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

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 non-invasive 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

into TBI clinical trials. In their follow-up study, Lu et al. [20] extended their previous investigation regarding the effects of non-differential dichotomous GOS misclassification in TBI clinical trials, to the effects of GOS misclassification on ordinal analysis in TBI clinical trials. Their results showed that given the specified misclassification distributions, misclassification with a random or upward pattern would have caused a slightly underestimated outcome of the observed data. However, misclassification with a downward pattern would have resulted in an inflated estimation. Thus, the sensitivity analysis suggests that non-differential misclassification can cause uncertainties about the primary outcome estimation in TBI trials. This uncertainty has also been demonstrated through evidence from other social media Big Data scenarios [21]. Sohlberg and Mateer [22] and Kowalczyk et al. [23] also investigated data mining techniques that have been used to build data-driven decision-making models in organizations, similar to those proposed in our TBI model. Similar to our canonical correlation research, 32 TBI patients of different ages and genders were studied. The authors found a significant relationship between the findings of neurologists and systems output for normal, mild, moderate, and severe electroencephalography tracing data.

2.3. K-means clustering

K-means clustering is another component at the heart of our knowledge extraction, and MacKay [24] provides us with an example of this algorithm. Pimentel and de Souza [25] demonstrated that clustering is the process of organizing objects into groups whose members are similar in some way and involves numeric data only. However, to model complex information that may be a histogram, distributions or intervals similar to those used in our research must be employed. SDA was developed, which provides clustering quality results that offer higher accuracy when variables have different variabilities. Krishnasamy et al. [26] proposed clustering as an important and popular technique in data mining. In their paper, they presented an efficient hybrid evolutionary data-clustering algorithm, similar to our method, whereby they combined K-means with modified cohort intelligence (MCI). Their proposed algorithm has been compared to other well-known algorithms such as K-means, K-means++, cohort intelligence (CI), MCI, genetic algorithm (GA), simulated annealing (SA), tabu search (TS), ant colony optimization (ACO), honeybee mating optimization (HBMO), and particle swarm optimization (PSO). Elango et al. [27] attempted to solve the multi-robot task allocation problem, and placed importance on balancing workloads among robots, similar to our medical workload distribution on ships. The paper proposed an algorithm that attempted to minimize the distance traveled by 'm' robots and balanced the workload between them equally, using a K-means clustering technique with the objective of minimizing the distance in a cost-effective manner. Yin et al. [28] used clustering to group data objects into sets of disjoint classes called clusters, so that objects within the same class were highly similar to each other and dissimilar from the objects in other classes. K-harmonic means (KHM) is one of the most popular clustering techniques, and has been widely and successfully applied to many fields, including medicine. Hadavand et al. [29] stated that success in forecasting and analyzing sales for given goods or services can mean the difference between profit and loss for an accounting period, and ultimately, the success or failure of the business itself. The reliable prediction of sales is a very important task, much like the life or death situations found in TBI forecasting. The article presented a novel sales forecasting approach by integrating genetic fuzzy systems (GFS) and data clustering to construct a sales forecasting expert system. The results showed that the proposed approach outperformed previous approaches. Dimoulas et al. [30] focused on the implementation of hybrid expert systems for audiovisual content description and management, by means of pattern analysis. Their proposed methodology combined audio detection–segmentation, motion detection surveillance, and hierarchical audio pattern recognition, using neural networks, statistical clustering, and syntactic pattern classification to deliver new potentials in non-invasive gastrointestinal motility (GIM) monitoring. Their current work introduced new hybrid techniques for content analysis in a medical application, with similarities to our TBI study. Celebi et al. [31] claimed that K-means is undoubtedly the most widely used partitioned clustering algorithm. They presented an overview of this method with an emphasis on computational efficiency, and then compared eight commonly used linear time complexity initialization methods on a large and diverse collection of datasets, using various performance criteria in a manner similar to our study. Higuera et al. [32] provided a medical example for studying the communities of microbial species, by searching for common metabolic characteristics that allowed common functional properties to be found that described the way of life of entire organisms or species. This approach parallels the common properties of TBI classification into functional clusters from the Big Data compilation of injuries on our ship. Birtolo and Ronca [33] investigated the application of model-based collaborative filtering (CF) techniques, and proposed a clustering CF framework and two clustering CF algorithms. In a similar manner to our research, they compared numerous approaches using several datasets with real customers. Lin et al. [34] investigated image retrieval databases in which color was the most important feature and was most commonly used with a K-means algorithm. To create a K-means algorithm for this study, first a level histogram of statistics for the image database was made, similar to the approach used in our study. Their results showed that the K-means algorithm was a more effective, faster, and convenient method for overcoming the problem of spending excessive amounts of time on re-training, caused by the continuous addition of images to the image database. This approach could be applied to our medical data, as the injuries in the battlefield are transferred to the ships. Bai et al. [35] introduced a multi-method multiple criteria approach for evaluating the performance of organizations. Their paper introduced the use of Fuzzy c-means (FCM) and used real company data to evaluate the predictive abilities of the technique in a manner similar to the one we employed to evaluate TBI mortality rates. Sancho-Asensio et al. [36]noted that data mining techniques are traditionally divided into two distinct disciplines, supervised and unsupervised learning, depending on the task to be performed by the algorithm. The latter method is aimed towards discovering regularly occurring patterns underlying the data, without making any a priori assumptions concerning their core structure. Our research used the unsupervised learning discipline to enhance the performance of data storage in the ship database of injuries, which is an approach similar to that used in the Smart Grids study.

2.4. Nearest Neighbor

Nearest Neighbor (NN) is an important player in our knowledge extraction process. The proposed model provides a set of past situations similar to the present situation within the parametric ranges set by the forecaster. The number of past similar situations gives a direct idea about the nature (unique or common) of the present situation. The model works on the following methodology. In an n-dimensional parametric space P with parameters P_i ($i=1,2,\ldots,n$), every n-dimensional record is represented by a point. The model selects a past record X_i ($i=1,2,\ldots,n$), if it satisfies the following condition:

$$\sum_{i=1}^{n}\left[(X_i - x_i)^2 / {R_i}^2\right] \quad <=1$$

where x_i ($i=1,2,\ldots,n$) is the present record and R_i ($i=1,2,\ldots,n$) is the range for parameter P_i around the value x_i.

Nearest neighbor search (NNS), also known as proximity search, similarity search, or closest point search, is an optimization problem for finding closest points in metric spaces. The problem is as follows: given a set S of points in a metric space M and a query point $q \in M$, find the closest point in S to q. In many cases, M is taken to be the d-dimensional Euclidean space, and distance is measured by the Euclidean distance or Manhattan distance [37]. In cluster analysis, single linkage, NN, or shortest distance are methods of calculating distances between clusters in hierarchical clustering. In single linkage, the distance between two clusters is computed as the distance between the two closest elements in the two clusters.

Mathematically, the linkage function – the distance $D(X,Y)$ between clusters X and Y – is described by the expression

$$D(X,Y) = \min_{x \in X, y \in Y} d(x,y),$$

where X and Y are any two sets of elements considered clusters, and $d(x,y)$ denotes the distance between the two elements x and y. A drawback of this method is the so-called *chaining phenomenon* where clusters may be forced together due to single elements being close to each other, although many of the elements in each cluster may be very distant to each other. The following algorithm is an agglomerative that erases rows and columns in a proximity matrix as old clusters are merged into new ones. The $N \times N$ proximity matrix D contains all distances $d(i,j)$. The clusterings are assigned sequence numbers (0,1,……, [n − 1]), and $L(k)$ is the level of the kth clustering. A cluster with sequence number m is denoted (m), and the proximity between clusters (r) and (s) is denoted $d[(r),(s)]$.

The algorithm is composed of the following steps:

1. Begin with disjoint clustering at level $L(0)=0$ and sequence number $m=0$.
2. Find the most similar pair of clusters in the current clustering; say pair (r), (s), according to $d[(r),(s)]=\min d[(i),(j)]$ where the minimum is over all pairs of clusters in the current clustering.
3. Increment the sequence number: $m=m+1$. Merge clusters (r) and (s) into a single cluster to form the next clustering m. Set the level of this clustering to $L(m)=d[(r),(s)]$
4. Update the proximity matrix, D, by deleting the rows and columns corresponding to clusters (r) and (s), and adding a row and column corresponding to the newly formed cluster. The proximity between the new cluster, denoted (r,s), and old cluster (k) is defined as $d[(k),(r,s)]=\mathbf{min\ d[(k),(r)],\ d[(k),(s)]}$.
5. If all objects are in one cluster, stop. Otherwise, go to step 2.

Qi et al. [38] stated that an adaptation phase is crucial for a good and reasonable case-based design (CBD) process, which is responsible for finding a solution to solve a new problem in the principle of k-nearest neighbor (KNN). Their paper presented a new adaptation method for solution feature values of retrieved cases, which could be adapted to our injury dataset. Chen et al. [39] presented an effective and efficient diagnosis system using fuzzy k-nearest neighbor (FKNN) for the diagnosis of Parkinson's disease. The proposed FKNN-based system was compared to support vector machine (SVM)-based approaches. This algorithm corresponds with our research attempt to further improve the medical diagnosis accuracy of TBI and other head wounds. García-Pedrajas and Ortiz-Boyer [40] adhered to the principle that the KNN classifier is one of the most widely used methods of classification due to several interesting features, such as good generalization and easy implementation. Although simple, it is usually able to match, and even beat more sophisticated and complex methods, so that the accurate classification of difficult instances is favored. This is one of the reasons that we chose to employ the NN method for investigating our ship injury databases. Govindarajan and Chandrasekaran [41] utilized text data mining as a process of exploratory data analysis. In a similar manner, we utilized Hive and Hadoop to help classify data into predefined groups or classes such as head, torso, extremities, and chest. This is often referred to as supervised learning, because the classes are determined before examining the data and complements. This segment of our analysis supports our unsupervised learning discipline, which was employed in the clustering component of our algorithm. The authors' paper described the proposed KNN classifier, which tests the feasibility of performing comparative cross-validation. The benefits of the proposed approach were demonstrated using data mining problems, similar to our medical injury mortality study. Aci et al. [42] formed a hybrid method using five UCI machine learning datasets: iris, breast cancer, glass, yeast, and wine datasets. In our study, we also propose a hybrid approach on multiple ship databases. Li et al. [43] acknowledged that partially missing datasets are a prevailing problem in clustering analysis. In their paper, missing attributes were represented as intervals, and a novel FCM algorithm for incomplete data based on NN intervals was proposed. In regard to future issues, our study could easily adapt this approach to handle missing values found in our medical TBI data. Lee et al. [44] recently investigated microarray technology to study gene expression in cancer diagnosis. In the past, researchers have always used parametric statistical methods to find significant genes. However, microarray data often do not follow some of the assumptions of parametric statistical methods, or type I errors can be overexpanded. They established a gene selection method without assumption restriction to reduce the dimension of the dataset and to ensure that all test samples could be correctly classified. This was a similar problem in our TBI misclassification of head injuries regarding both Type I and Type II errors for mortality rates. Jiang et al. [45] recognized that text categorization is a significant tool to manage and organize the surging text data. Many text categorization algorithms have been explored in previous studies such as KNN, Naïve Bayes, and SVM. They proposed an improved KNN algorithm for text categorization, which builds the classification model by combining constrained one pass clustering algorithm and KNN text categorization. The classification model constructed by the proposed algorithm can be updated incrementally, and has great scalability in many real-word applications such as text mining injuries from our Hive Hadoop database. Castillo et al. [46] described a hybrid intelligent system for classification of cardiac arrhythmias. The hybrid approach was tested with the ECG records of the MIT-BIH Arrhythmia Database. The samples considered for classification contained four types of arrhythmias. The signals of the arrhythmias were segmented and transformed for improving the classification results. Three methods of classification were used to combine the outputs of the individual classifiers, and a very high classification rate of 98% was achieved. This approach could be useful for decreasing the misclassification of TBI morbidity and survival rates in our study. Muthukaruppan and Er [47] presented a PSO-based fuzzy expert system for the diagnosis of coronary artery disease, which was based on Cleveland and Hungarian Heart Disease datasets and yielded 93.27% classification accuracy. Because the datasets consisted of many input attributes, a decision tree was used to unravel the attributes that contributed towards diagnosis. We used Hive Hadoop and SDA to mine the attributes in our ship TBI injury study.

2.5. Symbolic data analysis

Yang et al. [48] pointed out that Kohonen's self-organizing map (SOM) is a competitive learning neural network that uses a neighborhood lateral interaction function to discover the topological structure hidden in the dataset. Unsupervised learning has both visualization and clustering properties. Although there are different

SOM clustering methods for numeric data with real applications in the literature, there is less consideration in a SOM clustering for symbolic data. Their experimental results showed the feasibility and effectiveness of their proposed algorithm in these real applications, and provided evidence that this approach could be applied to finding symbolic injury data in our study. Cury et al. [49] conceded that structural health monitoring is a problem that can be addressed at many levels, and that one of the more promising approaches used in damage assessment problems is based on pattern recognition. The idea of this approach is to extract features from data that only characterize the normal condition, and to use them as a template or reference. During structural monitoring, data are measured and the appropriate features are extracted and compared to the reference. Any significant deviations from the reference are considered signal novelty or damage. Some SDA techniques are applied for data classification: on one hand, the body of SDA is applied for classifying different structural behaviors, and on the other hand, for comparing any structural behavior to the previous classification when new data become available.

The results of their study were based on experimental tests performed on a railway bridge in France to demonstrate the efficiency of the described methodology. The authors found that the SDA methods efficiently classified and discriminated among structural modifications, considering the vibration data or modal parameters. We applied similar SDA techniques to our injury dataset to show misclassification of TBI mortality rates. Evsukoff et al. [50] proposed fuzzy symbolic modeling as a framework for intelligent data analysis and model interpretation in classification and regression problems. Their resulting model was evaluated based on a set of benchmark datasets for classification and regression problems. Non-parametric statistical tests were performed on the benchmark results. These tests show how the rule weights provide additional information to help in data and model understanding, such that it can be used as a decision support tool for the prediction of new data. Le-Rademacher and Billard [51] claimed that likelihood functions are the foundation of many statistical methodologies in classic data analysis, and contended that for symbolic data, these functions must be introduced before the classic methods can be extended to data analysis. They proposed the likelihood function for symbolic data and illustrated its applications by finding the maximum likelihood estimators for the mean and variance of three common types of symbolic-valued random variables: interval-valued, histogram-valued, and triangular-distribution-valued variables. Fagundes et al. [52] presented a robust regression model that dealt with cases that had interval-valued outliers in the input dataset. Two applications with real-life interval datasets were considered. The prediction quality was assessed by the mean magnitude of relative error calculated from a test dataset. Baumert et al. [53] investigated how the dynamics of cardiovascular variables were modulated by respiration, with the aim of assessing baroreflex function in normal subjects based on the joint symbolic dynamics of heart rate, blood pressure, and respiration. Symbolic analysis showed a significant influence of the respiratory phase on the occurrence of baroreflex patterns. Symbolic dynamics provide a simple representation of cardiovascular dynamics and may be useful for assessing baroreflex function. Suyal et al. [54] used rank order statistics to analyze the effects of time series data on the fluctuations of slow solar wind velocity. First, they applied rank order statistics to time series from known nonlinear systems, and then extended the analysis to solar wind data. They found that the underlying dynamics governing the solar wind velocity remains almost unchanged during an activity cycle. De Carvalho [55] presented adaptive and non-adaptive FCM clustering methods for partitioning symbolic interval data. His proposed methods furnished a fuzzy partition and prototype for each cluster by optimizing an adequacy criterion based on suitable squared Euclidean distances between vectors of intervals. In the current study, experiments with real and synthetic datasets show the usefulness of these FCM clustering methods and the merit of cluster interpretation tools.

Table 1
Total number of diagnoses by specialty according to mission.

Specialty	USNS *ROBERT E. BYRD*		USS *BOXER*		USS *KEARSARGE*	
	N	%	*N*	%	*N*	%
HEENT	10129	27.38	2527	9.25	2738	4.36
Optometry	8221	22.22	4790	17.54	8368	13.34
Dental	7114	19.23	0	0.00	1433	2.28
Pulmonary	1803	4.87	2513	9.20	6431	10.25
Cardio	478	1.29	231	0.85	1331	2.12
Gastrointestinal	826	2.23	3008	11.01	4953	7.89
Gynecological	238	0.64	1130	4.14	2372	3.78
Musculoskeletal	3340	9.03	3178	11.64	7522	11.99
Skin	2606	7.04	2111	7.73	3915	6.24
Neurological	251	0.68	771	2.82	3228	5.15
Trauma	56	0.15	81	0.30	167	0.27
Infectious disease	187	0.51	438	1.60	2555	4.07
Ophthalmology	0	0.00	3722	13.63	8303	13.23
Nephrology	0	0.00	0	0.00	1571	2.50
General	1519	4.11	1904	6.97	7186	11.45
Miscellaneous	226	0.61	910	3.33	665	1.06
Totals	**36994**	**100.00**	**27314**	**100.00**	**62738**	**100.00**

2.6. Results

Table 1 provides the total number of patient encounters onboard the three ships. The USS Kearsarge had the most encounters, with a total of 62,738 patients seen, whereas the USS Boxer had the fewest encounters, with 27,314 patients seen. While most patient visits were routine, trauma accounted for 0.15%, 0.30%, and 0.27% of visits onboard the Byrd, Boxer, and Kearsarge respectively. We data mined TBI mortality rates from these datasets to determine the misclassification errors.

Table 2 provides some pertinent ship and injury descriptive statistics such as the means, standard deviations, and ranges of the patient encounters. In this step of the process, we used the informatics ship database to data mine the major trauma

Table 2
Descriptive statistics of ships and injuries.

	Byrd	Boxer	Kersage	Head	Torso	Extremity	Minimal	Dead
N Valid	160	160	160	160	160	160	160	160
Missing	18	18	18	18	18	18	18	18
Mean	788.17	600.48	1180.35	203.15	146.56	188.53	147.19	68.73
Std. Deviation	3316.265	2315.504	5228.439	469.598	320.137	638.409	307.691	287.951
Variance	10997610.707	5361558.515	27336573.147	220522.493	102487.544	407565.961	94673.776	82915.505
Range	36994	27314	62738	2707	1873	7114	1694	2581
Minimum	0	0	0	0	0	0	0	0
Maximum	36994	27314	62738	2707	1873	7114	1694	2581

Table 3
Example of hive selection of IED blast data.

Int ID	DTG - Date	DTG- Time	FOV	Model	Vehicle ID	Unit	Country	Crew ID	Injuries	Position	BR	Serv
304219	1/1/10	8:55:00	MRAP_OBJECTIVE	MRAP_FPI_COUGAR_4X4	302341	US PMT DEFIANCE WHITE 6	AFG	300871	PER CIDNE: HEAD WOUNDED PER EOD: BROKEN WRIST PER DCIPS, no medical record for this event. Possible broken hand, dizzy and nauseated. PER JTAPIC	GUNNER		ARMY
304219	1/1/10	8:55:00	MRAP_OBJECTIVE	MRAP_FPI_COUGAR_4X4	302341	US PMT DEFIANCE WHITE 6	AFG	30872	ARMS WOUNDED			
304326	1/5/10	17:32:00	M1114	M1151	302403	B/1-7CAV	IRQ	300897	PT DIAG LEG INJURY	PASSENGER	ISN1461	TRANSLATOR
304326	1/5/10	17:32:00	M1114	M1151	302403	B/1-7CAV	IRQ	300898	INITIALLY REPORTED WITH UNSPECIFIED WOUND MEDEVAC REPORT STATES: PT DIAG POSSIBLE LEG AMPUTATION. -MDC Per DCIPS and medical record. Traumatic LLE above the knee amputation and mangled RLE with open fracture and fragmentation injuries. PER JTAPIC 1X COPPER FRAG	GUNNER	CBC0966	ARMY
304236	1/4/10	17:32:00	M1114	M1151	302403	B/1-7CAV	IRQ	300899	INITIALLY REPORTED WITH UNSPECIFIED WOUND. MEDEVAC REPORT STATES: PT DIAG HEAD AND EYE INJURY. -MDC Per DCIPS and medical record. Ruptured right globe s/p enucleation, right thumb I&D with revision of traumatic amputation at mid-metacarpal level Multiple PW R Mandibular region; PER JTAPIC	DRIVER	CBL2410	ARMY
304327	1/5/10	11:54:00	M1114	M1151	302404	17th FiB	IRQ	300902	SCRAPES AND ABRASIONS - INITIAL Per DCIPS and medical record. Minor lacerations on his face and superficial burns to the cheek. PER JTAPIC	GUNNER		ARMY

physiological areas of the head, torso, and extremities, as well as the rates of minimal injuries and mortality. We found instances of misclassification errors such as misclassified mortality rates, head injuries being listed as dental or neurological, torso injuries being listed as cardio, and injuries to the extremities being classified as musculoskeletal. Through extensive data analysis and knowledge discovery utilizing Apache Hadoop Hive, we were able to gain additional insights into injuries, particularly with regard to TBI injuries and mortality rates. Important supplementary data such as the unique patient identifier number, date, time, field of view, vehicle model, unit, country, crew, injuries, position, and branch of service allowed for cross validation of ship records. This Big Data analytics approach to data mining the informatics database was very instrumental in discovering the misclassification of injuries, as well as evaluating TBI survival versus mortality. Specifically, Hadoop was used to connect the nodes of the medical data, after which various algorithms, methods, and methodologies were employed to data mine the TBI mortality rates to discover misclassifications. The dataset was the end result of extraction, cleaning, and verification of pertinent information from the three ships. This extraction eliminated all other data that were not related to TBI and the variables used to investigate the misclassification phenomenon. These two major databases gave us the basis for implementing our algorithm to analyze the resulting transcription errors as well as Type I and Type II errors that were present in the misclassifications.

Table 3 shows an example of the big data components of the database. Information resided in not only numeric data but also in qualitative word and pdf files. While Hadoop was used to link the various medical nodes together, Hive provided the data mining tool that gathered together the resulting hybrid TBI information.

Fig. 1 illustrates how we employed SODAS canonical chaining to gain insights into the TBI misclassifications. Until this point, we had collected data on three ship variables (Byrd, Boxer, Kearsage) and four physiological body region injuries (head, torso, extremities, and abrasions). Because we were interested in how the set of collection variables relates to body injuries, we performed further investigations using canonical correlation analysis. Two dimensions or canonical variables were necessary to understand the association between the two sets of variables (survival vs. mortality). For the ship variables, the first canonical dimension of survival was most strongly influenced by Byrd (0.086), Boxer (0.108), Kersarge (0.125) and for the second dimension of mortality was (−0.046), (−0.052), and (−0.045), respectively. For the physiological variables (the first dimension), survival was composed of the head (0.266), torso (0.308), and extremities (0.453). For the second dimension, mortality was (0.307), (0.209), and (0.115). We concluded that the ships did a good job of ensuring positive survival rates fairly equally and therefore they all also had negative mortality rates. For the head, torso, and extremities, the first-dimension survival rates were positive. However, the second-dimension mortality rates were also positive, indicating that there may be some discrepancies in the classification of these three variables and mortality rates due to the misdiagnosis of TBI, concussion, and other head, torso, and extremity wounds.

Since we suspected discrepancies in the classification of variables and mortality rates due to misdiagnosis of TBI, concussion, and other head, torso, and extremity wounds, we decided to further investigate by SDA. In Fig 2, SDA was used to compare the three symbolic ship dimensions to the 17 injuries that fit into four categories: head/neck, torso abdomen, extremities, and abrasion/burn, as illustrated in Fig 3.

As seen in Fig. 3, SDA confirmed that the TBI (brain) was classified as being part torso and part extremity. This misclassification was confirmed by the discriminant results, and may have resulting impacts on the reported morbidity and survival results for TBI.

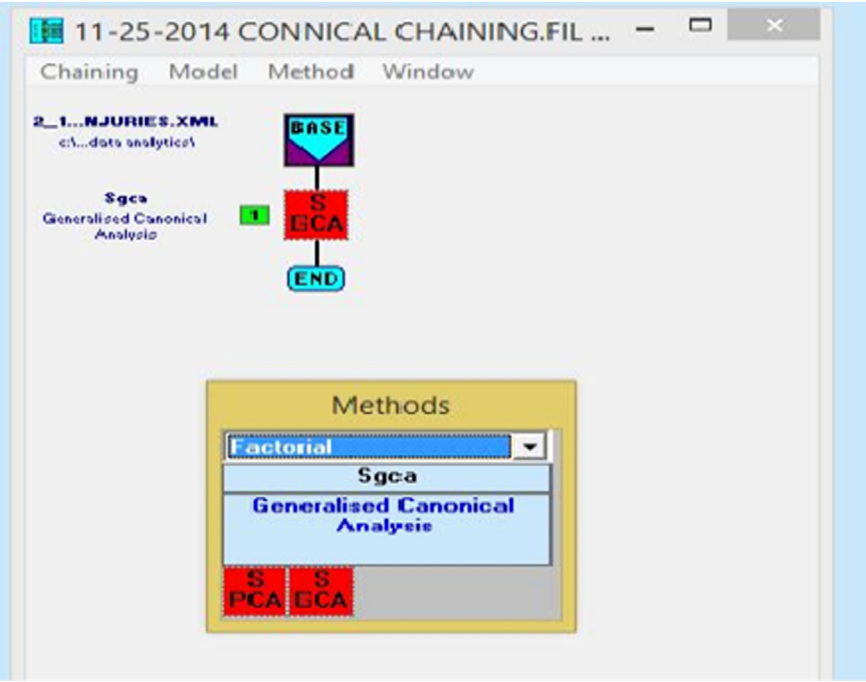

Fig. 1. SODAS Chaining.

	Brain	Head	Concussion	Eye	Neck	Teeth	Back	Chest	Torso	Hands	Arm	Leg	Fracture	Feet	Wrist	Minimal	Burn
Head Neck	[2.00 : 196.00]	[1.00 : 97.00]	[1.00 : 196.00]	[1.00 : 10.00]	[1.00 : 24.00]	[1.00 : 2.00]	[0.00 : 0.00]	[0.00 : 0.00]	[0.00 : 0.00]	[0.00 : 0.00]	[0.00 : 0.00]	[0.00 : 0.00]	[0.00 : 0.00]	[0.00 : 0.00]	[0.00 : 0.00]	[0.00 : 0.00]	[0.00 : 0.00]
Torso Abdomen	[1.00 : 38.00]	[0.00 : 0.00]	[0.00 : 0.00]	[0.00 : 0.00]	[0.00 : 0.00]	[0.00 : 0.00]	[1.00 : 38.00]	[1.00 : 20.00]	[1.00 : 5.00]	[0.00 : 0.00]	[0.00 : 0.00]	[0.00 : 0.00]	[0.00 : 0.00]	[0.00 : 0.00]	[0.00 : 0.00]	[0.00 : 0.00]	[0.00 : 0.00]
Extremeties	[3.00 : 249.00]	[0.00 : 0.00]	[0.00 : 0.00]	[0.00 : 0.00]	[0.00 : 0.00]	[0.00 : 0.00]	[0.00 : 0.00]	[0.00 : 0.00]	[0.00 : 0.00]	[1.00 : 2.00]	[1.00 : 249.00]	[1.00 : 57.00]	[1.00 : 13.00]	[1.00 : 3.00]	[1.00 : 8.00]	[0.00 : 0.00]	[0.00 : 0.00]
Abrasion Burn	[6.00 : 45.00]	[0.00 : 0.00]	[0.00 : 0.00]	[0.00 : 0.00]	[0.00 : 0.00]	[0.00 : 0.00]	[0.00 : 0.00]	[0.00 : 0.00]	[0.00 : 0.00]	[0.00 : 0.00]	[0.00 : 0.00]	[0.00 : 0.00]	[0.00 : 0.00]	[0.00 : 0.00]	[0.00 : 0.00]	[1.00 : 45.00]	[1.00 : 6.00]

Fig. 2. Symbolic data analysis.

It can be seen in Table 4 that the head and torso loaded more heavily on cluster two, whereas the extremities and minimal loaded more heavily on cluster 1, along with the ships. This provides evidence that good medical care and less extreme injuries were loaded on the survival factor, and that the more intensive injuries of the torso and head were loaded on the mortality factor. Simply put, there was a definite division between the cluster loads, with the more vital mortality factors loading separately from the less vital mortality body parts cluster.

Table 5 shows the high correlation between the ships and head injuries, although the correlation among the torso, extremities, and minimal were not significant. This may indicate that the majority of head injuries were transferred with equal success to the ships, whether they were TBI or less severe conditions such as concussions. In other words, the particular ship that the soldiers were transferred to did not impact the misclassification or the survival rate.

Table 6 compares those who died to those who survived. It indicates that of the 150,000 total patients, 2707 were originally classified as head injuries, with 138 TBI cases, which were only 50% of the original grouped cases, correctly classified as TBI mortalities. Only 35% of the cross-validated grouped cases were correctly classified. The large number of ungrouped cases may be due to misclassification of the TBI cases. It is worth noting that 13 cases were classified as dead when in fact they survived, and 7 cases were classified as surviving, when in fact they died. This raises the question of what criteria were employed to determine TBI versus a head or neck injury. In comparison, notice that the All Specialty

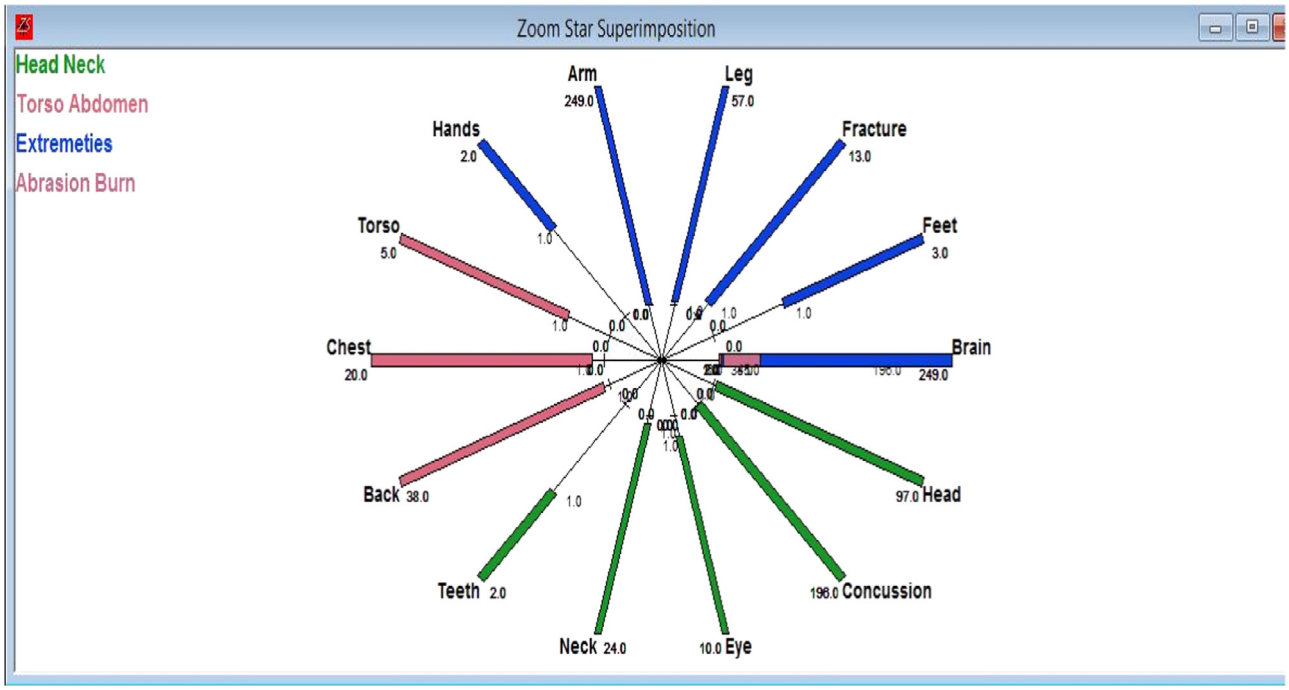

Fig. 3. Physiological variables.

Table 4
Final cluster centers.

	Cluster	
	Survival	Mortality
Byrd	36994	0
Boxer	27314	0
Kersage	62738	0
Head	1231	2707
Torso	0	145
Extremity	331	163
Minimal	713	75

Table 5
ANOVA.

Byrd	Cluster		Error		*F*	Sig.
	Mean square	df	Mean square	df		
Boxer	1319106632.959	1	2718439.680	158	485.244	.000
Kersage	718100546.340	1	850552.263	158	844.276	.000
Head	3813176627.444	1	3375560.145	158	1129.643	.000
Torso	1063120.123	1	215189.597	158	4.940	.028
Extremity	21613.821	1	102999.403	158	.210	.648
Minimal	20425.001	1	410016.221	158	.050	.824
	322157.268	1	93234.007	158	3.455	.065

Mortality Classification Results shown in Table 7 have no Type I or Type II errors, unlike the original TBI Mortality classification. This may provide evidence that the combination of our data mining of PIPSH[3] and our MMA algorithm led to an improved method with fewer transcription errors and misdiagnosis of TBI, than those that were reported in the original datasets. The difference between the mean percent of the original TBI Mortality Classification Results (50%) and the mean percent of the All Specialty Mortality Classification Results (100%) was significant with a t value of 12.838 and a p value of less than 0.000. Simply put, there was a significant difference, with 95% confidence, between the original and specialty results. Type I and Type II errors are the false positive and false negative errors, respectively, that detract from the correct classification. A comparative study with other algorithms was conducted (i.e., NN space). The results of both algorithms indicate that there were indeed misclassifications involving head injuries for TBI and concussions as false positives and false negatives, respectively.

Table 6
TBI mortality classification results[a,c].

Byrd		Dead	Predicted group membership		Total
Boxer			1	2	
Kersage head torso extremity	Count	1	5	13	18
		2	7	15	22
		Ungrouped cases	33	105	138
	%	1	27.8	72.2	100.0
		2	31.8	68.2	100.0
		Ungrouped cases	23.9	76.1	100.0
Byrd Boxer	Count	1	2	16	18
		2	10	12	22
	%	1	11.1	88.9	100.0
		2	45.5	54.5	100.0

Cross validation is done only for those cases in the analysis. In cross validation, each case is classified by the functions derived from all cases other than that case.

[a] Fifty percent of original grouped cases correctly classified.

[c] Thirty five percent of cross-validated grouped cases correctly classified.

Table 7
All specialty mortality classification results[a].

Mortality	**Predicted group membership**			
		1	2	Total
Original count	1	40	0	40
	2	0	80	80
Percent	1	100	0	100
	2	0	100	100
Cross-validated	Count 1	10	30	40

[a] Hundred percent of original grouped cases correctly classified.

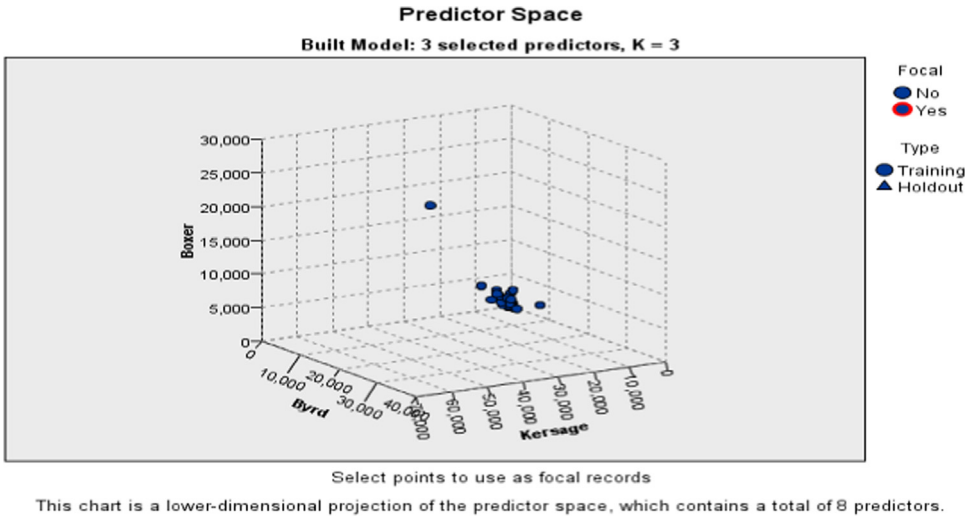

Fig. 4. This chart contains a total of 8 predictors within the predictor space.

Table 8
Case processing summary.

		N	Percent
Sample	Training	117	73.1
	Holdout	43	26.9
Valid		160	100.0
Excluded		18	
Total		178	

Table 8 shows the case-processing summary for the NN experiments. A total of 117 samples were trained, with 43 holdouts and 18 excluded samples. Fig. 4 illustrates that the predictor space gave excellent results, using the three ships of Byrd, Boxer, and Kersarge as predictors for all mortality and survival rates. It can clearly be seen that in conjunction with our other algorithm results, the predictor Head was outside the predictor space, possibly due to TBI cases being incorrectly classified as head wounds. These NN results seem to lend evidence to support the overall data mining results. Therefore, Fig. 4 demonstrates that with regard to the three ships, all of the affected body parts clustered together and had means within the predictor space. Only the head injuries lay outside the cluster, and were therefore susceptible to being misdiagnosed.

2.7. Conclusions, discussion, and future issues

This study used medical informatics to provide technical support to develop the necessary documentation that would ensure the correct classification of TBI cases onboard ships, applying the PIPSH[3] and MMA for data analysis. Our method provides analytical support for personnel to access and manipulate data within the three ship databases. By adopting this approach, medical informatics support personnel can develop new analytical solutions to improve TBI classifications in their medical objectives. This approach is different from low-technology solutions that capture patient data in support of missions, to assist in the certification of diagnosis for use in ship databases. The adopted medical informatics approach enables ship healthcare personnel to adopt technical and analytical applications to meet information standards. In addition, it provides recommendations and potential enhancements to the management of medical information, and provides novel management solutions. In addition, our method provides additional medical analytical support to ships through the development of the Apache Hadoop Hive injury search application and PIPS medical encounter form scan technology. This search application allows ships to provide a background search for various medical diagnoses and symptoms for health information data that may be inadvertently transcribed or incorrectly collected. This capability allows ships to ensure they are meeting the appropriate conduct standards in their research and reporting efforts. The form scan technology enhances the accurate collection of data and the capability to automate paper medical encounters in a large database. Scanning software is used to retrospectively collect handwritten patient data from ship missions. Our method facilitates the integration of blast-related data from multiple ship reports and activities to improve understanding of vulnerabilities to blast threats and enable the development of improved classification requirements to help with the TBI recovery process, ship materiel solutions, and projection models for future missions. This research could aid the development of leading-edge operational medicine capabilities to assist research operational modalities, and the development and enhancement of combat casualty care. These capabilities will continue to enhance operational medicine support to ships and improve TBI projections of mortality rates. Big Data analytical support could provide a foundation for operational medicine and help ships meet their objectives in the development of a wound surveillance analysis network and implementation of a personnel protective equipment (PPE) and materiel solution surveillance system. Our method further refines ship medical analysis capability by providing a “value added” construct to the clinical data (by injecting operational mission data elements), such as helping to correctly classify TBI data using a new algorithm. In addition, it provides direct support to ships, initial development of shipboard blast injury studies, and functional and technical support to health informatics processes. Our method also supports analysis through functional review, modeling, and simulation, and can establish a ship's strategic position in the forefront of expeditionary medical requirements.

In conclusion, our method transforms shipboard medical efforts using Big Data analytic capabilities and helps correct oversights of TBI cases that are incorrectly classified. Our work also integrates medical injuries databases that take emerging operational medicine concepts and align them with defined capabilities and identified capability gaps. It also facilitates capabilities-based assessments of whether TBI data are correctly classified. For planning purposes, a ship's future missions can be coordinated with this medical research, which can lead to the development of protocols that monitor mature and emerging medical technologies. The research can also be used as a model to evaluate and analyze other potential ship Big Data analytic technologies, analytics, and algorithms.

Conflict of interest statement

The author does not have any financial or personal relationships with other people or organizations that could have inappropriately influenced or biased this work.

References

[1] Hughes G. How big is 'big data' in healthcare? ⟨⟨http://blogs.sas.com/content/hls/2011/10/21/how-big-is-big-data-in-healthcare/⟩⟩; [accessed 21.10.11].
[2] Rodger JA. Utilization of data mining techniques to detect and predict accounting fraud: a comparison of neural networks and discriminant analysis. In: Kosrapohr, editor. Managing data mining technologies in organizations: techniques and applications. Hershey: Idea Group Publishing; 2003. p. 174–87.
[3] Bichler M, Heinzl A, Winter R. Diversity and quality in BISE research. Bus Inf Syst Eng 2014;6:313–20.
[4] Dhar V, Jarke M, Laartz J. Big data. Bus Inf Syst Eng 2014;6:257–62.
[5] Schermann M, Hemsen H, Buchmuller C, Bitter T, Krcmar H, Markl V, Hoeren T. Big data: an interdisciplinary opportunity for information systems research. Bus Inf Syst Eng 2014;6:261–75.
[6] Liang R, Wang Z, Zhang S, Feng Y, Jiang L, Ma X, Chen W, Tate DF. Visual exploration of HARDI fibers with probabilistic tracking. Inform Sci 2015;30:30.
[7] Seng JL, Chen TC. An analytic approach to select data mining for business decision. Expert Syst Appl 2010;37:8042–57.
[8] Debortoli S, Muller O, Brocke J. Comparing business intelligence and big data skills: a text mining study using job advertisements. Bus Inf Syst Eng 2014;6:289–95.
[9] Baars H, Felden C, Gluchowski P, Hilbert A, Kemper H-G, Olbrich S. Shaping the next incarnation of business intelligence: towards a flexibly governed network of information integration and analysis capabilities. Bus Inf Syst Eng 2014;6:11–21.
[10] Fisher RA. The use of multiple measurements in taxonomic problems. Ann Eugen 1936;7:179–88.
[11] Hotelling H. Relations between two sets of variates. Biometrika 1936;28:321–77.
[12] Griffiths JD, Williams JE, Wood RM. Modelling activities at a neurological rehabilitation unit. Eur J Oper Res 2013;226:301–12.
[13] Cruz AM, Rincon AMR. Medical device maintenance outsourcing: have operation management research and management theories forgotten the medical engineering community? A mapping review Eur J Oper Res 2012;221:186–97.
[14] Yang X, Zheng D, Sieminowski T, Paradi JC. A dynamic benchmarking system for assessing the recovery of inpatients: evidence from the neurorehabilitation process. Eur J Oper Res 2015;240:582–91.
[15] Kunene KN, Weistroffer HR. An approach for predicting and describing patient outcome using multicriteria decision analysis and decision rules. Eur J Oper Res 2008;185:984–97.
[16] Lin AP, Blüml S. Traumatic brain injury and concussion. In: Blüml S, Panigrahy A, editors. MR spectroscopy of pediatric brain disorders. Berlin: Springer; 2013. p. 67–75.
[17] Langlois JA, Rutland-Brown W, Wald MM. The epidemiology and impact of traumatic brain injury: a brief overview. J Head Trauma Rehabil 2006;21:375–8.
[18] Hoge CW, McGurk D, Thomas JL, Cox AL, Engel CC, Castro CA. Mild traumatic brain injury in U.S. soldiers, returning from Iraq. New Engl J Med 2008;358:15–27.
[19] Lu J, Murray GD, Steyerberg EW, Butcher I, McHugh GS, Lingsma H, Mushkudiani N, Choi S, Maas A, Marmarou IA. Effects of glasgow outcome scale misclassification on traumatic brain injury clinical trials. J Neurotrauma 2008;25:641–51.
[20] Lu J, Marmarou A, Lapane KL. Impact of GOS misclassification on ordinal outcome analysis of traumatic brain injury clinical trials. J Neurotrauma 2011;29:719–26.
[21] Bendler J, Wagner S, Brandt T, Neumann D. Taming uncertainty in big data: Evidence from social media in urban areas. Bus Inf Syst Eng 2014;6:279–89.
[22] Sohlberg MM, Mateer CA. Cognitive rehabilitation: an integrative neuropsychological approach. 1st ed.. New York: Guilford Press; 2001.
[23] Kowalczyk M, Buxmann P. Big data and information processing in organizational decision processes: a multiple case study. Bus Inf Syst Eng 2014;6:267–78.
[24] MacKay DJC. Information theory, inference and learning algorithms. London: Cambridge University Press; 2003.
[25] Pimentel BA, de Souza RM. A weighted multivariate Fuzzy C-Means method in interval-valued scientific production data. Expert Syst Appl 2014;41:3223–36.
[26] Krishnasamy G, Kulkarni AJ, Paramesran R. A hybrid approach for data clustering based on modified cohort intelligence and K-means. Expert Syst Appl 2014;41:6009–16.
[27] Elango M, Nachiappan S, Tiwari MK. Balancing task allocation in multi-robot systems using K-means clustering and auction based mechanisms. Expert Syst Appl 2011;38:6486–91.
[28] Yin M, Hu Y, Yang F, Li X, Gu W. A novel hybrid K-harmonic means and gravitational search algorithm approach for clustering. Expert Syst Appl 2011;38:9319–24.
[29] Hadavandi E, Shavandi H, Ghanbari A. An improved sales forecasting approach by the integration of genetic fuzzy systems and data clustering: case study of printed circuit board. Expert Syst Appl 2011;38:9392–9.
[30] Dimoulas CA, Papanikolaou GV, Petridis V. Pattern classification and audio-visual content management techniques using hybrid expert systems: a video-assisted bioacoustics application in abdominal sounds pattern analysis. Expert Syst Appl 2011;38:13082–93.
[31] Celebi ME, Kingravi HA, Vela PA. A comparative study of efficient initialization methods for the k-means clustering algorithm. Expert Syst Appl 2013;40:200–10.
[32] Higuera C, Pajares G, Tamames J, Morán F. Expert system for clustering prokaryotic species by their metabolic features. Expert Syst Appl 2013;40:6185–94.
[33] Birtolo C, Ronca D. Advances in clustering collaborative filtering by means of fuzzy C-means and trust. Expert Syst Appl 2013;40:6997–7009.
[34] Lin C, Chen C, Lee H, Liao J. Fast k-means algorithm based on a level histogram for image retrieval. Expert Syst Appl 2014;41:3276–83.
[35] Bai C, Dhavale D, Sarkis J. Integrating fuzzy c-means and TOPSIS for performance evaluation: an application and comparative analysis. Expert Syst Appl 2014;41:4186–96.
[36] Sancho-Asensio A, Navarro J, Arrieta-Salinas I, Armendáriz-Íñigo JE, Jiménez-Ruano V, Zaballos A, Golobardes E. Improving data partition schemes in smart grids via clustering data streams. Expert Syst Appl 2014;41:5832–42.
[37] Skiena SS. The algorithm design manual. 2nd ed.. Berlin: Springer; 2008.
[38] Qi J, Hu J, Peng Y. A new adaptation method based on adaptability under k-nearest neighbors for case adaptation in case-based design. Expert Syst Appl 2012;39:6485–502.
[39] Chen H, Huang C, Yu X, Xu X, Sun X, Wang G, Wang S. An efficient diagnosis system for detection of Parkinson's disease using fuzzy k-nearest neighbor approach. Expert Syst Appl 2013;40:263–71.
[40] García-Pedrajas N, Ortiz-Boyer D. Boosting k-nearest neighbor classifier by means of input space projection. Expert Syst Appl 2009;36:10570–82.
[41] Govindarajan M, Chandrasekaran RM. Evaluation of k-nearest neighbor classifier performance for direct marketing. Expert Syst Appl 2010;37:253–8.
[42] Aci M, Inan C, Avci M. A hybrid classification method of k nearest neighbor, Bayesian methods and genetic algorithm. Expert Syst Appl 2010;37:5061–7.
[43] Li D, Gu H, Zhang L. A fuzzy c-means clustering algorithm based on nearest-neighbor intervals for incomplete data. Expert Syst Appl 2010;37:6942–7.
[44] Lee C, Lin W, Chen Y, Kuo B. Gene selection and sample classification on microarray data based on adaptive genetic algorithm/k-nearest neighbor method. Expert Syst Appl 2011;38:4661–7.
[45] Jiang S, Pang G, Wu M, Kuang L. An improved K-nearest-neighbor algorithm for text categorization. Expert Syst Appl 2012;39:1503–9.
[46] Castillo O, Melin P, Ramírez E, Soria J. Hybrid intelligent system for cardiac arrhythmia classification with fuzzy k-nearest neighbors and neural networks combined with a fuzzy system. Expert Syst Appl 2012;39:2947–55.
[47] Muthukaruppan S, Er MJ. A hybrid particle swarm optimization based fuzzy expert system for the diagnosis of coronary artery disease. Expert Syst Appl 2012;39:11657–65.
[48] Yang M, Hung W, Chen D. Self-organizing map for symbolic data. Fuzzy Set Syst 2012;203:49–73.
[49] Cury A, Crémona C, Diday E. Application of symbolic data analysis for structural modification assessment. Eng Struct 2010;32:762–75.
[50] Evsukoff AG, Branco ACS, Galichet S. Intelligent data analysis and model interpretation with spectral analysis fuzzy symbolic modeling. Int J Approx Reason 2011;52:728–50.
[51] Le-Rademacher J, Billard L. Likelihood functions and some maximum likelihood estimators for symbolic data. J Stat Plan Infer 2011;141:1593–602.
[52] Fagundes RA, de Souza RMCR, Cysneiros FJA. Robust regression with application to symbolic interval data. Eng Appl Artif Intel 2013;26:564–73.
[53] Baumert M, Javorka M, Kabir MM. Joint symbolic analyses of heart rate, blood pressure, and respiratory dynamics. J Electrocardiol 2013;46:569–73.
[54] Suyal V, Prasad A, Singh HP. Symbolic analysis of slow solar wind data using rank order statistics. Planet Space Sci 2012;62:55–60.
[55] de Carvalho F. Fuzzy c-means clustering methods for symbolic interval data. Pattern Recogn Lett 2007;28:423–37.