49–e57. doi:10.1093/jamia/ocv124

Turaga, D. S. (2018). Introduction to the interfaces special issue: Applications of analytics and operations research in big data analysis. *Interfaces*, *48*(2), 93–93. doi:10.1287/inte.2018.0946

Waller, M. A., & Fawcett, S. E. (2013). Data science, predictive analytics, and big data: A revolution that will transform supply chain design and management. *Journal of Business Logistics*, *34*(2), 77–84. doi:10.1111/jbl.12010

Wamba, S. F., Akter, S., Edwards, A., Chopin, G., & Gnanzou, D. (2015). How 'big data'can make big impact: Findings from a systematic review and a longitudinal case study. *International Journal of Production Economics*, *165*, 234–246. doi:10.1016/j.ijpe.2014.12.031

Wamba, S. F., Anand, A., & Carter, L. (2013). A literature review of RFID-enabled healthcare applications and issues. *International Journal of Information Management*, *33*(5), 875–891. doi:10.1016/j.ijinfomgt.2013.07.005

Wang, G., Gunasekaran, A., Ngai, E. W., & Papadopoulos, T. (2016). Big data analytics in logistics and supply chain management: Certain investigations for research and applications. *International Journal of Production Economics*, *176*, 98–110. doi:10.1016/j.ijpe.2016.03.014

Ward, M. J., Marsolo, K. A., & Froehle, C. M. (2014). Applications of business analytics in healthcare. *Business Horizons*, *57*(5), 571–582. doi:10.1016/j.bushor.2014.06.003

West, V. L., Borland, D., & Hammond, W. E. (2014). Innovative information visualization of electronic health record data: A systematic review. *Journal of the American Medical Informatics Association*, *22*(2), 330–339.

Wills, M. J. (2014). Decisions through data: Analytics in healthcare. *Journal of Healthcare Management*, *59*(4), 254–262. doi:10.1097/00115514-201407000-00005

Yildirim, P., Majnarić, L., Ekmekci, O. I., & Holzinger, A. (2014). Knowledge discovery of drug data on the example of adverse reaction prediction. *BMC Bioinformatics*, *15*(6), S7.

Zhang, X., Yang, L. T., Liu, C., & Chen, J. (2014). A scalable two-phase top-down specialization approach for data anonymization using mapreduce on cloud. *IEEE Transactions on Parallel and Distributed Systems*, *25*(2), 363–373.

Zhang, Y., Sun, Y., & Xie, B. (2015). Quality of health information for consumers on the web: A systematic review of indicators, criteria, tools, and evaluation results. *Journal of the Association for Information Science and Technology*, *66*(10), 2071–2084. doi:10.1002/asi.23311