

BIG DATA ANALYTICS IN HEALTH CARE: PROMISE AND POTENTIAL

Foundation of Data Analytics



July 16, 2019

18BCE7244

SHANMUKH SAI THUMMALA

**Abstract**

**Objective:** To describe the promise and potential of big data analytics in healthcare.

**Methods:** The paper describes the nascent field of big data analytics in healthcare, discusses the benefits, outlines an architectural framework and methodology, describes examples reported in the literature, briefly discusses the challenges, and offers conclusions.

**Results:** The paper provides a broad overview of big data analytics for healthcare researchers and practitioners.

**Conclusions:** Big data analytics in healthcare is evolving into a promising field for providing insight from very large data sets and improving outcomes while reducing costs. Its potential is great; however there remain challenges to overcome.

**Keywords:** Big data, Analytics, Hadoop, Healthcare, Framework, Methodology

**Introduction**

The healthcare industry historically has generated large amounts of data, driven by record keeping, compliance & regulatory requirements, and patient care. While most data is stored in hard copy form, the current trend is toward rapid digitization of these large amounts of data. Driven by mandatory requirements and the potential to improve the quality of healthcare delivery meanwhile reducing the costs, these massive quantities of data (known as ‘big data’) hold the promise of supporting a wide range of medical and healthcare functions, including among others clinical decision support, disease surveillance, and population health management. Reports say data from the U.S. healthcare system alone reached, in 2011, 150 exabytes. At this rate of growth, big data for U.S. healthcare will soon reach the zettabyte (1021 gigabytes) scale and, not long after, the yottabyte (1024 gigabytes). Kaiser Permanente, the California-based health network, which has more than 9 million members, is believed to have between 26.5 and 44 petabytes of potentially rich data from EHRs, including images and annotations.

By definition, big data in healthcare refers to electronic health data sets so large and complex that they are difficult (or impossible) to manage with traditional software and/ or hardware; nor can they be easily managed with traditional or common data management tools and method. Big data in healthcare is overwhelming not only because of its volume but also because of the diversity of data types and the speed at which it must be managed. The totality of data related to patient healthcare and wellbeing make up “big data” in the healthcare industry. It includes clinical data from CPOE and clinical decision support systems (physician’s written notes and prescriptions, medical imaging, laboratory, pharmacy, insurance, and other administrative data); patient data in electronic patient records (EPRs); machine generated/sensor data, such as from monitoring vital signs; social media posts, including Twitter feeds (so-called tweets), blogs, status updates on Facebook and other platforms, and web pages; and less patient-specific information, including emergency care data, news feeds, and articles in medical journals. For the big data scientist, there is, amongst this vast amount and array of data, opportunity. By discovering associations and understanding patterns and trends within the data, big data analytics has the potential to improve care, save lives and lower costs. Thus, big data analytics applications in healthcare take advantage of the explosion in data to extract insights for making better informed decisions, and as a research category are referred to as, no surprise here, big data analytics in healthcare. When big data is synthesized and analysed and those aforementioned associations, patterns and trends revealed—healthcare providers and other stakeholders in the healthcare delivery system can develop more thorough and insightful diagnoses and treatments, resulting, one would expect, in higher quality care at lower costs and in better outcomes overall. The potential for big data analytics in healthcare to lead to better outcomes exists across many scenarios, for example: by analysing patient characteristics and the cost and outcomes of care to identify the most clinically and cost effective treatments and offer analysis and tools, thereby influencing provider behaviour; applying advanced analytics to patient profiles (e.g., segmentation and predictive modelling) to proactively identify individuals who would benefit from preventative care or lifestyle changes; broad scale disease profiling to identify predictive events and support prevention initiatives; collecting and publishing data on medical procedures, thus assisting patients in determining the care protocols or regimens that offer the best value; identifying, predicting and minimizing fraud by implementing advanced analytic systems for fraud detection and checking the accuracy and consistency of claims; and, implementing much nearer to real-time, claim authorization; creating new revenue streams by aggregating and synthesizing patient clinical records and claims data sets to provide data and services to third parties, for example, licensing data to assist pharmaceutical companies in identifying patients for inclusion in clinical trials. Many payers are developing and deploying mobile apps that help patients manage their care, locate providers and improve their health. Via analytics, payers are able to monitor adherence to drug and treatment regimens and detect trends that lead to individual and population wellness benefits. This article provides an overview of big data analytics in healthcare as it is emerging as a discipline. First, we define and discuss the various advantages and characteristics of big data analytics in healthcare. Then we describe the architectural framework of big data analytics in healthcare. Third, the big data analytics application development methodology is described. Fourth, we provide examples of big data analytics in healthcare reported in the literature. Fifth, the challenges are identified. Lastly, we offer conclusions and future directions.

The rapid development of the emerging information technologies, experimental technologies and methods, cloud computing, the Internet of Things, social networks supplies the amounts of generated data that is growing tremendously in numerous research fields.

On this point, contemporarily genomics and post-genomics technologies produce huge amounts of raw data about complex biochemical and regulatory processes in the living organisms. These high throughput – omics data provide comprehensive insight towards different kinds of molecular profiles, changes and interactions, such as knowledge allied to the genome, epigenome, transcriptome, proteome, metabolome, interactome, pharma-cogenome , diseasome, etc. These – omics data are heterogeneous and very often stored in different data formats.

**Big data analytics in healthcare**

Health data volume is expected to grow dramatically in the years ahead. In addition, healthcare reimbursement models are changing; meaningful use and pay for performance are emerging as critical new factors in today’s healthcare environment. Although profit is not and should not be a primary motivator, it is vitally important for healthcare organizations to acquire the available tools, infrastructure, and techniques to leverage big data effectively or else risk losing potentially millions of dollars in revenue and profits. What exactly is big data? A report delivered to the U.S. Congress in August 2012 defines big data as “large volumes of high velocity, complex, and variable data that require advanced techniques and technologies to enable the capture, storage, distribution, management and analysis of the information”. Big data encompasses such characteristics as variety, velocity and, with respect specifically to healthcare, veracity. Existing analytical techniques can be applied to the vast amount of existing (but currently unanalysed) patient-related health and medical data to reach a deeper understanding of outcomes, which then can be applied at the point of care. Ideally, individual and population data would inform each physician and her patient during the decision-making process and help determine the most appropriate treatment option for that particular patient.

Applications of big data analytics can improve the patient-based service, to detect spreading diseases earlier, generate new insights into disease mechanisms, monitor the quality of the medical and healthcare institutions as well as provide better treatment methods.

Data mining techniques employed on EHRs, web and social media data enable identifying the optimal practical guidelines in the hospitals, identifying the association rules in the EHRs and revealing the disease monitoring and health-based trends. Moreover, integration and analysis of the data with different nature, such as social and scientific, can lead to new knowledge and intelligence, exploring new hypothesis, identifying hidden patterns.

Nowadays, smart phones are excellent platforms to deliver personal messages to patients to involve them in behavioural changes to improve their wellbeing and health conditions. The mobile phone messages can substitute delivering of medical and motivational advices to the patients.

**Related Technologies**

**A. Big data platforms**

As in report, big data uses distributed storage technology based on cloud computing rather than local storage. Some big data cloud platforms are Google cloud services, Amazon S3 and Microsoft Azure. Google’s distributed file system GFS (Google File System) and its programming model Mapreduce are the lead in the field. The performance of mapreduce has received a valid amount of attention in large scale data processing. So many organizations use big data processing framework with map reduce. Hadoop, an influential aspect in big data was developed by Yahoo and it is an open-source version of GFS. Hadoop enables storing and processing big data in distributed environment on large clusters of hardware. Enormous data storage and faster processing are supported by hadoop. Hadoop Distributed File System (HDFS) provides reliable and scalable data storage. HDFS makes multiple copies of each data block and distributes them on systems on a cluster to enable reliable access. HDFS supports cloud computing through the use hadoop, a distributed data processing platform. Another one, ‘Big Table’ was developed by Google in 2006 that is used to process huge amount of structured data. It also supports map reduce.

Amazon developed Dynamo, a key-value pair storage system. It is a scalable distributed data store built for Amazon’s platform. It gives high reliability, cost effectiveness, availability and performance. Tom white elaborates various tools for big data analytics. Hive, a framework for data warehousing on top of handoop. It was built at Facebook. Hive with hadoop for storage and processing meets the scalability needs and is cost-effective. Hive uses a query language called HiveQL which is alike on SQL.

A scripting language for exploring large datasets is called ‘Pig’. An opinion of map reduce is that writing of mappers and reducers, compiling the package and code are tough and so the development cycle is long. Hence working with mapreduce needs experience. Pig overcomes this criticism by its simplicity. It allows the developers to write simple Pig Latin queries to process the big data and thereby save the time. A distributed column oriented database Hbase built on top of Hadoop Distributed File System. It can be used when we need random access of very large datasets. It speeds up the performance of operations. Hbase can be accessed through application programming interfaces (APIs) like REST (Representational State Transfer) and java. Hbase does not have its own queries, so it depends on Zookeeper. Zookeeper manages huge amount of data. This allows distributed process to manage through a namespace of data registers. This distributed service also has master and slave nodes like in hadoop. Another important tool is Mahout. It is a data mining and machine learning library. It can be categorized as collective filtering, categorization, clustering and mining. It can be executed by Mapreduce in a distributed mode. Big data analytics is not only based on platforms but also analytics algorithms plays a significant role.

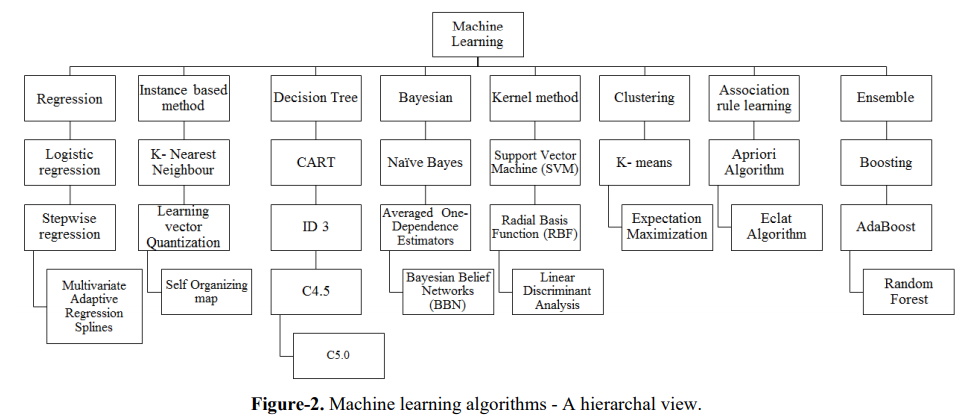
**B. Algorithmic techniques**

Unknown but useful information from massive amount of data. This information can be used to predict future situations as a help to decision making process. Helpful knowledge can be found by the usage of data mining techniques in healthcare applications like decision support system. The big data produced by healthcare organizations are very complicated and vast to be handled and analysed by usual methods. Data mining grants the procedure to transform those bundles of data into useful information for decision support. Big data mining in healthcare is about learning models to predict patients’ disease. For example, data mining can help healthcare insurance organizations to detect hypocrites and misuse, healthcare institutions make decisions of customer relationship management, doctors identify effective treatments and best practices, and patients get improved and more economical healthcare services. This predictive analysis is widely used in healthcare.

There are various data mining algorithms discussed in ‘Top 10 algorithms in data mining’ by Wu X et al. It discussed variety of algorithms along with their limitations. Those algorithms encompass clustering, classification, regression, statistical learning which are the issues in data mining investigation. The ten algorithms discussed include C4.5, k-means, Apriori, Support Vector Machines, Naïve Bayes, EM, CART, etc.

Big data analytics includes various methods such as text analytics, multimedia analytics and so on. But as given above, one of the crucial categories is predictive analytics which includes various statistical methods from modelling, data mining and machine learning that analyse current and historical facts to make prediction about future. In hospital context, there are predictive methods used to identify if someone may be at risk for readmission or is on a serious recession. This data helps therapists to make important care decisions. Here it is necessary to know about machine learning since it is widely employed in predictive analysis.

The process of machine learning is very much alike of data mining. Both of them hunt through data to look for patterns. But, rather than extracting data for human understanding as in data mining applications, machine learning model uses that data to improve the program's own understanding. Machine learning programs finds patterns in data and alters program functions respectively. Machine learning provides various algorithms. Jason Brownlee illustrates different machine learning strategies. The hierarchal structure of various algorithms is given below.



Hall et al. outlined a method for building learning the rules of the large dataset. The approach is to have a single decision scheme made from a huge subset of data. Meanwhile Patil et al. pursued a hybrid way pairing the two genetic algorithm and decision tree to make an advanced decision tree to improve performance and efficiency of computation. With the increasing knowledge in the area of big data, the variety of techniques for analysing data is represented in ‘Data Reduction Techniques for Large Qualitative Data Sets’. It describes that the selection for the particular technique is based on the type of dataset and the way the pattern are to be analysed. Jakrarin et al. [10] applied K-means clustering using Apache Hadoop. They aimed at efficiently analyzing large dataset in a minimal amount of time. They also explained that the accuracy and detection rate are affected by the number of fields in log files or not. Therefore, their tested results show the correct number of clusters and the correct amount of entries in log files, but the rate of accuracy reduces when the number of entries increases. The result shows that the accuracy needs to be improved.

Classification is one of the data mining techniques used to predict and classify the predetermined data for the specific class. There are different classifications methods propose by researchers. The widely used methods are described by Han et al. It includes the following:

* Bayesian classification
* Neural network algorithm
* Decision tree induction
* Rule based classification
* Support vector machine
* K-Nearest neighbour classifier
* Rough set approach
* Genetic algorithm
* Fuzzy set approach

Any one of the above mentioned classification techniques can be applied to classify the application oriented data. The applicable classification method is to be chosen according to the type of application and the dimensionality of the data. It is a very big challenge to the researchers to select and apply the appropriate data mining classification algorithm for diagnosing medical related problems. Choosing the correct method is a challenging task. The exact method can be chosen only after analysing all the available classification methods and checking its performance in term of accuracy. Various researches have been carried out in the area of medical diagnoses by using classification methodology. The most important fact in medical diagnosis system is the accuracy of the classifier. This research paper analyses the different classification methods applied in medical diagnoses and compares the performance of classification accuracy.

C4.5 is applied to analyse the SEER dataset for breast cancer and classify the patients either in the beginning stage or pre cancer stage. The records analysed are 500 and the accuracy achieved in testing datasets is 93%.

et al. used the Naïve Bayes, ANN, C4.5 and decision tree algorithms for diagnoses and prognoses of breast cancer. The results show that the decision trees give higher accuracy of 93.62 % where Naïve Bayes gives 84.5%, ANN produces 86.5% and C4.5 generates 86.7% of accuracy. Chaitrali et al. [4] used Decision Trees, Naïve Bayes and Neural Network algorithms for analysing heart disease. The results comparison tells that the Naïve Bayes achieves 90.74% of accuracy whereas Neural Network and Decision Trees give 100% and 99.62% of accuracy respectively.

Different data mining techniques were applied to predict heart disease in [24]. The accuracy of each algorithm is verified and stated as Naïve Bayes, Decision Tree and ANN are achieved 86.53%, 89% and 85.53% of accuracy respectively. The three different data mining algorithms, ANN, C4.5 and Decision Trees are used to analyse heart related diseases by using ECG signals. The analysis results clearly show the Decision Tree algorithm performs best and gives the accuracy of 97.5%. C4.5 algorithm gives 99.20% accuracy while Naïve Bayes algorithm gives 89.60 % of accuracy in [2]. Here these algorithms are used to estimate the supervision of liver disorder. Christobel et al. applied KNN method to diabetic dataset. It gives the accuracy of 71.94% with 10 fold cross validation.

C5.0 is the classification algorithm which is applicable for big data sets. It overcomes C4.5 on the speed, memory and the performance. C5.0 method works by splitting the sample based on the field that gives the maximum information gain. The C5.0 system can split samples regarding of the biggest information gain field. The sample subset that is got from the previous split will be split later. The action will continue until the sample subset cannot be split and is usually according to another field. Finally, consider the lowest level split, those sample subsets that do not have notable contribution to the model will be dropped. C5.0 approach easily handles the multi value attribute and missing attribute from data set. The C5.0 rule sets have noticeably lowers error rates on unseen cases for the sleep and forest datasets. The C4.5 and C5.0 rule sets have the same predictive accuracy for the income dataset, but the C5.0 rule set is smaller. The times are almost not comparable. For instance, C4.5 required nearly 15 hours finding the rule set for forest, but C5.0 completed the task in 2.5 minutes. C5.0 commonly uses an order of magnitude less memory than C4.5 during rule set construction. So it is clear that C5.0 approach is better than C4.5 in many aspects.

Hsi-Jen Chiang et al. proposed a method for analyzing prognostic indicators in dental implant therapy. They analyze 1161 implants from 513 patients. Data on 23 items are taken as impact factors on dental implants. These 1161 implants are analyzed using C5.0 method. Here 25 nodes are produced by using C5.0 approach. This model achieves the performance of 97.67% accuracy and 99.15% of specificity.

**Advantages to healthcare**

By digitizing, combining and effectively using big data, healthcare organizations ranging from single-physician offices and multi-provider groups to large hospital networks and accountable care organizations stand to realize significant benefits. Potential benefits include detecting diseases at earlier stages when they can be treated more easily and effectively; managing specific individual and population health and detecting health care fraud more quickly and efficiently. Numerous questions can be addressed with big data analytics. Certain developments or outcomes may be predicted and/or estimated based on vast amounts of historical data, such as length of stay (LOS); patients who will choose elective surgery; patients who likely will not benefit from surgery; complications; patients at risk for medical complications; patients at risk for sepsis, MRSA, C. difficile, or other hospital-acquired illness; illness/disease progression; patients at risk for advancement in disease states; causal factors of illness/disease progression; and possible comorbid conditions (EMC Consulting). McKinsey estimates that big data analytics can enable more than $300 billion in savings per year in U.S. healthcare, two thirds of that through reductions of approximately 8% in national healthcare expenditures. Clinical operations and R & D are two of the largest areas for potential savings with $165 billion and $108 billion in waste respectively [24]. McKinsey believes big data could help reduce waste and inefficiency in the following three areas:

**Clinical operations:** Comparative effectiveness research to determine more clinically relevant and cost-effective ways to diagnose and treat patients.

**Research & development:**

1) predictive modelling to lower attrition and produce a leaner, faster, more targeted R & D pipeline in drugs and devices.

2) statistical tools and algorithms to improve clinical trial design and patient recruitment to better match treatments to individual patients, thus reducing trial failures and speeding new treatments to market.

3) analysing clinical trials and patient records to identify follow-on indications and discover adverse effects before products reach the market.

**Public health:**

1) analysing disease patterns and tracking disease outbreaks and transmission to improve public health surveillance and speed response.

2) faster development of more accurately targeted vaccines, e.g., choosing the annual influenza strains.

3) turning large amounts of data into actionable information that can be used to identify needs, provide services, and predict and prevent crises, especially for the benefit of populations. In addition, suggests big data analytics in healthcare can contribute to.

**Evidence-based medicine:** Combine and analyse a variety of structured and unstructured data-EMRs, financial and operational data, clinical data, and genomic data to match treatments with outcomes, predict patients at risk for disease or readmission and provide more efficient care.

**Genomic analytics:** Execute gene sequencing more efficiently and cost effectively and make genomic analysis a part of the regular medical care decision process and the growing patient medical record.

**Pre-adjudication fraud analysis:** Rapidly analyse large numbers of claim requests to reduce fraud, waste and abuse.

**Device/remote monitoring**: Capture and analyse in real-time large volumes of fast-moving data from in-hospital and in-home devices, for safety monitoring and adverse event prediction.

**Patient profile analytics:** Apply advanced analytics to patient profiles (e.g., segmentation and predictive modelling) to identify individuals who would benefit from proactive care or lifestyle changes, for example, those patients at risk of developing a specific disease (e.g., diabetes) who would benefit from preventive care.

According to, areas in which enhanced data and analytics yield the greatest results include: pinpointing patients who are the greatest consumers of health resources or at the greatest risk for adverse outcomes; providing individuals with the information they need to make informed decisions and more effectively manage their own health as well as more easily adopt and track healthier behaviours; identifying treatments, programs and processes that do not deliver demonstrable benefits or cost too much; reducing readmissions by identifying environmental or lifestyle factors that increase risk or trigger adverse events and adjusting treatment plans accordingly; improving outcomes by examining vitals from at-home health monitors; managing population health by detecting vulnerabilities within patient populations during disease outbreaks or disasters; and bringing clinical, financial and operational data together to analyse resource utilization productively and in real time.

**The 4 “Vs” of big data analytics in healthcare**

Like big data in healthcare, the analytics associated with big data is described by three primary characteristics: volume, velocity and variety (http://www-01.ibm.com/soft ware/data/bigdata/). Over time, health-related data will be created and accumulated continuously, resulting in an incredible volume of data. The already daunting volume of existing healthcare data includes personal medical records, radiology images, clinical trial data FDA submissions, human genetics and population data genomic sequences, etc. Newer forms of big data, such as 3D imaging, genomics and biometric sensor readings, are also fuelling this exponential growth.

Fortunately, advances in data management, particularly virtualization and cloud computing, are facilitating the development of platforms for more effective capture, storage and manipulation of large volumes of data. Data is accumulated in real-time and at a rapid pace, or velocity. The constant flow of new data accumulating at unprecedented rates presents new challenges. Just as the volume and variety of data that is collected and stored has changed, so too has the velocity at which it is generated and that is necessary for retrieving, analysing, comparing and making decisions based on the output.

Most healthcare data has been traditionally static paper files, x-ray films, and scripts. Velocity of mounting data increases with data that represents regular monitoring, such as multiple daily diabetic glucose measurements (or more continuous control by insulin pumps), blood pressure readings, and EKGs. Meanwhile, in many medical situations, constant real-time data (trauma monitoring for blood pressure, operating room monitors for anesthesia, bedside heart monitors, etc.) can mean the difference between life and death.

Future applications of real-time data, such as detecting infections as early as possible, identifying them swiftly and applying the right treatments (not just broad-spectrum antibiotics) could reduce patient morbidity and mortality and even prevent hospital outbreaks. Already, real-time streaming data monitors neonates in the ICU, catching life-threatening infections sooner. The ability to perform real-time analytics against such high-volume data in Raghupathi and Raghupathi Health Information Science and Systems 2014, 2:3 Page 3 of 10 http://www.hissjournal.com/content/2/1/3 motion and across all specialties would revolutionize healthcare. Therein lies variety.

As the nature of health data has evolved, so too have analytics techniques scaled up to the complex and sophisticated analytics necessary to accommodate volume, velocity and variety. Gone are the days of data collected exclusively in electronic health records and other structured formats. Increasingly, the data is in multimedia format and unstructured. The enormous variety of data— structured, unstructured and semi-structured—is a dimension that makes healthcare data both interesting and challenging.

Structured data is data that can be easily stored, queried, recalled, analysed and manipulated by machine. Historically, in healthcare, structured and semi-structured data includes instrument readings and data generated by the ongoing conversion of paper records to electronic health and medical records. Historically, the point of care generated unstructured data: office medical records, handwritten nurse and doctor notes, hospital admission and discharge records, paper prescriptions, radiograph films, MRI, CT and other images.

Already, new data streams—structured and unstructured—are cascading into the healthcare realm from fitness devices, genetics and genomics, social media research and other sources. But relatively little of this data can presently be captured, stored and organized so that it can be manipulated by computers and analysed for useful information. Healthcare applications in particular need more efficient ways to combine and convert varieties of data including automating conversion from structured to unstructured data.

The structured data in EMRs and EHRs include familiar input record fields such as patient name, data of birth, address, physician’s name, hospital name and address, treatment reimbursement codes, and other information easily coded into and handled by automated databases. The need to field-code data at the point of care for electronic handling is a major barrier to acceptance of EMRs by physicians and nurses, who lose the natural language ease of entry and understanding that handwritten notes provide. On the other hand, most providers agree that an easy way to reduce prescription errors is to use digital entries rather than handwritten scripts.

The potential of big data in healthcare lies in combining traditional data with new forms of data, both individually and on a population level. We are already seeing data sets from a multitude of sources support faster and more reliable research and discovery. If, for example, pharmaceutical developers could integrate population clinical data sets with genomics data, this development could facilitate those developers gaining approvals on more and better drug therapies more quickly than in the past and, more importantly, expedite distribution to the right patients. The prospects for all areas of healthcare are infinite.

Some practitioners and researchers have introduced a fourth characteristic, veracity, or ‘data assurance’. That is, the big data, analytics and outcomes are error-free and credible. Of course, veracity is the goal, not (yet) the reality. Data quality issues are of acute concern in healthcare for two reasons: life or death decisions depend on having the accurate information, and the quality of healthcare data, especially unstructured data, is highly variable and all too often incorrect. (Inaccurate “translations” of poor handwriting on prescriptions are perhaps the most infamous example).

Veracity assumes the simultaneous scaling up in granularity and performance of the architectures and platforms, algorithms, methodologies and tools to match the demands of big data. The analytics architectures and tools for structured and unstructured big data are very different from traditional business intelligence (BI) tools. They are necessarily of industrial strength. For example, big data analytics in healthcare would be executed in distributed processing across several servers (“nodes”), utilizing the paradigm of parallel computing and ‘divide and process’ approach. Likewise, models and techniques—such as data mining and statistical approaches, algorithms, visualization techniques—need to take into account the characteristics of big data analytics. Traditional data management assumes that the warehoused data is certain, clean, and precise.

Veracity in healthcare data faces many of the same issues as in financial data, especially on the payer side: Is this the correct patient/hospital/payer/reimbursement code/dollar amount? Other veracity issues are unique to healthcare: Are diagnoses/treatments/prescriptions/procedures/outcomes captured correctly?

Improving coordination of care, avoiding errors and reducing costs depend on high-quality data, as do advances in drug safety and efficacy, diagnostic accuracy and more precise targeting of disease processes by treatments. But increased variety and high velocity hinder the ability to cleanse data before analysing it and making decisions, magnifying the issue of data “trust”.

The ‘4Vs’ are an appropriate starting point for a discussion about big data analytics in healthcare. But there are other issues to consider, such as the number of architectures and platforms, and the dominance of the open source paradigm in the availability of tools. Consider, too, the challenge of developing methodologies and the need for user-friendly interfaces. While the overall cost of hardware and software is declining, these issues have to be addressed to harness and maximize the potential of big data analytics in healthcare.

**Architectural framework**

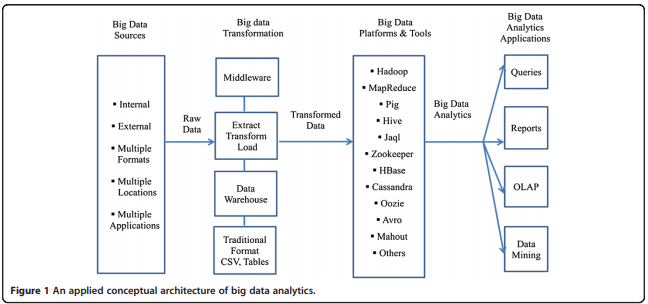
The conceptual framework for a big data analytics project in healthcare is similar to that of a traditional health informatics or analytics project. The key difference lies in how processing is executed. In a regular health analytics project, the analysis can be performed with a business intelligence tool installed on a stand-alone system, such as a desktop or laptop. Because big data is by definition large, processing is broken down and executed across multiple nodes. The concept of distributed processing has existed for decades. What is relatively new is its use in analysing very large data sets as healthcare providers start to tap into their large data repositories to gain insight for making better-informed health-related decisions. Furthermore, open source platforms such as Hadoop/MapReduce, available on the cloud, have encouraged the application of big data analytics in healthcare.

While the algorithms and models are similar, the user interfaces of traditional analytics tools and those used for big data are entirely different; traditional health analytics tools have become very user friendly and transparent. Big data analytics tools, on the other hand, are extremely complex, programming intensive, and require the application of a variety of skills. They have emerged in an ad hoc fashion mostly as open-source development tools and platforms, and therefore they lack the support and user-friendliness that vendor-driven proprietary tools possess. As Figure 1 indicates, the complexity begins with the data itself.

Big data in healthcare can come from internal (e.g., electronic health records, clinical decision support systems, CPOE, etc.) and external sources (government sources, laboratories, pharmacies, insurance companies & HMOs, etc.), often in multiple formats (flat files, .csv, relational tables, ASCII/text, etc.) and residing at multiple locations (geographic as well as in different healthcare providers’ sites) in numerous legacy and other applications (transaction processing applications, databases, etc.). Sources and data types include:

1. Web and social media data: Clickstream and interaction data from Facebook, Twitter, LinkedIn, blogs, and the like. It can also include health plan websites, smartphone apps, etc.
2. Machine to machine data: readings from remote sensors, meters, and other vital sign devices.
3. Big transaction data: health care claims and other billing records increasingly available in semi-structured and unstructured formats.
4. Biometric data: finger prints, genetics, handwriting, retinal scans, x-ray and other medical images, blood pressure, pulse and pulse-oximetry readings, and other similar types of data.
5. Human-generated data: unstructured and semi-structured data such as EMRs, physicians notes, email, and paper documents.

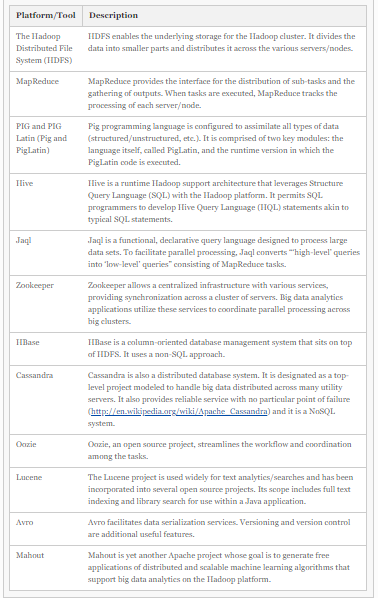
For the purpose of big data analytics, this data has to be pooled. In the second component the data is in a ‘raw’ state and needs to be processed or transformed, at which point several options are available. A service oriented architectural approach combined with web services (middleware) is one possibility [27]. The data stays raw and services are used to call, retrieve and process the data. Another approach is data warehousing wherein data from various sources is aggregated and made ready for processing, although the data is not available in real time. Via the steps of extract, transform, and load (ETL), data from diverse sources is cleansed and readied. Depending on whether the data is structured or unstructured, several data formats can be input to the big data analytics platform.



In this next component in the conceptual framework, several decisions are made regarding the data input approach, distributed design, tool selection and analytics models. Finally, on the far right, the four typical applications of big data analytics in healthcare are shown. These include queries, reports, OLAP, and data mining. Visualization is an overarching theme across the four applications. Drawing from such fields as statistics, computer science, applied mathematics and economics, a wide variety of techniques and technologies has been developed and adapted to aggregate, manipulate, analyse, and visualize big data in healthcare.

The most significant platform for big data analytics is the open-source distributed data processing platform Hadoop (Apache platform), initially developed for such routine functions as aggregating web search indexes. It belongs to the class “NoSQL” technologies—others include CouchDB and MongoDB—that evolved to aggregate data in unique ways. Hadoop has the potential to process extremely large amounts of data mainly by allocating partitioned data sets to numerous servers (nodes), each of which solves different parts of the larger problem and then integrates them for the final result. Hadoop can serve the twin roles of data organizer and analytics tool. It offers a great deal of potential in enabling enterprises to harness the data that has been, until now, difficult to manage and analyse. Specifically, Hadoop makes it possible to process extremely large volumes of data with various structures or no structure at all. But Hadoop can be challenging to install, configure and administer, and individuals with Hadoop skills are not easily found. Furthermore, for these reasons, it appears organizations are not quite ready to embrace Hadoop completely. The surrounding ecosystem of additional platforms and tools supports the Hadoop distributed platform. These are summarized in Table 1.

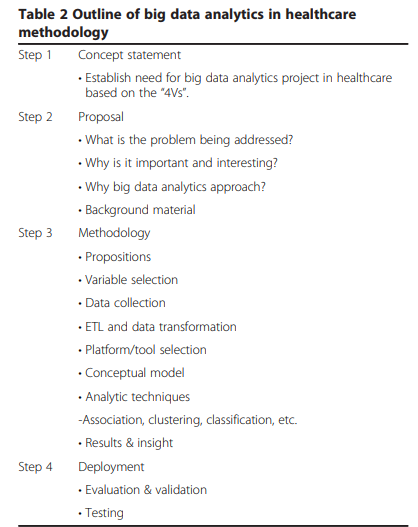
**Table 1 Platforms & tools for big data analytics in healthcare**



Numerous vendors—including AWS, Cloudera, Hortonworks, and MapR Technologies—distribute opensource Hadoop platforms [29]. Many proprietary options are also available, such as IBM’s Biglnsights. Further, many of these platforms are cloud versions, making them widely available. Cassandra, HBase, and MongoDB, described above, are used widely for the database component. While the available frameworks and tools are mostly open source and wrapped around Hadoop and related platforms, there are numerous trade-offs that developers and users of big data analytics in healthcare must consider. While the development costs may be lower since these tools are open source and free of charge, the downsides are the lack of technical support and minimal security. In the healthcare industry, these are, of course, significant drawbacks, and therefore the trade-offs must be addressed. Additionally, these platforms/tools require a great deal of programming, skills the typical end-user in healthcare may not possess. Furthermore, considering the only recent emergence of big data analytics in healthcare, governance issues including ownership, privacy, security, and standards have yet to be addressed. In the next section we offer an applied big data analytics in healthcare methodology to develop and implement a big data project for healthcare providers.

**Methodology**

While several different methodologies are being developed in this rapidly emerging discipline, here we outline one that is practical and hands-on. Table 2 shows the main stages of the methodology. In Step 1, the interdisciplinary big data analytics in healthcare team develops a ‘concept statement’. This is a first cut at establishing the need for such a project. The concept statement is followed by a description of the project’s significance. The healthcare organization will note that there are trade-offs in terms of alternative options, cost, scalability, etc. Once the concept statement is approved, the team can proceed to Step 2, the proposal development stage. Here, more details are filled in. Based on the concept statement, several questions are addressed: What problem is being addressed? Why is it important and interesting to the healthcare provider? What is the case for a ‘big data’ analytics approach? (Because the complexity and cost of big data analytics are significantly higher compared to traditional analytics approaches, it is important to justify their use). The project team also should provide background information on the problem domain as well as prior projects and research done in this domain.



Next, in Step 3, the steps in the methodology are fleshed out and implemented. The concept statement is broken down into a series of propositions. (Note these are not rigorous as they would be in the case of statistical approaches. Rather, they are developed to help guide the big data analytics process). Simultaneously, the independent and dependent variables or indicators are identified. The data sources, as outlined in Figure 1, are also identified; the data is collected, described, and transformed in preparation for for analytics. A very important step at this point is platform/tool evaluation and selection. There are several options available, as indicated previously, including AWS Hadoop, Cloudera, and IBM BigInsights. The next step is to apply the various big data analytics techniques to the data. This process differs from routine analytics only in that the techniques are scaled up to large data sets. Through a series of iterations and what-if analyses, insight is gained from the big data analytics. From the insight, informed decisions can be made. In Step 4, the models and their findings are tested and validated and presented to stakeholders for action. Implementation is a staged approach with feedback loops built in at each stage to minimize risk of failure.

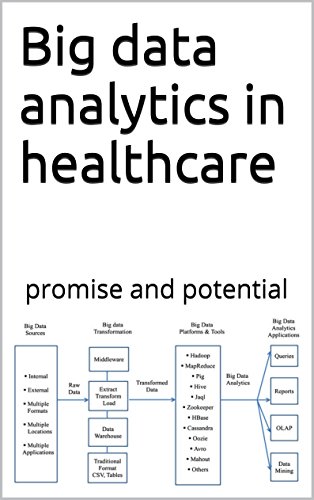
The next section describes several reported big data analytics applications in healthcare. We draw on publicly available material from numerous sources, including vendor sites. In this emerging discipline, there is little independent research to cite. These examples are from secondary sources. Nevertheless, they are illustrative of the potential of big data analytics in healthcare.

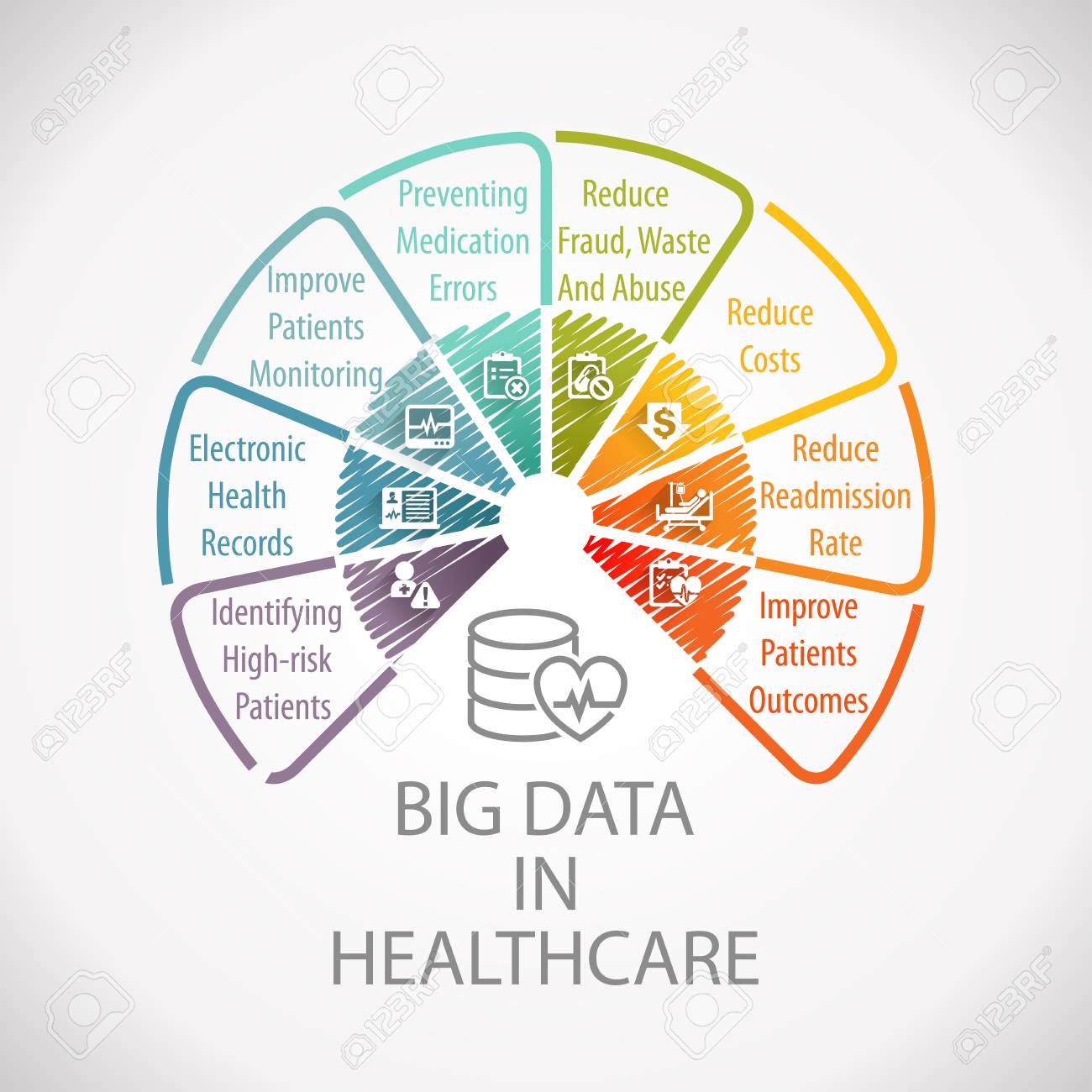
**Examples**

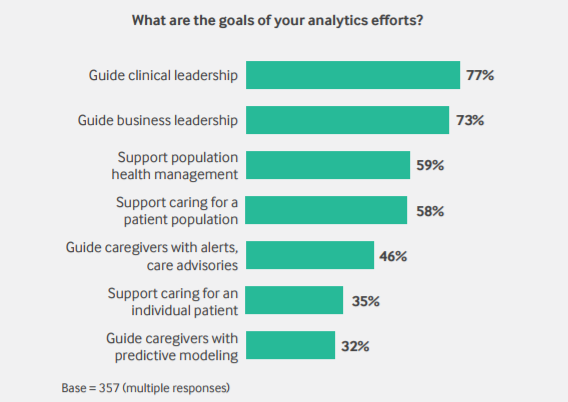
Premier, the U.S. healthcare alliance network, has more than 2,700 members, hospitals and health systems, 90,000 non-acute facilities and 400,000 physicians and is reported to have data on approximately one in four patients discharged from hospitals. Naturally, the network has assembled a large database of clinical, financial, patient, and supply chain data, with which the network has generated comprehensive and comparable clinical outcome measures, resource utilization reports and transaction level cost data. These outputs have informed decision-making and improved the healthcare processes at approximately 330 hospitals, saving an estimated 29,000 lives and reducing healthcare spending by nearly Table 2 Outline of big data analytics in healthcare methodology Step 1 Concept statement • Establish need for big data analytics project in healthcare based on the “4Vs”. Step 2 Proposal • What is the problem being addressed? • Why is it important and interesting? • Why big data analytics approach? • Background material Step 3 Methodology • Propositions • Variable selection • Data collection • ETL and data transformation • Platform/tool selection • Conceptual model • Analytic techniques -Association, clustering, classification, etc. • Results & insight Step 4 Deployment • Evaluation & validation • Testing Source: Adapted from [Raghupathi & Raghupathi. Raghupathi and Raghupathi Health Information Science and Systems 2014, 2:3 Page 7 of 10 http://www.hissjournal.com/content/2/1/3 $7 billion. North York General Hospital, a 450-bed community teaching hospital in Toronto, Canada, reports using real-time analytics to improve patient outcomes and gain greater insight into the operations of healthcare delivery. North York is reported to have implemented a scalable real-time analytics application to provide multiple perspectives, including clinical, administrative, and financial. Another example, reported by IBM, is that of the large, unnamed healthcare provider that is analysing data in the electronic medical record (EMR) system with the goal of reducing costs and improving patient care. (Data in the EMR include the unstructured data from physician notes, pathology reports and other sources). Big data analytics is used to develop care protocols and case pathways and to assist caregivers in performing customized queries. Another example of big data analytics in healthcare is Columbia University Medical Centre’s analysis of “complex correlations” of streams of physiological data related to patients with brain injuries. The goal is to provide medical professionals with critical and timely information to aggressively treat complications. The advanced analytics is reported to diagnose serious complications as much as 48 hours sooner than previously in patients who have suffered a bleeding stroke from a ruptured brain aneurysm. The Rizzoli Orthopedic Institute in Bologna, Italy, is reportedly using advanced analytics to gain a more “granular understanding” of the clinical variations within families whereby individual patients display extreme differences in the severity of their symptoms. This insight is reported to have reduced annual hospitalizations by 30% and the number of imaging tests by 60%. In the longterm, the Institute expects to gain insight into the role of genetic factors to develop treatments. The Hospital for Sick Children (Sick Kids) in Toronto is using analytics to improve the outcomes for infants prone to life threatening “nosocomial infections”. It is reported that Sick Kids applies advanced analytics to vital-sign data gathered from bedside monitoring devices to identify potential signs infection as early as 24 hours prior to previous methods. Additional examples are reported below.

A recent New Yorker magazine article by Atul Gawande, MD described how orthopedic surgeons at Brigham and Women’s Hospital in Boston relied on personal experience along with insight extracted from research on data based on a host of factors critical to the success of joint replacement surgery to systematically standardize knee joint-replacement surgery. The result: improved outcomes at lower costs. The University of Michigan Health System standardized the administration of blood transfusions using analytics in a similar fashion, combining experience with big data analytics research. This resulted in a 31% reduction in transfusions and $200,000 reduction in expenses per month (reported in). Another example is The National Institute for Health and Clinical Excellence (NICE) of the U.K.’s National Health Service. NICE is reportedly a leader in the analytics of large clinical datasets for exploring the effectiveness of clinical and cost factors in the use of new drugs and/or clinical treatments. The Italian Medicines Agency is also reported to collect and analyse clinical data on the use of expensive new drugs as one goal in a country-level cost-effectiveness program. Another leading example of big data analytics in healthcare is the Department of Veterans Affairs’ (VA) use of applications on its very large data set in an effort to comply with “performance-based accountability framework and disease management practice”. In one very famous example, California-based Kaiser Permanente associated clinical data with cost data to generate a key data set, the analytics of which led to the discovery of adverse drug effects and subsequent withdrawal of Vioxx from the market. Researchers at the Johns Hopkins School of Medicine discovered they could use data from Google Flu Trends to predict sudden increases in flu-related emergency room visits at least a week before warnings from the CDC. Likewise, the analysis of Twitter updates was as accurate as (and two weeks ahead of) official reports at tracking the spread of cholera in Haiti after the January 2010 earthquake. Also reported is an application developed by IBM that predicts the likely outcomes of diabetes patients using patients’ panel data linked to physicians, management protocols, and the overall relationship to population health management averages. In another diabetes application, physicians at Harvard Medical School and Harvard Pilgrim Health Care recently demonstrated the potential of analytics applications to EHR data to identify and group patients with diabetes for public health surveillance. Four years of worth of data based on numerous indicators from multiple sources was utilized. The analytics application also differentiated between Type 1 and Type II diabetes. Finally, at Blue Cross Blue Shield of Massachusetts (BCBSMA) there was a “need to embed analytics into business processes to help decision-makers across the business gain insight into financial and medical data and become more proactive”. Several benefits were reported. First, the analytics enabled medical directors to identify high-risk disease groups and act to minimize risk and improve patient outcomes. For example, new preventive treatment protocols could be introduced among patient groups with high cholesterol, thereby fending off heart problems. Also, complex health informatics reports were generated 300% faster than previously, helping BCBSMA service clients more effectively.

The next section briefly identifies some of the key challenges in big data analytics in healthcare.







Electronic Health Records (EHR) in conjunction with Electronic Medical Records (EMR) have been steadily increasing in use over the last 15 years.  In the time from 2001 to the end of 2014 EMR usage in physician offices rose from 20% to over 80% With the introduction of the Health Information Technology for Economic and Clinical Health (HITECH) Act of 2009, Meaningful Use incentives for higher billing and reimbursement rates from the federal government continue to drive adoption rates.

With the adoption of EMRs, the increase in EHRs has grown exponentially.  EHRs are a broader view of a patient’s collective EMR experience and contain a historical 360-degree view of a patient’s medical history.  While the exchange and sharing of this EHR data has been a primary focus in recent years (through Continuity of Care Documents (CCD) and Consolidated Clinical Document Architecture (C-CDA)), the massive collection of clinical data by large health systems and treatment centres (public, private, and academic) has moved into the realm of big data.

Coming from the healthcare space, one of the things that always fascinated me was the ability to use this wealth of data to do predictive analytics on treatment plans to improve patient outcomes. With big data, big answers and meaningful analytics can be extrapolated from the healthcare continuum.

**The Problem: Weighing Patient Outcomes and Operating Costs**

Health systems continually face the dichotomy of medicine: improving patient outcomes and maintaining effective operating costs.  It is difficult to maintain both and to give the highest level of satisfaction to a patient at the same time. By having to meander through treatment plans, trying one treatment after another, a physician continues to prolong the condition or symptoms of a condition leading to patient dissatisfaction.

Additionally, having to repeatedly see a patient without results may cause patient drift (the concept of a patient either going elsewhere or discontinuing to seek treatment), reducing revenue opportunities, and perhaps discouraging other patients from seeking treatment.

### The Solution: Selecting the Best Treatment Plans with Predictive Analytics

### Ultimately, the treatment plans for a patient are determined by the physicians and their patients based on what is known about the patient and their medical history.  But what if physicians could be given another tool in their arsenal of medicine to make more informed choices about a patient’s treatment based on patient population cohorts?  What if physicians could quickly and accurately determine the best treatment plan for any given patient based on their medical history and demographics?

### Take the example presented here:

### 1. A patient of a certain age bracket, with certain race characteristics, a given gender, specific medical history, problem lists, allergies, etc. presents at a clinic with a newly diagnosed condition.

2. Based on this condition, taking into account what is known about a patient’s existing conditions, medications, personal history, etc., search is used to find other patients within a population cohort that are similar.

3. Based on those specific facets (now aggregations) of data, the treatment plans given to these patients could be analysed by the physician to determine the most appropriate option for a patient.

4. This could then be used as an additional tool for the attending physician to make determinations about what their treatment plan should be, either one that they would have suggested reinforced, or another path that may not have been seen, with a level of certainty that this initial treatment plan would be best for this particular patient based on the population cohort.

A large portion of this relies on standardized, normalized, and centralized data from disparate data sources which can be done efficiently with content processing and data lakes (read more about how the data lake is leveraged in the biotech and health industries).

With the methodology outlined above, a patient could be directed to the most accurate treatment plan for their given conditions based on their existing conditions and the observed outcomes of other patients in the cohort.  With this knowledge in hand, a physician can provide a treatment plan that will have a better chance of improving a patient’s outcome.

### Use Cases for Predictive Analytics in Healthcare

**1. Aiding Diagnosis:**

In addition to identifying treatment plans, search can also be used in aiding diagnosis.  With a battery of tests performed, a host of symptoms, results, and observations, guidance could be provided for potential diagnosis.  Often the approach of “not knowing” or “it could be this” contribute to patient dissatisfaction and prolonged durations of being able to diagnose or misdiagnosing a condition.

**2. Applying to Elective Processes:**

Similarly, predictive analytics need not be limited to diagnosed chronic conditions.  Search can also be applied to elective processes like physician assisted weight loss clinics for example.  In this use case, a patient’s conditions are not only known, but additional data related to activity and diet are also recorded.

Other data gathering techniques like wearable health monitoring tools can be used to automatically populate EMRs or Personal Health Records (PHR) that can be consumed by the EMR.  With this information, a patient can immediately be given the best treatment plan for their age, race, gender, BMI, etc. that includes exercise plans, diet plans, and any assistive medications based not only on the population cohort, but on the highly visible and measurable results of those plans.  This is very crucial for something like weight loss, as patients will want to see effective results as quickly as possible to have faith in the plan they have been given, and to maintain participation in the clinic. This drives continued participation, as well as results-driven analytics encouraging recommendations to other participants.

### The Future: Integrating Machine Learning to Improve Healthcare Results over Time

While showing treatment plans for patients that most closely resemble the patient being attended to by a physician are valuable to a physician, the process can be taken a step further.  What if not only showing treatment plans for patients who have similar conditions as the patient being seen, the best treatment plan could be recommended based on how other patients that are in the same cohorts have responded to any given treatment?

This approach goes beyond identification of populations and treatments, and transitions into the natural extension of predictive analytics.  By identifying patients and their treatment plans, the related observations associated with that treatment plan can be utilized to make determinations about what plan is best.  This is a much more difficult problem to solve, as machine learning must be integrated to understand what a “positive outcome” is in relation to a treatment plan.

For example, if a patient presents with high blood pressure, the initial search would identify what treatments have been done for patients that are similar and have the same conditions.  However, this still leads the manual interpretive step of the physician deciding what is best.  Search can be extended beyond identification to interpretation by understanding that the condition is high blood pressure, and that a positive outcome is not only a reduction in overall blood pressure, but what other contributing factors caused that to happen.  This would include observations of:

* A patient’s related conditions that may or may not impact the presenting condition
* The vitals obtained on each visit (like their blood pressure readings, weight, etc.)
* The patient’s self-reporting of their activities (like exercise)
* The medications that they (may) have been prescribed

**The recommended treatment plan would not only be the best for the patient, but explained to the physician as to why the plan is the best.**

Understanding every facet of the treatment plan, the related observations, and what a positive outcome is, in conjunction with the presenting condition, is what truly makes machine learning and predictive analytics useful in improving healthcare for patients. This is the power of trending analysis on specific observations related to the condition and the results of those treatment plans.

Big data analytics in medicine and healthcare is very promising process of integrating, exploring and analysing of large amount complex heterogeneous data with different nature: biomedical data, experimental data, electronic health records data and social media data. Integration of such diverse data makes big data analytics to intertwine several fields, such as bioinformatics, medical imaging, sensor informatics, medical informatics, health informatics and computational biomedicine. As a further work, the big data characteristics provide very appropriate basis to use promising software platforms for development of applications that can handle big data in medicine and healthcare. One such platform is the open-source distributed data processing platform Apache Hadoop MapReduce that use massive parallel processing (MPP). These applications should enable applying data mining techniques to these heterogeneous and complex data to reveal hidden patterns and novel knowledge from the data.

Recent hardware innovations in processor technology, newer kinds of memories/network architecture will minimize the time spent in moving the data from storage to the processor in a distributed setting.

**Challenges**

At minimum, a big data analytics platform in healthcare must support the key functions necessary for processing Raghupathi and Raghupathi Health Information Science and Systems 2014, 2:3 Page 8 of 10 http://www.hissjournal.com/content/2/1/3 the data. The criteria for platform evaluation may include availability, continuity, ease of use, scalability, ability to manipulate at different levels of granularity, privacy and security enablement, and quality assurance. In addition, while most platforms currently available are open source, the typical advantages and limitations of open source platforms apply. To succeed, big data analytics in healthcare needs to be packaged so it is menu driven, user-friendly and transparent. Real-time big data analytics is a key requirement in healthcare. The lag between data collection and processing has to be addressed. The dynamic availability of numerous analytics algorithms, models and methods in a pull-down type of menu is also necessary for large-scale adoption. The important managerial issues of ownership, governance and standards have to be considered. And woven through these issues are those of continuous data acquisition and data cleansing. Health care data is rarely standardized, often fragmented, or generated in legacy IT systems with incompatible formats. This great challenge needs to be addressed as well.

Regarding collection of large amount data, some challenging issues should be considered. Obtaining high-throughput – omics data is tied to the cost of experimental measurements. Concerning heterogeneity of the data sources, the noise of the experimental – omics data and the variety of the experimental techniques, environmental conditions, biological nature should be considered, before integration of these heterogeneous data and before employing of the data mining methods. Different data mining techniques can be applied on these heterogeneous biomedical data sets, such as: anomaly detection, clustering, classification, association rules as well as summarization and visualization of those big data sets.

These shortcomings might lead to the unreliability of some of the data points, such as missing values or outliers. Despite of these drawbacks of the – omics data, EHRs data are very influenced by the staff who entered the patient’s data, which can lead to entering missing values, incorrect data as a result of mistakes, misunderstanding or wrong interpretation of the original data [[5](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6340124/#j_jib-2017-0030_ref_005)]. Integration of data from various databases and standardization for laboratory protocols and values still remain challenging issues.

High dimensionality of the – omics data means, that there have many more dimensions or features than the number of samples, and on the other side the EHRs data which regard to the individuals/patients, makes data mining techniques to be more challenging task.

The subsequent stage is the pre-processing of the data, which usually envelop handling noisy data, outliers, missing values, data transformation and normalization. This data pre-processing enables to be applied statistical techniques and data mining methods and thus the big data analytics quality and outcomes can improve and can result with discovering of novel knowledge. This novel knowledge obtained by integration of the – omics and EHRs data should results with improving of the implemented healthcare to the patients as well to advanced decision making by the healthcare decision policy makers.

Big data analytics not only provides charming opportunities but also faces lot of challenges. The challenge starts from choosing the big data analytics platform. While choosing the platform, some criteria like availability, ease of use, scalability, level of security and continuity should be considered. The other challenges of big data analytics are data incompleteness, scalability and security. Since cloud computing plays a major role in big data analytics, cloud security should be considered. Studies show that 90% of big data are unstructured data. But the representation, analytics and access of numerous unstructured data are still a challenge. Data timeliness is also critical in various healthcare areas like clinical decision support for making decisions or providing information that guides to take decisions. Big data can make decision support simpler, faster and more accurate because decisions are based on higher volumes of data that are more current and relevant. This needs scalable analytics algorithms to produce timely results. However, most of the current algorithms are inefficient in terms of big data analytics. So the availability of effective analytics algorithms is also necessary. Concerns about privacy and security are superior, although these are increasingly being attempted by new authentication approaches and policies that better protect patient identifiable data.

**Conclusions**

Big data analytics has the potential to transform the way healthcare providers use sophisticated technologies to gain insight from their clinical and other data repositories and make informed decisions. In the future we’ll see the rapid, widespread implementation and use of big data analytics across the healthcare organization and the healthcare industry. To that end, the several challenges highlighted above, must be addressed. As big data analytics becomes more mainstream, issues such as guaranteeing privacy, safeguarding security, establishing standards and governance, and continually improving the tools and technologies will garner attention. Big data analytics and applications in healthcare are at a nascent stage of development, but rapid advances in platforms and tools can accelerate their maturing process.

****

Large amounts of heterogeneous medical data have become available in various healthcare organizations. The rate of electronic health record (EHR) adoption continues to climb in both inpatient and outpatient aspects. Those data could be an enabling resource for deriving insights for improving patient care and reducing waste. Analyzing the massive amount of healthcare information that is newly available in digital format should enable advanced detection of powerful treatment, better clinical decision support and accurate predictions of who is likely to get sick. This requires high performance computing platforms and algorithms. This paper reviews the various big data analytics platforms and algorithms and challenges are discussed. Based on the study, although medical diagnoses applications use different algorithms, C4.5 algorithm gives better performance. But still the improvisation of C4.5 algorithm is required to maximize accuracy, handle large amount of data, reduce the space requirement for large amount of datasets and support new data types and to reduce the error rate.

C5.0 approach overcomes these criticisms by producing more accuracy; requiring less space when volume of data is increased from thousands to millions or billions. It also has lower error rate and minimizes the predictive error. C5.0 algorithm is the potentially suitable algorithm for any kind of medical diagnoses. In case of big data, the C5.0 algorithm works faster and gives the better accuracy with less memory consumption. In spite of the narrow work done on big data analytics so far, much effort is needed to beat its issues related to the above mentioned challenges. Also the rapid advances in platforms and algorithms can help to accelerate the performance.

**REFERENCES**

1. Raghupathi W: Data Mining in Health Care. In Healthcare Informatics: Improving Efficiency and Productivity. Edited by Kudyba S. Taylor & Francis; 2010:211–223.
2. Dembosky A: “Data Prescription for Better Healthcare.” Financial Times, December 12, 2012, p. 19; 2012. Available from: http://www.ft.com/intl/cms/ s/2/55cbca5a-4333-11e2-aa8f-00144feabdc0.html#axzz2W9cuwajK
3. Frost & Sullivan: Drowning in Big Data? Reducing Information Technology Complexities and Costs for Healthcare Organizations. http://www.emc.com/ collateral/analyst-reports/frost-sullivan-reducing-information-technologycomplexities-ar.pdf
4. jStart: “How Big Data Analytics Reduced Medicaid Re-admissions.” A jStart Case Study; 2012. http://www-01.ibm.com/software/ebusiness/jstart/portfolio/ uncMedicaidCaseStudy.pdf.
5. Knowledgent: Big Data and Healthcare Payers; 2013. http://knowledgent. com/mediapage/insights/whitepaper/482.
6. Intel: Leveraging Big Data and Analytics in Healthcare and Life Sciences: Enabling Personalized Medicine for High-Quality Care, Better Outcomes; 2012. <http://www.intel.com/content/dam/www/public/us/en/documents/whitepapers/healthcare-leveraging-big-data-paper.pdf>
7. Zikopoulos PC, Eaton C, DeRoos D, Deutsch T, Lapis G: Understanding Big Data – Analytics for Enterprise Class Hadoop and Streaming Data. McGraw-Hill: Aspen Institute; 2012.