Exploring the Potential Benefits of Big Data Analytics in Providing Smart Healthcare

Salma Al Mayahi, Ali Al-Badi, and Ali Tarhini(✉)

Information Systems Department, Sultan Qaboos University, Muscat, Oman
{salmam, aalbadi, alitarhini}@squ.edu.om

Abstract. Big Data is an emerging technology in different sectors. It refers to massive amount of heterogeneous data produced from many sources. Big data analytics is the process of analyzing a huge set of data to build and discover meaningful patterns, correlation that will add value to the corresponding business through predictive decisions and other useful information.

The health industry every years generate big data in different formats. The healthcare data need analysis to make decisions and forecasting but there is lack of understanding of the potential of big data in health industry.

This paper aims to explore the potential values of big data analytics in healthcare to enhance the efficiency and smartness of healthcare services. In addition, conducting an experiment on the dataset exported from an online healthcare research repository on a big data analytical topic.

The study has been carried out by conducting literature review big data analytic and referring datasets from online healthcare research repository.

The research concludes that providing evidence-based treatment, making predictive analysis and providing efficient healthcare service are the main potential benefits of applying analytics in healthcare. However, ensuring anonymity of patients' information and educating healthcare staff about the role of analytics in healthcare are essential steps before adopting such technologies.

This research is conceptual in nature based on existing literature reviews and secondary data. In future primary data would be used to understand the relevance of big data analytics in healthcare.

Keywords: Big data · Big data analytics · Healthcare · Predictive analysis
Biomedical informatics · Data mining

1 Introduction

Digitalization is becoming part of our lives penetrating every aspect of our normal life. Through the evolution of digitalization, a massive amount of data had been generated over the past years which remained beyond the capabilities of the available data storages and managements referred to as "Big Data" as described by IBM [1]. Big data includes the legacy enterprise data, machine generated data like sensors and weblogs and the data generated by social media like Twitter, Facebook and YouTube, etc. [2]. In 2001, big data defined by three main characteristics: (1) Volume which is the massive amount of data generated by industries, (2) Velocity which describes the fast movement of data among parties, (3) Varity which is the various type of data sources

M. H. Miraz et al. (Eds.): iCETiC 2018, LNICST 200, pp. 247–258, 2018.
https://doi.org/10.1007/978-3-319-95450-9_21

and types that includes social, mobile data, machine geographic data and biometrics [2]. In 2014, new characteristics of big data were proposed [3]:

- **Value**: How many values extracted from the data,
- **Veracity**: How accurate the data and how reliable is the source of data
- **Variability:** Consistency of the data and its continuity and availability
- **Viscosity:** Latency of data to the corresponding topic
- **Virality:** Rate of data spreading and how often the data is recurrent by another partner

This research aims to explore the potential benefits of big data analytics in healthcare by reviewing the previous studies and applications of this concept in health industry. The research is based on the study conducted between 2010 and 2017.

The paper is structured as follows: Sect. 2 presents the literature review. Section 3 describes the research methodology used. Section 4 presents the results. Section 5 provides a detailed discussion on the findings.

2 Literature Review

Big data will not be useful unless it provides information and meaningful and readable insights. Big data analytics is the process of constructing valuable patterns, useful information, and descriptive trends from the pool of data. Health industry is one of the major area that is growing exponentially and producing big data in the form of patient information, clinical notes, X-ray imaging and pharmaceutical data. In 2012, Bonnie Feldman said that it is expected that data over the world in healthcare will be 50% more than the current data and it will reach 25000 petabytes whereas in 2012 it was only 500 petabytes [4]. In addition, the unstructured data like written clinical notes, video and audio streams will increase 15 times more than the structured data along with the new forms of data beside the existing types like human genetics data, radiology images and biometric, and genomics readings [4].

There are tremendous benefits and huge advantages of big data analytics in healthcare sector will reach the level of disease investigation to the level of treatment. In US, there is a high demand to healthcare big data analytics since the expenses has been increasing rabidly in last decades [5]. In 2011, Manyika from McKinsey Global Institute proposed that if USA healthcare applied effective and innovative big data analytics, USA could save $300 billion every year [6]. Furthermore, big data analytics creates transparency and easy accessibility to relevant data in order to create more values and facilitates analytical experimentation to investigate needs and supports leader's decisions as an evidence [6]. The new innovative pathway of applying big data analytics in healthcare include the following benefits [5]:

- **Right living:** The advancing the lifestyle by engaging the end users in the health care process will eventually minimize the needed care by industry.
- **Right Care:** Care is provided to the patients based on evidence and this will ensure more safety and enhance the expected results.

- **Right provider:** The provider selection will be more accurate and give insights toward quality.
- **Right Value:** The analytics will help reducing the care cost and yet maintain the quality.
- **Right innovation:** The analytics will help encouraging innovative health solution, discovering new correlations and making new trends.

There are many promising applications of big data analytics proving the potential values of this technology in healthcare [2].

2.1 Challenges of Big Data Analytics

There are many challenges in implementing big data analytics due to the high complexity and diversity of the healthcare data sets [7]. As stated by Ward, Marsolo and Froehle in their study in 2104, the challenges to the applications of big data analytics include (1) no standard protocols of data structures (2) data collection obstacles (3) lack of qualified big data analysts [8]. A number of researchers from University of Otago, New Zealand, stated that they faced big challenge in managing the big data within the scope of the project in collaborative way [7]. Another challenge raised by LaValle from MIT Sloan Management Review is that the adoption process of data analytics might face data quality issues, unproductive data governance and management barriers [9].

2.2 Advantages of Big Data Analytics to Healthcare

LaValle from MIT Sloan Management Review partnered with the IBM stated that better utilization of the available technologies and tools is essential to leverage the healthcare data effectively and to help organization realize what is currently happening around and predicts what is most likely to happen to take proactive actions and be well prepared [9]. MIT Sloan Management Review conducted a wide survey among 3000 executives and managers in 100 countries with different sectors and interests and the key finding they confirmed that top-performing business utilizes the analytics in their data five times more than their opponents who perform less [9].

Raghupathi and Viju in their review in 2014 are supporting LaValle from MIT Sloan Management Review research outcomes of the survey by stating that big data analytics potential benefits are promising the healthcare industry with valuable outcomes and valuable results [10]. In addition, the use of analytics help organization converting challenges into opportunities by constructing future strategies and making day-to-day processes guide [9]. Manyika and colleagues from McKinsey Global Institute (MGI) stated in their report in June 2011 that big data analytics in healthcare have potential values and extraordinary results in reducing the cost, increasing the revenues and the efficiency and improving effectiveness of patient treatment [6]. Manyika proposed the benefits of big data analytics in healthcare into five categories. The first category is improving the use of clinical operations in better treatment and effective diagnosis of the diseases. The second category is reducing the treatments' cost and overall healthcare expenditure. The third category is raising the value of research and development by building predictive modules and developing new algorithms that

will improve the clinical design. The forth category is building new business models in healthcare industry. The fifth category is improving public health surveillance and responses [6].

In 2014, a group of researchers from Health Affairs in US conducted a research by applying big data analytics to identify the opportunity of reducing the healthcare costs. The research goals achieved by predicting the number of readmissions and high cost patients in six different use cases of inpatients records and they conclude that such analytics is considered as powerful tool to be adopted in healthcare industry [11]. Moreover, a recent literature review conducted in 2016 among 209 articles provided an obvious evidence of an exponential positive impact of big data analytics in increasing the accuracy and quality in healthcare services and reducing the costs of clinical analysis [12].

In 2014, a study conducted to experiment the potential benefits of big data analytics on electronic health records (HER) to build predictive models using three databases from different health systems and proved that the efficiency of building research models increased and the reuse of the health data helped in creating useful research objectives [13]. In 2016, Srinivasan Suresh stated that predictive analysis can be utilized to help healthcare providers to better prescribe medication to children and to give better awareness to the patients about their health which will lead to improvement in the collaboration between the physicians and patients and improve their life style habits [14].

2.3 Worldwide Applications of Big Data Analytics in Healthcare

Nowadays, health organizations around the world are starting to realize the value of big data analytics to enhance their productivity in healthcare. In 2013, United Kingdom initiated a project called care.data which aims to link all patients health records with the social healthcare in a centralized place owned by the Health and Social Care Information Centre (HSCIC) [15]. The main objectives of care.data project is to make predictive analysis for the various disease among UK citizens and to help the government to provide better healthcare services with evidence based treatments [15, 16]. In 2012, National institute of health in US lunched BigData to Knowledge (BD2K) and aimed to develop innovative tools and methods to utilize the biomedical big data into useful knowledge in healthcare. In 2015, a team from Oxford university conducted a project with the BD2K Center for Causal Discovery to test and train biomedical data consisting of cancer driver, lung disease information and human brain data by developing a new algorithm with a user-friendly system based on Bayesian algorithms "a datamining technique" [17]. In 2013, Ahumada (2013) proposed a medication alert fatigue application which aimed to discover and assess the medication alerts in Children's Hospital of Philadelphia [18]. This application implemented an analytical dashboard with a user-friendly interface to investigate drug-allergy and maximum dose alerts that helps in clinical decision supports and real time management.

In 2015, IBM Watson Health project is launched and considered to be one of the innovative approaches toward raising up the level of efficiency in healthcare in more simple and creative way [19]. IBM Watson is a cloud based project which is capable of storing a large size of structured and unstructured data, evaluate them and provide evidence based results [19]. Furthermore, Watson health is a cognitive system that play

a significant role to uncover the value hidden in the unstructured data in order to provide evidence based reasoning to support healthcare staff to make proper decisions [20]. As an evidence, more than 80% of Watson health's executives show positive influence in their running business [20]. In 2015, a practical study conducted by Househ, Hasman and Mantas from Germany Heilbronn university to predict the survival rate of the colon cancer from a predefined attribute based on medical experts' opinions. The objective of the project is to compare the data mining algorithms' accuracy with the physician's accuracy and their results shows that data mining algorithms an accuracy of about 67.7% comparing to 59% accuracy of the physician's [21]. There are many other projects and practical researches done in big data analytics nowadays around the world. Table 1 summarizes studies on valuable projects, areas, limitations and benefits in healthcare.

Table 1. Worldwide applications of big data analytics in healthcare

Project name	Project place	Technology used	Limitation	Benefits
Care.data	UK	NA	Mismanagement, miscommunications, inadequate protections for patient anonymity, conflicts with doctors	1. Predict diseases pattern 2. Future plans for health services for each area in UK 3. Propose better treatment 4. Make evidence based treatment
BD2K	US	Casual Bayesian algorithms	Might not fit in all biomedical domains	1. Predict the drivers of cancer disease 2. Discover factors leading to lung disease
Medication alert fatigue	Philadelphia	SQL queries of Children's Hospital of Philadelphia EHR database		1. Better surveillance on the medication alerts 2. Helps in Day-to-Day management
IBM Watson Health		Deep QA		1. Provide evidence based result 2. Support decision making for healthcare staff

(*continued*)

Table 1. (*continued*)

Project name	Project place	Technology used	Limitation	Benefits
Survival prediction	Germany	Data mining algorithms		1. Good prediction analysis 2. Support decision making 3. Can fit to other problems in health prediction 4. High accuracy

3 Research Methodology

The research started by conducting a thorough investigation using systematic literature review to explore the potential benefits achieved by applying big data analytics in healthcare domain. Next step after reviewing was building the conceptual understanding, discussion, and evaluation of the main benefits. This step is followed by preparing the dataset, the analytical tool and the techniques to be used in conducting the experiment. After that, a pilot study on a small amount of data to test the methodology and to make any adjustment in the experiment methodology before conducting the main study performed. This step helps to visualize constrains and obstacles that might show up during the main study. Finally, conducting the main study, analyze the results, discuss and map the output with the findings from the literature review.

This research focuses on the main potential benefits of big data analytics in healthcare collected from various research papers, reviews and projects. Figure 1 shows the considered benefits in the present study.

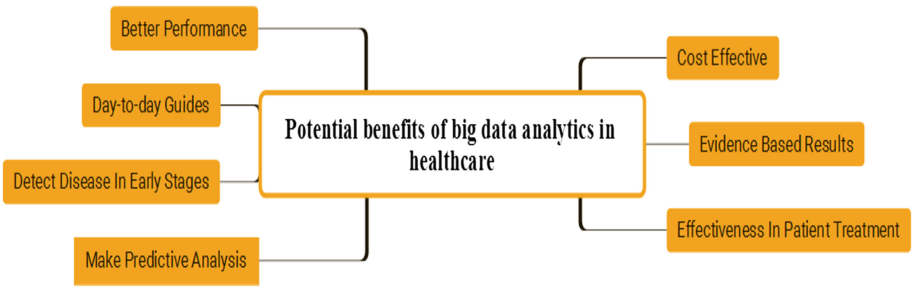

Fig. 1. Potential benefits of big data analytics in healthcare

4 Experimental Investigation on Big Data Analytics in Healthcare

This experiment aims to elaborate deeply on the idea of big data analytics in healthcare, to investigate the anticipated potential benefits in healthcare industry, and to support the findings from the literature review. The experiment is conducted using a real dataset about chronic kidney disease exported from UC Irvine Machine Learning Repository [22] and analyzed using an open source tool called Weka explorer version 3.8. This tool is used to apply two classification algorithms: naïve Bayesian and J48 trees. Classification is a data mining technique used to build a model that describes the data by analyzing and supervised learning of training data in which the class label is known and the resulted model in this case is then used for predicting whether the patient has a chronic kidney disease or not [23].

The exported dataset of chronic kidney disease is used to predict the disease and it consists of 25 attributes and 400 instances [22]. The dataset originally contains 400 as taken from the online repository, but when applied data mining techniques it reduced to 160 due to pre-possessing of the records. The list of attributes, description and the used abbreviations shown in the Table 2.

Table 2. Dataset of chronic kidney disease description

S. No	Abbreviation	Description
1	age	Age
2	bp	Blood Pressure
3	sg	Specific Gravity
4	al	Albumin
5	su	Sugar
6	rbc	Red Blood Cells
7	pc	Pus Cell
8	pcc	Pus Cell Clumps
9	ba	Bacteria
10	bgr	Blood Glucose Random
11	Bu	Blood Urea
12	Sc	Serum Creatinine
13	sod	Sodium
14	pot	Potassium
15	hemo	Hemoglobin
16	pcv	Packed Cell Volume
17	wc	White Blood Cell Count
18	rc	Red Blood Cell Count
19	htn	Hypertension
20	dm	Diabetes Mellitus
21	cad	Coronary Artery Disease

(*continued*)

Table 2. (*continued*)

S. No	Abbreviation	Description
22	appet	Appetite
23	pe	Pedal Edema
24	ane	Anemia
25	class	Class

The first classification algorithm to be applied is naïve Bayesian which is considered to be a very powerful algorithm which deals with each attribute independently and it performs well in the dataset that contains missing values [24]. After applying this algorithm, the resulted output is shown in the Fig. 2.

```
=== Summary ===

Correctly Classified Instances         152               95      %
Incorrectly Classified Instances         8                5      %
Kappa statistic                          0.8943
Mean absolute error                      0.0493
Root mean squared error                  0.2043
Relative absolute error                 10.5446 %
Root relative squared error             42.5665 %
Total Number of Instances              160

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
               0.922     0.000     1.000      0.922     0.960      0.899     1.000     1.000     ckd
               1.000     0.078     0.877      1.000     0.934      0.899     1.000     1.000     notckd
Weighted Avg.  0.950     0.028     0.956      0.950     0.951      0.899     1.000     1.000

=== Confusion Matrix ===

  a  b   <-- classified as
 95  8 |  a = ckd
  0 57 |  b = notckd
```

Fig. 2. Naïve Bayesian algorithm result

From the results of this algorithm, the achieved percentage of the corrected classified instances is 95%, which indicates a very good results and it gives the anticipated results using percentage split 60%. From the resulted confusion matrix, only 8 records were incorrectly classified. After increasing the percentage splits to explore more accuracy, the following Table 3 shows the results achieved. With 60% percentage split, the accuracy is more and the recall is better.

Table 3. Percentage spilt and achieved results

Percentage split	Accuracy	Error rate	Sensitivity	Specificity	Precision	Recall
60%	95%	5%	0.950	1 − 0.028 = 0.972	0.956	0.950
80%	92.5%	7.5%	0.925	1 − 0.040 = 0.960	0.938	0.925
90%	90%	10%	0.900	1 − 0.067 = 0.933	0.920	0.900

Figure 2 represents the readings when the splits percentage set to 60% only but the other readings of Table 3 shows the results when we change the split percentage to different one but at the end we achieve a better result in 60% split percentage which is represented in Fig. 2. In Fig. 2 also we refer to the readings of weighted average. The second classification method is J48 which is a decision tree algorithm that is constructed from a learning process of the given data and used widely in medicine, financial and biology [23]. After applying the algorithm to the chorionic kidney disease dataset, the following Fig. 3 shows the great result obtained by this algorithm. Obviously, this algorithm performed perfectly in chronic kidney disease dataset and 100% accuracy achieved. As shown from the confusion matrix, all the instances were classified correctly. Figure 4 shows resulted decision tree.

```
=== Summary ===

Correctly Classified Instances         160               100      %
Incorrectly Classified Instances         0                 0      %
Kappa statistic                          1
Mean absolute error                      0.0218
Root mean squared error                  0.0856
Relative absolute error                  4.6561 %
Root relative squared error             17.8392 %
Total Number of Instances              160

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     ckd
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     notckd
Weighted Avg.    1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000

=== Confusion Matrix ===

   a   b   <-- classified as
 103   0 |   a = ckd
   0  57 |   b = notckd
```

Fig. 3. J48 algorithm result

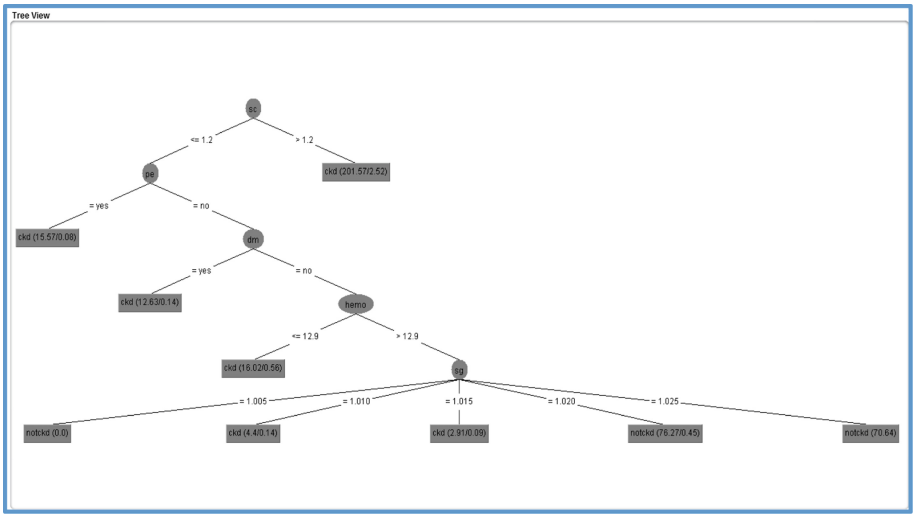

Fig. 4. Decision tree result

All the 25 attributes from Table 2 attributes are included in the data mining process but the tree have been created based on the significance of the attribute in the result whether to have kidney disease or not. From the tree, the main factors of the chronic kidney disease to be taken into consideration. The performance of the two algorithms is very strong and could be used to train different health problems datasets. As Patil and Sherekar proved in their research in 2013, both algorithms are efficient and give high accuracy [24].

5 Analysis and Discussion

From the exploration of big data analytics in healthcare in the literature review, the main benefits that most of the researchers have reached consensus are on providing evidence based treatment, improving patients' treatment, making predictive analysis and supporting decision making in healthcare. As a clear evidence, the experimental investigation conducted supports the findings of the literature review. The Table 4 explains the mapping between the benefits from the literature review and the experimental investigation of big data analytics using chronic kidney disease dataset.

Table 4. Mapping between the benefits from the literature review and the experimental investigation

SN	The benefit from literature review	References	Experimental investigation
1	Better performance	[6, 9, 19, 21]	The experiment shows high accuracy and the running time in that dataset was efficient. It took 0.01 s only but it may vary depending on the size of the dataset. Moreover, the graphical representation of the disease factors helps the doctor to grasp the knowledge in more efficient way
2	Day-to-Day guides	[6, 9, 18]	The medical staff could benefit from the classification results in their daily diagnosis process with different patients' cases
3	Detect disease in early stages	[14, 18, 21]	There is no obvious evidence from the experiment to achieve this benefit but it might be possible after using other type of analytics
4	Make predictive analysis	[6, 11, 12, 16, 21]	The dataset is trained by the classification algorithms to provide predictive analysis based on its attributes
5	Cost effective	[7, 11, 12]	The experiment shows very good performance comparing to reviewing the historical data using manual routines. This helps in saving medical staff time and hence, it is cost effective

(*continued*)

Table 4. (*continued*)

SN	The benefit from literature review	References	Experimental investigation
6	Evidence based results	[6, 15, 16, 19]	From the historical dataset used in the experiment, it might be used as an evidence to prove the achieved results and the decision taken by the medical staff
7	Effectiveness in patient treatment	[6, 10, 12, 21]	From the machine learning in this experiment, it shows high accuracy for prediction and this helps in reducing the errors while treating the patients

In addition, the literature review shows that big data analytics in healthcare is widely applied in different areas around the world. It shows that healthcare analytics provide positive impact and great benefits in healthcare sector. As highlighted in the review, there are some limitations that must be considered in adopting analytics in healthcare. The most important concern is maintaining the anonymity of the patients' information. Another concern is to avoid adoption's conflicts that can be resolved by developing well-managed project documentations and educating the healthcare staff on the potential benefits and application of big data analytics. In addition, analysts must have eligible skills and experience in analyzing healthcare datasets and medical expertise to understand the dataset's variables and correlations.

6 Conclusion

Big data analytics has a tremendous potential benefits in healthcare that will smarten the services and increase the services' efficiency. This research explored the anticipated benefits of applying the analytics in healthcare and looked into the various applications around the world. Moreover, an experiment conducted to investigate the potential benefits using a real dataset and it showed very promising results. Adopting the analytics in healthcare is essential but the highlighted limitations and challenges must be well addressed and resolved. Moreover, there are research topic, which has scope in healthcare such as apply different data mining algorithms and measure their accuracy; apply the experiment on different datasets in healthcare with supervision of a medical expert in the selected domain; analyze unstructured dataset like doctors' clinical notes in a local hospital; select different analytical tool and try to investigate other results.

References

1. Dewey, J.: Big Data. Salem Press Encyclopedia (2014)
2. Priyanka, K., Kulennavar, N.: A survey on big data analytics in health care. Int. J. Comput. Sci. Inf. Technol. **5**(4), 5865–5868 (2014)
3. Vorhies, W.: How many V's in big data? The characteristics that define big data. Data Science Central, October 2014. http://www.datasciencecentral.com/profiles/blogs/how-many-vs-in-big-data-the-characteristics-that-define-big-data

4. Feldman, B., Martin, E.M., Skotnes, T.: Big data in healthcare hype and hope. Dr. Bonnie **360** (2012)
5. Groves, P., et al.: The 'big data' revolution in healthcare. McKinsey Q. **2** (2013)
6. Manyika, J.: Big Data the Next Frontier for Innovation, Competition & Productivity. McKinsey & Company, S.L. (2011)
7. Frost, S.: Drowning in big data? Reducing information technology complexities and costs for healthcare organizations (2015)
8. Ward, M.J., Marsolo, K.A., Froehle, C.M.: Applications of business analytics in healthcare. Bus. Horiz. **57**(5), 571–582 (2014)
9. LaValle, S., et al.: Big data, analytics and the path from insights to value. MIT Sloan Manag. Rev. **52**(2), 21 (2011)
10. Raghupathi, W., Raghupathi, V.: Big data analytics in healthcare: promise and potential. Health Inf. Sci. Syst. **2**(1), 1 (2014)
11. Bates, D.W., et al.: Big data in health care: using analytics to identify and manage high-risk and high-cost patients. Health Aff. **33**(7), 1123–1131 (2014)
12. de la Torre Díez, I., et al.: Big data in health: a literature review from the year 2005. J. Med. Syst. **40**(9), 209 (2016)
13. Ng, K., et al.: PARAMO: a PARAllel predictive MOdeling platform for healthcare analytic research using electronic health records. J. Biomed. Inform. **48**, 160–170 (2014)
14. Suresh, S.: Big data and predictive analytics: applications in the care of children. Pediatr. Clin. North Am. **63**(2), 357–366 (2016)
15. Patient: Care.data - sharing your information (2014)
16. Sterckx, S., et al.: "You hoped we would sleep walk into accepting the collection of our data": controversies surrounding the UK care.data scheme and their wider relevance for biomedical research. Med. Health Care Philos. **19**(2), 177–190 (2016)
17. Cooper, G.F., et al.: The center for causal discovery of biomedical knowledge from big data. J. Am. Med. Inform. Assoc. **22**(6), 1132–1136 (2015). https://doi.org/10.1093/jamia/ocv059
18. Ahumada, L.M., et al.: Medication alert fatigue: the design and use of a medication alert dashboard as part of a comprehensive approach to drug-drug interaction alerts. In: Anesthesia and Analgesia. Lippincott Williams & Wilkins, Philadelphia (2013)
19. Aggarwal, M., Madhukar, M.: IBM's Watson analytics for health care: a miracle made true. In: Cloud Computing Systems and Applications in Healthcare, pp. 117–134 (2016)
20. IBM: IBM Watson Health: A New Partnership Between Humanity and Technology (2016). http://www.ibm.com/watson/health/. Accessed 12 June 2016
21. Househ, M., Hasman, A., Mantas, J.: Enabling Health Informatics Applications. Studies in Health Technology and Informatics. IOS Press, Amsterdam (2015)
22. UCI: UC Irvine Machine Learning Repository, Chronic_Kidney_Disease Data Set (2015). http://archive.ics.uci.edu/ml/datasets/Chronic_Kidney_Disease. Accessed 10 Dec 2016
23. Marston, S., et al.: Cloud computing—the business perspective. Decis. Support Syst. **51**(1), 176–189 (2011)
24. Patil, T.R., Sherekar, S.: Performance analysis of Naive Bayes and J48 classification algorithm for data classification. Int. J. Comput. Sci. Appl. **6**(2), 256–261 (2013)