Septic Shock Prediction for Patients with Missing Data

JOYCE C. HO, CHENG H. LEE, and JOYDEEP GHOSH, University of Texas at Austin

Sepsis and septic shock are common and potentially fatal conditions that often occur in intensive care unit (ICU) patients. Early prediction of patients at risk for septic shock is therefore crucial to minimizing the effects of these complications. Potential indications for septic shock risk span a wide range of measurements, including physiological data gathered at different temporal resolutions and gene expression levels, leading to a nontrivial prediction problem. Previous works on septic shock prediction have used small, carefully curated datasets or clinical measurements that may not be available for many ICU patients. The recent availability of a large, rich ICU dataset called MIMIC-II has provided the opportunity for more extensive modeling of this problem. However, such a large clinical dataset inevitably contains a substantial amount of missing data. We investigate how different imputation selection criteria and methods can overcome the missing data problem. Our results show that imputation methods in conjunction with predictive modeling can lead to accurate septic shock prediction, even if the features are restricted primarily to noninvasive measurements. Our models provide a generalized approach for predicting septic shock in any ICU patient.

Categories and Subject Descriptors: J.3 [**Computer Application**]: Life and Medical Sciences; I.2.1 [**Artificial Intelligence**]: Applications and Expert Systems; I.2.6 [**Artificial Intelligence**]: Learning

General Terms: Algorithms, Performance

Additional Key Words and Phrases: Missing data, imputation methods, sepsis, data mining

ACM Reference Format:
Ho, J. C., Lee, C. H., and Ghosh, J. 2014. Septic shock prediction for patients with missing data. *ACM Trans. Manage. Inf. Syst.* 5, 1, Article 1 (April 2014), 15 pages.
DOI:http://dx.doi.org/10.1145/2591676

1. INTRODUCTION

Sepsis is a systemic response to infection that is a common and life-threatening in-hospital complication, causing more deaths than prostate cancer, breast cancer, and HIV/AIDS combined [WSD Coalition 2012]. Despite modern medical advancements, the number of sepsis cases has doubled over the last 10 years, with an estimated $14.6 billion spent on hospitalizations for sepsis in 2008 alone [Hall et al. 2011]. Sepsis is one of the leading causes of mortality in intensive care unit (ICU) patients [Lukaszewski et al. 2008]. Severe cases of sepsis often lead to septic shock, a condition characterized by hypotension (low blood pressure) despite treatment that dramatically increases mortality risk [Bone et al. 1992]. Early intervention and therapy have been shown to improve the outcome of patients with severe sepsis and septic shock [Kumar et al. 2006; Nguyen et al. 2007; Rivers et al. 2001], thus making accurate identification of patients at risk for developing these conditions crucial to improving standards of clinical care.

Development of highly accurate predictive models for medical applications is often complicated by the nature of clinical data, which are typically noisy and inconsistently gathered. For example, while physiological variables such as heart rate may

This work is supported by the Schlumberger Centennial Chair in Engineering and Army Research Office under grant no. W911NF-11-1-0258.

Authors' address: J. C. Ho (corresponding author), C. H. Lee, and J. Ghosh, University of Texas at Austin, 1 University Station, C0803 Austin, TX 78712; email: joyceho@utexas.edu.

DOI:http://dx.doi.org/10.1145/2591676

be electronically monitored, these measurements must often be manually recorded in a patient's chart by a healthcare provider, potentially leading to erroneous or irregularly sampled data. Furthermore, highly accurate measurements for some physiological variables require invasive techniques that would place patients at unnecessary risk (e.g., accurate blood pressure measurements require arterial catheterization), and therefore cannot be ethically gathered. In such cases, the only available data are less accurate measurements obtained by noninvasive means.

As a result, clinical studies must often deal with missing data. A commonly used solution for 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. Furthermore, the reduction may hinder the development of accurate models and only generalize to a small population. Previous works on predicting the onset of septic shock have generally avoided the missing data problem by restricting themselves to very modest, highly curated datasets with small number of patients and limited sets of features.

To build a predictive model for septic shock onset generalizable to larger groups of ICU patients, we make use of the MIMIC-II database [Saeed et al. 2011], one of the largest publicly available clinical datasets, with data for >30,000 patients and >40,000 ICU admissions. As with any large database, missing data are a pervasive problem. To fully utilize these data, our work investigates the role and impact of imputation methods while building predictive models for septic shock. We limit our features to commonly observed, mostly noninvasive clinical measurements that are continually monitored across the entire patient population. We demonstrate that imputation methods allow us to build better predictive models for septic shock risk that are generalizable to broader groups of ICU patients and allow for earlier diagnosis and intervention for at-risk patients.

The main contributions of our work are as follows.

- We use simple and accessible approaches to handle patients with partially missing observations.
- We utilize noisy and/or intermittently gathered noninvasive measurements as proxies for their invasive and potentially risky counterparts.
- We develop a model that can identify high-risk septic shock patients for additional monitoring using invasively gathered techniques.
- We introduce a customizable performance-oriented imputation (POI) algorithm to optimize performance objectives beyond traditional metrics such as AUC.

2. BACKGROUND AND RELATED WORK

This work uses the definitions for sepsis and septic shock established during the 1991 American College of Chest Physicians/Society of Critical Care Medicine Consensus Conference.

Definition 2.1 (*Sepsis*). 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 nonspecific systemic inflammatory response syndrome (SIRS) abnormalities: (1) abnormal body temperature, (2) increased heart rate, (3) increased respiratory rate, or (4) abnormal white blood cell counts [Bone et al. 1992].

Definition 2.2 (*Septic Shock*). Septic shock is "sepsis-induced hypotension, persisting despite adequate fluid resuscitation, along with the presence of hypoperfusion abnormalities or organ dysfunction" [Bone et al. 1992]. Septic shock onset occurs when a patient has systolic blood pressure <90 mmHg and >600 mL of fluid input over the last hour [Bone et al. 1992; Shavdia 2007].

Prior research has focused largely on predicting the outcome of sepsis patients. A study conducted to assess the accuracy of mortality prediction systems on ICU patients with severe sepsis and septic shock showed that several of these systems had adequate accuracy but poor calibration [Arabi et al. 2003]. Fuzzy models and knowledge-based neural networks were used to predict the outcome of 121 patients with abdominal septic shock [Fialho et al. 2010]. Lagu et al. introduced a multilevel mixed-effects logistic regression model using patient demographics, presence of comorbidities, treatment, and ICU admission status as features to predict sepsis mortality [Lagu et al. 2011]. Another paper proposed the use of a logistic regression using extracted latent factors to predict mortality in severe sepsis patients [Ribas et al. 2012].

More recent work has concentrated on early prediction of septic shock. A septic shock early warning system (EWS) was developed using multivariate logistic regression with commonly measured clinical variables as features to predict septic shock one hour prior to onset [Shavdia 2007]. The EWS used data from 250 sepsis patients, 65 of whom developed septic shock, and achieved an area under the receiver operating characteristic curve (AUC) of 0.928. However, the model used invasively gathered features, such as central venous pressure and five laboratory results, data that may not be commonly available in ICU patients. Another study used a rule-based approach to notify clinicians of patients requiring specific treatment and monitoring [Nguyen et al. 2007]. Thiel et al. developed a predictive model using a Recursive Partitioning and Regression Tree (RPART) to identify early predictions from clinical data of 1,864 hospitalized non-ICU septic patients [Thiel et al. 2010]. The model used eleven routine laboratory tests and certain vital signs but only correctly identified 55% of septic shock patients. However, these models all failed to address the missing data problem.

Some models have been developed to predict septic shock in the absence of full feature data. Paetz inserted randomly sampled data from a suitable normal distribution to deal with incomplete data and prevent the model from erroneous learning via missing values [Paetz 2003]. The outcome was then predicted using a trapezoidal function neural network to classify 874 patients. The model required at least 10 of the 12 variables to be present and could only correctly classify $\sim$70% of the test data with a sensitivity of 15.01%. Another study proposed the use of a modified Fuzzy C-Means algorithm with Partial Distance Strategy (FCM-PDS) that does not require any imputation of the missing values by means of product-space clustering [Pereira et al. 2011]. The authors also suggested the combination of Zero-Order-Hold (ZOH), which holds the measurement value until a new observation is available, to deal with incomplete data and FCM-PDS to predict abdominal septic shock. Although the model performance obtained an AUC of 0.899 on 121 patients, the features rely heavily on laboratory results and other invasive measurements. Additionally, it is uncertain whether this model is predictive for all septic shock patients or is limited to only those patients with abdominal sepsis.

Various imputation methods have been developed to address the missing value problem in medical settings. One study conducted a comparison of statistical and machine learning techniques on predicting breast cancer [Jerez et al. 2010]. The results showed that machine learning algorithms—in particular, multilayer perceptrons, self-organization maps, and k-nearest neighbors (KNN)—generally outperformed statistical techniques such as hot deck, mean, and multiple imputation. Another paper proposed the combination of a multiple imputation approach based on fuzzy clustering and an ensemble of weak classifiers trained on random subspaces that performed well across a host of medical problems [Nanni et al. 2012].

We have recently shown that combining matrix-factorization-based imputation techniques and standard classification models can be used to predict septic shock in ICU patients [Ho et al. 2012]. This article is an extension of that work and focuses on further minimizing the use of invasive clinical measurements and introducing prediction-aware imputation selection criteria.

3. MISSING DATA IMPUTATION

Traditional approaches decouple predictive modeling with incomplete data into two stages; first, missing data are imputed, and second, predictive models are built using the imputed data. We use three simple and accessible approaches to estimate the missing observations.

(1) *Mean (or median) imputation* uses either the mean (or median) value of nonmissing observations to fill in missing values. This method is surprisingly effective on recommendation systems.

(2) *Matrix-factorization-based techniques* for missing values use a specific feature matrix decomposition. We used three estimators provided by the BioConductor `pcaMethods` package [Stacklies et al. 2007].

- Singular-value-based decomposition imputation (SVDImpute) uses a linear combination of k-eigenvalues to predict the missing value [Troyanskaya et al. 2001].
- Probabilistic Principal Component Analysis (PPCA) combines an Expectation Maximization (EM) approach to Principal Component Analysis (PCA) with a probabilistic model [Roweis 1998].

(3) *Neighborhood-based imputation* finds the k-nearest neighbors (KNN) with nonmissing observations and uses their averages to impute missing values [Hastie et al. 2012].

The methods listed previously span the “global to local” imputation space. Mean imputation is the most global approach available where a patient’s missing observation is influenced by measurements from all of the other patients in the population. Neighborhood-based imputation is on the opposite end of the spectrum, using local information (small subset of the patient population) to determine the missing value. Matrix factorization methods can be viewed as a combination of the two approaches, imposing a global structure where the individual matrix values are then influenced by a smaller dimensional space. Matrix factorization and neighborhood-based imputation methods have a parameter k which controls the resolution or locality of the imputation.

Imputation methods are primarily evaluated by randomly removing observations and comparing the values of imputed versus original using a difference metric such as root mean squared error (RMSE) or mean absolute error (MAE). However, these evaluation measures may not always be appropriate. A recent study of recommendation systems, such as Amazon and Netflix, showed that improvements in RMSE do not necessarily translate to the task of recommending the top-N items and proposed alternative metrics for these systems [Cremonesi et al. 2010]. Similarly, we expect imputation methods selected using RMSE or MAE can be suboptimal for clinical predictive models that are evaluated by other metrics. Thus, we propose a performance-oriented imputation (POI) algorithm which selects the imputation method via the performance metric (e.g., F-score and lift) used to evaluate the predictive model.

The POI algorithm finds the optimal number, k, for SVDImpute, PPCA, and KNN dependent on the selection criterion, sc. The algorithm creates J copies of the data, randomly removing some percentage of the data. For each copy, grid search over a predefined set K is performed to evaluate the imputed feature matrix according to *sc*. POI selects the *k* that yields the optimal evaluation measure. The general POI framework is outlined in Algorithm 1.

ALGORITHM 1: POI algorithm for matrix-factorization or clustering based approaches.

```
Input: Data set, imputation method, selection criterion, and $J$
Output: Optimal $k$
Randomly create J train/test splits for the data set for $k$ in $K$ do
    for $j$ in 1:$J$ do
        $Data_j$ = Remove x% of the data
        $Impute_{Data}$ = Impute($Data_j$, $k$)
        Split $Impute_{Data}$ into I folds
        for $i$ in 1:$I$ do
            Separate into $Train_i$ and $Test_i$
            $Model$ = BuildModel($Train_i$)
            $Error_{ij}$ = EvaluateSelectionCriteria($Model$, $Test_i$, $s$)
        end
        $Evalue_j$ = Mean($Error_{1j}$, ..., $Error_{Ij}$) - Standard Deviation($Error_{1j}$, ..., $Error_{Ij}$)
    end
    $Value_k$ = Mean($Evalue_1$, ..., $Evalue_J$) - Standard Devation($Evalue_1$, ..., $Evalue_J$)
end
Select $k$ = Max($Value_1$, ..., $Value_K$) or Min($Value_1$, ..., $Value_K$)
```

4. EXPERIMENTS

4.1. Data

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 [Saeed et al. 2011]. 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 study was conducted on septic adults ($>$18 years of age at time of admission) from the MIMIC-II database with a single ICU admission over the course of their hospital stay. Patients younger than 18 years of age were omitted to avoid: (1) confounding factors arising from different physiologies and (2) complications associated with obtaining informed consent from such parents. To ensure sufficient data, we only used patients with at least ten observations of blood pressure (BP) taken noninvasively, heart rate (HR), respiratory rate (RR), blood oxygen saturation (SpO_2), temperature (TEMP), and two observations of white blood cell count (WBC) during their ICU stay. Potential septic patients were identified based on their ICD-9 codings ("995.91" or "995.92"). Sepsis patients must have at least one interval that met the SIRS criteria listed in Definition 2.1.

Septic shock patients were identified through their ICD-9 code ("785.52"). Time of septic shock onset was determined using modified criteria based on Definition 2.2. Any time point where the systolic blood pressure (SBP) was $<$90 mmHg and within a SIRS interval was marked as a hypotension observation. Consecutive hypotensive observations were aggregated together to define a hypotension region. Total fluid intake was calculated starting one hour prior to the hypotension observation to halfway

Table I. Features for Predicting Septic Shock

Last measurement, min, mean, and max of observations in the last 8 hours
Cardiac: non-invasive systolic and diastolic blood pressure, heart rate, pulse pressure
Respiratory: respiratory rate, SpO_2
Other: temperature
Last measurement only
White blood cell count
SOFA
SAPS-I
Shock index

through the hypotension region [Shavdia 2007]. Shock onset was defined as the start time of any hypotension region with total fluid intake >600 mL.

From 27,542 adult ICU stays in the MIMIC-II database, 2,175 had the sepsis diagnosis code and 1,027 also had the septic shock ICD-9 code. However, we could only identify at least one SIRS interval for 1,353 of the potential septic patients. Moreover, we were unable to detect the time of shock onset for a significant portion of the patients with the septic shock ICD-9 code. Thus, only 213 of the potential septic shock patients were used as positive cases in the study, with the remaining 814 adults omitted from the study. Our study was conducted on 1,353 septic patients where 213 ($\sim$15.7%) transitioned to septic shock.

4.2. Septic Shock Prediction Features

As explained before, we focus on using common, mostly noninvasive measurements as features for predicting septic shock onset. Basic patient information in the form of demographic data (gender and ages at hospital and ICU admission), medical history (flags indicating previous hospital and ICU admissions), and ICU care unit are included. Physiological and laboratory features are chosen based on prior studies, invasiveness, and measurement frequency. White blood cell count is the only invasive clinical feature and can be used to diagnose septic patients (Definition 2.1). The feature set includes two derived features: (1) pulse pressure, calculated as the difference between systolic and diastolic blood pressure, and (2) shock index [Allgöwer and Burri 1967], based on heart rate and systolic blood pressure; and two daily mortality scores: (1) SOFA [Ferreira et al. 2001] and (2) SAPS-I [Le Gall et al. 1984]. Unlike previous work that focused on using the last measurement, difference between last several measurements, and mean of the last several measurements [Shavdia 2007; Ho et al. 2012], our current model uses summary statistics (minimum, mean, and maximum) in addition to the last recorded measurement from the last 12 hours before onset for physiological variables. The inclusion of summary statistics allows the model to capture variability in a patient's physiological state [Kennedy and Turley 2011]. Table I lists the features used to predict septic shock in ICU patients.

4.3. Evaluation

A feature matrix of physiological and laboratory values was generated for data available at reference time, defined as a specified time prior to shock onset. SIRS patients that did not transition to septic shock were assigned a random evaluation time during the first SIRS interval. Each feature matrix contains the basic patient information and the list of features shown in Table I. All features, except shock index, were normalized to the $[0, 1]$ range following the procedure outlined for the septic shock early warning system (EWS) [Shavdia 2007]. Feature matrices were created for reference times 30, 60, 90, 120, and 180 minutes prior to shock onset or evaluation time.

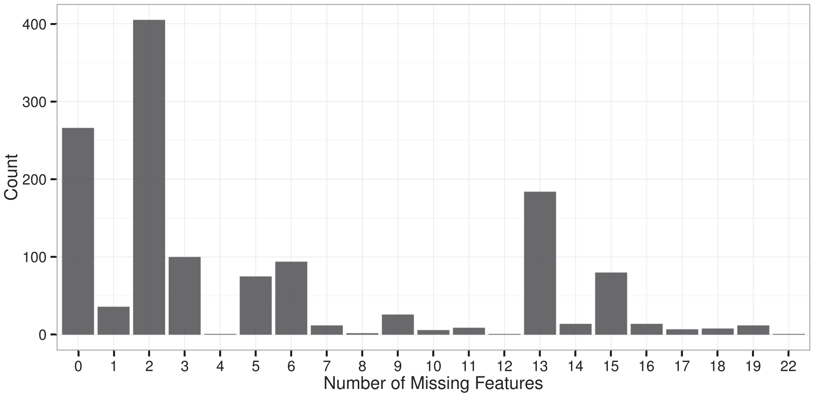


Fig. 1. Histogram of the number of missing features for each patient at reference time t = 30 minutes.

Table II. The Availability of Clinical Measurements

	% Missing				
	Reference time				
Feature	30 mins	60 mins	90 mins	120 mins	180 mins
Respiratory rate	0.67	0.68	0.71	0.64	0.67
SpO_2	0.81	0.68	0.55	0.64	0.42
Temperature	1.70	2.05	1.80	1.77	1.85
White Blood Cells	15.30	14.69	14.67	14.16	14.61
Systolic/Diastolic	23.28	23.44	23.69	23.73	24.01
SOFA	60.24	59.74	59.22	58.81	58.10
SAPS-I	63.05	62.48	62.04	61.54	60.96

Patients with >40% missing features were omitted from the study. Figure 1 demonstrates the importance of missing value imputation due to the lack of consistent sampling of clinical data. The problem is heightened by restricting the physiological measurements to the last 12 hours. At reference time t = 30 minutes, only 176 of the 1,353 patients have complete data. Table II shows the percentage of missing observations per feature.

Multivariate logistic regression, linear kernel support vector machine, and decision trees (RPART) were used to predict septic shock on the feature matrices. These models were selected based on their usage in previous septic shock work. Variable selection for logistic regression was performed using three different methods:

(1) Lasso (least absolute shrinkage and selection operator) [Hastie et al. 2008];
(2) Ridge regression [Hastie et al. 2008];
(3) Elastic net with different penalties on ℓ_1 and ℓ_2 [Friedman et al. 2010].

The predictive performance of the various models was evaluated using 10 stratified bootstrap samples, with 60% of the data used for training. The evaluation metrics selected for our study are: (1) AUC, (2) lift, (3) F_1, or F-score, (4) F_2 which emphasizes recall higher than precision, and (5) $F_{0.5}$ which gives higher weight to precision over recall. Note that for lift and the F-measures, a windowed average is used to increase stability of the metric instead of a single, potentially noisy evaluation point. The selection criteria for k, the optimal number of principal components for SVDImpute, and PPCA or the number of nearest neighbors in kNN, was: (1) MAE, (2) RMSE, (3) AUC, (4) F_1, (5) F_2, (6) $F_{0.5}$, and (7) lift. For both mean and median imputation, the conditional mean or median value based on the patient's age and gender group was used.

Table III. Wilcoxon-Mann Whitney Test Results for Determining if the Complete and Missing Population Distributions are Different

	Missing		Complete		
Reference Time	Sepsis only	Shock	Sepsis only	Shock	P-value
30	749	79	199	110	4.56e-26
60	723	79	196	106	6.99e-24
90	705	79	196	103	4.63e-22
120	685	74	193	103	7.06e-23

Table IV. AUC Comparison with the Septic Shock EWS Model 60 Minutes Prior to Shock Onset

Feature Selection	EWS Features	POI for Septic Shock Features
All	0.722±0.060	0.789±0.074
Forward stepwise	0.780±0.078	0.796±0.091
Backward stepwise	0.752±0.062	0.772±0.090

Our feature set uses a smaller and simpler set of measurements, thus making it applicable to a broader patient population (see text for details).

4.4. Feature Set Evaluation

The first series of experiments first compared the POI features with the septic shock early warning system (EWS) [Shavdia 2007] feature set second, verified the need for imputation even with common clinical measurements, and lastly motivated the use of noninvasive features to cater to a broader patient population.

The EWS feature matrices [Shavdia 2007] were replicated to compare the distribution of sepsis patients progressing to septic shock. The Wilcoxon-Mann Whitney test was performed to test the null hypothesis that the prevalence of shock in the complete dataset is stochastically greater than in the patients with at least one missing feature dataset. Table III summarizes the results for the reference times and shows the difference in prevalence is statistically significant. Thus, imputation helps us generalize to a larger and more representative ICU patient population.

We evaluated our septic shock features against the replica EWS features. For comparison purposes, patient demographics were excluded from our feature set and arterial blood pressure measurements were used instead of their noninvasive counterparts. Additionally, EWS features utilized three measurements of arterial pH while our feature set contained a measurement each of SAPS-I and SOFA scores. Any patient without complete data for either EWS or our modified features was omitted. The dataset was reduced to 149 sepsis patients, of which 86 transitioned to septic shock. Three logistic regression models were trained for each set of features: (1) all the features, (2) forward stepwise regression, and (3) backward stepwise regression.

Table IV summarizes the mean AUC and the 95% confidence interval for predicting septic shock 60 minutes before onset. The results demonstrate a noticeable performance degradation from the reported 0.928 to 0.780. This can be attributed to differences in the two patient cohort studies, as the original EWS dataset consisted only of 110 patients. Our septic shock features, which use a smaller and simpler set of measurements (omitting arterial pH), had a higher mean AUC across all three selection types. However, the improvement is not statistically significant, as the EWS performance is within the confidence range.

Table V illustrates the benefit and pitfall of substituting noninvasive blood pressure measurements for the arterial blood pressure measurements. Although more patients have at least one noninvasive measurement, the average time between measurements increases by a factor of 1.5 and results in less total blood pressure measurements.

Table V. Patient Statistics for the Two Types of Blood Pressure Measurements

Measurement	Patients (#)	Avg. Interval (mins)	Avg. measurements (#)
Arterial	682	45.79±28.27	60.96±87.00
Non-invasive	1323	56.14±72.28	25.18±40.85

Table VI. AUC Comparison of Forward Selection Logistic Regression Model Trained on Complete versus Mean Imputed Data

		Reference time t before onset			
Train	Test	30 min	60 min	90 min	120 min
Complete	Complete	0.796±0.065	0.777±0.050	0.763±0.034	0.731±0.073
Complete	Mean Impute	0.815±0.033	0.800±0.053	0.803±0.036	0.786±0.067
Mean Impute	Mean Impute	0.834±0.025	0.829±0.030	0.834±0.023	0.801±0.051
Mean Impute	Complete	0.839±0.044	0.828±0.047	0.809±0.033	0.783±0.062

Table VII. AUC Comparison of Different Models at t = 120 min

		Model Type				
Impute	Selection	ℓ_1 LogR	Elastic-net ($\alpha = \frac{1}{2}$)	ℓ_2 LogR	RPART	SVM
Mean	None	0.739±0.031	0.744±0.031	0.759±0.020	0.689±0.052	0.744±0.032
Median	None	0.746±0.036	0.754±0.034	0.770±0.019	0.690±0.054	0.742±0.031
SVD	POI	0.723±0.056	0.737±0.040	0.704±0.040	0.634±0.090	0.697±0.054
SVD	MAE	0.741±0.028	0.749±0.024	0.765±0.021	0.716±0.045	0.747±0.024
SVD	RMSE	0.650±0.037	0.657±0.034	0.695±0.035	0.635±0.042	0.649±0.038
PPCA	POI	0.698±0.040	0.687±0.040	0.672±0.064	0.585±0.092	0.650±0.072
PPCA	MAE	0.706±0.072	0.717±0.073	0.742±0.060	0.645±0.063	0.695±0.065
PPCA	RMSE	0.701±0.069	0.713±0.073	0.734±0.061	0.641±0.059	0.685±0.064
KNN	POI	0.733±0.033	0.746±0.029	0.758±0.022	0.673±0.037	0.733±0.031
KNN	MAE	0.735±0.033	0.747±0.029	0.759±0.022	0.671±0.041	0.744±0.026
KNN	RMSE	0.735±0.033	0.746±0.029	0.759±0.022	0.671±0.040	0.744±0.026

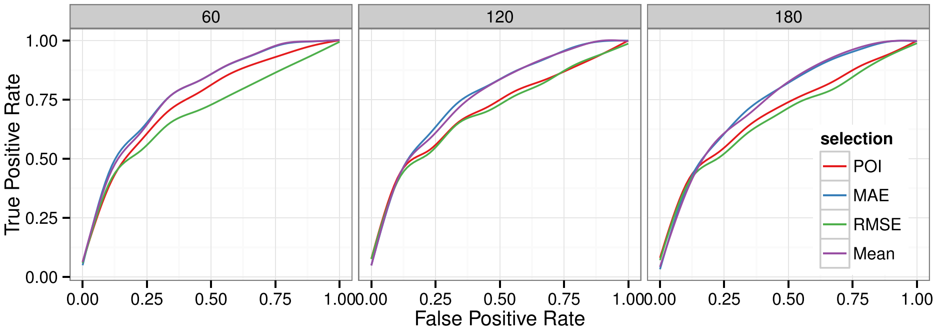
Additionally, invasive blood pressure measurement requires arterial catheterization, a procedure that places patients at risk for severe complications and is therefore not medically necessary in all cases. To address the larger ICU population, our feature set will only use noninvasive blood pressure measurements.

One approach for predicting septic shock on all patients is to train on patients with complete data and use mean imputation to perform prediction for patients with missing observations. Table VI illustrates the potential benefit of training on imputed data. Models trained on the mean imputed data generally outperform the models trained only on complete data. An additional benefit of imputation is lower AUC variable compared to the model trained and tested only on complete data. Furthermore, the results suggest that imputation does not bias the models and can be used to generalize to a broader population.

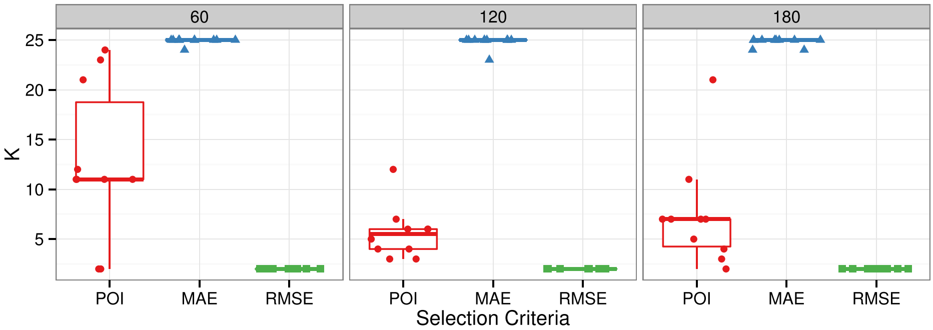
4.5. Results on Noninvasive POI

The remaining studies use only noninvasive features listed in Table I. Thus, the models are applicable to the broader ICU patient population.

Table VII summarizes the imputation effect on AUC performance across the different models at reference time $t = 120$ mins. Mean and median imputation generally outperform the matrix-factorization- and neighborhood-based imputation methods. However, SVD imputation using the MAE selection criteria achieves the best AUC performance for the decision trees and SVM models. The results also show that KNN has similar performance to mean and median imputation.



(a) average ROC curves



(b) box plot of the optimal k

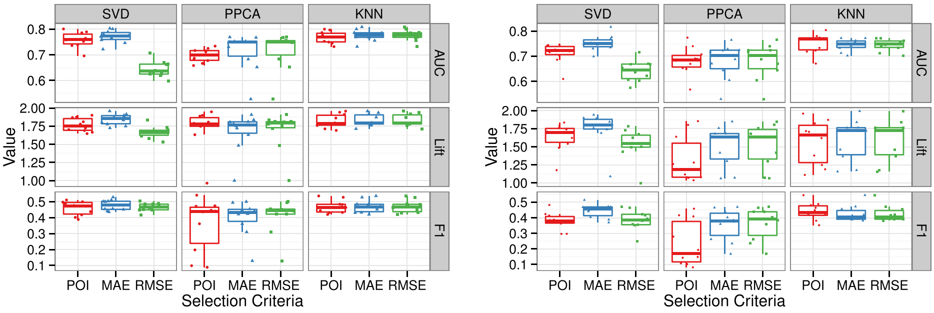
Fig. 2. ℓ_2 regularized logistic regression model results trained on SVD imputed data at the various reference times t.

Figure 2 summarizes the average ROC curves and the optimal k for the ℓ_2 regularized models trained on SVD imputed data. The mean imputation ROC curve is also included in Figure 2(a) for comparative purposes. Mean imputation and the MAE selection criteria consistently outperform both the RMSE and POI selection criteria. Note that MAE is marginally better than mean imputation for false positive rates below 0.50. The optimal k plot, Figure 2(b), helps explain some of the performance differences between MAE, RMSE, and POI. MAE repeatedly chooses the largest k while RMSE selects a small k. The large variability of k for the POI selection criteria at $t = 60$ results in an ROC curve halfway between MAE and RMSE.

A septic shock alert system focuses on reducing the false positives. Thus, the $F_{0.5}$ measure may be a more applicable performance metric. Table VIII summarizes the results across four of the predictive models three hours (180 minutes) prior to shock onset. These results demonstrate the potential of our POI algorithm. SVD imputation using the POI selection criterion achieves better predictive perfomance for ℓ_1 and ℓ_2 regularized logistic regression, decision tree, and SVM. For the ℓ_2 regularized logistic regression model, the POI selection criterion performs the best compared to MAE, RMSE, and mean imputation. However, the POI selection criterion does not always result in the best performance as illustrated by the PPCA imputed models. Except for the ℓ_2 regularized logistic regression model, the $F_{0.5}$ values are significantly below the mean imputation and PPCA imputation using MAE and RMSE. This suggests that

Table VIII. $F_{0.5}$ Measure Comparison of Different Models at t = 180 min

		Model Type			
Impute	Selection	ℓ_1 LogR	ℓ_2 LogR	RPART	SVM
Mean	None	0.397±0.028	0.351±0.040	0.253±0.072	0.379±0.037
SVD	POI	0.403±0.028	0.373±0.047	0.291±0.107	0.383±0.034
SVD	MAE	0.402±0.025	0.357±0.033	0.280±0.084	0.389±0.038
SVD	RMSE	0.388±0.033	0.368±0.045	0.343±0.091	0.360±0.028
PPCA	POI	0.265±0.140	0.363±0.040	0.233±0.078	0.235±0.110
PPCA	MAE	0.383±0.041	0.347±0.029	0.195±0.086	0.337±0.031
PPCA	RMSE	0.384±0.041	0.350±0.028	0.188±0.084	0.342±0.031
KNN	POI	0.376±0.032	0.353±0.038	0.237±0.071	0.380±0.041
KNN	MAE	0.380±0.032	0.354±0.040	0.255±0.053	0.383±0.037
KNN	RMSE	0.375±0.037	0.353±0.042	0.244±0.051	0.381±0.042



(a) ℓ_1 regularized logistic regression

(b) SVM

Fig. 3. Predictive performance across three imputation methods using POI, RMSE, and MAE selection criteria at reference time t = 60.

during the selection of the optimal k, the models may be overtrained and thus do not generalize well to the test data.

Figure 3 shows the results for two predictive models, the ℓ_1 regularized logistic regression and SVM, evaluated using AUC, lift, and F-score. MAE generally has the best performance across both models, but POI outperforms MAE and RMSE on the three measures for logistic regression models trained on PPCA imputed data and on AUC and F-score for SVM trained on KNN imputed data. These results suggest no single imputation method and selection criterion consistently yields the best performance.

Table IX illustrates the effect of the various selection criteria on feature ranking pertaining to the physiological measurements. PPCA imputation is used in conjunction with an elastic-net regularized logistic regression model. The most recent systolic blood pressure is the single most important feature across all the different models, as the magnitude of the coefficient associated with systolic blood pressure is the largest. Noticeable differences between the mean imputation and the PPCA imputations are the mean systolic blood pressure, mean respiratory rate, mean temperature, and the last diastolic blood pressure values. Even within the same imputation method (PPCA), selection criteria influences feature ranking. The last temperature reading is absent from the RMSE model and both MAE and RMSE place higher value on the shock index compared to the mean systolic blood pressure. Thus, the selection of k based on the selection criteria influences which measurements will be more important.

The selection of k for the elastic-net regularized logistic regression with $\alpha = \frac{1}{4}$ is shown in Figure 4. Both SVD and PPCA imputation are heavily impacted by the

Table IX. Mean Feature Ranking of the Physiological Measurements by Magnitude for an Elastic-Net ($\alpha = \frac{1}{4}$) Regularized Logistic Regression

Feature	Mean	AUC	Lift	F_1	MAE	RMSE
Systolic BP	1.50	1.70	1.80	1.70	2.70	2.40
Systolic BP (min)	4.00	3.00	3.40	3.50	2.56	2.89
SpO_2	2.22	3.00	3.00	3.22	3.22	2.56
Systolic BP (mean)	6.11	3.50	4.30	4.70	5.43	5.00
Shock Index	4.40	4.40	4.80	4.60	3.20	3.30
Temp	5.00	5.00	7.50	7.50	6.00	
Resp Rate (mean)	7.33	5.67	7.67	7.67	6.25	6.67
Temp (min)	5.17	6.00	6.40	6.33	7.50	8.00
Temp (mean)	9.33	7.00	6.50	6.50	11.50	6.00
Pulse Pressure (min)	9.25	7.00	7.00	7.00	9.50	8.00
$SpO_2$2 (mean)	8.67	7.67	7.00	7.00	6.67	7.00
Resp Rate	7.57	7.75	6.33	6.50	8.25	7.00
Diastolic BP	11.00	8.00	8.75	8.25	7.00	5.00
WBC	8.75	8.33	9.75	9.50	8.75	7.33
SOFA	9.50	9.00	9.20	9.60	9.00	7.80
$SpO_2$2 (min)	9.38	9.50	9.33	9.33	7.40	7.00
SAPS-I	11.50	12.00	11.40	11.40	10.44	8.25

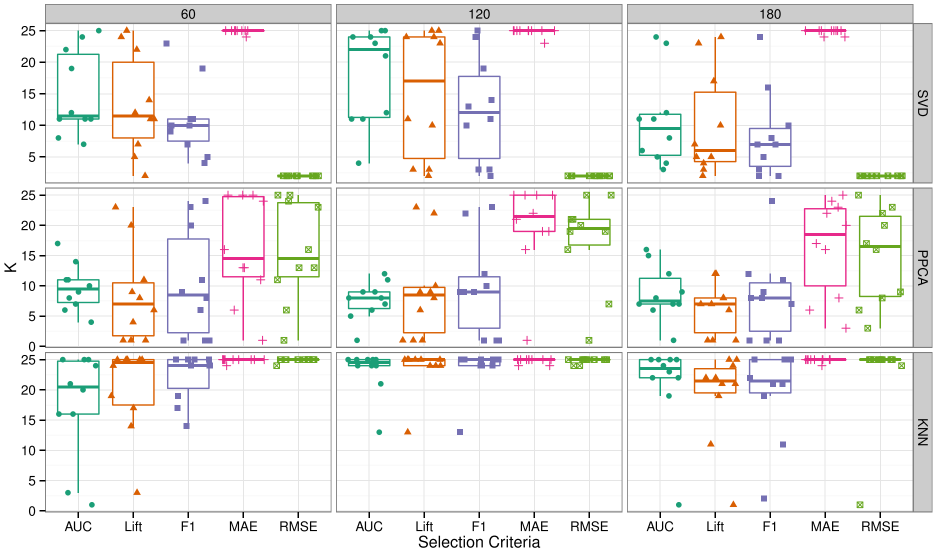


Fig. 4. Box plot of the optimal k for the different selection criteria and imputation methods using elastic net with $\alpha = \frac{1}{4}$.

selection criteria. For SVD, MAE and RMSE favor the two extremes and have less variability, while the POI selection criterion tends to span the entire search space. The pattern does not hold for PPCA as the POI selection criterion consistently results in smaller k. KNN imputation favors a larger k across all the criteria, thus the performance results are generally similar to mean imputation.

The focus on noninvasive measurements allows us to apply our models to a broader population of patients, especially those for whom arterial catheterization is not medically necessary. The models can be used as a screening process to identify patients

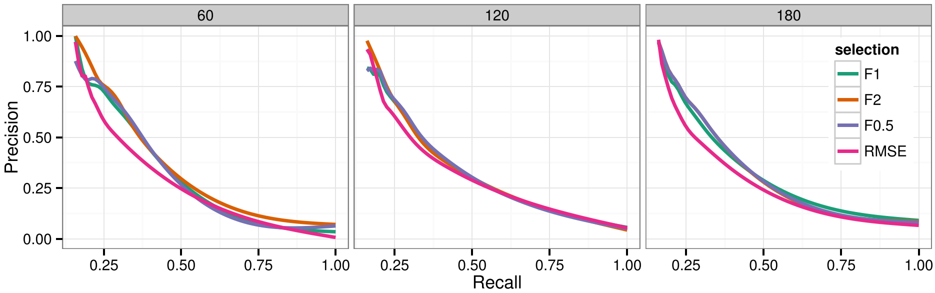


Fig. 5. Linear SVM model precision recall curves trained on SVD imputed data at various reference times t.

in need of closer monitoring. Figure 5 shows the precision recall plots comparing the three F-measures and standard RMSE selection criterion for the SVM trained on SVD imputed data. The results show that the F_2 selection criterion results in a higher precision for recall below 0.25. This results in fewer false positives while also identifying patients at high risk of developing septic shock.

4.6. Discussion

The experimental results demonstrate that imputation methods can help provide an accurate assessment of septic shock risk when features are restricted primarily to noninvasive clinical measurements and patients have partially missing observations. Simple mean imputation generally outperformed the matrix-factorization- and neighborhood-based imputation techniques. However, the POI algorithm achieved better results for some classification models and performance metrics.

The following factors may have impacted the performance of POI. First, the features were normalized following the procedure outlined for EWS [Shavdia 2007], favoring simple mean imputation. Second, the robustness of several of the performance metrics makes it easy to overfit the training data. Although windowed averaging was used to stabilize the lift and F-measures, additional model regularization may improve the POI algorithm. Finally, the small number of case patients for testing (85 patients) potentially resulted in high variance of performance metrics. Minor perturbations of the classification output can cause drastic changes for lift and the three F-measures. Further exploration of the POI algorithm on various datasets can help differentiate under what conditions simple imputation- (mean or median), traditional matrix-factorization-, and neighborhood-based approaches (MAE and RMSE), and POI should be used.

Our results show that our models can identify patients to be more closely monitored using systems, such as the septic shock early warning system, which rely on invasive measurements. However, there are open questions relating to the practicality and feasibility of our models. Where is the appropriate trade-off between precision and recall? What are acceptable detection rates in actual practice? Thus, one could design a study to determine the applicability of our models. However, any external evaluation of our models is beyond the current scope of this work.

Our current models use summary statistics to capture variability in a patient's physiological state. There is additional information embedded in the temporal patterns found in a patient's clinical measurements [Batal et al. 2012; Wang et al. 2012]. Furthermore, assigning a random evaluation time during the first SIRS interval for control patients is not practical from an application perspective. As future work, we can incorporate time-series models into our framework to improve the prediction of septic shock.

5. CONCLUSION

We presented a novel approach to accurately predict septic shock from noisy and/or intermittently gathered clinical data. The features we chose minimize the use of laboratory tests and invasive features. We also proposed alternate criteria for the imputation selection process by optimizing the predictive performance objective.

Our results show the importance of training classifier models on imputed data. Although the performance of our models does not outperform previous works, they can more readily handle predictions for patients with partially missing observations, a common scenario in most "real world" clinical settings. The results also show that alternate selection criteria of k can improve predictive performance. Missing data imputation allows us to apply the models to larger, noisier, and more incomplete datasets encountered in modern clinical studies.

Our septic shock prediction models can provide healthcare providers with patients in need of closer monitoring, decreasing the medical response time to an adverse event and improving their outcomes. Future work will focus on incorporating time-series models into the current framework to further help septic shock prediction.

ACKNOWLEDGMENTS

We would like to thank Yubin Park and Sanmi Koyejo for their comments.

REFERENCES

Allgöwer, M. and Burri, C. 1967. Schockindex. *Deutsche Medizinische Wodenschrif 46*, 1–10.

Arabi, Y. Y., Al Shirawi, N. N., Memish, Z. Z., Venkatesh, S. S., and Al-Shimemeri, A. A. 2003. Assessment of six mortality prediction models in patients admitted with severe sepsis and septic shock to the intensive care unit: A prospective cohort study. *Critical Care 7*, 5, R116–R122.

Batal, I., Fradkin, D., Harrison, J., Moerchen, F., and Hauskrecht, M. 2012. Mining recent temporal patterns for event detection in multivariate time series data. In *Proceeding of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'12).* ACM Press, New York, 280–288.

Bone, R. C., Balk, R. A., Cerra, F. B., Dellinger, R. P., Fein, A. M., Knaus, W. A., Schein, R. M., and Sibbald, W. J. 1992. 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.

Cremonesi, P., Koren, Y., and Turrin, R. 2010. Performance of recommender algorithms on top-n recommendation tasks. In *Proceedings of the 4th ACM Conference on Recommender Systems.*

Ferreira, F., Bota, D., Bross, A., Mélot, C., and Vincent, J. 2001. Serial evaluation of the SOFA score to predict outcome in critically ill patients. *J. Amer. Med. Assoc. 286*, 14, 1754.

Fialho, A. S., Cismondi, F., Vieira, S. M., Sousa, J. M. C., Reti, S. R., and Howell, M. D. 2010. Predicting outcomes of septic shock patients using feature selection based on soft computing techniques. In *Proceedings of the 13th Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems*. Springer, 65–74.

Friedman, J., Hastie, T., and Tibshirani, R. 2010. Regularization paths for generalized linear models via coordinate descent. *J. Statist. Softw. 33*, 1, 1–22.

Hall, M. J., Williams, S. N., DeFrances, C. J., and Golosinskiy, A. 2011. Inpatient care for septicemia or sepsis: A challenge for patients and hospitals. *NCHS Data Brief 62*, 1–8.

Hastie, T., Tibshirani, R., and Friedman, J. H. 2008. *The Elements of Statistical Learning. Data Mining, Inference, and Prediction*. Springer.

Hastie, T., Tibshirani, R., Narasimhan, B., and Chu, G. 2012. Impute: Imputation for microarray data. R package version 1.30.0.

Ho, J. C., Lee, C. H., and Ghosh, J. 2012. Imputation-enhanced prediction of septic shock in ICU patients. In *Proceedings of the ACM SIGKDD Workshop on Health Informatics (HI-KDD'12).*

Jerez, J. M., Molina, I., García-Laencina, P. J., Alba, E., Ribelles, N., Martín, M., and Franco, L. 2010. Missing data imputation using statistical and machine learning methods in a real breast cancer problem. *Artif. Intell. Med. 50,* 2, 11–11.

Kennedy, C. E. and Turley, J. P. 2011. Time series analysis as input for clinical predictive modeling: Modeling cardiac arrest in a pediatric ICU. *Theor. Biol. Med. Model. 8,* 40.

Kumar, A., Roberts, D., Wood, K. E., Light, B., Parrillo, J. E., Sharma, S., Suppes, R., Feinstein, D., Zanotti, S., Taiberg, L., Gurka, D., Kumar, A., and Cheang, M. 2006. Duration of hypotension before initiation of effective antimicrobial therapy is the critical determinant of survival in human septic shock. *Critical Care Med. 34*, 6, 1589–1596.

Lagu, T., Lindenauer, P. K., Rothberg, M. B., Nathanson, B. H., Pekow, P. S., Steingrub, J. S., and Higgins, T. L. 2011. Development and validation of a model that uses enhanced administrative data to predict mortality in patients with sepsis. *Critical Care Med. 39*, 11, 2425–2430.

Le Gall, J. R., Loirat, P., Alperovitch, A., Glaser, P., Granthil, C., Mathieu, D., Mercier, P., Thomas, R., and Villers, D. 1984. A simplified acute physiology score for ICU patients. *Critical Care Med. 12,* 11, 975–977.

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

Nanni, L., Lumini, A., and Brahnam, S. 2012. A classifier ensemble approach for the missing feature problem. *Artif. Intell. Med. 55*, 1.

Nguyen, H. B., Corbett, S. W., Steele, R., Banta, J., Clark, R. T., Hayes, S. R., Edwards, J., Cho, T. W., and Wittlake, W. A. 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 Med. 35*, 4, 1105–1112.

Paetz, J. 2003. Knowledge-based approach to septic shock patient data using a neural network with trapezoidal activation functions. *Artif. Intell. Med. 28*, 2, 207–230.

Pereira, R., Almeida, R., Kaymak, U., Vieira, S., Sousa, J., Reti, S., Howell, M., and Finkelstein, S. 2011. Predicting septic shock outcomes in a database with missing data using fuzzy modeling: Influence of preprocessing techniques on real-world data-based classification. In *Proceedings of the IEEE International Conference on Fuzzy Systems (FUZZ'11).* 2507–2512.

Ribas, V. J., Vellido, A., Ruiz-Rodríguez, J. C., and Rello, J. 2012. Severe sepsis mortality prediction with logistic regression over latent factors. *Expert Syst. Appl. Int. J. 39*, 2.

Rivers, E., Nguyen, B., Havstad, S., Ressler, J., Muzzin, A., Knoblich, B., Peterson, E., Tomlanovich, M., and Early Goal-Directed Therapy Collaborative Group. 2001. Early goal-directed therapy in the treatment of severe sepsis and septic shock. *New England J. Med. 345*, 19, 1368–1377.

Roweis, S. 1998. EM algorithms for PCA and SPCA. In *Proceedings of the Conference on Advances in Neural Information Processing Systems (NIPS'98).* 626–632.

Saeed, M., Villarroel, M., Reisner, A. T., Clifford, G., Lehman, L.-W., Moody, G., Heldt, T., Kyaw, T. H., Moody, B., and Mark, R. G. 2011. Multiparameter intelligent monitoring in intensive care ii (mimic-ii): A public-access intensive care unit database. *Critical Care Med. 39*, 5, 952–960.

Shavdia, D. 2007. Septic shock: Providing early warnings through multivariate logistic regression models. Tech. rep., Harvard-MIT Division of Health Sciences and Technology.

Stacklies, W., Redestig, H., Scholz, M., Walther, D., and Selbig, J. 2007. PcaMethods–A bioconductor package providing PCA methods for incomplete data. *Bioinf. 23*, 9, 1164–1167.

Thiel, S. W., Rosini, J. M., Shannon, W., Doherty, J. A., Micek, S. T., and Kollef, M. H. 2010. Early prediction of septic shock in hospitalized patients. *J. Hospital Med. 5,* 1, 19–25.

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

Wang, F., Lee, N., Hu, J., Sun, J., and Ebadollahi, S. 2012. Towards heterogeneous temporal clinical event pattern discovery: A convolutional approach. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'12).*

WSD Coalition. 2012. The world sepsis day fact sheet. http://www.world-sepsis-day.org/.

Received December 2012; revised November 2013; accepted November 2013