Heterogeneous Postsurgical Data Analytics for Predictive Modeling of Mortality Risks in Intensive Care Units

Yun Chen and Hui Yang*, *Member, IEEE*

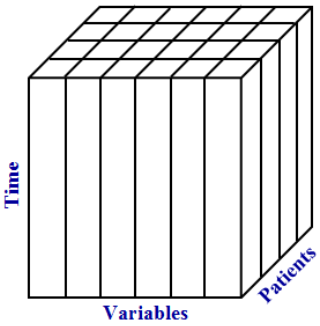
***Abstract*—The rapid advancements of biomedical instrumentation and healthcare technology have resulted in data-rich environments in hospitals. However, the meaningful information extracted from rich datasets is limited. There is a dire need to go beyond current medical practices, and develop data-driven methods and tools that will enable and help (i) the handling of big data, (ii) the extraction of data-driven knowledge, (iii) the exploitation of acquired knowledge for optimizing clinical decisions. This present study focuses on the prediction of mortality rates in Intensive Care Units (ICU) using patient-specific healthcare recordings. It is worth mentioning that postsurgical monitoring in ICU leads to massive datasets with unique properties, e.g., variable heterogeneity, patient heterogeneity, and time asyncronization. To cope with the challenges in ICU datasets, we developed the postsurgical decision support system with a series of analytical tools, including data categorization, data pre-processing, feature extraction, feature selection, and predictive modeling. Experimental results show that the proposed data-driven methodology outperforms traditional approaches and yields better results based on the evaluation of real-world ICU data from 4000 subjects in the database. This research shows great potentials for the use of data-driven analytics to improve the quality of healthcare services.**

I. Introduction

US healthcare spending is approximately 17% of GDP (i.e., $2.5 trillion), and will continue the historical upward trend, reaching 19.5% by 2017 [1]. The rapid advancements of biomedical instrumentation and healthcare technology have led to data-rich environments in hospitals. Nevertheless, the meaningful information extracted from rich datasets is limited. Laboratory tests and patient monitoring are two of primary information sources for monitoring critical conditions of postsurgical patients in the Intensive Care Units (ICU). Commonly, physicians make inferences about patient conditions based on most recent test results, ignoring important factors such as historical test results and the relationships among different types of tests. In the general practice of medicine, physicians lack decision-support tools that can help them delineate hidden interactions among different lab tests, identify temporal variations of patient conditions, and predict mortality risks.

With massive data readily available, it becomes a challenge for healthcare providers to improve the current utilization of common measures, e.g., lab test results and patient monitoring signals. This is even more critical for patients suffering from chronic diseases (e.g., heart diseases) or patients in intensive care units. It is estimated that more than five million patients are admitted to ICUs yearly in the US and 10% to 20% of them die in hospitals [2]. There is a dire need to go beyond current medical practices, and develop data-driven methods and tools that will enable and help (i) the handling of big data, (ii) the extraction of data-driven knowledge, (iii) the exploitation of acquired knowledge for optimizing clinical decisions.

Predicting ICU mortality is critical for the improvement of the quality of healthcare services (e.g., promoting effectiveness of surgical procedures, medication usages, care guidelines, treatment plans). Further, it will provide data-driven performance measures to compare the differences of healthcare facilities and services, thereby eliminating healthcare disparities in the country. In the state of the art, APACHE and SAPS scores are widely used to describe the acuity levels of ICU patients. However, they have yielded limited success due to the fewer variables and shorter time period considered.



Variable heterogeneity:
- Quantitative (continuous, discrete), qualitative (categorical, factors, dummy variables)

Patient heterogeneity:
- Various types of diseases, ages, surgical procedures, populations

Time asynchronization:
- Data collection per physician's discretion, missing data, non-uniform sampling

Fig. 1. Characteristics of postsurgical datasets in ICU.

As shown in Fig. 1, postsurgical monitoring in ICU leads to massive datasets with unique properties, e.g., variable heterogeneity, patient heterogeneity, and time asynchronization. There are significant challenges to extract useful knowledge from heterogeneous postsurgical datasets for the optimization of clinical decision making.

(1) ***Variable heterogeneity***: In order to capture a complete picture of the recovery process of postsurgical patients, ICU monitoring includes a large number of variables (e.g., lab test results, pulse oximetry, blood pressure, and heart rate). Most importantly, there are different types of variables. Some are quantitative (continuous, discrete), while others may be qualitative (categorical, factors, dummy variables). As opposed to the conventional univariate analysis, it is critical to discover risk factors and interactions hidden in heterogeneous types of variables, reducing them to a parsimonious set of sensitive biomarkers that will help in the diagnosis, monitoring and prediction.

This work is supported by the National Science Foundation (IOS-1146882 and CMMI-1266331).

Yun Chen and Hui Yang are with the Complex Systems Monitoring, Modeling and Analysis Lab, University of South Florida, Tampa, FL 33620 USA (e-mail of corresponding author: huiyang@usf.edu).

978-1-4244-7929-0/14/$26.00 ©2014 IEEE

(2) ***Patient heterogeneity***: Further, it may be noted that there are also heterogeneous types of patient populations, which may be classified by ages, gender, diseases, surgical types, or ICU types (e.g., coronary care unit, cardiac surgery recovery unit, medical ICU, surgical ICU). This also provides an opportunity to investigate mortality rates for different patient populations.

(3) ***Time asynchronization***: It should also be noted that the data collection procedures are not standardized in ICU. It is common that the frequency of data measurements is at the physician's discretion. In particular, each variable has an associated time stamp indicating the time point of data recording. However, time stamps are often not uniformly distributed along the time axis. During 48-hour ICU monitoring, some variables may be recorded in an extremely low sampling rate while others may be in a high sampling rate. Missing data is also a common property of ICU datasets.

Hence, there is an urgent need to address the issues of variable heterogeneity, patient heterogeneity, and time asynchronization and develop analytical methods for patient-specific prediction of in-hospital mortality. This present paper focuses on the prediction of mortality rates in Intensive Care Units using patient-specific and heterogeneous postsurgical datasets. To cope with the challenges in ICU datasets, we developed the postsurgical decision support system with a series of analytical tools, including data categorization, data pre-processing, feature extraction, feature selection, and predictive modeling.

This paper is organized as follows: Section II will introduce the methodology of postsurgical data analytics. Section III will detail the materials used and experimental results. Section IV discusses and concludes this study.

II. Postsurgical ICU Monitoring and Analytics

Fig. 2 shows the overall flowchart of the proposed data-driven postsurgical ICU decision support system. Notably, healthcare technology in 21st century has given rise to the big data in the ICU that involves a greater level of complexity and challenge, including variable heterogeneity, patient heterogeneity, and time asynchronization. The proposed decision support system is embodied by five core components (i.e., data categorization, data pre-processing, feature extraction, feature selection, and predictive modeling) that are effectively integrated to improve patient-specific prediction of in-hospital mortality.

First, we categorize various types of variables into 4 groups (namely general descriptors, low-sampling variables, med-sampling variables and high-sampling variables) based on the missing percentage in databases and the average number of observations per variable. Second, these four categories of variables will be pre-processed to ensure the data quality with various imputation and derivation methods (see details in Table I). Third, we transform variables into features that contain critical clinical information, and then use feature selection techniques to reduce high-dimensional features into a sparse set of sensitive biomarkers. Finally, we construct the predictive models with sensitive biomarkers that predict the clinical outcomes for ICU patients. These five components are detailed in the following sections.

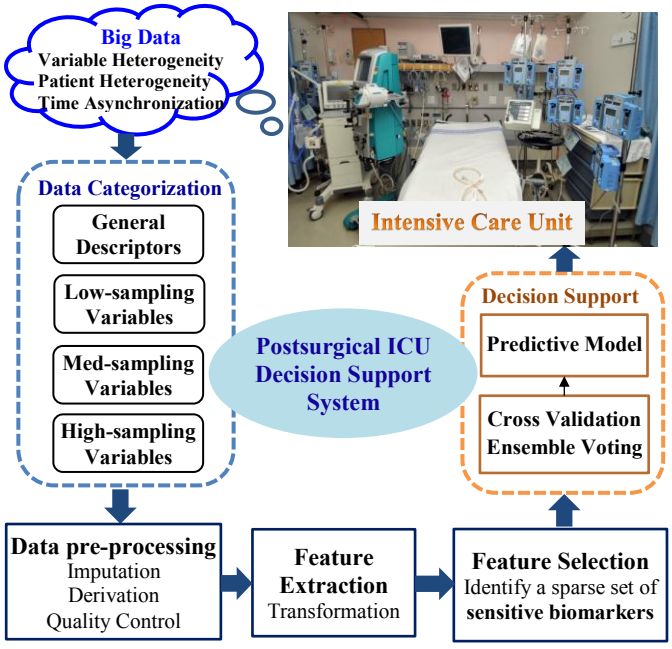


Fig. 2. Flow chart of postsurgical ICU decision support system.

A. Data Categorization

The common measurements in ICU consist of 44 variables (see details of variable names in Table I). Over the course of 48 hours, certain variables were measured at different time points with physicians' discretion due to different conditions of patients. It is very often that not all the 44 variables are recorded for each patient. Each patient may be monitored with a subset of variables at non-uniformly sampled time points. Variables may be recorded once, more than once, or not at all within 48 hours of ICU stay.

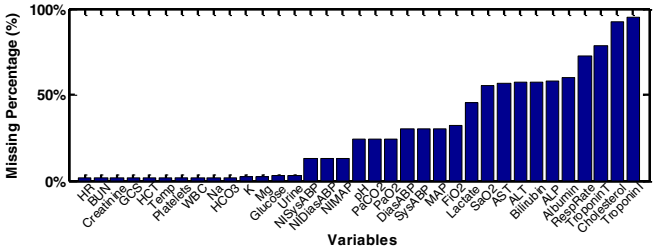


Fig. 3. The percentage of missing data for variables.

For example, Fig. 3 shows the percentage of missing data for common variables in one ICU database. Here, six general descriptors are excluded because they are recorded once in the beginning of ICU stay. It can be seen that none of variables is completely recorded for all patients. Also, some variables have more than 50% missing in the database. Based on the percentage of missing data and the average number of observations per variable, we categorize these 44 variables into 4 groups as shown in the following Table I.

- *General descriptors*: This group of variables includes general properties of a patient that are collected when the patient is first admitted into the ICU, e.g., RecordID, Age, Gender, Height, ICUType, MechVent.
- *Low-sampling variables*: more than 50% missing for the patients in the database.

TABLE I: POSTSURGICAL ICU DATA CHARACTERISTICS, CATEGORIZATION, PRE-PROCESSING AND FEATURE EXTRACTION

Data Category	Variables	Normal Range	Missing Percentage (%)	# Observations (median±std)	Data Processing		Feature Extraction
General Descriptor	RecordID		All available	Recorded once at the beginning			Remain unchanged
	Age						
	Gender						
	Height				Imputation method 1[a]		
	ICUType						
	MechVent		35.87	7±7.56			1: ventilation required 0: otherwise
Low-sampling Variables	TroponinI	0~10	94.87	0±0.55	Imputation method 2[b]		0: not recorded; 1: within normal range; 2: abnormal
	TroponinT	0~0.1	78.42	0±1.19			
	Cholesterol	200~1000	92.37	0±0.27			
	RespRate	10~20	72.47	0±23.55			
	Albumin	3.5~5.4	59.62	0±0.9			
	ALP	44~147	57.75	0±1.26			
	Bilirubin	0.2~1.9	57.05	0±1.28			
	ALT	F: 10~50; M: 5~38	56.97	0±1.28			
	AST	F: 8~40; M: 6~34	56.87	0±1.28			
	SaO2	94~100	55.2	0±3.46			
Med-sampling Variables	Lactate	3.7~5.2	45.42	1±3.15		Imputation method 4[f]	Compute the mean for each variable
	BUN	6~20	1.6	3±1.68			
	Creatinine	F: 0.6~1.1; M: 0.7~1.3	1.6	3±1.7	Replaced by CreatinineClearance		
	Glucose	70~100	2.82	3±1.8			
	HCO3	23~29	1.9	3±1.7			
	K	0.5~2.2	2.4	3±1.92			
	Mg	1.7~2.2	2.57	3±1.77			
	Na	135~145	1.87	3±1.86			
	Platelets	150~450	1.7	3±1.91			
	WBC	4.5~10	1.82	3±1.57			
	HCT	F: 35~48; M: 40~53	1.6	4±2.58			
	PaCO2	35~45	24.42	5±5.72			
	PaO2	75~100	24.42	5±5.71			
	pH	7.38~7.42	24	5±5.91			
High-sampling Variables	FiO2	0.21~0.5	32.07	8±7.34			Extract the following statistics for each variable: minimum maximum mean median the first observation the last observation linear trend in the first 24 hrs linear trend in the second 24 hrs linear trend over 48 hrs
	GCS	0~3	1.6	13±7.88			
	Temp	36~40	1.6	14±17.45			
	Urine	1500	2.92	37±12.49	Replaced by Urine.Sum[d]		
	Weight		All available	37±26.43	Imputation method 1[a]		
	HR	60~100	1.57	55±16.05			
	MAP	70~100	30.2	42±30.14	Imputation method 3[e]		
	NIMAP	70~100	12.97	21±20.48			
	DiasABP	60~90	30.02	43±29.57	Imputation method 3[e]		
	NIDiasABP	60~90	12.92	21±20.7			
	SysABP	100~140	30.02	43±29.59	Imputation method 3[e]		
	NISysABP	100~140	12.67	21±20.7			

a. Erroneous values were removed, and missing values were replaced using linear regression based on the most common values by gender.
b. Combine TroponinI and 100*TroponinT as a new variable - Troponin.
c. $\text{CreatinineClearance} = (140 - \text{Age}) \times \text{Weight} \times (0.85 + 0.15 \times \text{Gender}) / (72 \times \text{Creatinine})$.
d. Urine.Sum is the cumulative sum of Urine.
e. Combine two variables together and add a new descriptor as: 1 if the majority observations were from invasive procedure; 0 otherwise.
f. Missing variables were imputed by a random value from the Gaussian distribution representing the normal physiology of each variable.

- *Med-sampling variables*: The average number of observations is less than 15 per patient per variable.
- *High-sampling variables*: variables that do not meet with the above criteria.

B. Data Pre-processing

The step of data pre-processing is to ensure the data quality with various imputation and derivation methods that are detailed in Table I.

First, erroneous weight and height values were removed, and missing height/weight values were replaced using a simple linear regression based on the most common height/weight values by gender.

Second, TroponinT was multiplied by 100 and then combined with TroponinI as a new variable, Troponin. If one is missing, then the new variable Troponin takes the value of the other. Otherwise, it will take the average value.

Third, Creatinine is replaced by CreatinineClearance, which is calculated based using the Cockcroft Gault equation:

$$\text{CreatinineClearance} = (140 - \text{Age}) \times \text{Weight} \times (0.85 + 0.15 \times \text{Gender}) / (72 \times \text{Creatinine})$$

Fourth, Urine is replaced by a new variable Urine.Sum, which is the cumulative sum of the Urine measurements.

Fifth, three pairs of variables, i.e., DiasABP and NIDiasABP, MAP and NIMAP, SysABP and NISysABP, were combined respectively as 3 new time series and add a binary variable that will be assigned 1 if the majority observations were from the invasive procedure, 0 otherwise.

Finally, missing values for med-sampling and high-sampling variables are imputed by a random value from the gender-specific Gaussian distribution representing the normal physiology of each variable.

C. Feature Extraction

After the data pre-processing, we transform variables into features with meaningful clinical information (see Table I). First, MechVent is transformed into a new binary variable: 1 if the patient required mechanical ventilation at any time during the 48 hour observation period, 0 otherwise. Second, low-sampling variables are transformed into the new categorical variables (0: not recorded; 1: within normal range; 2: abnormal). Third, we computed the mean of med-sampling variables as new predictors. Finally, for high-sampling variables, a number of features were extracted (including minimum, maximum, mean, median, the first observation, the last observation, linear trend over the first 24 hours, linear trend over the second 24 hours, and linear trend over 48 hours. Trends are not computed for variables having less than 10 observations in 48 hours or 5 observations in either 24-hour period. In total, there are 112 features extracted from the ICU datasets (see *Section III*). Any statistics with spurious medical meaning, e.g., Urine.Sum.Min, were excluded. In addition, the change of weight over 48 hours is recorded as Weight.Delta.

D. Feature Selection

Note that a large amount of features are extracted. As a result, this may bring the "curse of dimensionality" issues for predictive models, e.g., model sensitivity and overfitting problems [3]. In this study, we used a filtering method, namely minimum redundancy and maximum relevance (mRMR) [4], to reduce high-dimensional features into a sparse set of sensitive biomarkers. The mRMR method selects features that are maximally relevant to the response variable while minimizing redundancy between select variables. The mutual information of two variables x and y is computed as using their joint probabilistic distribution $\mathrm{p}(\mathrm{x},\mathrm{y})$ such that

$$I(x,y) = \sum_{i,j} p(x_i, y_j)\, log\, \frac{p(x_i,y_j)}{p(x_i)p(y_j)}$$

where $\mathrm{p}(\mathrm{x})$ and $\mathrm{p}(\mathrm{y})$ are marginal probabilities and $\mathrm{p}(\mathrm{x},\mathrm{y})$ is the joint probabilistic distribution. Redundancy and relevancy are defined respectively according to the following equations:

$$W_I = \frac{1}{|S|^2}\sum_{j\in S} I(i,j)\,, V_I = \frac{1}{|S|}\sum_{i\in S} I(h,i)$$

where $|\mathrm{S}|$ represents the number of features in the feature set S, i and j denote the ith and jth features, and h is response variable. Minimizing $\mathrm{W_I}$ and maximizing $\mathrm{V_I}$ ensures minimal redundancy and maximum relevancy, and the Mutual Information Difference (MID) is $\mathrm{MID} = \max(V_I - W_I)$.

The higher the MID score, the most significant the feature is. In order to further minimize redundancy, we add the constraint to select only the most important feature for high-sampling variables. For example, we choose either mean or median, minimum or maximum values, first observation or last observation, and finally, only one of trend values.

E. Predictive model and cross validation

Furthermore, we construct the predictive models that associate the input feature pattern $\mathcal{s}$ to one of the $\mathcal{K}$ classes of outcomes $\mathcal{C}_1, \dots, \mathcal{C}_{\mathcal{K}}$. In this study, clinical outcomes are binary ($\mathcal{K} = 2$), i.e., survival or in-hospital death. The whole dataset $\mathcal{D}$ is partitioned into the training dataset $\mathcal{D}_1 = \{\langle y(i), \mathcal{s}(i)\rangle | i = 1, \cdots, N_1\}$ and testing dataset $\mathcal{D}_2 = \{\langle y(i), \mathcal{s}(i)\rangle | i = N_1 + 1, \cdots, N_1 + N_2\}$, where N_1 and N_2 are the size of training and testing datasets, $y(i)$ takes values in the output sets $\mathcal{C}_1, \dots, \mathcal{C}_{\mathcal{K}}$, $\mathcal{s}(i) = \{\mathcal{s}_{i1}, \mathcal{s}_{i2}, \dots, \mathcal{s}_{i\ell}\}$ is the set of ℓ selected features for the ith record in $\mathcal{D}$.

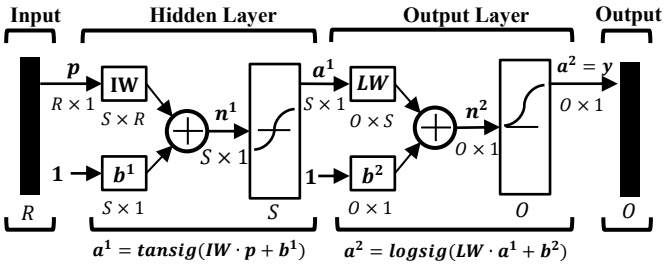


Fig. 4. The structure diagram of multilayer neural network.

Fig. 4 shows the multilayer neural network (NN) model used in this present study, in which Hyperbolic tangent sigmoid transfer function (tansig) is used in the hidden layer and log-sigmoid transfer function (logsig) in the output layer. The hidden layer includes $S = 40$ neurons and the output layer contains $O = 2$ neurons. To reduce the bias in NN models, we have utilized both K-fold cross-validation and random subsampling [5] in this present investigation. In addition, it may be noted that the class sizes are often not equal. The classification models, in general, will favor the larger (majority) class, thereby affecting the performance for testing datasets. For example, we have a highly-imbalanced datasets, i.e., 3446 survivals and 554 in-hospital deaths in this present study. Therefore, we adopted the ensemble-based learning that statistically bootstraps the minority class with random replacements [6]. For K-fold cross validation, $K - 1$ folds are used for training, denoted as $\mathcal{D}_1^{(\mathrm{i})}$ and the rest one for testing, denoted as $\mathcal{D}_2^{(\mathrm{i})} (= \mathcal{D} - \mathcal{D}_1^{(\mathrm{i})})$. Here, the training dataset $\mathcal{D}_1^{(\mathrm{i})}$ is augmented by bootstrapping additional samples form the minority class in $\mathcal{D}_1^{(\mathrm{i})}$ so as to reconstruct the balanced training dataset. The classification models are trained with this balanced training set. This procedure is replicated n times to obtain n different bootstrapped training datasets $\mathcal{D}_1^{(\mathrm{i})}$. The final prediction results are based on the majority voting from n classifiers trained. As such, this ensemble voting approach provides more balanced estimates of performance metrics.

III. Materials and Results

In this present study, we used real-world ICU dataset to evaluate and validate the proposed methodology. This dataset is extracted from Multiparameter Intelligent Monitoring in Intensive Care (MIMIC) II Clinical Database [7, 8], which was developed to advance intelligent patient monitoring research in the critical care environment. This dataset is divided into two groups, i.e., Set A and Set B, and each of

them consists of 4,000 patient records from 48 hours of ICU stays (including coronary care unit, cardiac surgery recovery unit, medical ICU and surgical ICU). Clinical outcomes (i.e., in-hospital death or survival) are made available for Set A, but not for Set B. The training of predictive models is only based on 4,000 subjects in Set A.

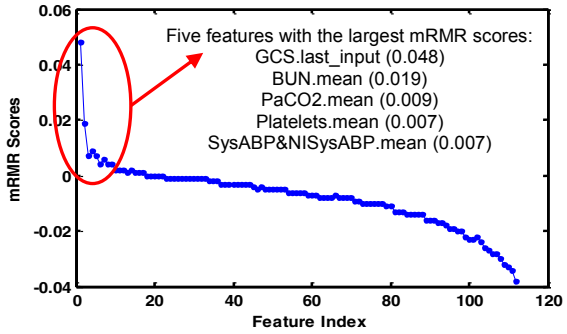


Fig. 5. The sorted mRMR scores for all the extracted features.

Fig. 5 shows the sorted mRMR scores of all features extracted (see Table I). We empirically selected 47 features with the mRMR score greater than -0.005 from the entire set of 112 features.. Note that the mean and the last input of variables are shown to be more significant than other features, and contain sensitive information for the prediction of mortality risks. Fig. 6 shows the average performance metrics of ensemble NN models (i.e., sensitivity, specificity, PPV, NPV and accuracy) that are computed from 100 random replications of 4-fold cross-validation of Set A. Note that the final score is the minimum of sensitivity and PPV. Fig. 6 also shows the receive operating characteristic (ROC) curve, and the area under the curve (AUC) reaches 0.8755 for the NN model.

Set A Results	
Sensitivity	0.8339
Specificity	0.9170
PPV	0.6176
NPV	0.9717
Accuracy	0.9055
Score	0.6176

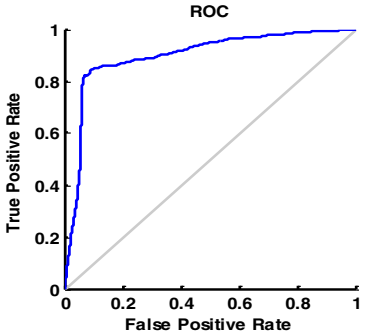


Fig 6. Performance measures of the ensemble NN model

Furthermore, Table II shows the comparison of the proposed method with various methods in the state of the art [9]. The proposed method achieves the score of 0.617, indicating data-driven models can not only effectively extract the sensitive biomarkers, but also provide accurate prediction of ICU mortality risks. In addition, it may be noted that the final score for Set B with undisclosed outcomes is 0.50, which was evaluated with the help of Dr. Ikaro Silva at the Harvard-MIT Division of Health Sciences and Technology.

TABLE II. PERFORMANCE COMPARISONS OF PREDICTIVE MODELS

Methods	Random Classifier	SOFA	SAPS-I	Fuzzy Rule	Cascaded AdaBoost
Scores	0.15	0.28	0.32	0.36	0.38
Methods	Time Series Motifs	LR & HMM	Neural Network	Bayesian Ensemble	**Proposed Method**
Scores	0.50	0.50	0.51	0.53	**0.617**

IV. CONCLUSION AND DISCUSSION

Currently, physicians have access to a great deal of data to evaluate patient conditions, but these data are not processed to be easily interpretable and then be useful to perform a proper assessment of patient conditions. The development of patient-specific data analytical methods and tools will help healthcare providers to better use massive healthcare recordings for clinical decision support.

This present study developed the data-driven ICU decision support system with a series of analytical tools, including data categorization, data pre-processing, feature extraction, feature selection, and predictive modeling. As opposed to current clinical practice of visual inspection and univariate analysis, this investigation specifically considered addressing the challenges of ICU data, including variable heterogeneity, patient heterogeneity, and time asynchronization. Experimental results on real-world data show great potentials of data-driven analytics for improving the prediction of ICU mortality risks. Advances in postsurgical monitoring practices for patients who undergo surgical procedures will significantly decrease the mortality rates in ICU and lead to broader social impacts.

ACKNOWLEDGMENTS

The authors would like to thank Dr. Ikaro Silva, Harvard-MIT Division of Health Sciences and Technology, for his kind help on the evaluation and scoring of the proposed methodology presented in this paper.

REFERENCES

[1] S. Keehan, A. Sisko, C. Truffer, S. Smith, C. Cowan, J. Poisal, M. K. Clemens and National Health Expenditure Accounts Projections Team, "Health spending projections through 2017: the baby-boom generation is coming to Medicare," *Health Affairs,* vol. 27, pp. w145-155, 2008.

[2] P. J. Pronovost, D. M. Needham, H. Waters, C. M. Birkmeyer, J. R. Calinawan, J. D. Birkmeyer and T. Dorman, "Intensive care unit physician staffing: financial modeling of the Leapfrog standard," *Critical Care Medicine,* vol. 32, pp. 1247-1253, 2004.

[3] H. Yang, "Multiscale recurrence quantification analysis of spatial cardiac vectorcardiogram (VCG) signals," *IEEE Trans. Biomed. Eng.,* vol. 58, pp. 339-347, February, 2011.

[4] C. Ding and H. Peng, "Minimum redundancy feature selection from microarray gene expression data," *Journal of Bioinformatics and Computational Biology,* vol. 3, pp. 185-205, 2005.

[5] Y. Chen and H. Yang, "Multiscale recurrence analysis of long-term nonlinear and nonstationary time series," *Chaos, Solitons & Fractals,* vol. 45, pp. 978-987, 2012.

[6] Y. Chen and H. Yang, "Self-organized neural network for the quality control of 12-lead ECG signals," *Physiological Measurement,* vol. 33, pp. 1399, 2012.

[7] M. Saeed, M. Villarroel, A. T. Reisner, G. Clifford, L. Lehman, G. Moody, T. Heldt, T. H. Kyaw, B. Moody and R. G. Mark, "Multiparameter Intelligent Monitoring in Intensive Care II (MIMIC-II): A public-access intensive care unit database," *Critical Care Medicine,* vol. 39, pp. 952-960, 2011.

[8] A. L. Goldberger, L. Amaral, L. Glass, J. Haussdorff, P. C. Ivanov, R. Mark, J. Mietus, G. Moody, C. -. Peng and H. E. Stanley, "PhysioBank, physiotoolkit, and physionet: Components of a new research resource for complex physiologic signals," *Circulation,* vol. 23, pp. e215-e220, June. 13, 2000.

[9] I. Silva, G. Moody, J. Scott, L. A. Celi and R. G. Mark, "Predicting In-Hospital Mortality of ICU Patients : The PhysioNet / Computing in Cardiology Challenge 2012," *Computing in Cardiology,* vol. 39, pp. 245-248, 2012.