International Journal of Information Management 50 (2020) 206–216

Contents lists available at ScienceDirect

International Journal of Information Management

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

Big data analytics in health sector: Theoretical framework, techniques and prospects

Panagiota Galetsi[a], Korina Katsaliaki[a], Sameer Kumar[b,*]

[a] School of Economics, Business Administration & Legal Studies, International Hellenic University, 14th km Thessaloniki-N. Moudania, Thessaloniki, 57001, Greece
[b] Opus College of Business, University of St. Thomas Minneapolis Campus, 1000 LaSalle Avenue, Schulze Hall 435, Minneapolis, MN 55403, USA

ARTICLE INFO

Keywords:
Big data analytics
Health-Medicine
Decision-making
Machine learning
Operations research (OR) techniques

ABSTRACT

Clinicians, healthcare providers-suppliers, policy makers and patients are experiencing exciting opportunities in light of new information deriving from the analysis of big data sets, a capability that has emerged in the last decades. Due to the rapid increase of publications in the healthcare industry, we have conducted a structured review regarding healthcare big data analytics. With reference to the resource-based view theory we focus on how big data resources are utilised to create organization values/capabilities, and through content analysis of the selected publications we discuss: the classification of big data types related to healthcare, the associate analysis techniques, the created value for stakeholders, the platforms and tools for handling big health data and future aspects in the field. We present a number of pragmatic examples to show how the advances in healthcare were made possible. We believe that the findings of this review are stimulating and provide valuable information to practitioners, policy makers and researchers while presenting them with certain paths for future research.

1. Introduction

The healthcare industry is data intensive and could use interactive dynamic big data platforms with innovative technologies and tools to advance patient care and services (Ali, Shrestha, Soar, & Wamba, 2018; Carvalho, Rocha, Vasconcelos, & Abreu, 2019). The healthcare industry manages a wide amount of data every day from clinical and operational information systems, such as Electronic Health Records (EHR) (Brooks, El-Gayar, & Sarnikar, 2015) and Laboratory Information Library Systems (LIMS) (Groves, Kayyali, Knott, & Van Kuiken, 2013). Practitioners are developing new applications in order to assist healthcare stakeholders to increase opportunities for a greater value (Groves et al., 2013).

Business Analytics include the techniques, technologies, systems, practices, methodologies, and applications for the analysis of the vast amount of data and help organizations better understand its business, market, and make timely decisions (Chen, Chiang, & Storey, 2012; De Camargo Fiorini, Seles, Jabbour, Mariano, & de Sousa Jabbour, 2018; Srinivasan & Swink, 2018; Wamba et al., 2017). Big Data Analytics (BDA) in healthcare involve the methods of analysing the wide amount of electronic data related to patient healthcare and well-being. This data is so diverse and difficult to measure by traditional software or hardware. There are various forms of health data such as, clinical and lab data, medical notes, machine generated data from medical equipment or from at home monitoring sensors, health services financial data, hospital bills, literature data from medical journals, social media posts blogs in health subjects, etc. These data may be available internally in health services (e.g. EHR, LIMS) or come from external sources (e.g. insurance companies, pharmacies, government) and could be in a structured format (e.g. tables with laboratory results) or unstructured (e.g. text of medical notes in EHR) (Raghupathi & Raghupathi, 2013).

To illustrate data volume magnitude, the health data explosion from 500 petabytes in 2012 (Feldman, Martin, & Skotnes, 2016) will reach 163 zetabytes in 2025[1] . Big data are recognised by four characteristics, the so called 4Vs: 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) and veracity, which refers to reliable data (Gandomi & Haider, 2015). The methods of BDA refer to techniques such as forecasting, optimization, simulation, and others which assist decision-making and provide insights to managers and policy-makers (Doumpos & Zopounidis, 2016; Duan, Edwards, & Dwivedi, 2019).

Computer practitioners constantly develop new applications to help

* Corresponding author.
E-mail addresses: p.galetsi@ihu.edu.gr (P. Galetsi), k.katsaliaki@ihu.edu.gr (K. Katsaliaki), skumar@stthomas.edu (S. Kumar).

[1] Andrew Cave, "What Will We Do When The World's Data Hits 163 Zettabytes In 2025?," https://www.forbes.com/sites/andrewcave/2017/04/13/what-will-we-do-when-the-worlds-data-hits-163-zettabytes-in-2025/#38fa3aae349a (19/01/2019).

https://doi.org/10.1016/j.ijinfomgt.2019.05.003
Received 3 March 2019; Received in revised form 23 April 2019; Accepted 8 May 2019
Available online 14 June 2019
0268-4012/ © 2019 Elsevier Ltd. All rights reserved.

healthcare stakeholders increase opportunities for greater value. Organizations also develop infrastructure with big data capabilities to help improve manager decision-making (Groves et al., 2013). It is said that 80% of the growth of information and communication technology will be about cloud services, big data analytics, mobile technology and social media technologies (Andreu-Perez, Poon, Merrifield, Wong, & Yang, 2015).

Several studies have contributed in different ways to the understanding of BDA in healthcare. Baro, Degoul, Beuscart, and Chazard (2015) and Wamba, Akter, Edwards, Chopin, and Gnanzou (2015) are literature reviews that discuss the meaning of big data in healthcare. The studies of Raghupathi and Raghupathi (2013) and Ward, Marsolo, and Froehle (2014) provide a general overview through the analysis of examples in the health analytics area, concentrating in certain aspects of the field. The study of Zhang and Li (2017) reviewed literature for a specialized healthcare domain, in HIV self-management. Wang, Kung, Wang, and Cegielski (2018) identified the relationships among big data analytics capabilities, IT-enabled transformation practices and benefits, using the healthcare sector as their case study. Jacofsky (2017), in his overview, raised concerns to physicians about the pitfalls of analytics reports from large metadata sets in healthcare.

In this article, we conduct a systematic literature review study to map the scientific field. Our theoretical framework draws upon the resource-based theory and aims to identify the created organisational values along with a special interest in the used data, the applied analysis techniques and the information technology innovations. We believe that there is a need for a deeper analysis of the "state of the art status in the subject field" in order to connect the technological accomplishments of BDA in healthcare with the achieved values and the call for future work. We believe that this paper contributes to the global literature because it attempts to classify basic analytic terms followed by representative results and examples. As a result of an increasing interest in health analytics, the synthesis of the current literature, through a theoretical framework and the presentation of its outcomes, is beneficial to researchers and the industry itself.

2. Research framework

"Resources", such as data and IT infrastructure solutions and "activities", such as big data analysis, are described as the essential mechanisms that contribute to the value creation of organizations (Lim et al., 2018). It is important though for an organization to recognize and understand the factors of data-based value creation to gain competitive advantage and to provide better services. The resource-based view states that a firm, by acquiring valuable resources and synthesizing them appropriately, can create unique values/capabilities that provide their competitive advantage (Barney, 1991). This is the most commonly used organizational theory to big data research (see, De Camargo Fiorini et al., 2018; Gunasekaran et al., 2017). The data gathered from IT infrastructure is reported as an important organizational resource for gaining competitive advantage (Jaklič, Grublješič, & Popovič, 2018; Mamonov & Triantoro, 2018). Success in business analytics depends on the firm's ability to simultaneously utilize multiple resources (including data) and capabilities within a business context, and make decisions to deliver a valued output (Dubey, Gunasekaran, Childe, Roubaud et al., 2019; Dubey, Gunasekaran, Childe, Fosso Wamba et al., 2019; Srinivasan & Swink, 2018; Vidgen, Shaw, & Grant, 2017).

The last decades, medical scientists rely more and more on automation and cooperate with IT specialists for the creation of new software solutions to manage the vast amount of patient and other related data. Therefore, the healthcare sector is an appropriate application of the resource-based view theory for examining the value chain created from the analysis of the vast amount of data. In the case of the healthcare industry, data comes from clinical and operational information systems. Scientists use this data to address healthcare problems (reduced budgets, demand for faster turnaround times, etc.) and to gain value from better decision making. Data resources in healthcare (Table 1), such as clinical, patient, pharmaceutical data, etc., must be appropriately processed and analyzed in order to create capabilities translated into business values, which are going to be thoroughly discussed in Section 4.3 (and Table 3). Their analysis is based on OR techniques, such as modeling, simulation, machine learning, visualization, data mining and others (Chen & Zhang, 2014; Yaqoob et al., 2016) (Table 2). These techniques develop models which are fed with raw big data and to cope with their volume and their processing time utilize computing applications, such as Apache Hadoop. These applications allow the distributed processing of large data sets across clusters using simple programming models. The effective use of data analytics tools or models can reach organizations' "agility" only when there is continuous cooperation of various bundled resources (Ghasemaghaei, Hassanein, & Turel, 2017). These models are useful for interested parties to offer solutions for observed problems based on quantifiable

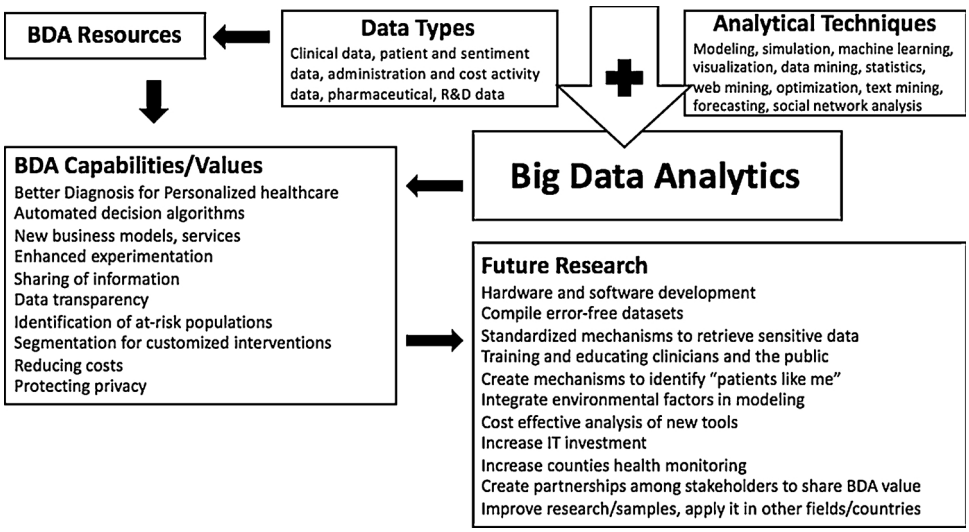

Fig. 1. Schematic Research Framework.

measures and propose alternatives which can lead to improved performance (Katsaliaki, Mustafee, & Kumar, 2014). This study aims to identify the use of the big data resources and their analysis techniques and examine the capabilities and values that are created for the healthcare industry (Wamba, Anand, & Carter, 2013). These values lead to the need of further developments and therefore future research is essential in terms of technological and organizational improvements that big data analytics will bring in healthcare. According to the above description, we summarize the research framework of BDA in health in Fig. 1.

3. Methodology

For the literature review process, a methodology comprised of two steps have been applied following the systematic processes of Dubey, Gunasekaran, and Papadopoulos (2017) and Tranfield, Denyer, and Smart (2003). Fig. 2 presents a descriptive scheme of the research steps.

The first step refers to the collection of relevant articles. To identify them, we ran searches in two well-known electronic databases, Web of Science® and Scopus, for the years 2000–2016 inclusive, based on specific terms. As search keywords, we used the combination of the terms a) "business intelligence" b) "analytics" and c) "big data", which have been used to describe the BDA field in other studies too (Chen et al., 2012; Duan & Xiong, 2015) and appear in the literature in the late 1990s beginning of 2000s. Given that our goal was to expand the research in the healthcare industry, we also used additionally the keywords: "health*" and its derivatives (healthcare, health industry, health records, health datasets, etc.), "medical" (medical records, medical data, etc.) and "clinical" (Liberatore & Nydick, 2008).

The articles were selected based on the following inclusion-exclusion criteria agreed by all authors. The dataset is comprised only of:

(1) "articles" and "reviews"
(2) studies written in the English language
(3) studies relevant to the health sector
(4) studies relevant to big data analytics

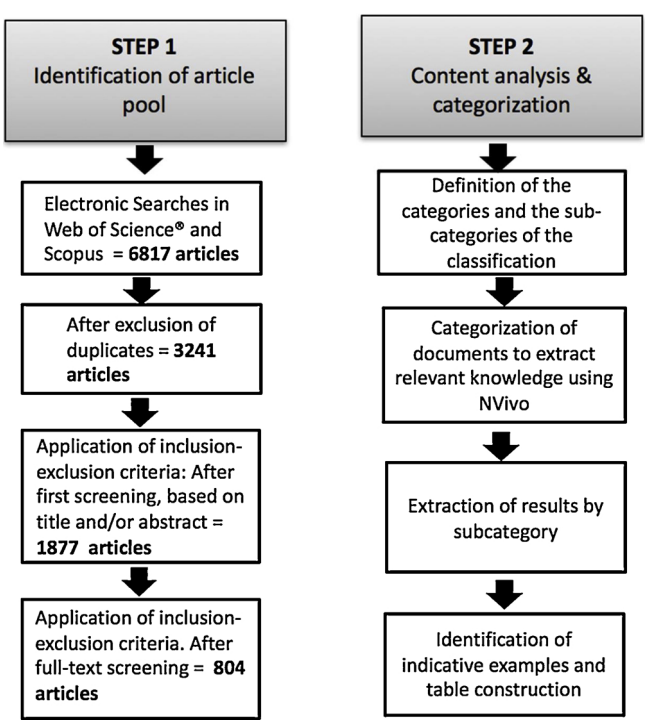

Fig. 2. Scheme of review methodology.

The search results numbered 6817 articles and after excluding duplicates, were reduced to 3241 papers. We conducted a first screening, based on the title and/or abstract and we excluded 1364 papers. From the full-text screening of the 1877 papers we excluded another 1073 papers, as they were found to be irrelevant to the health industry and the big data analytics field. We ended with a final dataset of 804 articles. All authors assessed all abstracts and full-text screening independently discussing cases of disagreement upon the inclusion or exclusion criteria.

The second step of the methodology refers to defining the categories and subcategories for the classification, content analysis and categorization of documents to the specific subcategories. Some of the selected categories and subcategories were based on the existing literature and were enhanced with additional groups from the knowledge generated by reading the articles in our dataset. We must acknowledge here that many of the 804 studies, during the allocation process, were categorized in more than one subcategory. The second step also refers to the outputs of the classification process. We present tables with the classifications, articles' frequency per dimension and indicative research examples. All authors were involved in the identification of the categories and the construction of tables and one author was responsible for classifying the articles under the subcategories with the use of the Nvivo software and with advice from the other authors when needed.

We strongly believe that the particular sample of this research is large enough to be considered as representative of the health analytics field and therefore the presented results can enhance researchers' and practitioners' knowledge.

4. Research context

4.1. Data types

The healthcare data resources which are used in the collection of our papers are categorized under five groups as follows: 1. Clinical, 2. Patient and Sentiment 3. Administration and Cost Activity, 4. Pharmaceutical and R&D and 5. Data derived from Databases. The first 4 categories are adopted from the literature (Gaitanou, Garoufallou, & Balatsoukas, 2014; Groves et al., 2013) and the last (Databases) was our own addition which refers to data retrieved from bibliographic databases, such as Medline, PubMed and databases from public health systems such as WHO, OECD and FDA.

Table 1 presents a definition of these types of data and the allocation of the 804 papers based on the data type which was analyzed in each study. Of course, there are papers that have analyzed more than one data type. The frequencies are followed by a short description of one or two studies which are indicative of each data category and/or highly cited. With these provisions (description and example), we hope to improve the comprehension of the different data types and how these are used in BDA studies in the healthcare sector. It is not surprising that the majority of studies manipulate clinical data with almost 70% representation in our dataset (562 papers out of 804). In particular, electronic health records (EHRs) seem to be the most common input in models of BDA (Gaitanou et al., 2014).

Table 1
Data Types related to health.

Data Type Definition	% N	Indicative Research Example
Clinical data: Patient data such as EHR & medical images	69.9% (562)	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 hospitalization among children under the age of 15 years, affecting 9.3% of children in the US. Their results assist in the improvement of health care quality (Basole et al., 2015)
Patient behavior and sentiment data: Data collected from wearable sensors and social sites	16.5% (133)	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 (Boulos, Sanfilippo, Corley, & Wheeler, 2010) Described a wearable-technology, called "fitness MOOC" that tracks physical activity of humans in order to gather knowledge and promote healthy ageing. As part of an R&D – interaction of seniors with wearable fitness trackers (Buchem, Merceron, Kreutel, Haesner, & Steinert, 2014)
Administrative and cost data: Data that describe costs, bills, reimbursement categories and other patient characteristics	7.3% (59)	Used a vast number of individuals' administrative and clinical data to create a cloud based solution (Software as a Service) that provides personalized recommendations about the health insurance plans according to the user specified criteria (Abbas, Bilal, Zhang, & Khan, 2015)
Pharmaceutical R&D data: Drug therapeutic mechanisms, R&D data from target behavior in the body such as effects of toxicity etc.	4.7% (38)	Described "Pharmachosychrony" as a new concept of analytical pharmacy solutions to improve care coordination. Pharmacy data are elaborated for the effective and safe use of medication (Calabrese, Minkoff, & Rawlings, 2014)

4.2. Big data techniques

In this section, we present the OR techniques which were used in the studies for the analysis of the data types to assist decision-making. The created categories are adopted from the literature (Chen & Zhang, 2014; Waller & Fawcett, 2013). The groups are not mutually exclusive, and overlaps are apparent. For example, *text-mining* and *web-mining* can be considered as special cases of *data-mining*, however we refer to them separately as they indicate a hype and present a more analytic view of the mining techniques. On the other side is the *modeling* technique which is a generic term and can incorporate many of the other methods. Articles were distributed to the *modeling* category if mathematical formulations of variable relationships were described in a static form, whereas the *simulation* category incorporated data variability when

Table 2
Big Data Techniques.

Techniques	% N	Indicative Research Example
Modeling: Methods of fast and cost-effective mathematical analysis with approximate relationships between variables (Waller & Fawcett, 2013)	42.8% (344)	Developed a linear predictive Bayesian model to designing patient centered medical homes indicating that risk adjustment for patient health conditions can improve the prediction power. Data for this study were assessed from the Veteran Health Administration (Ajorlou, Shams, & Yang, 2015)
Machine Learning: Artificial intelligence aimed to design algorithms that allow computers to evolve behaviors based on empirical data. (Chen & Zhang, 2014)	40.7% (327)	Presented a software application where the user can input the appropriate information such as the symptoms and get the diagnosis of the disease and the drug related to that treatment. This application collects, stores, and analyses massive amounts of indicative data and provides knowledge at little cost (Mohan, Vigneshwaran, Vineeth Raj, & Harlin Jesuva Prince, 2016)
Data mining: A set of techniques to extract information from data (Chen & Zhang, 2014)	24.9% (200)	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 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 (Delen, 2009)
Visualization Approaches: The techniques used to create tables, images, diagrams and other intuitive display ways to understand data (Chen & Zhang, 2014)	19% (153)	Presented a visualization 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 (Angelelli et al., 2014)
Statistics: The technique of organizing, and interpreting data applying statistical techniques. (Chen & Zhang, 2014)	16.4% (132)	Used Stanford's clinical data warehouse from Lucile Packard Children's Hospital to analyse patient characteristics associated with chronic uveitis in a large juvenile idiopathic arthritis cohort and through logistic regression the results indicated a new association between allergic conditions and chronic uveitis in juvenile idiopathic arthritis patients (Cole et al., 2015)
Simulation: Quantitative analysis of a system in a stochastic setting (Waller & Fawcett, 2013)	6.8% (55)	Developed a simulation model, FVGWAS for analysis of "whole-genome brain data" suggesting that their approach could be a valuable statistical toolbox for fast large-scale imaging genetic analysis (Huang et al., 2015)
Web mining: The process of information discovery from sources across the World Wide Web. (Cooley, Mobasher, & Srivastava, 1997)	6.7% (54)	Developed an analytics platform, called "Cytobank", for community cytometry data analysis (to track cells and subsets in blood and tissue) and collaboration (Chen & Kotecha, 2014)
Optimization Methods: Methods that improve the accuracy of forecasting and quantitative problems following computational strategies (Chen & Zhang, 2014)	6.1% (49)	Built machine learning tools using web mining methods, statistical predictions and mathematical optimization methods for selecting the chemotherapy regimens to be tested in phase II and phase III clinical trials of advanced gastric cancer with the primary objective of improving the quality of regimens tested in phase III trials compared to current practice. A database of 414 clinical trials was used for this purpose (Bertsimas et al., 2016)

(*continued on next page*)

Table 2 (*continued*)

Techniques	% N	Indicative Research Example
Text mining: Techniques based on machine learning and data mining to find useful patterns in text data (Holzinger, Schantl, Schroettner, Seifert, & Verspoor, 2014)	5.2% (42)	Presented a new software to manage and transform Big Data in a new comprehensive format based on text indexing system for mammographic images retrieval and classification (Farruggia, Magro, & Vitabile, 2014)
Forecasting: Using predictive techniques for evaluating what would have happened under different circumstances (Waller & Fawcett, 2013)	2.7% (22)	Developed and evaluated a web-based forecast tool that predicts the daily bed need for admissions from the cardiac catheterization laboratory using available 6384 clinical data from catheterization patients (Toerper et al., 2015)
Social Network Analysis: A technique that views and analyses data from social networks	2.5% (20)	Provided a brief introduction to commonly used Social Web tools such as mashups and aggregators as a way to observe people's collective health status and create a clear picture of epidemiological outbreaks (Boulos et al., 2010)

having multiple runs of the model. We have included them all in an attempt to address as many approaches as possible. Table 2 presents the allocation of the papers to the categories which are followed by a short definition and an indicative research example. It is not surprising, therefore, that *Modeling* surfaces is the most popular of the techniques. The first most specific technique is "machine learning", which automates the execution of rules in modeling through algorithms. This is an emerging technique with many successful applications in the healthcare sector (Chen, Hao, Hwang, Wang, & Wang, 2017; Khalaf et al., 2017). *Data mining* and *Visualization* follow with a substantial number of studies using the specific methods.

4.3. Gained values/capabilities from the use of BDA in the health sector

The benefits from the analytics in healthcare have been summarized in the ability to provide comparative effectiveness research to determine more clinically relevant and cost-effective ways to diagnose and treat patients (V. Raghupathi & Raghupathi, 2014; W. Raghupathi & Raghupathi, 2014). More specifically, in order to identify the full range of the emerging capabilities in the healthcare sector from the use of big data analytics, we identified them in the 804 papers of our dataset and classified them under 10 categories of value creation. Table 3 presents these values sorted by popularity with a short explanation and the frequency of papers from the dataset that refer to one or more of these gains based on the research they present. The first five values are similar to those identified in the study of Wamba et al. (2017).

The most popular value "Better diagnosis for provision of more personalized healthcare" refers to the BDA capability to direct to better case diagnosis from the collection of more data and therefore offer more targeted therapy or health service to the individual. This, for example, could be the analysis of the numerous relationships of specific patient's biomarkers which can lead disease therapy to precision medicine (Alyass, Turcotte, & Meyre, 2015), or the investigation of patient health metrics and behavior through wearables and the Internet of Things leading to specific interventions based on the collected data (Banos et al., 2016).

The second value "Supporting/replacing professionals' decision-making with automated algorithms" is about mining knowledge from large data sets and training algorithms to pattern matching. This means better automatic categorization of new information entering the analysis process and improved decision-making when it comes to diagnosis and choice of therapeutic scheme, for example.

The third value "New business models, products and services" refers to the development of new business models, products, and services through the capabilities offered by BDA, such as a new visualization software with real-time statistical analyses of brain images for better patient diagnosis (Angulo, Schneider, Oliver, Charpak, & Hernandez, 2016) or a mobile application in which people can enter symptoms and get possible diagnoses and recommended medication.

The fourth value "Enabling experimentation, expose variability and improve performance" from the use of BDA, is for researchers to acquire a deeper understanding of all possible interrelationships between variables and develop scenarios for further experimentation with their models and expose new health information.

Table 3
Created Values from the use of BDA.

Value	Types	Definition	N	%
V1	Better diagnosis for provision of more personalized healthcare (P)	Analytic approaches for better patient diagnosis which lead to provision of more personalized therapeutic schemes or services to the users	286	35.6
V2	Supporting/replacing professionals' decision-making with automated algorithms (P)	Through adaptive rules/algorithms for fast categorization of symptoms/medical results and pattern matching, analytics can provide recommendations for diagnosis and remedy/actions.	206	25.6
V3	New business models, products and services (P)	BDA enables companies to create new products and services. e.g. new software for analysis of data/images, enhance existing ones, and invent entirely new business models, new ways of reaching to patients.	197	24.5
V4	Enabling experimentation, expose variability and improve performance (A)	Analytics create conditions for enhanced experimental applications of large datasets for testing "what-if" scenarios and assisting performance and decision-making	144	17.9
V5	Healthcare information sharing and coordination (A)	BDA can organize the selection and sharing of information and data analysis among stakeholders to gain operational efficiency	122	15.2
V6	Creating data transparency (A)	BDA can collect/convert data in a standardized format and treat data in the same way for reducing time, cost of search and processing while maintaining clarity and quality	115	14.3
V7	Identifying patient care-risk (P)	BDA create enhanced opportunities of health risk prediction for acting proactively to patient care-risk	79	9.8
V8	Offering customized actions by segmenting populations (M)	BDA through high exploitation capabilities of big data can discover specific segmentations and tailor products and services to meet patients or health professionals' needs.	72	9
V9	Reducing expenditure while maintaining quality (M)	BDA enables new, cost-effective ways to intervene on the determinants of health, aiming at reducing expenditures while sustaining health outcomes.	72	9
V10	Protecting privacy (A)	BDA can identify ways of securing privacy of health-related data to support the ethical principles and people respect.	41	5.1

The fifth value "Healthcare information sharing and coordination" is gained by the coordination and sharing of health information across healthcare services or even countries to improve of health professionals' decision-making.

The sixth value "Creating data transparency" is about the ability of BDA to collect big data and format them in a standardized way. This capability reduces data identification and analysis time and assists the previous value of coordinating meaningful and comprehensive health-related information.

The next value "Identifying patient care-risk" refers to the capability of running the big data in advanced statistical techniques, such as logistic regression models and regressions trees which can identify scenarios of risk patterns and send an alert for areas of health risk prevention. For example, identifying high risk populations for a particular disease helps policy-makers to decide on giving earlier access to screening to these populations.

The following value "Offering customized actions by segmenting populations" refers to the use of BDA to identify new factors, through clustering and other methods, for segmenting populations differently or in more categories and offer more targeted health services or products.

Value 9, that is "Reducing expenditure while maintaining quality" focuses on the capability of analytics, through process mining, visualization techniques and collaborative tools, to propose ways for reducing health organizations' costs from better resource utilization, elimination of non-value-added actions, capturing hospital underpayments, etc., while maintaining the quality level. An example could be the use of visualization tools for identifying non-value-added processes in patients with chronic diseases by tracking patient data over time during home, ambulatory and hospital care.

The last value, "Protecting privacy", is about how BDA can offer data security in ways such as the identification of privacy breaches, the capability to extract data by eliminating ID recognition from electronic medical records and others. This has become a big issue, especially for organizations that use cloud computing as their main processing platform in which privacy and security are difficult to be controlled (Larson & Chang, 2016). Overall, we can see that the majority of health data analytics studies attempt to direct their efforts to patient benefit. Needless to say, almost all studies have this ultimate goal but their direct focus may be at the intermediate stage for improving the way of doing it.

Gaining an overall picture of the identified values, we can say that values 1, 2 3 and 7 directly relate to patient wellbeing (P), values 4, 5, 6, 10 relate to analysts (A) for better data handling and values 8 and 9 relate to management (M) for better positioning their products/services and gaining management efficiencies respectively. The identification codes (P), (A), (M) are presented under the "Types" column in Table 3.

4.4. The association of the selected analytical techniques with the data types and capabilities

Using the NVivo analysis software, we performed comparisons to define the techniques that have most popularly been applied for each data type and presented value. During this procedure, we assigned each one of the 804 articles to the data types (as presented in Table 1) and performed a breakdown by technique (Table 4). Table 4 shows that out of all studies that deal with clinical data (562), 41.5% have used machine learning for their analysis, from studies with patient behavior/sentiment data 51% have used machine learning, etc. We also repeated this process to the 10 identified values (Tables 5). It seems that the most popular techniques scientists need or prefer to use are: modeling, machine learning, data mining, visualization, and statistical analysis (Table 2). The same techniques are also popular, in the same order, with the exception that machine learning comes first, for all data types (Table 4). Overall, machine learning and modeling are the most applied

Table 4
Classification of dataset articles based on the analytical techniques by data type.

Data types→	Clinical	Patient behavior & sentiment	Administrative (activity & cost)	Pharmaceutical & R&D data
% (n)	100 (562)	100(133)	100(59)	100(38)
BDA Techniques↓	% (n)	% (n)	% (n)	% (n)
Machine learning	**41.5 (233)**	**51 (68)**	32(19)	**31.6(12)**
Modeling	28.6 (161)	33.8(45)	**49(29)**	**44.7(17)**
Data mining	24(135)	21.8 (29)	25.4(15)	15.8(6)
Visualization	22.2(125)	15.8 (21)	13.5(8)	5.3(2)
Statistical analysis	20.3(114)	11.3(15)	15.2 (9)	13(5)
Simulation	7.1(40)	6.7 (9)	8.5(5)	13(5)
Optimization	7 (36)	3(4)	6.8(4)	5.3(2)
Web mining	4.5(25)	16.5(22)	5(3)	2.6(1)
Text mining	4.5(25)	6 (8)	3.4(2)	2.6(1)
Forecasting	2 (14)	2.2(3)	1.5(2)	0
Social net. Analysis	1(6)	12 (16)	3.4(2)	2.6(1)

Table 5
Classification of dataset articles based on the analytical techniques by created value.

Values →	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10
% (n)	100(286)	100 (206)	100 (197)	100 (144)	100 (122)	100(115)	100(79)	100 (72)	100(72)	100(41)
BDA Techniques↓	% (n)	% (n)	% (n)	% (n)	% (n)	% (n)	% (n)	% (n)	% (n)	% (n)
Machine learning	**48(137)**	**39 (80)**	**50 (98)**	**51.4 (74)**	**47.59(58)**	**31.3 (36)**	**38 (30)**	**32 (23)**	**26.4 (19)**	**61 (25)**
Modeling	**46.5 (133)**	**49 (101)**	**39.6 (78)**	**43 (62)**	**37.7 (46)**	**54.8 (63)**	**60.8 (48)**	**32 (23)**	**43 (31)**	**34.1 (14)**
Data mining	**28 (80)**	**25.2 (52)**	**20 (39)**	**23 (33)**	23.8 (29)	23.5 (27)	13.9 (11)	**30.5 (22)**	**33.3 (24)**	**41.5 (17)**
Visualization	20.3 (58)	18.4 (38)	16.7 (33)	20.9 (30)	**28.7 (35)**	**29.5 (34)**	17.7 (14)	11 (8)	16.6 (12)	7.3 (3)
Data mining	**28 (80)**	**25.2 (52)**	**20 (39)**	**23 (33)**	23.8 (29)	23.5 (27)	13.9 (11)	**30.5 (22)**	**33.3 (24)**	**41.5 (17)**
Statistical analysis	19.2 (55)	17 (35)	9.6 (19)	13.9 (20)	13.9 (17)	13 (15)	**38 (30)**	**30.5 (22)**	8.3 (6)	12.2 (5)
Simulation	6 (17)	8.2(17)	7.1 (14)	14 (20)	1.6 (2)	12.2 (14)	3.8 (3)	7 (5)	9.7 (7)	2.4 (1)
Optimization	5.6 (16)	6.3 (13)	5 (10)	13.2 (19)	3.3 (4)	7.8 (9)	3.8 (3)	4 (3)	9.7 (7)	0
Web mining	5.6 (16)	4.4 (9)	7.1 (14)	9.7 (14)	11.5 (14)	18.3 (21)	7.6 (6)	11 (8)	4.2 (3)	4.9 (2)
Text mining	2.5 (7)	4.9 (10)	5 (10)	3.5 (5)	7.4 (9)	1.7 (2)	7.6 (6)	8.3 (6)	1.4 (1)	0
Forecasting	2.09	2 (4)	4.5 (9)	3.5 (5)	0	6 (7)	1.2 (1)	0	6.9 (5)	2.4 (1)
Social net. analysis	1.7 (5)	1.4 (3)	2 (4)	2.8 (4)	4.9 (6)	0.9 (1)	1.3 (1)	7 (5)	0	0

techniques amongst all data types with a variance of 29%–49% and across almost all values, with a variance of presence between 32% and 61%. Noticing the percentages on both tables, we ascertain a level of uniform distribution throughout each technique. This is consistent with most values as indicated in Table 3, and with most data types as shown in Table 1.

4.5. The use of Machine Learning in the health field

Machine learning, the most preferred of the analytical techniques for the variety of data types, offers immense potential in the healthcare predictive analytics arena for improving outcomes in many domains of research (López-Martínez, Schwarcz, Núñez-Valdez, & García-Díaz, 2018). It facilitates the development of patient-centric models for improving diagnoses and intervention. Machine learning is a data analysis method that automates analytical model building. As a branch of artificial intelligence refers to analytical algorithms that iteratively learn from data, identify patterns and allow computers to make inferences and find insights without being explicitly programmed where to look (Breiman, 1996). Machine learning techniques can be used to integrate and interpret complex health data in scenarios where traditional statistical methods cannot perform (Shameer, Johnson, Glicksberg, Dudley, & Sengupta, 2018). Usually, a plethora of machine learning models for risk prediction are evaluated to choose the most accurate one. The use of machine learning based methods is important during data collection, dimension reduction, etc. to achieve different value creation objectives (ur Rehman, Chang, Batool, & Wah, 2016).

Machine learning algorithms are proving convenient in medical diagnosis that require more accurate prognostic models, such as detecting diabetic retinopathy (Gulshan, Peng, & Coram, 2016) and in medical disciplines such as oncology, and where pattern recognition is of ultimate importance, such as radiology and pathology (Cabitza, Rasoini, & Gensini, 2017).

Based on our path-to-value theoretical framework and through content analysis of our articles, we targeted some excellent examples of the application of the machine learning algorithm. In relation to the value for the diagnosis of the personalized health (V1), Bertsimas, O'Hair, Relyea, and Silberholz (2016) developed models that use machine learning and optimization which identify a better combination of chemotherapy drugs and improve the outcome of chemotherapy regimens tested in clinical trials without changing toxicity levels. In line with V1 too, Voisin, Pinto, Morin Ducote, Hudson, and Tourassi (2013) identified the best performing machine learning algorithm to predict diagnostic error in mammography by merging gaze behavior characteristics from the radiologist and image features.

To support the business value of "supporting/replacing human decision-making with automated algorithms (V2)", Lary, Woolf, Faruque, and LePage (2014), used the machine learning algorithms to analyze geospatial data of populations (e.g. smoking-obesity rates, education level, air pollution, existing health and social-support services) and to construct tools for public health data-driven decisions (budget allocation on health interventions based on best return on investment).

A paradigm of a new innovative product (V3) that creates value to the healthcare business is the Wiki-Health service platform that collects, stores, and analyses personal health sensor data which are used for tracking existing health conditions and most importantly predicting them, through the use of machine learning algorithms, encouraging a pro-active approach to healthcare (Li & Guo, 2016). Moreover, for improving the performance of the model (V4), Breiman (1996) used new approaches at that time, such as bagging (i.e. Bootstrap Aggregation) to decrease the variance of the prediction.

5. Big data analytics

Because big data are large, processing cannot be performed by traditional health informatics such as "a standalone system" with just a simple analytic software. What is required is a more complex, programming intensive system with a variety of skills (V. Raghupathi & Raghupathi, 2014; W. Raghupathi & Raghupathi, 2014). That is in many cases the Hadoop open-source platform. Hadoop, released by Apache in 2011, consists of mainly the Hadoop Distributed File System (HDFS/ a way to divide large data sets in smaller types and store them across multiple servers) and MapReduce (a computational paradigm using two sequences of execution - parallel processing) which includes: a) the map phase that produces interposed key value pairs from initial key-value pairs and b) the reduce phase where the interposed key-value pairs are aggregated by a key and the values are combined together to a final reduction output. HBase is a distributed database built on top of HDFS to provide storage for Hadoop Distributed Computing using ZooKeeper as a coordination service (McClay et al., 2015). First, Google introduced MapReduce allowing big data processing on clusters with Mapping and Reducing. Yahoo developed Hadoop as an open source implementation of MapReduce (Van Poucke et al., 2016). Map/reduce jobs on Hadoop, which can also be developed on Hive (a runtime Hadoop support Architecture), provide a mechanism to project structure on this data and query them allowing MapReduce jobs in other languages when required (Van Poucke et al., 2016).

Business analytic tools are faced with many challenges and researchers evaluate them in terms of availability, continuity, ease of use, scalability, ability to manipulate at different levels of granularity, privacy and security enablement or quality assurance (V. Raghupathi & Raghupathi, 2014; W. Raghupathi & Raghupathi, 2014). For example, in order to overcome the major disadvantage of Hadoop that is tight coupling between the programming model and the resource management infrastructure, a new architecture was developed called YARN. YARN decouples the programming model from the resource management infrastructure and delegates many scheduling functions (Van Poucke et al., 2016). Further, the Apache Pig dataflow system was developed to allow users to easily compose multiple data processing functions because Hadoop MapReduce was restricted to practitioners with advanced technical skills due to the complexity of parallel operations and multi-step data flows (Sahoo et al., 2016).

So overall, the computing platform most often used for the BDA tools in general and for the healthcare in particular is Apache Hadoop (De Silva, Burstein, & Jelinek, 2015; Dinov, 2016). Additionally, MapReduce is a programming paradigm that provides scalability across many servers in a Hadoop cluster with a broad variety of real-world applications (Belle et al., 2015; Berger & Doban, 2014; Khan et al., 2014; Luo, Wu, Gopukumar, & Zhao, 2016). From the screening of our article pool, we identified 36 papers published in 2016 that present applications based on the Hadoop ecosystem with different applications and capabilities. From the most recent literature we present examples of a few representative studies with a reference to their data types and techniques and the achieved value. Along with this, we also discuss the technical restrictions that each case brings up and attempts to overcome.

An example of the use of Hadoop ecosystem is the one presented in the research of Batarseh and Latif (2016), that introduced a "user friendly" tool called CHESS and has been developed in Visual studio for C# to read EHR and provide means for analysts to run queries and experiments. CHESS moves the uploaded datasets to Hadoop and aggregated data, with much fewer rows, are settled to a SQL server for analysis. Then, users access them through the statistical software of their choice (e.g. excel, Tableau, R), and after re-organizing the data in the necessary format can run statistical tests to examine, for example, the importance of some factors (e.g. demographics) over certain health conditions. The application relies on Hadoop for handling big data issues, and the users can query only smaller amounts of data to the statistical software. The application could benefit from more advanced clustering methods to allow for running statistical significance tests to identify important healthcare factors in a more automated way.

In the post-genomic era, as the focus of biology has started to shift from mapping genomes to analyzing the vast amount of information resulting from functional genomics research Bodenreider and Burgun (2005); Cui, Tao, and Zhang (2016) describe the evolution of using Hadoop and MapReduce in the scalable and computational powerful cloud computing environment to perform biomedical ontology quality assurance (OQA). This capability has made it possible to reduce the standard sequential approach for implementing OQA methods from weeks to hours. With this speed, more exhaustive structural analysis of large ontological hierarchies can be performed and structural changes between versions for evolutional analysis can be systematically tracked. Areas of further research are around the development of better user interfaces for reviewing OQA results and visualizing ontological alignment and evolution while also increasing the performance of the visual interface by automatically pre-computing intensive jobs while in interaction with the user.

Istephan and Siadat (2016) presented a new approach of unleashing the content of unstructured medical data and enabling queries and processing of both structured and unstructured health data for the diagnosis of personalized health. This is a step forward as most applications are limited to being able to query only from structured medical data, such as part of the EHR datasets of a population.

For example, when it comes to medical image and EHR processing, there are cloud based software and platforms, such as LifeImage, Nuance mPower for sharing and retrieving big data medical images and other health records, but they are limited to using structured data (e.g. run a query on patient gender) to retrieve all relevant images and records and cannot handle unstructured data (e.g. query based on volume of a brain structure).

Other developments incorporate models, even in a Hadoop/MapReduce environment (Yao et al., 2014), that are related to pattern matching in data medical images. This means that an image is uploaded as an input and feature extraction and similarity pattern matching techniques are used to retrieve similar images (Toews, Wachinger, Estepar, & Wells, 2015)

Some technology restrictions that are apparent from the papers in our dataset with regards to the Hadoop and MapReduce environment is that they cannot always handle unstructured content from health data and medical images in the desired way. In order to overcome the problem, researchers create customized tools (instead for example of a Hadoop component like Hive) (Istephan & Siadat, 2016). Such approaches further aid medical experts in getting support for decision-making with automated algorithms.

6. Future research

The adopted research framework concludes with the future research directions. Table 6 presents in terms of popularity the classification of the future directions in 19 categories, which are further classified under three standpoints: 1) technological improvements (T), 2) healthcare organization processes improvements (O) and 3) research improvements (R)". Although the research presented in each paper was different, the future perspectives were in many cases similar. The reported frequencies show in how many papers, out of the 804, the particular future direction was identified and therefore gives a signal of the importance of pursuing this research. In terms of popularity, we note that the first 3 future directions refer to the technological aspects and the majority of the organizational improvements come last.

The first future direction, which is mentioned in 16.5% of the studies in our pool, refers to the need for developing a more *advanced version* of the presented IT method to e.g. increase running time, reduce bugs, etc. Similar to this, the second direction is about developing *new mechanisms* for handling the data more effectively and their required analysis, such as new OR techniques for large dataset exploitation, new platforms for running specific big data types' analyses, etc. The 3rd direction is about the development of computational methods for *extending system capabilities*, like the tracking of patients across service sites, the provision of more standardized and comprehensive outcome from the analysis of data, and the accessibility of EHR to all caregivers (Barkley, Greenapple, & Whang, 2013).

The next direction is about researchers who want to further and better *prove their hypothesis* with a bigger or a more diverse dataset. This is followed by the need to investigate potential IT solutions for *reducing errors* when merging and standardizing databases, and when compiling information from multiple sources. The first direction concerning the organizations comes next, requesting the development of a system that will standardize and secure the process of *extracting anonymized healthcare datasets* (Al-Shaqi, Mourshed, & Rezgui, 2016) which re-asrchers can manipulate without limits. The following direction comes from researchers who want to *apply their presented approach to other healthcare applications* to test if similar values are gained. The 8th direction is about *increasing health data accuracy*, by developing new technological methods for efficiently handling issues of missing values, correction of wrong data entries, etc.

The next important direction is the call for appropriately *educating the public and training health professionals* in the use of BDA techniques to gain knowledge of their capabilities and their limitations. The 10th direction is about researchers improving the tested approach in terms of *proposing new modalities* to successfully provide more sufficient results. Another emerging need is the creation of secure mechanisms to pool patient data from across health services around the world so that the clinician can find 'patients like mine' to help with real-time clinical decision-making (Broughman & Chen, 2016). Following from the previous, the next two directions are about researchers' need to expand the scope of their models by examining their applied BDA techniques either with *another pool of patient demographics (e.g. from other countries)* or in *other scientific areas* as, for example, the interpretation of post genomics data in nutrigenomics, pharmacogenomics, vaccinomics, etc. (Ben-Ari Fuchs et al., 2016).

The 14th direction refers to the *integration of environmental factors* into the BDA models in order to improve and secure patient living environment, such as enabling automatic operation of corridor/toilet lights to minimize patient falls. The next direction is the anticipation that these new technologies, such as the computer-aided diagnosis

Table 6
The Directions of Future Research of BDA in the Health Sector.

	Future Research	% N
1	Development of the specific approach (T)	16.5% (133)
2	Creation of new mechanisms to accrue maximum value of data (T)	12.6% (101)
3	Hardware and software development – extend systems capabilities (T)	10.9% (87)
4	More studies to prove the hypothesis (R)	7.1% (57)
5	Compile error-free datasets (T)	5.1% (41)
6	Need of standardized mechanism to retrieve sensitive data and securing privacy (O)	4.6% (37)
7	Propose an approach that can be used in other healthcare applications (R)	3.9% (31)
8	Identify data elements that can be automatically corrected (T)	2.6% (21)
9	Training and education of clinicians and public (O)	2.5% (20)
10	Alternate the proposed approaches (R)	2.1% (17)
11	Create mechanisms to identify "patients like me" (T)	1.9% (15)
12	To replicate same methods in other countries (R)	1.7% (14)
13	Create value to other scientific areas (R)	1.5% (12)
14	Integrating environmental factors in analytics for decision making (O)	1.2% (10)
15	Change the protocols and define policy purposes (O)	1.1% (9)
16	More investment in infrastructure (O)	0.6% (5)
17	National investments on health monitoring (O)	0.5% (4)
18	Cost effective analysis of the new tool (O)	0.4% (3)
19	Create partnerships among stakeholders to establish the value of BD (O)	0.2% (2)

system, will *challenge old practices* and will create new protocols of treatments. In line with this there are also the expectations that *higher budgets will be given to IT infrastructure* and to BDA experts working in health as well as from nations towards *healthcare monitoring*, such as automated bio-surveillance systems. The 18th direction is about better understanding the *cost-effectiveness of the design of new tools/policies* like monitoring drug prescription patterns. The least mentioned direction is the *development of partnerships* among manufacturers, health providers, payers, and regulators to communicate within each other the values of BDA.

In particular about machine learning in health, for which we took a special focus in this paper as it was identified to be the most used technique, future directions that derived from our dataset should focus on the following perspectives: use of unsupervised learning techniques to more precisely phenotype complex disease; the development of automated risk prediction algorithms which can be used to guide clinical care; and the implementation of reinforcement learning algorithms to intelligently augment healthcare providers. From the technological perspective, an important issue is the strain between accuracy and interpretability. Studies should be directed towards the development of machine learning decision support systems which will automatically provide clarifications, and offer doctors' interactive visualization tools to examine the implications of potential exposure variables (Batarseh & Latif, 2016; Cabitza et al., 2017). Finally, at the organizational level, an important issue, which was also mentioned in Table 6, is the training of doctors to assessing the value of machine learning–based aids in practice and avoid the reduction of the skill for diagnosis or the loss in judgement of the accuracy of the decision-support systems results. This further requires knowledge of how these machine learning algorithms work in practice, therefore, it requires the acquisition of statistical and data analysis skills.

Of course, as the development of technology is a step ahead of its presentation in academic papers, we can assume that recent technological innovations in analytical techniques are creating further opportunities deriving from hidden, up to date information. Novel analytic fields, for instance, the analysis of data gathered from social media or data retrieved from mobile applications, will likely lead to new information systems for the healthcare sector. A limitation of profiling studies, such as ours, is the time-lag between the year of publication of the reviewed papers and the time of the presentation of the synthesis of their findings, due to the long process of literature reading and synthesis. We hope that future research will discuss more recent advances in the investigated field.

7. Discussion – conclusions

Given the large numbers and frequently updating healthcare publications, systematic reviews assist healthcare practitioners to make decisions as they provide summarized research on a given topic of interest (Ali et al., 2018). In this study, we aimed to present a systematic overview of the literature in order to determine the way BDA has managed to improve the healthcare domain. We followed the resource–based theory to identify the big data sources and the analytics techniques which allow big data capacities to create values which will continue to fuel through new research in the field. We mapped the existing literature on the field of BDA in Healthcare using content analysis, while we aimed to provide explanatory definitions of the categorisation through representative examples.

The most popular analytical techniques that scientists use to make meaningful interpretations of data are: modeling, machine learning, data mining, visualization and statistical analysis. In particular, machine learning is the most applied technique across almost all created values and data types that offer immense potential in the healthcare predictive analytics arena to improve outcomes in many domains of research (López-Martínez et al., 2018). Machine learning is described as a complex field that provides numerous kinds of tools, techniques, and frameworks that can be exploited to address challenges created by the fusion of data (Chowriappa, Dua, & Todorov, 2014). Moreover, it is apparent that all applied techniques are, more or less, equally used across the data types, as well as equally create the different capabilities in the healthcare sector. Clinical data was the most used source of data analysis (70%).

From the presentation of BDA software, based on Hadoop ecosystem or the MapReduce process, this research confirms that most users use clinical or medical structured or unstructured data for their studies to build new approaches for the diagnosis of personalized healthcare and to invent entirely new business models to reduce time, cost of search or processing while maintaining quality.

From the articles in our dataset it is clear, that there is demand for research in health analytics to focus on improving the technological aspects. There is a definite need in healthcare for systems that support or improve the decision-making ability of clinical experts, specifically, to diagnose complex diseases or pathologies (López-Martínez et al., 2018). Progress that has been made via Hadoop and MapReduce has increased performance by reducing time and pre-computing computationally intensive jobs (Cui et al., 2016). The main difficulty with big data in healthcare is that most data are often unstructured, which means that there are obstacles to computationally processing the largest part of them (Dinov, 2016). That is why scientists are in a continuous effort to advance infrastructure in order to achieve the greatest possible analysis and to further develop computational methods in order to extend systems' capabilities. It is expected that more investment will be given to IT infrastructure and to BDA experts in the healthcare sector, or from nations for health monitoring, or for the development of systems that can track patients' health-related data across health services and home and make these accessible to relevant professionals.

Therefore, firms in the healthcare industry, in the private and public sector, started incorporating BDA for strategic decision-making (Gandomi & Haider, 2015). However, the major reason behind BDA non-adoption is that first firms do not realize their strategic value and their managers are not prepared to bring the changes because of technological or organizational difficulties (Gupta, Kar, Baabdullah, & Al-Khowaiter, 2018). We hope that this profiling study will act as a trigger for health organizations to redesign their strategies towards a greater adoption of BDA and harness their capabilities to improve service, mitigate risks, reduce costs and grasp new opportunities.

References

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.

Ajorlou, S., Shams, I., & Yang, K. (2015). An analytics approach to designing patient centered medical homes. *Health Care Management Science, 18*(1), 3–18.

Ali, O., Shrestha, A., Soar, J., & Wamba, S. F. (2018). Cloud computing-enabled healthcare opportunities, issues, and applications: A systematic review. *International Journal of Information Management, 43*, 146–158.

Al-Shaqi, R., Mourshed, M., & Rezgui, Y. (2016). Progress in ambient assisted systems for independent living by the elderly. *SpringerPlus, 5*(1), 624.

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

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.

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*(1) 1–1.

Angulo, D. A., Schneider, C., Oliver, J. H., Charpak, N., & Hernandez, J. T. (2016). A multi-facetted visual analytics tool for exploratory analysis of human brain and function datasets. *Frontiers in Neuroinformatics, 10*, 36.

Banos, O., Amin, M. B., Khan, W. A., Afzal, M., Hussain, M., Kang, B. H., ... Lee, S. (2016). The Mining Minds digital health and wellness framework. *Biomedical Engineering Online, 15*(1), 76.

Barkley, R., Greenapple, R., & Whang, J. (2013). Actionable data analytics in oncology: Are we there yet? *Cancer Center Business Development Group,* 93–96.

Barney, J. B. (1991). Firm resources and sustained competitive advantage. *Journal of Management, 17*, 99–120.

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

Basole, R. C., Braunstein, M. L., Kumar, V., Park, H., Kahng, M., Chau, D. H., ... Lesnick, B. (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.

Batarseh, F. A., & Latif, E. A. (2016). Assessing the quality of service using Big Data analytics: With application to healthcare. *Big Data Research, 4*, 13–24.

Belle, A., Thiagarajan, R., Soroushmehr, S. M., Navidi, F., Beard, D. A., & Najarian, K. (2015). Big data analytics in healthcare. *BioMed Research International, 2015*.

Ben-Ari Fuchs, S., Lieder, I., Stelzer, G., Mazor, Y., Buzhor, E., Kaplan, S., ... Kohn, A. (2016). GeneAnalytics: An integrative gene set analysis tool for next generation sequencing, RNAseq and microarray data. *OMICS A Journal of Integrative Biology, 20*(3), 139–151.

Berger, M. L., & Doban, V. (2014). Big data, advanced analytics and the future of comparative effectiveness research. *Journal of Comparative Effectiveness Research, 3*(2), 167–176.

Bertsimas, D., O'Hair, A., Relyea, S., & Silberholz, J. (2016). An analytics approach to designing combination chemotherapy regimens for cancer. *Management Science, 62*(5), 1511–1531.

Bodenreider, O., & Burgun, A. (2005). *Biomedical ontologies. Medical informatics*. Boston, MA: Springer211–236.

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.

Breiman, L. (1996). Bagging predictors. *Machine Learning, 24*(2), 123–140.

Brooks, P., El-Gayar, O., & Sarnikar, S. (2015). A framework for developing a domain specific business intelligence maturity model: Application to healthcare. *International Journal of Information Management, 35*(3), 337–345.

Broughman, J. R., & Chen, R. C. (2016). Using big data for quality assessment in oncology. *Journal of Comparative Effectiveness Research, 5*(3), 309–319.

Buchem, I., Merceron, A., Kreutel, J., Haesner, M., & Steinert, A. (2014). Wearable enhanced learning for healthy ageing: Conceptual framework and architecture of the 'Fitness MOOC'. *Interaction Design and Architecture(s) Journal, 24*, 111–124.

Cabitza, F., Rasoini, R., & Gensini, G. F. (2017). Unintended consequences of machine learning in medicine. *JAMA, 318*(6), 517–518.

Calabrese, N., Minkoff, B., & Rawlings, K. (2014). Pharmacosynchrony: Road map to transformation in pharmacy benefit management. *The American Journal of Pharmacy Benefits*.

Carvalho, J. V., Rocha, Á., Vasconcelos, J., & Abreu, A. (2019). A health data analytics maturity model for hospitals information systems. *International Journal of Information Management, 46*, 278–285.

Chen, T. J., & Kotecha, N. (2014). *Cytobank: Providing an analytics platform for community cytometry data analysis and collaboration. High-dimensional single cell analysis*. Berlin, Heidelberg: Springer127–157.

Chen, H., Chiang, R. H. L., & Storey, V. C. (2012). Business intelligence and analytics: From big data to big impact. *MIS Quarterly, 36*(4), 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: Practical Innovations, Open Solutions, 5*, 8869–8879.

Chen, C. L. P., & Zhang, C.-Y. (2014). Data-intensive applications, challenges, techniquesand technologies: A survey on Big Data. *Information Sciences, 275*, 314–347.

Chowriappa, P., Dua, S., & Todorov, Y. (2014). *Introduction to machine learning in healthcare informatics. Machine learning in healthcare informatics*. Berlin, Heidelberg: Springer1–23.

Cole, B. K., Simmers, M. B., Feaver, R., Qualls, C. W., Jr, Collado, M. S., Berzin, E., ... Manka, D. (2015). An in vitro cynomolgus vascular surrogate system for preclinical drug assessment and human translation. *Arteriosclerosis, Thrombosis, and Vascular Biology, 35*(10), 2185–2195.

Cooley, R., Mobasher, B., & Srivastava, J. (1997). Web mining: Information and pattern discovery on the world wide web. November *Tools With Artificial Intelligence, 1997. Proceedings., Ninth IEEE International Conference on,* 558–567.

Cui, L., Tao, S., & Zhang, G. Q. (2016). Biomedical ontology quality assurance using a big data approach. *ACM Transactions on Knowledge Discovery from Data (TKDD), 10*(4), 41.

De Camargo Fiorini, P., Seles, B. M. R. P., Jabbour, C. J. C., Mariano, E. B., & de Sousa Jabbour, A. B. L. (2018). Management theory and big data literature: From a review to a research agenda. *International Journal of Information Management, 43*, 112–129.

De Silva, D., Burstein, F., & Jelinek, H. (2015). Addressing the complexities of big data analytics in healthcare: The diabetes screening case. *Australasian Journal of Information Systems, 19*, 99–115.

Delen, D. (2009). Analysis of cancer data: A data mining approach. *Expert Systems with Applications, 26*(1), 100–112.

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

Doumpos, M., & Zopounidis, C. (2016). Editorial to the special issue "business analytics". *Omega, 59*, 1–3.

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

Duan, Y., Edwards, J. S., & Dwivedi, Y. K. (2019). Artificial intelligence for decision making in the era of Big Data–evolution, challenges and research agenda. *International Journal of Information Management, 48*, 63–71.

Dubey, R., Gunasekaran, A., & Papadopoulos, T. (2017). Green supply chain management: Theoretical framework and further research directions. *Benchmarking an International Journal, 24*(1), 184–218.

Dubey, R., Gunasekaran, A., Childe, S. J., Fosso Wamba, S., Roubaud, D., & Forupon, C. (2019). Empirical investigation of data analytics capability and organizational flexibility as complements to supply chain resilience. *International Journal of Production Research,* 1–19.

Dubey, R., Gunasekaran, A., Childe, S., Roubaud, D., Fosso Wamba, S., Giannakis, M., ... Forupon, C. (2019). Big data analytics and organizational culture as complements to swift trust and collaborative performance in the humanitarian supply chain. *International Journal of Production Economics, 210*, 120–136.

Farruggia, A., Magro, R., & Vitabile, S. (2014). A text based indexing system for mammographic image retrieval and classification. *Future Generation Computer Systems, 37*, 243–251.

Feldman, B., Martin, E. M., & Skotnes, T. (2016). Big data in healthcare hype and hope. *Dr. Bonnie, 360*(2012).

Gaitanou, P., Garoufallou, E., & Balatsoukas, P. (2014). The effectiveness of big data in health care: Systematic review. *Metadata and Semantics Research,* 141–153.

Gandomi, A., & Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management, 35*(2), 137–144.

Ghasemaghaei, M., Hassanein, K., & Turel, O. (2017). Increasing firm agility through the use of data analytics: The role of fit. *Decision Support Systems, 101*, 95–105.

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

Gulshan, V., Peng, L., Coram, M., et al. (2016). Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA, 316*(22), 2402–2410.

Gunasekaran, A., Papadopoulos, T., Dubey, R., Wamba, S. F., Childe, S. J., Hazen, B., ... Akter, S. (2017). Big data and predictive analytics for supply chain and organizational performance. *Journal of Business Research, 70*, 308–317.

Gupta, S., Kar, A. K., Baabdullah, A., & Al-Khowaiter, W. A. (2018). Big data with cognitive computing: A review for the future. *International Journal of Information Management, 42*, 78–89.

Holzinger, A., Schantl, J., Schroettner, M., Seifert, C., & Verspoor, K. (2014). *Biomedical text mining: State-of-the-art, open problems and future challenges. Interactive knowledge discovery and data mining in biomedical informatics*. Berlin, Heidelberg: Springer271–300.

Huang, M., Nichols, T., Huang, C., Yu, Y., Lu, Z., Knickmeyer, R. C., ... Alzheimer's Disease Neuroimaging Initiative (2015). FVGWAS: Fast voxelwise genome wide association analysis of large-scale imaging genetic data. *Neuroimage, 118*, 613–627.

Istephan, S., & Siadat, M. R. (2016). Unstructured medical image query using big data–an epilepsy case study. *Journal of Biomedical Informatics, 59*, 218–226.

Jacofsky, D. J. (2017). The myths of 'big data' in health care. *The Bone & Joint Journal, 99*(12), 1571–1576.

Jaklič, J., Grublješič, T., & Popovič, A. (2018). The role of compatibility in predicting business intelligence and analytics use intentions. *International Journal of Information Management, 43*, 305–318.

Katsaliaki, K., Mustafee, N., & Kumar, S. (2014). A game-based approach towards facilitating decision making for perishable products: An example of blood supply chain. *Expert Systems with Applications, 41*(9), 4043–4059.

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.

Khan, N., Yaqoob, I., Hashem, I. A. T., Inayat, Z., Ali, M., Kamaleldin, W., ... Gani, A. (2014). Big data: Survey, technologies, opportunities, and challenges. *The Scientific World Journal, 2014*.

Larson, D., & Chang, V. (2016). A review and future direction of agile, business intelligence, analytics and data science. *International Journal of Information Management, 36*(5), 700–710.

Lary, D. J., Woolf, S., Faruque, F., & LePage, J. P. (2014). Holistics 3.0 for health. *ISPRS International Journal of Geo-information, 3*(3), 1023–1038.

Li, Y., & Guo, Y. (2016). Wiki-health: From quantified self to self-understanding. *Future Generation Computer Systems, 56*, 333–359.

Liberatore, M. J., & Nydick, R. L. (2008). The analytic hierarchy process in medical and health care decision making: A literature review. *European Journal of Operational Research, 189*(1), 194–207.

Lim, C., Kim, K. H., Kim, M. J., Heo, J. Y., Kim, K. J., & Maglio, P. P. (2018). From data to value: A nine-factor framework for data-based value creation in information-intensive services. *International Journal of Information Management, 39*, 121–135.

López-Martínez, F., Schwarcz, A., Núñez-Valdez, E. R., & García-Díaz, V. (2018). Machine learning classification analysis for a hypertensive population as a function of several risk factors. *Expert Systems with Applications*.

Luo, J., Wu, M., Gopukumar, D., & Zhao, Y. (2016). Big data application in biomedical research and health care: A literature review. *Biomedical Informatics Insights, 8* BII-S31559.

Mamonov, S., & Triantoro, T. M. (2018). The strategic value of data resources in emergent industries. *International Journal of Information Management, 39*, 146–155.

McClay, W. A., Yadav, N., Ozbek, Y., Haas, A., Attias, H. T., & Nagarajan, S. S. (2015). A real-time magnetoencephalography brain-computer interface using interactive 3D visualization and the Hadoop ecosystem. *Brain Sciences, 5*(4), 419–440.

Mohan, M., Vigneshwaran, B., Vineeth Raj, G., & Harlin Jesuva Prince, S. (2016). Disease diagnosis for personalized health care using map reduce technique. *An International Journal of Optimization and Control Theories & Applications, 9*(5), 2153–2164.

Raghupathi, W., & Raghupathi, V. (2013). An overview of health analytics. *Journal of Health Medical Information, 4*(132), 2.

Sahoo, S. S., Wei, A., Valdez, J., Wang, L., Zonjy, B., Tatsuoka, C., ... Lhatoo, S. D. (2016). NeuroPigPen: A scalable toolkit for processing electrophysiological signal data in neuroscience applications using apache pig. *Frontiers in Neuroinformatics, 10*, 18.

Shameer, K., Johnson, K. W., Glicksberg, B. S., Dudley, J. T., & Sengupta, P. P. (2018). Machine learning in cardiovascular medicine: Are we there yet? *Heart, 104*(14), 1156–1164.

Srinivasan, R., & Swink, M. (2018). An investigation of visibility and flexibility as complements to supply chain analytics: An organizational information processing theory perspective. *Production and Operations Management, 27*(10), 1849–1867.
Toerper, M. F., Flanagan, E., Siddiqui, S., Appelbaum, J., Kasper, E. K., & Levin, S. (2015). Cardiac catheterization laboratory inpatient forecast tool: A prospective evaluation. *Journal of the American Medical Informatics Association, 23*(e1), e49–e57.
Toews, M., Wachinger, C., Estepar, R. S. J., & Wells, W. M. (2015). A feature-based approach to big data analysis of medical images. June *International Conference on Information Processing in Medical Imaging,* 339–350.
Tranfield, D., Denyer, D., & Smart, P. (2003). Towards a methodology for developing evidence-informed management knowledge by means of systematic review. *British Journal of Management, 14*(3), 207–222.
ur Rehman, M. H., Chang, V., Batool, A., & Wah, T. Y. (2016). Big data reduction framework for value creation in sustainable enterprises. *International Journal of Information Management, 36*(6), 917–928.
Raghupathi, V., & Raghupathi, W. (2014). An unstructured information management architecture approach to text analytics of Cancer blogs. *International Journal of Healthcare Information Systems and Informatics (IJHISI), 9*(2), 16–33.
Van Poucke, S., Zhang, Z., Schmitz, M., Vukicevic, M., Vander Laenen, M., Celi, L. A., ... De Deyne, C. (2016). Scalable predictive analysis in critically ill patients using a visual open data analysis platform. *PloS One, 11*(1) e0145791.
Vidgen, M., Shaw, S., & Grant, D. B. (2017). Management challenges in creating value from business analytics. *European Journal of Operational Research, 261*(2), 626–639.
Voisin, S., Pinto, F., Morin Ducote, G., Hudson, K. B., & Tourassi, G. D. (2013). Predicting diagnostic error in radiology via eye tracking and image analytics: Preliminary investigation in mammography. *Medical Physics, 40*(10).
Raghupathi, W., & Raghupathi, V. (2014). Big data analytics in healthcare: Promise and potential. *Health Information Science and Systems, 2*(1), 3.
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.
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.
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.
Wamba, S. F., Gunasekaran, A., Akter, S., Ren, S. J. F., Dubey, R., & Childe, S. J. (2017). Big data analytics and firm performance: Effects of dynamic capabilities. *Journal of Business Research, 70*, 356–365.
Wang, Y., Kung, L., Wang, W. Y. C., & Cegielski, C. G. (2018). An integrated big data analytics-enabled transformation model: Application to health care. *Information & Management, 55*(1), 64–79.
Ward, M. J., Marsolo, K. A., & Froehle, C. M. (2014). Applications of business analytics in healthcare. *Business Horizons, 57*(5), 571–582.
Yao, Q. A., Zheng, H., Xu, Z. Y., Wu, Q., Li, Z. W., & Lifen, Y. (2014). Massive medical images retrieval system based on Hadoop. *Journal of Multimedia, 9*(2), 216.
Yaqoob, I., Hashem, I. A. T., Gani, A., Mokhtar, S., Ahmed, E., Anuar, N. B., ... Vasilakos, A. V. (2016). Big data: From beginning to future. *International Journal of Information Management, 36*(6), 1231–1247.
Zhang, Y., & Li, X. (2017). Uses of information and communication technologies in HIV self-management: A systematic review of global literature. *International Journal of Information Management, 37*(2), 75–83.