Imputation-Enhanced Prediction of Septic Shock in ICU Patients

Joyce C. Ho
University of Texas at Austin
1 University Station, C0803
Austin, TX 78712, USA
joyceho@utexas.edu

Cheng H. Lee
University of Texas at Austin
1 University Station, C0800
Austin, TX 78712, USA
chlee@utexas.edu

Joydeeph Ghosh
University of Texas at Austin
1 University Station, C0803
Austin, TX 78712, USA
ghosh@ece.utexas.edu

ABSTRACT

Sepsis and septic shock are potentially fatal complications that frequently occur in intensive care unit patients. The ability to predict which patients are at risk for sepsis and septic shock is therefore crucial to limiting the effects of these complications. Potential indications for sepsis risk are scattered in a wide range of clinical measurements, including high-temporal resolution physiological waveforms, X-rays and gene expression levels, etc., leading to a non-trivial prediction problem. Thus previous works on sepsis prediction have used very small, carefully curated datasets, with limited applicability. Recently however, a large, rich ICU dataset called MIMIC-II has been made publicly available, providing opportunity for more extensive modeling of this problem. However, such a large dataset inevitably comes with a substantial higher amount of missing data. In this paper, we investigate how different imputation methods can overcome the handicap of missing information while leveraging such a large dataset. Our results show that imputation approaches in conjunction with predictive modeling lead to a decent boost in accuracy of sepsis risk prediction and a huge improvement in prediction of septic shock, even when one is restricted to only using non-invasive measurements. Our models can be applied to any ICU patient and lead to a generalized approach for predicting sepsis related complications.

Categories and Subject Descriptors

J.3 [**Computer Application**]: Life and medical sciences; I.2.6 [**Artificial Intelligence**]: Learning

Keywords

missing data, imputation methods, logistic regression, sepsis, data mining

HI-KDD '12, August 12, 2012, Beijing, China

1. INTRODUCTION

Sepsis is a systemic response to infection that is a common and often life-threatening complication in intensive care unit (ICU) patients [4]. Cases of severe sepsis often lead to septic shock, a complication characterized by low blood pressure despite treatment that dramatically increases mortality risk [1]. Early intervention and therapy have been shown to improve the outcome of patients with severe sepsis and septic shock [8], thus making accurate identification of patients at risk for developing these conditions crucial to improving standards of clinical care.

However, the nature of clinical data complicates the development of highly accurate predictive models. Clinical data are typically noisy and inconsistently gathered. For example, while physiological measures such as heart rate are electronically monitored, they are often manually recorded in a patient's chart by a healthcare provider, and such recordings may be done at irregular intervals. Furthermore, accurate measures for physiological variables like blood pressure require the use of invasive techniques that would place patients at unnecessary risk and therefore cannot be ethically gathered. In such cases, less accurate measurements obtained by non-invasive means are the only available data.

Thus, clinical studies must often deal with large amounts of missing data. A typical approach to addressing this problem is to simply ignore subjects or features that have missing data. However, doing so can cause dramatic decreases in sample sizes or feature spaces, potentially hindering the development of accurate models. In addition, data that is not missing (completely) at random can introduce a bias in the results. We also note that previous work on sepsis prediction have largely been restricted to very modest, highly curated datasets with small number of samples and limited sets of features.

The issue of missing data is also encountered in the MIMIC II database [10], which is one of the most extensive publicly available data collected to date on ICU patients. To fully exploit the potential of such available data, this paper investigates the role and impact of imputation methods while building predictive models for sepsis risk and septic shock. We limit our features to commonly observed, mostly non-invasive clinical measures, so that the results are easily actionable on large numbers of patients. We demonstrate that even with such restrictions, one can achieve accuracies comparable to those models that rely on more invasively-gathered measures and that are much more limited in scope. Our approaches easily generalize to all ICU patients and allow for early diagnosis and intervention for septic patients.

2. BACKGROUND AND RELATED WORK

2.1 Clinical Definitions

Sepsis is a severe, systemic inflammatory response and is diagnosed when a patient has an infection (or evidence of an infection) that is associated with two or more of the following critera: (1) abnormal body temperature, (2) increased heart rate, (3) increased respiratory rate, or (4) abnormal white blood cell counts [1]. Severe sepsis is defined as "sepsis associated with organ dysfunction, hypoperfusion, or hypotension" [1]. We do not distinguish between these two conditions in this study and use the term "sepsis" to refer to both sepsis and severe sepsis in the remainder of the paper.

Septic shock is diagnosed when a septic patient has a systolic blood pressure (SBP) <90 mmHg despite treatment by >600 mL of fluid inputs in the last hour [1, 11].

2.2 MIMIC II Database

The Multiparameter Intelligent Monitoring in Intensive Care II (MIMIC II) database is a publicly available resource developed to support research in clinical decision support and critical care medicine [10]. Version 2.6 of this database provides data on >30,000 patients in the ICUs of Boston's Beth Israel Deaconess Medical Center between 2001 and 2007. The clinical records include charted physiological measures, medication records, fluid input and output records, laboratory test results, procedure orders, and free-form text notes produced for each of the >40,000 ICU stays recorded in the database.

The database also provides high temporal-resolution physiological waveform data gathered on $\sim 2,000$ patients using bedside electronic monitoring units. However, this study did not use any of these data due to ambiguities associating some waveforms with specific subjects as well as the lack of such data for the most of the patients in the database.

2.3 Sepsis and severe sepsis prediction

Several approaches for identifying patients at risk for developing sepsis have been proposed. Lukaszewski et. al [4] used several clinical factors and the expression levels of seven immune system related genes to train a multilayered perceptron and reported an 83% predictive accuracy for patients who became septic. Tang et. al [13] applied PCA and a nonlinear support vector machine (SVM) to high temporal-resolution physiological waveform data and achieved an 84% accuracy for predicting sepsis onset. Gwadry-Sridhar et. al [3] used twenty clinical variables and a decision tree to achieve nearly 100% predictive accuracy.

Despite their high accuracies, these methods all suffer from similar faults. The studies were all performed on extremely small samples of 30 or fewer patients, and the reproducibility of their results has yet to be determined. Additionally, each method required data that are difficult or overly invasive to gather in regular clinical settings; for example, Lukaszewski's approach required daily blood draw and time-consuming quantification protocols, and Gwadry-Sridhar's required chest X-rays. Finally, none of these proposed methods could reliably deal with missing data.

2.4 Septic shock prediction

Other recent work has focused on early prediction of septic shock. A septic shock early warning system (EWS) was developed using multivariate logistic regression on commonly measured clinical variables [11]. Using a dataset with 65 septic shock and 185 sepsis only patients, the model could predict the onset of septic shock one hour in advance with an area under the receiver operating characteristic curve (AUC) of 0.928. However, the system used invasively-gathered data and extracted features from the MIMIC waveform data, which provide higher time-resolution than data more commonly available in most ICUs. Another study performed Recursive Partitioning and Regression Tree (RPART) analysis on 1864 septic patients to identify early predictors from clinical data of hospitalized non-ICU patients [15]. The model required results from eleven basic, routine laboratory tests and certain vital signs, and correctly identified only 55% of the septic shock patients. Furthermore, neither papers addressed problems with missing data.

Several other models have been developed to predict septic shock in the absence of full featured data. Pereira et. al used Zero-Order-Hold (ZOH) method to deal with incomplete data and Fuzzy C-Means algorithm with Partial Distance Strategy (FCM-PDS) to predict the outcome on 121 patients with abdominal septic shock [7] with an AUC of 0.899. Another paper also used the ZOH method to deal with incomplete data and compared the accuracy of fuzzy models and neural networks on the same abdominal shock dataset with 121 patients [2]. Using 28 features, the fuzzy model with bottom-up feature selection obtained an AUC of 0.818. Although the features are the most frequently measured variables found in the abdominal shock database, they rely heavily on laboratory results and invasive measurements. Additionally, it is uncertain whether the model can generalize to all ICU patients. Paetz uses the centers of gravity to replace missing observations and a trapezoidal function neural network to classify data from 874 patients [6]. However, the experiment used results from several laboratory tests, utilized invasive catheter measurements, and required at least 10 of the 12 variables to be present. The model only correctly classifies $\sim$70% of the test data and has a sensitivity of 15.01%.

3. METHODS

3.1 Patient selection

This study examined adults (>18 years of age at time of admission) from the MIMIC II database who had only one ICU admission over the course of their hospital stay. We exclude patients younger than 18 years of age to avoid (1) confounding factors arising from their different physiologies and (2) complications arising from the need to obtain informed consent from parents when migrating our model to other clincial data sets. To ensure that we had enough data for our septic shock models, we only included those patients who had at least ten observations of blood pressure (BP), heart rate (HR), respiration rate (RR), blood oxygen saturation measured by pulse oximeter (SpO_2), temperature (TEMP), and two observations of white blood cell count (WBC). Septic patients were identified based on their ICD-9 codings ("995.91" or "995.92") assigned after hospital discharge.

Patients with septic shock were identified by examining their clinical chart records, with time of onset determined

using criteria modified from those used by Shavdia [11]. Due to the low frequency of blood pressure observations in our data (>30 min between measurements), we defined a hypotension observation as any time point where systolic blood pressure (SBP) was <90 mmHg. Consecutive hypotension observations were then aggregated to define a hypotension region. Total fluid intake from one hour prior to the first hypotension observation in a hypotension region to halfway through the region was calculated. Any hypotension region that had a total fluid intake >600 mL was classified as septic shock, with onset defined as the start time of this region.

Of the 12,179 patients included in our study, 1,310 ($\sim$ 10.8%) were diagnosed with sepsis or severe sepsis. Of the septic patients, 586 ($\sim$44.7%) were diagnosed with septic shock according to their ICD-9 codings. Noisy, inconsistent, and infrequent measurements for fluid intake and blood pressure prevent us from identifying the precise time of onset for all septic shock patients. Consequently, only 328 of the patients diagnosed with septic shock ($\sim$25.0% of all septic patients) were used in the study.

3.2 Predictive models for sepsis

We generated two sets of features for each patient. The first set represented a patient's clinical history and consisted only of information available when the patient was first admitted into the ICU; these features included demographic data (gender and ages at hospital and ICU admission), medical history (flags indicating previous hospital and ICU admissions), basic health data (weight and physician calculated SOFA and SAPSI scores), and the ICU unit into which the patient was entering.

The second set included the first non-invasive measurement of four physiological variables (BP, HR, RR, and SpO_2). These features were meant to capture the patient's initial physiological state upon entering the ICU, and therefore, only data taken within the first six hours of ICU admission were considered. Measurements taken after the first six hours were treated as missing data that would later have to be imputed.

We developed two sets of multivariate logistic regression models to predict which patients would develop sepsis at some point in their ICU stay. The first set of models used only the clinical history features and included models that: (1) used all available features, (2) performed forward stepwise regression, and (3) performed backwards stepwise regression. The second set used both the clinical history and the initial physiological state features, and again, models that used all available features, performed forward stepwise regression, and performed backward stepwise regression were developed.

3.3 Predictive models for septic shock

Sepsis patients without an ICD-9 code for septic shock were assigned an onset time halfway between their first and last available BP measurement. A feature matrix of physiologic and laboratory values was generated based on a specified time prior to the onset or the reference time. Each data matrix contained basic information available at ICU admission, the most recent non-missing observation at reference time, and a baseline calculated from the mean of the previous 5 or fewer non-missing observations. A list of all the physiological features extracted from the clinical database is found in Table 1. Feature matrices were created at reference times of 30, 60, 90, and 120 minutes prior to the onset of septic shock.

Table 1: Physiologic and laboratory features for predicting septic shock

Baseline and recent value
Cardiac: SBP, HR, PP (= SBP - DBP)
Respiratory: RR, SpO_2
Other: TEMP
where DBP = diastolic BP, PP = pulse pressure

Recent value only
WBC
Arterial pH
Shock Index = HR / SBP

Table 2: Availability of the physiological measurements for septic shock prediction

Feature	% missing	Feature	% missing
RR	0.30	Mean RR	1.71
TEMP	1.10	Mean TEMP	6.43
SpO_2	0.40	Mean SpO_2	2.11
Mean HR	1.20	Mean SBP	1.20
Mean PP	1.20	Weight	4.92
APH	29.62	WBC	5.82
SOFA	25.20	SAPS-I	28.92

We used three different models, multivariate logistic regression, a linear kernel support vector machine, and regression trees to predict septic shock on the feature matrices. For logistic regression, we trained a set of three models at each reference time: one incorporating all available features, the second using forward stepwise regression, and the third using backward stepwise regression.

Our choice for using logistic regression for predicting sepsis risk as well as septic shock onset was motivated by the fact that such models are relatively simple to implement, can handle non-linear effects, and return readily interpretable results. Support vector machines and regression trees were also included because they have been used in previous septic shock work.

3.4 Missing value imputation

The lack of consistent sampling found in the MIMIC II database resulted in missing physiological, laboratory, and severity score values. Missing data were more prominent in the sepsis shock dataset, where physiological measurements were often not available in the required time windows. Table 3.4 shows the percentage of patients missing a particular feature. Although the missing values only comprised 6.7% of all observations, excluding patients without the full feature matrix would have reduced the size by of our dataset 47.2%. Consequently, techniques for imputing missing values were necessary.

Simple and accessible approaches were used to estimate the missing observations. The effectiveness of using the conditional mean for recommendation systems led us to our first two imputation methods: mean feature values derived from patient gender and mean feature values derived from the patients' gender and age group. Matrix factorization-based

Table 3: AUC comparison for sepsis models using non-imputed and imputed feature sets

	None	Imputed
All H	0.791±0.004	0.792±0.003
Stepwise H	0.790±0.003	0.791±0.004
All H ∪ P	0.821±0.002	0.822±0.002
Stepwise H ∪ P	0.823±0.002	0.823±0.002

H = clinical history feature set
P = initial physiological state feature set

approaches were also used to impute missing values. The BioConductor `pcaMethods` package [12] offered three missing value estimators. SVDImpute used a linear combination of k-eigenvalues to predict the missing value [16]. Probabilistic Principal Component Analysis (PPCA) combined an Expectation-Maximization (EM) approach to Principal Component Analysis (PCA) with a probabilistic model [9]. Bayesian PCA (BPCA) used a likelihood function that penalizes data "far" from the training set with an EM PCA to estimate the missing values [5].

The optimal number of principal components (k) for SVDImpute, PPCA, and BPCA was calculated using cross-validation on the complete training dataset with no missing observations. 10% of the data was removed at random and the root mean square error (RMSE) was calculated between the predicted matrix and actual matrix. The k resulting in the lowest RMSE for the imputation method was used on the full training dataset with missing entries.

4. RESULTS

4.1 Sepsis

Since no previous studies on predicting sepsis risk used a similar set of features, we had no "gold standard" with which we could evaluate our models' performance and instead used cross validation to obtain estimates for our models' performance.

Table 3 shows the average cross-validated AUC for the various models using non-imputed and imputed methods. Imputation by means (feature values derived from the patient's gender and age group), SVD, and BPCA all yielded identical mean AUC estimates. We note that for a given feature set, both forward stepwise and backward stepwise regression selected the same subset of features (shown in Table 4), and that models that incorporated all of the available features outperformed models that performed feature selection. More importantly, we note that models that incorporated initial physiological state information performed significantly better than models that relied only on clinical history. This suggests that a patient's trajectory toward sepsis-related complications may be determined soon after admission into the ICU.

These results also show that the choice of imputation method has no significant impact on performance of the models. However, this observation is not entirely surprising, as the data used to predict sepsis risk are relatively complete, with <7% of all the data missing.

4.2 Septic Shock

The septic shock EWS [11] was used as the "gold standard" for evaluating predictive performance for septic shock

Table 4: Logistic regression coefficients for four imputation models predicting sepsis

	None	Mean	SVD	BPCA
Admit age	0.013	0.013	0.013	0.013
Weight	0.601	0.595	0.590	0.589
Prev. hosp.	0.370	0.371	0.372	0.375
In MICU	-0.643	-0.570	-0.571	-0.565
CSRU service	-1.109	-1.032	-1.028	-1.022
MICU service	2.122	2.047	2.048	2.036
SICU service	0.769	0.758	0.761	0.757
HR	0.012	0.011	0.011	0.011
RR	0.026	0.025	0.025	0.027
SBP	-0.017	-0.018	-0.018	-0.018
Temp	0.069	0.074	0.072	0.076
SOFA	0.147	0.153	0.147	0.152

Table 5: AUC comparison with the septic shock EWS model

Feature Selection	EWS Features	Our Features
All	0.716±0.094	0.775±0.092
Forward stepwise	0.799±0.078	0.786±0.097
Backward stepwise	0.786±0.087	0.773±0.090

onset. A majority of the features were replicated except for cardiac output estimate and total peripheral resistance which derive from waveform data. Any patient without data for all the features from both models was omitted from the test. Consequently, the comparison was performed on 406 patients, of which 62 transition to septic shock. Three logistic regression models were trained for each set of model features: (1) all the features, (2) forward stepwise regression, and (3) backward stepwise regression. Table 5 summarizes the performance comparison for predicting septic shock 60 minutes before onset. The results show that use of noisy, inconsistently gathered clinical measurements instead of high-resolution waveform data causes a noticeable degradation in AUC of the septic shock EWS. Although the septic shock EWS slightly outperforms our model, it relies on frequent and invasive measurements, thereby reducing its utility in many clincial settings.

The logistic regression models illustrate the effect of imputation for predicting septic shock. Table 6 presents the comparison of AUC to the baseline model which discards the missing data. Imputation approaches show improvement across every reference time. In particular, BPCA generally has the highest AUC in comparison to the other methods. Furthermore, there is a degradation in accuracy as t (time before onset) grows. Figure 1 shows the ROC curves obtained for three of the imputation methods compared to the baseline model, and clearly bring out the benefits of imputation for increased predictive performance.

Table 7 demonstrates the effect imputation has on the logistic regression coefficients for predicting septic shock 60 minutes before onset. The coefficients are quite consistent across the different imputation techniques, but for some variables differ substantially from the values obtained without imputation. In particular, the non-imputed model focused on mean TEMP, SpO_2, and mean SpO_2 features. Although they are prominent features in the imputed models, the effects are dampened in conjunction with an increase of the

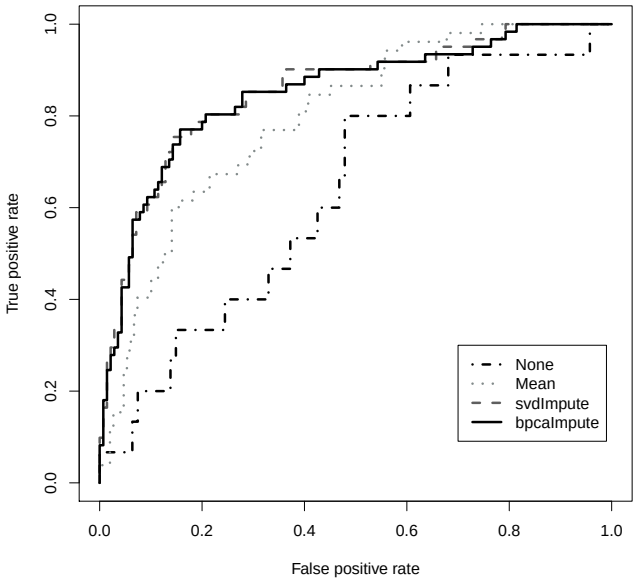

Figure 1: ROC curves for the full featured logistic regression model predicting septic shock 60 minutes before onset comparing 3 of the imputation methods with the baseline.

Table 6: AUC comparison of the full-featured logistic regression model using various imputation methods to predict septic shock

	Time t before onset		
Method	30 min	60 min	120 min
None	0.764±0.050	0.728±0.041	0.699±0.045
Mean	0.822±0.026	0.798±0.025	0.751±0.029
SVD	0.854±0.010	0.848±0.007	0.831±0.018
PPCA	0.855±0.009	0.852±0.007	0.828±0.016
BPCA	0.853±0.010	0.847±0.007	0.829±0.017

shock index measurement. The coefficients also illustrate the importance of imputation to provide a proper interpretation of the roles of all the variables. Restricting the dataset to only patients with all the features eliminated almost half of the data and decreased the coverage of the feature space. Consequently, the non-imputed model suffered from overfitting. However, the majority of the discarded data was only missing several features. Imputation allowed the prediction models to leverage these patients to obtain a more accurate representation of the weights of the features.

Table 8 presents an AUC comparison of septic shock prediction 60 minutes before onset for three types of models: (1) logistic regression (LogR), (2) linear kernel support vector machine (SVM), (3) regression trees (RPART) [14]. All three models show an increase in AUC for the various imputation methods. In particular, BPCA and PPCA perform comparably across the different classifier types.

Table 7: Logistic regression coefficients for four imputation models predicting septic shock

	Imputation			
Feature	None	Mean	SVD	BPCA
SBP	-4.035	-0.404	-0.343	-0.281
Mean SBP	0.721	-1.726	-1.775	-1.692
PP	-1.464	-0.615	-0.630	-0.600
Mean PP	-0.262	0.829	0.956	0.673
HR	1.158	-1.129	-1.169	-1.244
Mean HR	-1.038	-0.517	-0.531	-0.507
RR	1.722	1.961	1.930	1.890
WBC	-0.023	-0.461	-0.354	-0.191
Mean TEMP	-34.182	-13.595	-13.967	-12.277
SpO_2	-14.898	-6.508	-6.150	-6.154
Mean SpO_2	-10.709	-2.069	-1.829	-1.512
APH	-3.598	-3.417	-3.482	-3.455
SI	1.523	3.710	3.744	3.800

Table 8: AUC comparison of various model types using different imputation methods to predict septic shock

	Model Type		
Method	LogR	SVM	RPART
None	0.728±0.041	0.599±0.326	0.613±0.063
Mean	0.798±0.025	0.861±0.209	0.731±0.026
SVD	0.848±0.007	0.793±0.247	0.869±0.026
PPCA	0.852±0.007	0.874±0.218	0.873±0.018
BPCA	0.847±0.007	0.882±0.190	0.867±0.020

5. CONCLUSION

This paper presented a novel approach to accurately predicting both sepsis risk and septic shock from noisy, intermittently gathered clinical data. The features we chose minimize the use of laboratory tests and invasive procedures while still allowing us to maintain comparable performance to other models. Moreover, they can be applied to large datasets that could have substantial amounts of noise and missing values.

The combined results of our experiments show the effects that imputing missing data can have on improving model performance, especially when dealing with larger, noisier, and more incomplete datasets encountered in modern clinical studies. These results also indicate that imputing using matrix factorization methods like BPCA lead to models with better predictive accuracy than imputing with simpler approaches like group means.

The models presented in this work show promise in their ability to decrease morbidity rates resulting from septic shock and improve the outcome for sepsis patients. Future work will focus more on directly incorporating time series analyses into the current system to further help in predicting the onset of sepsis and septic shock.

6. REFERENCES

[1] R. C. Bone, R. A. Balk, F. B. Cerra, R. P. Dellinger, A. M. Fein, W. A. Knaus, R. M. Schein, and W. J. Sibbald. Definitions for sepsis and organ failure and guidelines for the use of innovative therapies in sepsis. The ACCP/SCCM Consensus Conference Committee. American College of Chest Physicians/Society of

Critical Care Medicine. *Chest*, 101(6):1644–1655, 1992.

[2] A. S. Fialho, F. Cismondi, S. M. Vieira, J. a. M. C. Sousa, S. R. Reti, M. D. Howell, and S. N. Finkelstein. Predicting Outcomes of Septic Shock Patients Using Feature Selection Based on Soft Computing Techniques. In *Information Processing and Management of Uncertainty in Knowledge-Based Systems. Applications*, volume 81 of *Communications in Computer and Information Science*, pages 65–74. Springer Berlin Heidelberg, 2010.

[3] F. Gwadry-Sridhar, B. Lewden, S. Mequanint, and M. Bauer. Comparison of Analytic Approaches for Determining Variables - A Case Study in Predicting the Likelihood of Sepsis. In *HEALTHINF 2009*, pages 90–96, 2009.

[4] R. A. Lukaszewski, A. M. Yates, M. C. Jackson, K. Swingler, J. M. Scherer, A. J. Simpson, P. Sadler, P. McQuillan, R. W. Titball, T. J. G. Brooks, and M. J. Pearce. Presymptomatic prediction of sepsis in intensive care unit patients. *Clinical and Vaccine Immunology*, 15(7):1089–1094, 2008.

[5] S. Oba, M.-a. Sato, I. Takemasa, M. Monden, K.-i. Matsubara, and S. Ishii. A Bayesian missing value estimation method for gene expression profile data. *Bioinformatics*, 19(16):2088–2096, 2003.

[6] J. Paetz. Knowledge-based approach to septic shock patient data using a neural network with trapezoidal activation functions. *Artificial Intelligence in Medicine*, 28(2):207–230, 2003.

[7] R. Pereira, R. Almeida, U. Kaymak, S. Vieira, J. Sousa, S. Reti, M. Howell, and S. Finkelstein. Predicting septic shock outcomes in a database with missing data using fuzzy modeling: Influence of pre-processing techniques on real-world data-based classification. In *2011 IEEE International Conference on Fuzzy Systems (FUZZ)*, pages 2507–2512, 2011.

[8] E. Rivers, B. Nguyen, S. Havstad, J. Ressler, A. Muzzin, B. Knoblich, E. Peterson, and M. Tomlanovich. Early Goal-Directed Therapy in the Treatment of Severe Sepsis and Septic Shock. *New England Journal of Medicine*, 345(19):1368–1377, 2001.

[9] S. Roweis. EM Algorithms for PCA and SPCA. In *Advances in Neural Information Processing Systems*, pages 626–632, 1998.

[10] M. Saeed, M. Villarroel, A. T. Reisner, G. Clifford, L.-W. 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*, 39(5):952–960, 2011.

[11] D. Shavdia. *Septic shock: providing early warnings through multivariate logistic regression models.* Harvard-MIT Division of Health Sciences and Technology, 2007.

[12] W. Stacklies, H. Redestig, M. Scholz, D. Walther, and J. Selbig. pcaMethods–a bioconductor package providing PCA methods for incomplete data. *Bioinformatics*, 23(9):1164–1167, 2007.

[13] C. H. H. Tang, P. M. Middleton, A. V. Savkin, G. S. H. Chan, S. Bishop, and N. H. Lovell. Non-invasive classification of severe sepsis and systemic inflammatory response syndrome using a nonlinear support vector machine: a preliminary study. *Physiological Measurement*, 31(6):775–793, 2010.

[14] T. M. Therneau and B. A. R. port by Brian Ripley. Note that maintainers are not available to give advice on using a package they did not author. *rpart: Recursive Partitioning*, 2012. R package version 3.1-51.

[15] S. W. Thiel, J. M. Rosini, W. Shannon, J. A. Doherty, S. T. Micek, and M. H. Kollef. Early prediction of septic shock in hospitalized patients. *Journal of Hospital Medicine*, 5(1):19–25, 2010.

[16] O. G. Troyanskaya, M. Cantor, G. Sherlock, P. O. Brown, T. Hastie, R. Tibshirani, D. Botstein, and R. B. Altman. Missing value estimation methods for DNA microarrays. *Bioinformatics*, 17(6):520–525, 2001.