Informatics in Medicine Unlocked 1 (2015) 17–26

Contents lists available at ScienceDirect

Informatics in Medicine Unlocked

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

Discovery of medical Big Data analytics: Improving the prediction of traumatic brain injury survival rates by data mining Patient Informatics Processing Software Hybrid Hadoop Hive

James A. Rodger [1,2,3]

MIS and Decision Sciences, Eberly College of Business & Information Technology, Indiana University of Pennsylvania, Indiana, PA 15705, USA

ARTICLE INFO

Article history:
Received 18 November 2015
Received in revised form
1 January 2016
Accepted 13 January 2016
Available online 23 February 2016

Keywords:
Decision support
Traumatic brain injuries
Apache hive
Symbolic data analysis
Informatics
Data mining

ABSTRACT

Entering medical encounter data by hand is time-consuming. In addition, data are often not entered into the database in a timely enough fashion to enable their use for subsequent mission planning. The Patient Informatics Processing Software semi-automates the data collection process onboard ships. Then data within these images are captured and used to populate a database, after which multiple ship databases are used for reporting and analysis. In this paper, we used the Patient Informatics Processing Software Hybrid Hadoop Hive to orchestrate database processing via various ships, by marshaling the distributed servers, running the various tasks in parallel, managing all of the communications and data transfers between the various parts of the system, and providing for redundancy and fault tolerance. Then we employed the Apache Hive as a data warehouse infrastructure built on top of Hadoop for data summarization, query, and analysis to identify traumatic brain injury (TBI) as well as other injury cases. Finally, a proposed Misdiagnosis Minimization Approach method was used for data analysis. We collected data on three ship variables (Byrd, Boxer, Kearsage) and injuries to four body regions (head, torso, extremities, and abrasions) to determine how the set of collected variables relates to the body injuries. Two dimensions or canonical variables (survival vs. mortality) were necessary to understand the association between the two sets of variables. Our method improved data classification and showed that survival, mortality, and morbidity rates can be derived from the superset of Medical Operations data and used for future decision-making and planning. We suggest that an awareness of procedural errors as well as methods to reduce misclassification should be incorporated into all TBI clinical trials.

1. Introduction

1.1. Misdiagnosis minimization approach

The study of medical diseases and injuries often generates huge amounts of data [1]. Hughes [1] stated that “estimates are that the global size of Big Data in Healthcare stands at roughly 150 Exabytes in 2011, increasing at a rate between 1.2 and 2.4 Exabytes per year. Big Data isn't simply about the volume, velocity and variety of the data in storage, it is also about the potential value of those data that already exist but are poorly coordinated and stored in widely disparate formats across industries that haven't typically shared data openly.”

While the outcomes of this process are well documented, little has been written about the collection and dissemination of these data and their correct classification. To fill this gap, we looked at hospital ships, which are a medical asset that supports military operations (MOs) worldwide. This requires the hospital ship to provide medical care to various military populations, under a varying set of medical conditions. It is becoming increasingly common for informatics data to be collected from multiple sources or represented by multiple views, where different views describe distinct perspectives of the data. MO medical personnel collect hundreds of thousands of completed medical encounter forms from missions each year. Previously, these data were entered by hand into a database for reporting and analysis. The United States military applies findings developed from the use of Patient Informatics Processing Software (PIPS) data to support the logistics of planning activities for future missions as they can help save

E-mail address: jrodger@iup.edu
[1] Tel.: +724 357 5944; fax: +724 357 4831.
[2] Submitted to: IIMU.
[3] **James A. Rodger** is a Professor of Management Information Systems at Indiana University of Pennsylvania (IUP). He received his Doctorate in MIS from Southern Illinois University at Carbondale in 1997. Dr. Rodger has published several journal articles related to these subjects. His work has appeared in Annals of Operations Research, Communications of ACM, Computers & Operations Research, Decision Support Systems, Expert Systems with Applications, Lecture Notes in Computer Science, International Journal of Human-Computer Studies as well as several other journals.

http://dx.doi.org/10.1016/j.imu.2016.01.002
2352-9148/

money, reduce waste, improve levels of preparedness for future missions, and save lives. In medical diagnosis, it is important to not only maximize correct classifications, but to also minimize false positive Type I and false negative Type II errors. While these errors go hand in hand with classification, they are not equal. Traditional classification systems, such as linear discriminant analysis and neural networks, do not take into consideration all of the ramifications of the impacts of misdiagnosis [2]. In our research, we contend that predicting a patient who does not have traumatic brain injury (TBI), will survive, when in fact the patient does have TBI, is a bigger error than misdiagnosing a patient with TBI. While traditional systems do not incorporate the impact of misdiagnosis on TBI survival rates, our approach minimizes these medical misclassifications.

2. Literature review

2.1. Big Data analytics, informatics and data mining, discriminant analysis, and canonical correlation

In this section, we discuss basic concepts, widely used algorithms, and some real-world applications in Big Data analytics for healthcare. We also show how the diversity and quality of research have changed due to these factors [3,4]. In addition, we relay how Big Data has impacted information systems in an interdisciplinary way, and how informatics has provided an opportunity to investigate this concept [5]. Liang et al. [6] proposed a novel visual analytics approach for studying brain fiber paths that allows users to explore fiber bundles, revealing the probability that fiber paths use a new visual classification method. In a similar manner, our paper illustrates how analyzing a large number of diverse user-generated content, on healthcare media platforms, can be used to make informed decisions. Various scalable machine-learning algorithms have been successfully deployed in many domains, particularly in the field of business. Similar to our model utilizing symbolic data access (SDA), canonical correlation, and discriminant analysis, Seng and Chen [7] postulated that data mining is a powerful method for extracting knowledge from data by handling various data types in all formats for enhancing business intelligence [8,9]. This paper was also relevant because it emphasized the fact that data mining works in the context of knowledge extraction from medical data, and provided some guidelines to help medical practitioners. Discriminant analysis is at the center of our knowledge extraction. Fisher [10] first utilized linear discriminant analysis (LDA), and postulated that two classes of observations have means $\overrightarrow{\mu_0}, \overrightarrow{\mu_1}$ and covariances $\sum_0, \sum_1$. Canonical correlation is also a valuable tool in our knowledge extraction. Hotelling [11] proposed that given two column vectors $X=(x_1, \cdots, x_n)'$ and $Y=(y_1, \cdots, y_m)'$ of random variables with finite second moments, one may define the cross-covariance $\sum_{XY}=\mathrm{cov}(X, Y)$ to be the $n \times m$ matrix whose i,j entry is the covariance $\mathrm{cov}(x_i, y_j)$.

2.2. Traumatic brain injuries

Griffiths et al. [12] studied a queuing model of a specialist neurological rehabilitation unit and employed the concept that treatment intensity affects a patient's length of stay. A Coxian phase-type distribution was fitted to the length of time from admission until discharge readiness, and some hypothetical scenarios were considered and compared on the grounds of a number of performance measures and cost implications. Cruz and Rincon [13] examined the large body of existing research on outsourcing, and assessed the research status on outsourcing the maintenance of medical devices such as the magnetic resonance imaging (MRI) used in diagnosing TBI. The authors concluded that, "research into the outsourcing of medical device maintenance services in hospitals is still in its infancy stages, and that further progress in this field would benefit from additional empirical study grounded in management theory." Our study extends this research as it applies to the outsourcing of devices onboard medical ships. Yang et al. [14] reported that the shortage of medical resources (mainly beds) is a critical and increasingly prevalent problem affecting hospitals. This fact was true in our ship hospital study as well. The authors found that the factors contributing to these shortages, including the ambiguity and insufficiency of the criteria used to identify whether an inpatient should be discharged, were among the most detrimental. To address this issue, the study applied data envelopment analysis (DEA) and categorized the dynamic model inpatient's discharge status as rejected, under observation, or approved. Their results provided insight into the potential causes of medical resource shortages. Much like our TBI study, their method allows clinicians to treat inpatients more effectively based on the discharge categories.

Kunene and Weistroffer [15] demonstrated that "patient outcome in brain trauma patients is affected by a multiplicity of factors, beginning with ambulatory transportation and routing, to the grade of the receiving facility and treatment therein, and finally the treatment and monitoring in definitive care (the brain trauma intensive care unit). Factors and events in each of these phases can be modeled as a multicriteria problem, where the objective is to optimize patient outcome; moreover, a more comprehensive model can embody the interactions of all three phases." Their study focused on modeling the factors that affect patient outcomes in a definitive way to better describe or predict them using data mining tools.

Lin and Blüml [16] suggested that acute and chronic injuries at the cellular level are sometimes difficult to discern from normal features by anatomical imaging, which often leads to misclassification, similar to our findings. The authors suggested that magnetic resonance spectroscopy (MRS) offers a unique noninvasive approach to assess injury at microscopic levels by quantifying cellular metabolites. Their findings obtained with MRS for concussion and more severe head trauma were heterogeneous, reflecting the different times after injury, degrees of injury, and different physiologic and pathologic responses of the brain to injury. Langlois et al. [17] suggest that the estimated 5.3 million Americans living with TBI-related disabilities face numerous challenges in their efforts to return to a full and productive life. The authors also provide evidence that supports our findings; namely, that routinely reported data underestimates the number of persons who receive medical care when TBI is not diagnosed, or who sustain a TBI but do not seek care. In their study of TBI, Hoge et al. [18] reported that the differences among diagnosis of TBI, stroke, acquired brain injury, anoxic brain injury, and other head and neck injuries need clarification. They further stated that the epidemiology of combat-related mild TBI is poorly understood. Much like the results of our study, the authors reported misclassification of TBI and concluded that mild TBIs, such as concussion, are important mediators of the relationship between mild TBI and physical health problems. Lu et al. [19] investigated the results utilizing the Glasgow Outcome Scale (GOS) as the primary endpoint for analysis of the efficacy of clinical trials on TBI. They postulated that the accurate and consistent assessment of outcome after TBI is essential to the evaluation of treatment results, particularly in the context of multicenter studies and trials, as found onboard ships. They further presumed that the effects of inconsistent measurement or interobserver variation on GOS outcome, or for that matter, on any outcome scales, could adversely affect the sensitivity for detecting treatment effects in clinical trials. Their research concluded that non-differential misclassification directly reduces the power of finding the true treatment effect, and that an awareness of this procedural error as well as methods to reduce misclassification should be incorporated