International Journal of Healthcare Information Systems and Informatics
Volume 14 • Issue 3 • July-September 2019

Using Data Analytics to Predict Hospital Mortality in Sepsis Patients

Yazan Alnsour, University of Illinois at Springfield, Springfield, USA
Rassule Hadidi, University of Illinois at Springfield, Springfield, USA
Neetu Singh, University of Illinois at Springfield, Springfield, USA

ABSTRACT

Predictive analytics can be used to anticipate the risks associated with some patients, and prediction models can be employed to alert physicians and allow timely proactive interventions. Recently, health care providers have been using different types of tools with prediction capabilities. Sepsis is one of the leading causes of in-hospital death in the United States and worldwide. In this study, the authors used a large medical dataset to develop and present a model that predicts in-hospital mortality among Sepsis patients. The predictive model was developed using a dataset of more than one million records of hospitalized patients. The independent predictors of in-hospital mortality were identified using the chi-square automatic interaction detector. The authors found that adding hospital attributes to the predictive model increased the accuracy from 82.08% to 85.3% and the area under the curve from 0.69 to 0.84, which is favorable compared to using only patients' attributes. The authors discuss the practical and research contributions of using a predictive model that incorporates both patient and hospital attributes in identifying high-risk patients.

KEYWORDS

Health Analytics, Health IT, In-Hospital Mortality, Predictive Models, Sepsis

INTRODUCTION

The Centers for Medicare and Medicaid Services (CMS) estimated that the healthcare expenditure in the United States in 2016 alone was $3.4 trillion, an increase of about 4.8% from 2015. According to the CMS, this trend will continue, with an average growth rate of 5.6% per year until 2025 ("National Health Expenditure Data," 2018), and this significant growth is partially the result of an aging population, increased lifespans, and growing costs associated with repeated hospital visits for patients with serious infectious diseases.

Sepsis is an inflammatory condition caused by infection and results in a relatively high mortality rate (Amland & Hahn-Cover, 2016; Singer et al., 2016). Sepsis and septic shock are common causes of morbidity and mortality (Raghavan & Marik, 2006). The condition of patients with sepsis can change from stable to near death in a very short period, from days to even just several hours (Taneja et al., 2017). Early diagnosis and prompt treatment have been associated with improved outcomes and

DOI: 10.4018/IJHISI.2019070104

alleviated risks (Nachimuthu & Haug, 2012). During the past decade, cases of sepsis have increased significantly (Raghavan & Marik, 2006), and sepsis is considered to be one of the main causes for admission to intensive care (Vincent et al., 1996). In the United States, for example, sepsis accounts for more cases of death than prostate cancer, breast cancer, and AIDS combined ("Sepsis Fact Sheet," 2016). The Agency for Healthcare Research and Quality (AHRQ) indicated that sepsis is the most expensive condition treated in U.S. hospitals, costing more than $20 billion in 2011, with an average annual increase of 11.9%. Fleischmann et al. (2016) argued that reducing the burden of sepsis is a global challenge, and many countries suffer from a high level of mortality and morbidity from it. Given the current state of sepsis complications and the significant social and economic benefits of better treatment of this disease, it is paramount for health care providers to find more effective ways to deal with sepsis and improve the health outcomes of sepsis patients.

According to the Centers for Disease Control and Prevention, sepsis can be a challenge for medical providers and health care professionals (HCPs) because there is no standard diagnostic test for it (Epstein, 2016). Sepsis diagnosis relies on the judgment of HCPs. Furthermore, it has been estimated that if a country like the United States can achieve an earlier sepsis identification and evidence-based treatment, there will be 92 thousand fewer deaths and savings of more than 1.5 billion in medical cost annually (Shorr, Micek, Jackson, & Kollef, 2007). Predictive analytics in healthcare facilities has mostly been limited to the simple heuristics and scoring systems. HCPs can leverage predictive analytics and machine-learning (ML) techniques to harness the variables available through electronic health records (EHRs) to better predict patient outcomes (Bhattacharjee, Edelson, & Churpek, 2017).

In this study, we explored how predictive models can be used to identify patients with a high mortality risk while they are hospitalized. We empirically developed and examined our model using a large archival dataset from the National Inpatient Sample (NIS), which was created by the AHRQ through a federal-state-industry partnership. A total of 1,048,575 sepsis patients' records were collected between the years 2008 and 2012. This sample was subdivided randomly into training, testing, and validation data partitions. Death during hospitalization (in-hospital mortality) was defined as a target variable. The independent predictors of mortality were identified using the Chi-square Automatic Interaction Detector (CHAID) model in testing. Additionally, the model was validated using receiver-operating characteristic (ROC) curves and accuracy metrics.

BACKGROUND

The healthcare literature indicates that patients' mortality risk is higher when patients do not receive appropriate medical interventions and follow-ups (Schmitt et al., 2013). HCPs need the appropriate tools to make the right decisions and take the needed actions at the right time to enhance healthcare outcomes (Koh & Tan, 2011; Lee et al., 2003; Martin et al., 2009; Mehta et al., 2002; Varlamis et al., 2017). A major benefit anticipated from advancements in healthcare information systems and big data is the ability to create new insights by applying advanced data analytics to enhance health care delivery and quality (Hagland, 2011; Kociol et al., 2012; Varlamis et al., 2017).

Clinical support tools can help overwhelmed medical professionals better manage a patient's health (Chen, Chiang, & Storey, 2012; Srinivas, Rani, & Govrdhan, 2010). One of the distinct new features of clinical support tools is the ability to provide HCPs with some type of outcome predictions. Clinical support tools that have predictive capabilities can provide HCPs with early alerts to enhance clinical decisions and actions for a better level of care and a reduction of unnecessary risk (Mitchell et al., 2016; Varlamis et al., 2017). Insights regarding the prediction of mortality risks, for example, can help HCPs and hospital management identify patients at high risk of in-hospital mortality. This will allow HCPs to be proactive in their actions and apply early interventions to improve patients' health outcomes, alleviate unnecessary risks, and unnecessary cost (Mitchell et al., 2016; Roshanov et al., 2013; Varlamis et al., 2017). Further, predictive models

and clinical decision support tools will help HCPs make better decisions and lower the risk of in-hospital mortality.

When surveying the literature, we found that existing predictive models are mostly based on limited data rather than a large dataset derived from multiple sources (Mitchell et al., 2016; Sommerfeld, Althouse, Prince, & Hickey, 2016). Most prior studies either used explanatory rather than predictive statistics or used traditional statistical models to predict mortality (Amarasingham et al., 2015; Bardhan, Oh, Zheng, & Kirksey, 2014; Bardhan & Thouin, 2013; Futoma, Morris, & Lucas, 2015). Nonetheless, models focusing on predicting mortality using data-mining tools have recently emerged (Krumholz et al., 2000; Tsugawa et al., 2017; van Walraven et al., 2010; Wang et al., 2014). Such tools require large and broad data sets that are sometimes challenging to obtain. Some recent studies focusing on predictive analytics are presented in Table 1, and some studies which are employing data mining to predict mortality are presented in Appendix A (see Table 9).

Scholars such Rezaei Hachesu, Moftian, Dehghani, and Samad-Soltani (2017) developed a framework to assess the risk factors and outcomes of mortality, but they did not report the accuracy of their developed models, proposing that future research do so. At the same time, other studies such as Paoin (2011) paper that assessed the accuracy of its prediction model using a large dataset of more than 1,000,000 records from the World Health Organization (WHO) database, but the developed models had an accuracy rate of less than 40%; the researcher pointed out that the reason for such low accuracy is the inadequacy of the used variables. In another more recent study by Varlamis et al. (2017), researchers found that decision trees can predict cancer survival with an accuracy of 80%, outperforming other algorithms. Although the results of the study are promising, the model was tuned toward cancer patients and may not be applicable to other diseases. Another study by Delen, Walker, and Kadam (2005) reported a higher accuracy percentage of 93.6% using the C5 decision tree. This model was also tuned for breast cancer survivability, and it may not produce the same results for other diseases. Examples of recent studies that focused on predicting sepsis mortality are shown in Table 2.

According to Lin et al. (2017), predictive analytics is an emerging research area with significant practical value. Researchers have also highlighted how modeling risks of health events can provide clinical intelligence for preventive interventions. Lee and Yoon (2017) argued that the need to improve health care quality and patient outcomes is driving ongoing research in health care predictive analytics and big data. Bandyopadhyay et al. (2015) argued that most current and commonly used

Table 1. Example of recent studies focusing on predictive analytics in healthcare

Study	Methodology	Findings/Results
(Veith & Steele, 2018)	Applying multiple ML algorithms to 58,000 encounters	LazyKStar was the best model and was ranked number 1 according to the ROC curve value compared to other models.
(Walczak & Okuboyejo, 2017)	The use of artificial neural networks (ANN) to predict medication non-adherence	ANN models were able to predict 63% of nonadherence reasons. After removal of two highly common nonadherence reasons, new models are able to identify 73% of the remaining nonadherence reasons.
(Lin, Chen, Brown, Li, & Yang, 2017)	Multitask learning strategy on big EHR data to develop a novel approach to simultaneously model and predict patient's risk	Bayesian multitask learning approach can augment health care providers' capability in identifying high-risk patients.
(Bandyopadhyay et al., 2015)	ML approach based on Bayesian networks trained on electronic health database to predict a cardiovascular event	The ML approach can lead to better predictive performance than the Cox proportional hazards model or a BN with ad hoc approaches to right censoring.

Table 2. Example of recent studies focusing on predicting Sepsis mortality

Study	Data	Method	Findings/Results
(Ghosh, Li, Cao, & Ramamohanarao, 2017)	Experimental dataset	Experiment	Single and multi-channel patterns coupled with HMM have better accuracy compared to baseline models.
(Ford et al., 2016)	Administrative data from 5 U.S. states	Retrospective cohort study	Maximum likelihood estimation logistic regression (LR) was used to develop a predictive model for in-hospital mortality.
(Widgren & Jourak, 2011)	8,695 patients from ED referred to hospital	Developing a new protocol	Medical Emergency Triage and Treatment System triage method is a sensitive tool to find those in need of immediate medical attention.
(Lee et al., 2008)	525 sepsis adult patients admitted to ED	Observational study	The main outcome was early and late mortality predictions.
(Nguyen et al., 2007)	330 patients from sepsis registry	LR to identify significant indicators	The sepsis LR model can be used among ED physicians and nurses.
(Howell, Donnino, Clardy, Talmor, & Shapiro, 2007)	1,287 adults admitted to ED	A model to assess 28-day in-hospital mortality	The venous lactate level predicts 28-day in-hospital mortality independent of blood pressure and adds significant prognostic information.

risk-prediction models had been built from carefully selected patient cohorts of limited size. The authors explained how the homogeneity and limited size of such cohorts restrict the predictive power and generalizability of such models to other populations.

Multiple studies aimed to predict some key medical outcomes using a large dataset (I. Bardhan et al., 2014). Although there are a plethora of studies in the literature that aim to develop mortality-risk prediction models, to our knowledge these studies do not account for hospitals' or health care providers' different characteristics. Thus, most of the models in the literature do not have adequate prediction accuracy to be deployed in a medical setting (Yu et al., 2015). Sepsis is considered to be a type of health-care-acquired infection; that is, an infection that is acquired while patients are in a healthcare facility ("Sepsis and Healthcare-Acquired Infections," 2017). Thus, some health care facilities or health care providers' characteristics may help in predicting death due to septic shock. In this study, we used a large national dataset that included attributes at both the patient level and health care provider level. This allowed us to investigate whether health care provider attributes can increase the accuracy of mortality prediction among sepsis patients and bridge the existing gap in the literature.

DATA AND VARIABLES

This retrospective study used a dataset of 1,048,575 sepsis patients who were hospitalized between the years 2008 and 2012. The data was extracted from the NIS. The NIS is a set of longitudinal hospital inpatient databases included in the Healthcare Cost and Utilization Project family. These databases were created by the AHRQ through a federal-state-industry partnership. In addition to patient attributes (Table 3), we extracted hospital and provider-related information and attributes (Table 4). Hospital information is general information regarding the size of the hospital, whether the hospital is in a rural or urban area, whether the hospital is a teaching hospital or not, and other variables related to the healthcare facility. The dataset used for developing the predictive model consisted of both patient and healthcare facility attributes.

Table 3. Patient-related attributes

Variable	Description	Values
DIED	Died during hospitalization, target variable	0 = did not die, 1 = died
GENDERE	Indicator of gender	0 = male, 1 = female
AGE	Age in years at admission	(0–124) age in years
RACE	Race	1 = White, 2 = Black, 3 = Hispanic, 4 = Asian or Pacific Islander, 5 = Native American, 6 = other
YEAR	The calendar year	4-digit calendar year
QTR	Quarter	1 = first quarter (Jan–Mar), 2 = second quarter (Apr–Jun), 3 = third quarter (Jul–Sep), 4 = fourth quarter (Oct–Dec)
AMONTH	Admission month	(1–12) admit month
AWEEKEND	Admission day is on a weekend	0 = admitted Monday–Friday, 1 = admitted Saturday–Sunday
ATYPE	Indicates the type of admission	1 = emergency, 2 = urgent, 3 = elective, 4 = newborn, 5 = trauma center, 6 = other
TRAN_IN	An indicator of a transfer to the hospital	0 = not transferred in, 1 = transferred in from a different acute care hospital, 2 = transferred in from another type of health facility
ELECTIVE	Elective versus nonelective admission	0 = nonelective admission, 1 = elective admission
LOS	Length of stay in days, continuous variable	(0–365) days
CM_X	AHRQ comorbidity measure for ICD-9-CM codes*	0 = comorbidity is not present, 1 = comorbidity is present
APRDRG Severity	All patient refined DRG: Severity of illness subclass	0 = no class specified, 1 = minor loss of function, 2 = moderate loss of function, 3 = major loss of function, 4 = extreme loss of function
NCHRONIC	ICD-9-CM number of chronic conditions	(0–nn) ICD-9-CM number of chronic conditions
NDX	Number of ICD-9-CM diagnoses	(0–nn) Number of diagnoses
NPR	Number of ICD-9-CM procedures	(0–nn) Number of procedures
PAY1	Expected primary payer	1= Medicare, 2 = Medicaid, 3 = private insurance, 4 = self-pay, 5 = no charge, 6 = other
PAY2	Expected secondary payer	1= Medicare, 2 = Medicaid, 3 = private insurance, 4 = self-pay, 5 = no charge, 6 = other

*AHRQ comorbidity measure for ICD-9-CM codes for acquired immune deficiency syndrome, alcohol abuse, deficiency anemia, rheumatoid arthritis/collagen vascular diseases, chronic blood loss anemia, congestive heart failure, chronic pulmonary disease, coagulopathy, depression, uncomplicated diabetes mellitus, diabetes with chronic complications, drug abuse, hypertension both uncomplicated and complicated, hypothyroidism, liver disease, lymphoma, fluid and electrolyte disorders, metastatic cancer, other neurological disorders, obesity, paralysis, peripheral vascular disorder, psychoses, pulmonary circulation disorders, renal failure, solid tumor without metastasis, peptic ulcer disease excluding bleeding, valvular disease, and weight loss.

DATA PREPARATION

The data were uploaded to a relational database. We conducted exploratory data analysis (EDA) as a first step to analyze the data (Martinez, Martinez, Martinez, & Solka, 2010; Velleman & Hoaglin, 1981). In this step, we checked for preliminary selection of different models, determined the relationships among the explanatory variables, and assessed the direction and rough size of the

Table 4. Hospital-related attributes

Variable	Description	Values
HOSP_BEDSIZE	Bed size of hospital	1 = small, 2 = medium, 3 = large
HOSP_CONTROL	Control/ownership of hospital	0 = government, 1 = government nonfederal, 2 = private not-for-profit, 3 = private investor owned
HOSP_LOCATION	Location (urban/rural) of hospital	0 = rural, 1 = urban
HOSP_LOCTEACH	Location/teaching status of hospital	1 = rural, 2 = urban nonteaching, 3 = urban teaching
HOSP_TEACH	Teaching status of hospital	0 = nonteaching, 1 = teaching
HOSP_REGION	Region of hospital	1 = northeast, 2 = Midwest, 3 = south, 4 = west
HOSP_RNPCT	Percentage of registered nurse (RN) among all nurses (RNs and licensed practical nurse [LPNs])	(0–100%) Percentage
HOSP_RNFTEAPD	RN full-time employees (FTEs) per 1,000 adjusted inpatient days	Continuous decimal number
HOSP_LPNFTEAPD	LPN FTEs per 1,000 adjusted inpatient days	Continuous decimal number
HOSP_NAFTEAPD	Nurse aides per 1,000 adjusted inpatient days	Continuous decimal number
HOSP_OPSURGPCT	Percentage of all surgeries performed in the outpatient setting	(0–100%) Percentage
MDNUM1_R	Physician 1 number	9-digit identifier
MDNUM2_R	Physician 2 number	9-digit identifier

relationships between explanatory and outcome variables (Hartwig & Dearing, 1979; Martinez et al., 2010). Table 5 shows the descriptive statistics for patient-related attributes, and Table 6 shows the descriptive statistics for healthcare facility-related attributes. Only a small number of records were missing complete information. There were 272 records missing the APRDRGSEVERITY, 100 records missing the ELECTIVE variable, 2,781 records missing the QTR data, and 7,726 missing the admission month. None of the other variables in our sample were a concern during the EDA.

DESIGN AND DEVELOPMENT

We used SPSS Modeler 18.1 Premium Edition to build the predictive model. The data was first installed on a local relational database. A link was created between the modeler and the relational database to obtain the data needed for the modeling. In the first phase, we filtered the dataset and included only patient-related attributes (45 attributes) to build different datamining models. After that, variables were defined according to their type (continuous, ordinal, nominal, and binary). We partitioned the data into a training set (70% of the data), a testing set (20%), and a validation set (10%). The training set was used by the modeler to adjust the model, the testing set was used to assess the performance of the model, and validation was used to tune the model and avoid overfitting.

We applied multiple ML models that are commonly used in the healthcare literature, such as logistic regression (LR) (Celi et al., 2012; Howell et al., 2007; Nguyen et al., 2007), random tree (RT) (Austin, 2010), Bayesian network (BN), neural networks (NN) (Celi et al., 2012), support vector machine (SVM) (Wang, Wu, & Wang, 2010), and Quest (Sut & Simsek, 2011). Additionally, we used

Table 5. Descriptive statistics of patient-related attributes

Variable	Measure	Minimum	Maximum	Mean	Std. Dev.
DIED	Binary	0	1	0.18	0.39
FEMALE	Binary	0	1	0.51	0.50
AGE	Continuous	0	100	66.78	18.00
RACE	Nominal	1	6	1.57	1.11
YEAR	Nominal	2008	2012	2009.83	1.277
QTR	Nominal	1	4	2.51	1.125
AMONTH	Nominal	1	12	6.53	3.477
AWEEKEND	Binary	0	1	0.25	0.43
ATYPE	Nominal	1	6	1.36	0.666
TRAN_IN	Nominal	0	3	1.21	0.84
ELECTIVE	Binary	0	1	0.09	0.279
LOS	Continuous	0	364	10.732	13.52
APRDRG Severity	Continuous	1	4	3.436	0.73
NCHRONIC	Continuous	0	27	6.11	3.15
NDX	Continuous	1	72	15.20	5.89
NPR	Continuous	0	39	2.91	3.48
PAY1	Nominal	1	6	1.69	1.121
PAY2	Nominal	1	6	2.55	1.181
CM_X		Binary		*See Table 10, Appendix B*	

Table 6. Descriptive statistics of healthcare facility attributes

Variable	Measure	Minimum	Maximum	Mean	Std. Dev.
HOSP_BEDSIZE	Continuous	1	3	2.49	0.72
HOSP_CONTROL	Nominal	0	4	2.31	0.81
HOSP_LOCATION	Nominal	0	1	1.00	0.00
HOSP_LOCTEACH	Nominal	1	3	2.33	0.68
HOSP_REGION	Nominal	1	4	2.51	1.02
HOSP_TEACH	Nominal	0	1	0.40	0.49
HOSP_RNPCT	Continuous	0%	100%	78.24%	34.52%
HOSP_RNFTEAPD	Continuous	0	10.90	3.41	1.92
HOSP_LPNFTEAPD	Continuous	0	3.90	0.24	0.28
HOSP_NAFTEAPD	Continuous	0	3.90	0.86	0.60
HOSP_OPSURGPCT	Continuous	0%	100%	52.44%	23.58%

the auto classifier from SPSS. The auto classifier estimates and compares models for either nominal or binary targets using a number of different methods, allowing the modeler to try out a variety of approaches in a single modeling run ("IBM Knowledge Center," 2018). The auto classifier built a total of 7 models. The accuracies among the algorithms were compared and contrasted as shown in Table 7. In addition to accuracies, the ROC curve was used to revalidate the best model for prediction.

In the second phase, we added the healthcare provider's attributes to the data used to build the models and set the type of each attribute accordingly. We partitioned those data the same way we did in phase 1. We used the same algorithms from phase 1 in addition to the SPSS auto classifier. The accuracies between the algorithms were compared and contrasted, as shown in Table 8. In addition to accuracies, the ROC curve was used to revalidate the best model for prediction. Detailed analysis of each model is conducted in the next section.

We applied Principal component analysis (PCA) for dimensionality reduction. The PCA/Factor node produced 5 factors that were used as inputs to the models that were identified in the previous phases. The results were consistent with phase 1 and 2, but the accuracies of the models were slightly lower. This was expected given that the PCA identifies factors based on the variance between the original attributes and perform linear transformation combining the original variables into new ones (Clarke, Fokoue, & Zhang, 2009; Fodor, 2002). We observed that the presence of different types of predictor variables (binary, nominal and continuous) would not make PCA a better solution because it is usually performed on continuous variables. In addition, the attributes in our data have low correlation between them which indicates low or even no redundancy (see Table 11, Appendix C).

Table 7. Models performance with patient attributes

	Accuracy			ROC		
	Training	**Testing**	**Validation**	**Training**	**Testing**	**Validation**
CHAID	82.22%	82.10%	82.08%	0.70	0.69	0.69
Quest	81.82%	81.73%	81.71%	0.50	0.50	0.50
SVM	81.83%	81.73%	81.73%	0.68	0.68	0.68
Neural Net	51.79%	51.56%	51.71%	0.51	0.50	0.50
Random Tree	69.37%	69.28%	69.14%	0.70	0.69	0.70
Bayes Net	50%	50%	50%	0.50	0.50	0.50
Logistic Regression	50%	50%	50%	0.50	0.50	0.50

Table 8. Models performance with all attributes

	Accuracy			ROC		
	Training	**Testing**	**Validation**	**Training**	**Testing**	**Validation**
CHAID	85.5%	85.45%	85.33%	0.85	0.84	0.84
Quest	83.73%	83.64%	83.59%	0.73	0.73	0.73
SVM	80.04%	79.625%	79.21%	0.51	0.50	0.50
Neural Net	78.84%	75.94%	75.94%	0.73	0.73	0.73
Random Tree	70.72%	70.68%	70.38%	0.81	0.81	0.81
Bayes Net	51%	50%	50%	0.51	0.50	0.50
Logistic Regression	51%	50%	50%	0.51	0.50	0.50

Lastly, the factors produced by the PCA/Factor node do not have a real meaning that can be used in the context of our research.

RESULTS

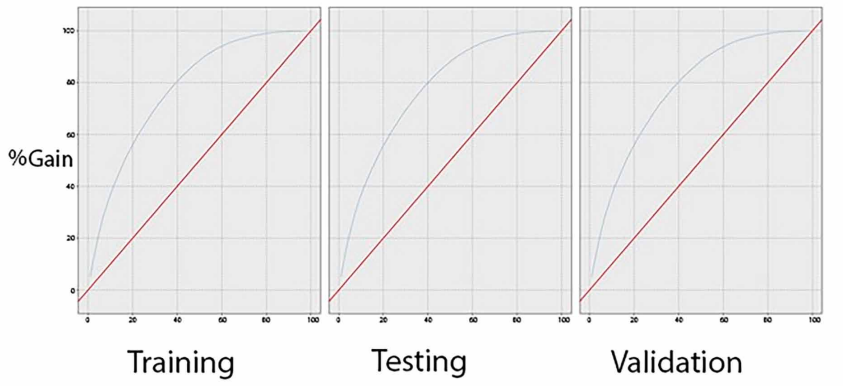
In general, accuracy is calculated based on the class distribution of a test dataset. Such a ratio may change when applied to real-life data. ROC describes the discriminative power of a classifier independent of class distribution. The true positive (TP) and false positive (FP) rates, which are used to construct the AUC, will not be affected by class-distribution shifting. Thus, the AUC is a more desirable metric for comparing different models for accuracy.

In terms of performance, the LR model had the lowest accuracy in the three different data partitions: training, testing, and validation (Table 8). BN was the second lowest performer, followed by RT. NN performed slightly better than these models, but its performance was still weak in terms of accuracy. Although both the NN and the SVM models have performed better in accuracy, their performance was worse than the RT in terms of the area under the curve (AUC) of their respective ROC curves. The Quest model was the second-best performer in terms of accuracy but was worse than RT in terms of the AUC. The auto classifier identified the CHAID model as the best performer, with a consistent accuracy of 85.3% and AUC of 0.84 across the different data partitions (see Table 8). The model's performance was superior compared to all the other models we tested. Figure 1 shows the ROC curves for the CHAID model across the different used datasets. The ROC is a plot of the TP rate against the FP rate for different possible predictors. The results show consistency among the different data partitions.

DISCUSSION AND CONCLUSION

Sepsis is considered a medical challenge, especially among elderly patients. Because of the lack of a standardized diagnostic test, the medical decision of the HCP plays a crucial role in the diagnosis of sepsis (Joynes, 2016). Researchers are placing more emphasis on finding a faster diagnosis method for Sepsis condition. Clinical predictive models can help HCPs intervene in a timely manner and make informed decisions which may lower the risk of in-hospital mortality (Koh & Tan, 2011; D. S. Lee et al., 2003; Martin et al., 2009; Mehta et al., 2002). Healthcare-related use of decision support systems has been documented in various publications. For example, Saoud, Boubetra, and Attia (2016) propose a decision support system based on simulation and knowledge extraction to improve

Figure 1. ROC curves for the CHAID model



hospital emergency room operation. The healthcare literature is full of examples of how mortality can occur when patients do not receive appropriate interventions and follow-ups (Schmitt et al., 2013).

In this study, we used various datamining techniques to develop a predictive model for in-hospital mortality among septic patients using big data. At the beginning, we only used patient-related attributes as inputs to the SPSS auto classifier. The auto classifier identified CHAID as the best model to predict in-hospital mortality, with an accuracy rate of approximately 82% but when adding healthcare provider-related attributes the accuracy of CHAID prediction went up to approximately 85%. Further, using the ROC curve as a performance metric for evaluation of the model the AUC went from approximately 0.70 to 0.84 across different data partitions. This provides an evidence that provider-related attributes help in predicting mortality among sepsis patients in our dataset. For example, hospital bedside was among the top 10 attributes used by CHAID to predict mortality and hospital location (urban vs. rural) was among the top 20.

The CHAID was developed by Kass (1980) based on automatic interaction detection and theta automatic interaction detection. CHAID builds decision trees by using chi-square statistics to identify optimal splits. CHAID can be applied to detect an interaction between variables and classification as well. CHAID has been applied in areas like medical (Chan, Cheing, Chan, Rosenthal, & Chronister, 2006; Ture, Tokatli, & Kurt, 2009), psychiatric (Kobayashi, Takahashi, Arioka, Koga, & Fukui, 2013), marketing (Legohérel, Hsu, & Daucé, 2015), and other applications (Ramaswami & Bhaskaran, 2010). CHAID is nonparametric, which gives it some advantage over models such as multiple regression. In addition, unlike the C&R tree and QUEST models, CHAID uses multi-way splits, meaning that some of its splits have more than two branches. It, therefore, tends to create a wider tree than binary growing methods. CHAID works for all types of inputs (nominal, ordinal, and continuous data), and it accepts both case weights and frequency variables (Kass, 1980). The previous make CHAID more efficient and explainable and thus have higher utility for deployment comparing to other datamining techniques. To our knowledge this is the first study that applies CHAID using both patient-related and provider-related attributes to predict in-hospital mortality among sepsis patients. Our study can benefit medical practitioners to better identify risky patients and apply timely interventions to lower the risk of in-hospital mortality. In addition, the study can help decision makers and regulators to better identify provider-related attributes that may be associated with higher risk of in-hospital mortality.

This study has some limitations. First, this study was conducted considering one medical condition; future studies may replicate for other diseases and conditions. Additionally, this study used data for patients that were hospitalized between the years 2008 and 2012; future studies can use a wider range of data by considering a longer time frame. The data we used for our model contained ICD9 codes; future studies can use ICD10 codes, which provide a more detailed diagnosis. Lastly future studies can investigate how to deploy the CHAID model in EHRs and incorporate the outcomes in the HCPs workflow to identify high-risk patients.

ACKNOWLEDGMENT

The authors want to express their gratitude to Sepan S. Desai, MD, Ph.D., MBA, for his cooperation and for providing access to the dataset used in this study.

REFERENCES

Amarasingham, R., Velasco, F., Xie, B., Clark, C., Ma, Y., Zhang, S., & Halm, E. A. et al. (2015). Electronic medical record-based multicondition models to predict the risk of 30 day readmission or death among adult medicine patients: Validation and comparison to existing models. *BMC Medical Informatics and Decision Making*, *15*(1), 39. doi:10.1186/s12911-015-0162-6 PMID:25991003

Amland, R. C., & Hahn-Cover, K. E. (2016). Clinical decision support for early recognition of sepsis. *American Journal of Medical Quality*, *31*(2), 103–110. doi:10.1177/1062860614557636 PMID:25385815

Austin, J. E. (2010). *The collaboration challenge: How nonprofits and businesses succeed through strategic alliances* (Vol. 109). John Wiley & Sons.

Bandyopadhyay, S., Wolfson, J., Vock, D. M., Vazquez-Benitez, G., Adomavicius, G., Elidrisi, M., & O'Connor, P. J. et al. (2015). Data mining for censored time-to-event data: A Bayesian network model for predicting cardiovascular risk from electronic health record data. *Data Mining and Knowledge Discovery*, *29*(4), 1033–1069. doi:10.1007/s10618-014-0386-6

Bardhan, I., Oh, J., Zheng, Z., & Kirksey, K. (2014). Predictive analytics for readmission of patients with congestive heart failure. *Information Systems Research*, *26*(1), 19–39. doi:10.1287/isre.2014.0553

Bardhan, I. R., & Thouin, M. F. (2013). Health information technology and its impact on the quality and cost of healthcare delivery. *Decision Support Systems*, *55*(2), 438–449. doi:10.1016/j.dss.2012.10.003

Bhattacharjee, P., Edelson, D. P., & Churpek, M. M. (2017). Identifying patients with sepsis on the hospital wards. *Chest*, *151*(4), 898–907. doi:10.1016/j.chest.2016.06.020 PMID:27374948

Celi, L. A., Galvin, S., Davidzon, G., Lee, J., Scott, D., & Mark, R. (2012). A database-driven decision support system: Customized mortality prediction. *Journal of Personalized Medicine*, *2*(4), 138–148. doi:10.3390/jpm2040138 PMID:23766893

Chan, F., Cheing, G., Chan, J. Y. C., Rosenthal, D. A., & Chronister, J. (2006). Predicting employment outcomes of rehabilitation clients with orthopedic disabilities: A CHAID analysis. *Disability and Rehabilitation*, *28*(5), 257–270. doi:10.1080/09638280500158307 PMID:16492620

Chen, H., Chiang, R. H., & Storey, V. C. (2012). Business intelligence and analytics: From big data to big impact. *Management Information Systems Quarterly*, *36*(4), 1165. doi:10.2307/41703503

Clarke, B., Fokoue, E., & Zhang, H. H. (2009). *Principles and theory for data mining and machine learning*. Springer Science & Business Media. doi:10.1007/978-0-387-98135-2

Delen, D., Walker, G., & Kadam, A. (2005). Predicting breast cancer survivability: A comparison of three data mining methods. *Artificial Intelligence in Medicine*, *34*(2), 113–127. doi:10.1016/j.artmed.2004.07.002 PMID:15894176

Epstein, L. (2016). Varying estimates of sepsis mortality using death certificates and administrative codes—United States, 1999–2014. *MMWR. Morbidity and Mortality Weekly Report*, 65. PMID:27054476

Fann, W. C., Chiang, I. J., Hsiao, C. T., Hong, Y. C., & Chen, I. (2012). Predicting the mortality of necrotizing fasciitis with blood pressure and white blood cell count. *Surgical Practice*, *16*(3), 103–108. doi:10.1111/j.1744-1633.2012.00598.x

Fleischmann, C., Scherag, A., Adhikari, N. K., Hartog, C. S., Tsaganos, T., Schlattmann, P., & Reinhart, K. et al. (2016). Assessment of global incidence and mortality of hospital-treated sepsis. Current estimates and limitations. *American Journal of Respiratory and Critical Care Medicine*, *193*(3), 259–272. doi:10.1164/rccm.201504-0781OC PMID:26414292

Fodor, I. K. (2002). A survey of dimension reduction techniques. Center for Applied Scientific Computing, Lawrence Livermore National Laboratory.

Ford, D. W., Goodwin, A. J., Simpson, A. N., Johnson, E., Nadig, N., & Simpson, K. N. (2016). A severe sepsis mortality prediction model and score for use with administrative data. *Critical Care Medicine*, *44*(2), 319–327. doi:10.1097/CCM.0000000000001392 PMID:26496452

Futoma, J., Morris, J., & Lucas, J. (2015). A comparison of models for predicting early hospital readmissions. *Journal of Biomedical Informatics*, *56*, 229–238. doi:10.1016/j.jbi.2015.05.016 PMID:26044081

Ghosh, S., Li, J., Cao, L., & Ramamohanarao, K. (2017). Septic shock prediction for ICU patients via coupled HMM walking on sequential contrast patterns. *Journal of Biomedical Informatics*, *66*, 19–31. doi:10.1016/j.jbi.2016.12.010 PMID:28011233

Hagland, M. (2011). Mastering readmissions: Laying the foundation for change. *Healthcare Informatics*, *28*(4), 10–16. PMID:21560716

Hartwig, F., & Dearing, B. E. (1979). *Exploratory data analysis* (Vol. 16). Sage. doi:10.4135/9781412984232

Howell, M. D., Donnino, M., Clardy, P., Talmor, D., & Shapiro, N. I. (2007). Occult hypoperfusion and mortality in patients with suspected infection. *Intensive Care Medicine*, *33*(11), 1892–1899. doi:10.1007/s00134-007-0680-5 PMID:17618418

IBM Knowledge Center. (2018). Retrieved from https://www.ibm.com/support/knowledgecenter/en/SS3RA7_15.0.0/com.ibm.spss.modeler.help/binary_classifier_node.htm

Joynes, E. (2016). More challenges around sepsis: Definitions and diagnosis. *Journal of Thoracic Disease*, *8*(11), E1467–E1469. doi:10.21037/jtd.2016.11.10 PMID:28066632

Kass, G. V. (1980). An exploratory technique for investigating large quantities of categorical data. *Applied Statistics*, *29*(2), 119–127. doi:10.2307/2986296

Kobayashi, D., Takahashi, O., Arioka, H., Koga, S., & Fukui, T. (2013). A prediction rule for the development of delirium among patients in medical wards: Chi-Square Automatic Interaction Detector (CHAID) decision tree analysis model. *The American Journal of Geriatric Psychiatry*, *21*(10), 957–962. doi:10.1016/j.jagp.2012.08.009 PMID:23567433

Kociol, R. D., Lopes, R. D., Clare, R., Thomas, L., Mehta, R. H., Kaul, P., & Armstrong, P. W. et al. (2012). International variation in and factors associated with hospital readmission after myocardial infarction. *Journal of the American Medical Association*, *307*(1), 66–74. doi:10.1001/jama.2011.1926 PMID:22215167

Koh, H. C., & Tan, G. (2011). Data mining applications in healthcare. *Journal of Healthcare Information Management*, *19*(2), 65. PMID:15869215

Krumholz, H. M., Chen, Y.-T., Wang, Y., Vaccarino, V., Radford, M. J., & Horwitz, R. I. (2000). Predictors of readmission among elderly survivors of admission with heart failure. *American Heart Journal*, *139*(1), 72–77. doi:10.1016/S0002-8703(00)90311-9 PMID:10618565

Le Duff, F., Muntean, C., Cuggia, M., & Mabo, P. (2004). Predicting survival causes after out of hospital cardiac arrest using data mining method. *Paper presented at the Medinfo*.

Lee, C.-C., Chen, S.-Y., Tsai, C.-L., Wu, S.-C., Chiang, W.-C., Wang, J.-L., & Hsueh, P.-R. et al. (2008). Prognostic value of mortality in emergency department sepsis score, procalcitonin, and C-reactive protein in patients with sepsis at the emergency department. *Shock (Augusta, Ga.)*, *29*(3), 322–327. PMID:17724429

Lee, C. H., & Yoon, H.-J. (2017). Medical big data: Promise and challenges. *Kidney Research and Clinical Practice*, *36*(1), 3–11. doi:10.23876/j.krcp.2017.36.1.3 PMID:28392994

Lee, D. S., Austin, P. C., Rouleau, J. L., Liu, P. P., Naimark, D., & Tu, J. V. (2003). Predicting mortality among patients hospitalized for heart failure: Derivation and validation of a clinical model. *Journal of the American Medical Association*, *290*(19), 2581–2587. doi:10.1001/jama.290.19.2581 PMID:14625335

Legohérel, P., Hsu, C. H., & Daucé, B. (2015). Variety-seeking: Using the CHAID segmentation approach in analyzing the international traveler market. *Tourism Management*, *46*, 359–366. doi:10.1016/j.tourman.2014.07.011

Lin, Y.-K., Chen, H., Brown, R. A., Li, S.-H., & Yang, H.-J. (2017). Healthcare predictive analytics for risk profiling in chronic care: A Bayesian multitask learning approach. *Management Information Systems Quarterly*, *41*(2), 473–495. doi:10.25300/MISQ/2017/41.2.07

Martin, L. T., Ruder, T., Escarce, J. J., Ghosh-Dastidar, B., Sherman, D., Elliott, M., & Culbert, A. et al. (2009). Developing predictive models of health literacy. *Journal of General Internal Medicine*, *24*(11), 1211–1216. doi:10.1007/s11606-009-1105-7 PMID:19760299

Martinez, W. L., Martinez, A. R., Martinez, A., & Solka, J. (2010). *Exploratory data analysis with MATLAB*. CRC Press. doi:10.1201/b10434

Mehta, R. L., Pascual, M. T., Gruta, C. G., Zhuang, S., Chertow, G. M., & Group, P. S. (2002). Refining predictive models in critically ill patients with acute renal failure. *Journal of the American Society of Nephrology*, *13*(5), 1350–1357. doi:10.1097/01.ASN.0000014692.19351.52 PMID:11961023

Mitchell, S. E., Martin, J., Holmes, S., van Deusen Lukas, C., Cancino, R., Paasche-Orlow, M., & Jack, B. et al. (2016). How hospitals reengineer their discharge processes to reduce readmissions. *Journal for Healthcare Quality: Official Publication of the National Association for Healthcare Quality*, *38*(2), 116–126. doi:10.1097/JHQ.0000000000000005 PMID:26042743

Nachimuthu, S. K., & Haug, P. J. (2012). Early detection of sepsis in the emergency department using Dynamic Bayesian Networks. *Paper presented at the AMIA Annual Symposium Proceedings*.

National Health Expenditure Data. (2018). Retrieved from https://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/NationalHealthExpendData/NationalHealthAccountsHistorical.html#

Nguyen, H. B., Corbett, S. W., Steele, R., Banta, J., Clark, R. T., Hayes, S. R., & Wittlake, W. A. et al. (2007). Implementation of a bundle of quality indicators for the early management of severe sepsis and septic shock is associated with decreased mortality. *Critical Care Medicine*, *35*(4), 1105–1112. doi:10.1097/01.CCM.0000259463.33848.3D PMID:17334251

Paoin, W. (2011). Lessons learned from data mining of WHO mortality database. *Methods of Information in Medicine*, *50*(4), 380–385. doi:10.3414/ME10-02-0019 PMID:21691674

Raghavan, M., & Marik, P. E. (2006). Management of sepsis during the early "golden hours". *The Journal of Emergency Medicine*, *31*(2), 185–199. doi:10.1016/j.jemermed.2006.05.008 PMID:17044583

Ramaswami, M., & Bhaskaran, R. (2010). A CHAID based performance prediction model in educational data mining. arXiv:1002.1144

Rezaei Hachesu, P., Moftian, N., Dehghani, M., & Samad-Soltani, T. (2017). Analyzing a Lung Cancer Patient Dataset with the Focus on Predicting Survival Rate One Year after Thoracic Surgery. *Asian Pacific Journal of Cancer Prevention*, *18*(6), 1531–1536. PMID:28669163

Roshanov, P. S., Fernandes, N., Wilczynski, J. M., Hemens, B. J., You, J. J., Handler, S. M., & Van Spall, H. G. et al. (2013). Features of effective computerised clinical decision support systems: Meta-regression of 162 randomised trials. *BMJ (Clinical Research Ed.)*, *346*, f657. PMID:23412440

Saoud, M. S., Boubetra, A., & Attia, S. (2016). A simulation knowledge extraction-based decision support system for the healthcare emergency department. *International Journal of Healthcare Information Systems and Informatics*, *11*(2), 19–37. doi:10.4018/IJHISI.2016040102

Schmitt, S., McQuillen, D. P., Nahass, R., Martinelli, L., Rubin, M., Schwebke, K., & Slama, T. et al. (2013). Infectious diseases specialty intervention is associated with decreased mortality and lower healthcare costs. *Clinical Infectious Diseases*, *58*(1), 22–28. doi:10.1093/cid/cit610 PMID:24072931

Sepsis and Healthcare-Acquired Infections. (2017). Retrieved from https://www.sepsis.org/sepsis-and/healthcare-acquired-infections/

SheetSepsis Fact. (2016).

Shorr, A. F., Micek, S. T., Jackson, W. L. Jr, & Kollef, M. H. (2007). Economic implications of an evidence-based sepsis protocol: Can we improve outcomes and lower costs? *Critical Care Medicine*, *35*(5), 1257–1262. doi:10.1097/01.CCM.0000261886.65063.CC PMID:17414080

Singer, M., Deutschman, C. S., Seymour, C. W., Shankar-Hari, M., Annane, D., Bauer, M., & Coopersmith, C. M. et al. (2016). The third international consensus definitions for sepsis and septic shock (sepsis-3). *Journal of the American Medical Association*, *315*(8), 801–810. doi:10.1001/jama.2016.0287 PMID:26903338

Sommerfeld, A. J., Althouse, A. D., Prince, J., & Hickey, G. W. (2016). Obstructive Sleep Apnea is Associated with Increased Readmissions in CHF Patients. *Journal of Cardiac Failure*, *22*(8), S88. doi:10.1016/j.cardfail.2016.06.282

Srinivas, K., Rani, B. K., & Govrdhan, A. (2010). Applications of data mining techniques in healthcare and prediction of heart attacks. *International Journal on Computer Science and Engineering*, *2*(02), 250–255.

Sut, N., & Simsek, O. (2011). Comparison of regression tree data mining methods for prediction of mortality in head injury. *Expert Systems with Applications*, *38*(12), 15534–15539. doi:10.1016/j.eswa.2011.06.006

Taneja, I., Reddy, B., Damhorst, G., Zhao, S. D., Hassan, U., Price, Z., & Wachspress, S. et al. (2017). Combining Biomarkers with EMR Data to Identify Patients in Different Phases of Sepsis. *Scientific Reports*, *7*(1), 10800. doi:10.1038/s41598-017-09766-1 PMID:28883645

Tsugawa, Y., Jena, A. B., Figueroa, J. F., Orav, E. J., Blumenthal, D. M., & Jha, A. K. (2017). Comparison of hospital mortality and readmission rates for Medicare patients treated by male vs female physicians. *JAMA Internal Medicine*, *177*(2), 206–213. doi:10.1001/jamainternmed.2016.7875 PMID:27992617

Ture, M., Tokatli, F., & Kurt, I. (2009). Using Kaplan–Meier analysis together with decision tree methods (C&RT, CHAID, QUEST, C4. 5 and ID3) in determining recurrence-free survival of breast cancer patients. *Expert Systems with Applications*, *36*(2), 2017–2026. doi:10.1016/j.eswa.2007.12.002

van Walraven, C., Dhalla, I. A., Bell, C., Etchells, E., Stiell, I. G., Zarnke, K., & Forster, A. J. et al. (2010). Derivation and validation of an index to predict early death or unplanned readmission after discharge from hospital to the community. *Canadian Medical Association Journal*, *182*(6), 551–557. doi:10.1503/cmaj.091117 PMID:20194559

Varlamis, I., Apostolakis, I., Sifaki-Pistolla, D., Dey, N., Georgoulias, V., & Lionis, C. (2017). Application of data mining techniques and data analysis methods to measure cancer morbidity and mortality data in a regional cancer registry: The case of the island of Crete, Greece. *Computer Methods and Programs in Biomedicine*, *145*, 73–83. doi:10.1016/j.cmpb.2017.04.011 PMID:28552128

Veith, N., & Steele, R. (2018). Machine Learning-based Prediction of ICU Patient Mortality at Time of Admission.

Velleman, P. F., & Hoaglin, D. C. (1981). *Applications, basics, and computing of exploratory data analysis*. Duxbury Press.

Vincent, J.-L., Moreno, R., Takala, J., Willatts, S., De Mendonça, A., Bruining, H., & Thijs, L. (1996). *The SOFA (Sepsis-related Organ Failure Assessment) score to describe organ dysfunction/failure*. Springer.

Walczak, S., & Okuboyejo, S. R. (2017). An Artificial Neural Network Classification of Prescription Nonadherence. *International Journal of Healthcare Information Systems and Informatics*, *12*(1), 1–13. doi:10.4018/IJHISI.2017010101

Wang, H., Robinson, R. D., Johnson, C., Zenarosa, N. R., Jayswal, R. D., Keithley, J., & Delaney, K. A. (2014). Using the LACE index to predict hospital readmissions in congestive heart failure patients. *BMC Cardiovascular Disorders*, *14*(1), 97. doi:10.1186/1471-2261-14-97 PMID:25099997

Wang, S.-L., Wu, F., & Wang, B.-H. (2010). Prediction of severe sepsis using SVM model. In *Advances in Computational Biology* (pp. 75–81). Springer. doi:10.1007/978-1-4419-5913-3_9

Widgren, B. R., & Jourak, M. (2011). Medical Emergency Triage and Treatment System (METTS): A new protocol in primary triage and secondary priority decision in emergency medicine. *The Journal of Emergency Medicine*, *40*(6), 623–628. doi:10.1016/j.jemermed.2008.04.003 PMID:18930373

Yu, S., Farooq, F., Van Esbroeck, A., Fung, G., Anand, V., & Krishnapuram, B. (2015). Predicting readmission risk with institution-specific prediction models. *Artificial Intelligence in Medicine*, *65*(2), 89–96. doi:10.1016/j.artmed.2015.08.005 PMID:26363683

APPENDIX A: EXAMPLE OF STUDIES EMPLOYING DATA MINING FOR MORTALITY/SURVIVAL PREDICTION

Table 9. Example of studies employing data mining for mortality/survival prediction

Study	Data	Target	Findings/Results
(Varlamis et al., 2017)	3,763 patients from a cancer registry	Cancer survival	C4.5 classification algorithm manages to predict survival with an accuracy rate of 80%.
(Rezaei Hachesu et al., 2017)	470 records from the Polish National Cancer Registry	Lung cancer (LC) survival	Development of a framework based on the key influencers of LC survival for one year after thoracic surgery.
(Ghosh et al., 2017)	Experimental dataset	Septic shock	Single and multi-channel patterns coupled with a hidden Markov model (HMM)demonstrate a strong competitive accuracy in the prediction of septic shock.
(Fann, Chiang, Hsiao, Hong, & Chen, 2012)	272 patients from university-affiliated community hospital	Necrotizing fasciitis in-hospital mortality	C4.5 decision tree accuracy with cross-validation was 84.2%.
(Celi et al., 2012)	ICU database and subset of patients ≥ 80 in a cardiac registry	Mortality	LR, BN, and ANN models for specific patient subsets perform better than general models like EuroSCORE.
(Paoin, 2011)	1,109,537 records from WHO Mortality Database	Mortality	Classification tools produced poor results in predicting the cause of death in the WHO mortality database.
(Austin, 2010)	7,889 cases from 86 hospital corporations and 11,506 cases from 81 other hospitals	Acute myocardial infarction and CHF mortality	Bagged RT, random forests, and boosted RT may result in a superior prediction of 30-day mortality.
(Delen et al., 2005)	433,272 cases from cancer database	Breast cancer survivability	Decision tree (C5) is the best predictor with 93.6% accuracy.
(Le Duff, Muntean, Cuggia, & Mabo, 2004)	533 patients from emergency medical services	Survival after heart failure	The probability of remaining alive is directly associated with the resuscitation techniques employed.

LR: Logistic Regression, BN: Bayesian Network, ANN: Artificial Neural Network, RT: Random Tree, LC: Lung Cancer

APPENDIX B: DESCRIPTIVE STATISTICS OF COMORBIDITIES

Table 10. Descriptive statistics of comorbidities

Variable	Description	Measure	Min.	Max.	Mean	S.D.
CM_AIDS	Acquired immune deficiency syndrome	Binary	0	1	0.00	0.04
CM_ALCOHOL	Alcohol abuse	Binary	0	1	0.04	0.19
CM_ANEMDEF	Deficiency anemia	Binary	0	1	0.30	0.46
CM_ARTH	Rheumatoid arthritis/collagen vascular diseases	Binary	0	1	0.03	0.18
CM_BLDLOSS	Chronic blood loss anemia	Binary	0	1	0.01	0.12
CM_CHF	Congestive heart failure	Binary	0	1	0.22	0.41
CM_CHRNLUNG	Chronic pulmonary disease	Binary	0	1	0.24	0.43
CM_COAG	Coagulopathy	Binary	0	1	0.13	0.34
CM_DEPRESS	Depression	Binary	0	1	0.10	0.30
CM_DM	Uncomplicated diabetes mellitus	Binary	0	1	0.23	0.42
CM_DMCX	Diabetes with chronic complications	Binary	0	1	0.08	0.27
CM_DRUG	Drug abuse	Binary	0	1	0.03	0.16
CM_HTN_C	Hypertension both uncomplicated and complicated	Binary	0	1	0.52	0.50
CM_HYPOTHY	Hypothyroidism	Binary	0	1	0.12	0.32
CM_LIVER	Liver disease	Binary	0	1	0.05	0.21
CM_LYMPH	Lymphoma	Binary	0	1	0.02	0.13
CM_LYTES	Fluid and electrolyte disorders	Binary	0	1	0.55	0.50
CM_METS	Metastatic cancer	Binary	0	1	0.05	0.21
CM_NEURO	Other neurological disorders	Binary	0	1	0.15	0.36
CM_OBESE	Obesity	Binary	0	1	0.10	0.30
CM_PARA	Paralysis	Binary	0	1	0.07	0.25
CM_PERIVASC	Peripheral vascular disorder	Binary	0	1	0.09	0.28
CM_PSYCH	Psychoses	Binary	0	1	0.05	0.22
CM_PULMCIRC	Pulmonary circulation disorders	Binary	0	1	0.05	0.22
CM_RENLFAIL	Renal failure	Binary	0	1	0.24	0.43
CM_TUMOR	Solid tumor without metastasis	Binary	0	1	0.04	0.19
CM_ULCER	Peptic ulcer disease excluding bleeding	Binary	0	1	0.00	0.02
CM_VALVE	Valvular disease	Binary	0	1	0.06	0.24
CM_WGHTLOSS	Weight loss	Binary	0	1	0.18	0.38

APPENDIX C: CORRELATION TABLE BETWEEN KEY VARIABLES

Table 11. Correlation table between key variables

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	27	28	29
DIED	1																											
AGE	.111**	1																										
APRDRGSeverity	.206**	-.073**	1																									
ATYPE	-.034**	-.074**	.074**	1																								
AMONTH	0.007	-0.004	0.011	-.048**	1																							
AWEEKEND	0.006	-0.001	-0.022	-.136**	0.016	1																						
CM_ANEMDEF	-.082**	-0.014	.049**	.034**	0.000	0.002	1																					
CM_CHF	.041**	.176**	.146**	.041**	-0.008	-0.009	.079**	1																				
CM_CHRNLUNG	0.013	.028*	.067**	-0.004	-.022*	0.012	.048**	.158**	1																			
CM_DEPRESS	-.049**	-.038**	-.049**	-.052**	.029**	0.003	.058**	-0.008	.053**	1																		
CM_DM	-.038**	0.010	0.000	0.011	0.006	-0.020	.098**	.069**	.040**	0.005	1																	
CM_DRUG	-.026*	-.181**	.023*	-0.001	-0.009	-0.001	0.014	-.033**	-0.012	.041**	-0.013	1																
CM_HTN_C	-.054**	.113**	-.035**	-0.004	0.017	0.007	.139**	.095**	.027*	.044**	.159**	-.030**	1															
CM_HYPOTHY	-0.013	.109**	-0.013	-.038**	0.021	0.008	.070**	.046**	0.013	.067**	.034**	-0.019	.091**	1														

continued on following page

Table 11. Continued

	YEAR	HOSPID	RACE	NPR	NCHRONIC	MDNUM1_R	LOS	HOSP_REGION	HOSP_LOCATION	HOSP_CONTROL	HOSP_BEDSIZE	FEMALE	NECODE	DQTR	CM_LYTES
1	-0.002	-0.021	-0.004	.121**	-0.012	0.013	-.130**	-.031**	0.009	-.031**	0.001	-0.009	-.034**	0.001	.073**
2	-.083**	.111**	-.038**	-.193**	.033**	-.054**	-.100**	-.072**	-.068**	0.008	-0.011	.080**	-.074**	-0.004	-0.013
3	.064**	-.028*	.064**	.354**	.229**	.043**	.219**	0.011	.129**	.060**	-.067**	-.091**	.074**	0.001	.152**
4	-.055**	.122**	.200**	0.022	-0.011	-.136**	.379**	-.073**	.111**	.288**	-.335**	0.000	-.035**	-.083**	-.184**
5	-.045**	-.048**	-.058**	-0.019	.031**	.025*	-0.010	0.004	-0.022	-.040**	.047**	-0.004	0.003	.762**	.027*
6	.023*	-.030**	-.046**	0.017	-0.009	.024*	-.105**	.029**	-.043**	-.065**	.123**	0.003	0.001	0.017	.047**
7	.022*	-.067**	.068**	.085**	.283**	.028*	.059**	-.028*	.063**	.067**	-0.005	0.021	-.024*	0.016	.054**
8	-0.005	0.003	-0.012	.043**	.326**	-0.004	.031**	-.070**	0.016	.058**	.025*	.035**	-.035**	-.022*	0.010
9	.035**	-.051**	-.075**	.035**	.263**	.032**	-0.020	-0.021	-0.006	0.006	0.009	-.035**	-.037**	-0.015	0.016
10	.056**	-.056**	-.067**	-.040**	.180**	.055**	-.045**	-.036**	-0.018	0.017	.040**	.067**	-0.012	.028*	.025*
11	.031**	-0.020	.050**	0.009	.247**	0.014	-0.010	-0.011	0.011	0.005	-0.014	.029**	-0.021	0.005	-0.012
12	0.008	-.046**	0.004	.035**	.036**	.040**	0.006	.028*	0.016	-.024*	0.019	-0.010	0.022	-0.006	.022*
13	.029**	-.029**	-0.003	.047**	.408**	.027*	-.022*	-.050**	.030**	0.015	.057**	.023*	-.022*	0.017	0.012
14	0.011	0.008	-.057**	-0.002	.212**	0.010	-.038**	-0.008	0.004	-0.003	.053**	.142**	-0.011	.026*	.026*
15	.074**	-.103**	-.051**	.129**	.083**	.101**	-.058**	-0.006	0.006	-.080**	.098**	0.012	-0.004	.031**	1
16	-.092**	-.076**	-.091**	-.026*	.033**	.031**	-.067**	0.012	-.044**	-.058**	.086**	-0.001	0.013	1	
17	0.015	-.073**	0.008	.120**	-0.006	-.113**	.070**	.071**	-.076**	0.010	-0.009	-.039**	1		
18	-.037**	.038**	-0.003	-.062**	-.029**	0.000	-.024*	-.036**	-0.015	-0.008	.033**	1			
19	-.042**	-0.008	-.278**	.063**	.043**	0.003	-.274**	-0.006	-.160**	-.214**	1				
20	-.059**	0.016	.166**	.025*	.095**	-.206**	.175**	-.356**	-.050**	1					
21	-.089**	.104**	.122**	.132**	.091**	.112**	.151**	.431**	1						
22	-.043**	-.090**	-.042**	.038**	-.059**	-0.011	-.030**	1							
23	-.043**	.113**	.181**	.350**	.079**	-.037**	1								
24	.341**	-.027*	-.029*	0.013	.048**	1									
25	.111**	-.146**	-.029*	.212**	1										
27	.040**	-.057**	.052**	1											
28	-.042**	.145**	1												
29	-.168**	1													