Developing Predictive Models Using Electronic Medical Records: Challenges and Pitfalls

Chris Paxton
Computer Science Department
Johns Hopkins University
Baltimore, MD 21218

Alexandru Niculescu-Mizil
NEC Labs America
Princeton, NJ 08540

Suchi Saria
Computer Science Department
Johns Hopkins University
Baltimore, MD 21218

Abstract

While Electronic Medical Records (EMR) contain detailed records of the patient-clinician encounter — vital signs, laboratory tests, symptoms, caregivers' notes, interventions prescribed and outcomes — developing predictive models from this data is not straightforward. These data contain systematic biases that violate assumptions made by off-the-shelf machine learning algorithms, commonly used in the literature to train predictive models. In this paper, we discuss key issues and subtle pitfalls specific to building predictive models from EMR. We highlight the importance of carefully considering both the special characteristics of EMR as well as the intended clinical use of the predictive model and show that failure to do so could lead to developing models that are less useful in practice. Finally, we describe approaches for training and evaluating models on EMR using early prediction of septic shock as our example application.

1. Introduction

Large investments by the government and the Meaningful Use Act have accelerated the adoption of Electronic Medical Records (EMR). These information systems contain granular measurements made over the course of a patient's stay in the hospital, including detailed records of symptoms, test measurements, data from monitoring devices, clinicians' observations and billing data. Potential uses of these data include risk stratification or early prediction [1, 2], biomarker discovery [3], cohort detection for clinical trial recruitment [4], and optimizing resource use. In this paper, we study issues surrounding the development and evaluation of early prediction systems from such *observational* data (i.e. data collected in the process of providing care) for early identification of complications. Systems that can detect complications early can help identify opportunities to intervene and allow for more cost-effective allocation of resources. For trauma patients, early diagnosis and rapid intervention to treat shock and organ dysfunction was found to decrease health-care resource utilization and improve outcomes [5]. In the intensive care unit, delays in identification and management of critically ill patients have been associated with higher mortality rates [6] and increased utilization of hospital resources [7]. Using observational EMR for training and evaluating early detection systems has several benefits when compared with using data from controlled studies, chief among which is the abundance of available data. Collecting data is virtually zero cost and therefore, baring privacy issues, the only restriction on the amount of data available is the incidence of the condition in question. In the inpatient and the outpatient setting, as EMRs collect visit data near real-time, the opportunity to deploy real-time prediction systems for effective patient management is enormous. It can influence decisions as wide-ranging as whether the patient should be transferred to a more resource intensive setting to whether a more aggressive treatment course should be pursued.

In contrast with much of the work on building predictions systems from EMR, the goal of this paper is *not* to engineer a better feature set, or develop a new algorithm or achieve better performance. The main goal of this paper is to highlight and discuss a number of key issues that arise when trying to build practical early prediction systems from observational EMR. We discuss how *confounding medical interventions* can mask the ground truth labels needed to train and evaluate a prediction system, and show that not taking this masking into account can lead to models that are less useful in practice. We highlight the importance of considering the clinical end use of a prediction system and how the training and evaluation of the predictive model depends on the desired clinical use. We present two different clinical applications of early prediction systems: a *stand-alone monitoring* or *diagnostic* system that would be used by caregivers to decide whether or not to intervene, and an *assisted monitoring* or *alert* system that works in parallel with caregivers and is used to detect at-risk patients who would otherwise be missed. This distinction of the two use cases is important in understanding the generalization performance of the predictive system. We show that, in general, it is not possible to use EMR to train and evaluate stand-alone monitoring systems, but assisted monitoring systems can be trained and evaluated using data contained in EMRs. Finally we discuss and evaluate some choices available to the modeler in dealing with data containing confounding medical interventions. We exemplify these issues in the context of building an early prediction system for septic shock using the MIMIC-II EMR database.

2. Septic Shock: Background, Data and Features

We use the development of a clinically useful early prediction system for septic shock as an example application that illustrates the challenges of predictive modeling from observational EMR.

Sepsis defined as a severe inflammatory response to infection is the leading cause of death in the inpatient setting. The Surviving Sepsis Campaign has established a set of guidelines for treatment which require immediate intervention with pressors, antibiotics, and fluid resuscitation [8]. Early detection and intervention is critical [9]; if the disease progresses to septic shock mortality rates increase to 40-80% [10,11]. Thus, early prediction systems can have a significant impact in reducing the mortality rates. Consequently, information models for prediction of sepsis have received significant attention [2, 12–20].

2.1 Data and Features

We use data from adult ICU patients contained in the MIMIC-II clinical data [21]. This dataset has been previously used in modeling prediction of septic shock [2, 15]. Rather than viewing each patient as a single value or a number of discrete bins, each patient is viewed as a set of time points, where the time of each entry in the EMR defines a time point. Since data collection is undertaken by ICU staff, we can assume more measurements will be taken during episodes of interest [22]. For an early prediction system to be clinically useful, it needs to identify patients at risk before treatment has started, thus we discard all time points after treatment has began. Since pressors and antibiotics are the standard treatment for severe sepsis [8], we take administration of pressors and antibiotics to represent beginning of treatment. We also discard all time points after septic shock has been observed.

Prior works have demonstrated improved performance by using a combination of chart data, lab results, patient demographics, and medications [2, 20]. We use the same subset of 77 chart variables that were previously used for septic shock prediction from MIMIC-II [2]. We check if values fall within an acceptable range in order to account for errors in data entry [22]. Maximum, variance, least-squares fit line over the last 6, 12, and 24 hours were computed for features derived from chart data in order to capture patient change over time preceding each time stamp. This is important as variability in vital signs such as heart rate is a known predictor of sepsis [17, 18]. Other values were taken from patient fluid I/O events, a list of fluids provided via IV drip and patient urine output. Sums for total volume of red blood cells, urine output, and all fluid input were recorded for 1 and 24 hours previous to each time point. Similar information recorded as Total Balance Events was also included directly. Medications were included as features whenever they were determined not to be either pressors or antibiotics. We include demographic information for each patient, looking at which ethnic group the patient falls in to. We also include binary features indicating if a culture or medical procedure that may be correlated with sepsis was performed on the patient within 24, 48, or 72 hours. In all, we compute 1011 features. At each time point, the most current value of a particular feature is used, with values older that a certain time window discarded [2].

3. Challenges of using EMR Data

The goal of an early prediction system is to detect when a patient is at high risk of developing an adverse condition as early as possible to increase the chances of successful treatment. To be useful in practice, the detection should happen prior to the caregiver intervening to treat the condition. Prior works train supervised classifiers to predict whether the patient will have an adverse condition (e.g. septic shock) in the future based on the past clinical history. More specifically, given the set of clinical events $(x_1, x_2, \cdots, x_k, \cdots, x_n)$ the goal is to predict at time k whether any of the future events $x_{k+1}, ..., x_n$ will be an adverse condition. When training and evaluating such a system, the "future" is available. Under some conditions, one could observe or infer whether the adverse event has occurred or not. At deployment, only the history up to time k is available. Learning and evaluating early prediction systems using observational data alone encounters three main challenges.

Incomplete Observation EMR can be left or right censored because the patient was transferred too late or discharged too early. This problem is relatively well understood and our treatment of this bias is similar to that in previous studies. We exclude samples where less than 12 hours of data are present prior to the septic shock.

Selection Bias The set of recorded patients in an institutional EMR is not a random sample from the population. Instead, it varies depending on the nature of particular practice, the care unit, and the geographical location of the medical institution. These factors impose biases on the patient demographics and the health condition of admitted patients. In turn this can impose restrictions on where the trained early prediction systems can be deployed: models trained on data from one practice might not generalize to another practice. In this paper we assume that the system is deployed in the same practice where the EMR were

collected, thus alleviating this issue. Since EMR are continuously collected in the process of regular care, significant amounts of EMR have already been collected for many practices, so this assumption is not too restrictive.

Confounding Medical Interventions (CMIs) These are interventions performed by the caregivers that will affect the risk of the outcome of interest. In management of sepsis for instance, the relevant interventions are administration of pressors and broad-spectrum antibiotics. From a model training point of view, a CMI often hides the true label of a patient's trajectory. After a CMI, one cannot distinguish between a patient who was at risk but is now treated due to antibiotic treatment versus a patient that received unnecessary treatment due to conservative judgment by the caregiver. This means data containing CMIs must be handled differently.

In table 1, we describe the four categories of data that are present in observational data. The data in the categories A and B contain no CMIs and the observed presence or absence of the adverse condition within a time frame is a direct indication whether the patient is at risk during their hospital stay. We refer to these as *clear data* samples. Category D contains patients upon whom a CMI has been performed. In this case the absence of the outcome (septic shock) is not indicative of whether the patient was truly at risk or not. We refer to these as *confounded data* samples. For patients in category C, even though CMIs were performed, the adverse condition is still observed. Depending on the studied adverse condition, it is possible that its presence was an unintended consequence of the CMI itself, in which case samples in category C should also be considered confounded data samples. In the case of sepsis, however, this is unlikely and the presence of the adverse condition indicates that the patient was truly at risk and the CMI was possibly insufficient or applied too late. Thus we still consider these to be clear samples. Whether a patient ends up in categories A or B, or C or D depends on the caregivers that makes the treatment decision, which can vary from practice to practice or even patient to patient.

To understand how CMIs can introduce unwanted biases in learning consider the following example. Consider a care unit where all children with $102\,\mathrm{deg}\,F$ are prescribed treatment for flu and are subsequently cured. A natural approach (and one that is frequently used in the literature) might be to use the adverse condition (column 4 in table 1) as the ground truth labels: the sample is marked as at risk (positive) if the adverse condition is observed (categories B and C) during the hospital stay and not at risk (negative) otherwise (categories A and D). In this case, samples from children with $102\,\mathrm{deg}\,F$ who were prescribed treatment and were cured will fall in category D. Since cases from D are considered not to be at risk, a learning algorithm trained using this data will rightfully learn a model that predicts that children with high temperature have a low risk of developing flu, and will direct the doctor not to prescribe treatment. Worse, standard cross-validation techniques would not detect anything wrong with the model as in the validation set cases from D are also considered low risk. The presence of CMIs and confounded data constrains how an early prediction system can be used in practice.

Cohort	Underlying risk (ground truth)	CMI	adverse condition
A	Not at risk	no	unobserved
B	At risk	no	observed
C	At risk	yes	observed
D	**Unknown**	yes	unobserved

Table 1: Types of samples present in EMR data

3.1 Clinical Application and Model Evaluation

When developing an early prediction system, one needs to understand how the system will be ultimately applied in the clinical setting and evaluate it accordingly. We make a distinction here between two different settings in which predictive models for early identification are applied. In the first case, the system is a *stand-alone monitoring* system that is trained to maximize detection accuracy on all patients. For example, a diagnostic system is optimized to correctly diagnose any patient based on their signs and symptoms. In the second case, the system serves to assist the clinician in monitoring patients. Such a system is not guaranteed to detect all patients with the given condition, rather it is optimized to detect patients that the clinicians currently miss. We call this system an *assisted monitoring* system. The source of observational data used in training our early prediction system affects the applicability of the derived system.

To evaluate the early prediction system, one must obtain ground truth data for a cohort that is representative of the population on whom the system will be deployed. In our application, the ground truth labels are whether or not the patient will develop septic shock in the absence of a CMI (prescription of antibiotics or pressors). If the goal is to do stand-alone monitoring our training data must contain representative samples of all possible patients. CMIs introduce systematic bias with regard to the types of patients on whom ground truth data are not available. To see why, note that most care units have a practice culture and caregivers ascribe to that culture to decide when to intervene. Therefore, caregivers are likely to systematically

intervene on the same set of patients, thereby adding CMIs that hide the ground truth labels on that subgroup of patients. In other words, patient populations on whom we have *clear* data with ground truth (categories A, B and C) likely differ from patient populations on whom we have *confounded* data that is missing ground truth labels. Therefore, we cannot draw conclusions on the generalizability of our model to patients in category D. In the absence of large variability in practice patterns, data acquired from single institution EMRs or from a small group of care providers are likely to be systematically biased in the patients on whom we are able to extract ground truth data and generalize our model to[1]. Therefore, patients on whom we have clear versus confounded data are not similar and therefore no conclusions can be made about generalizability as a stand-alone monitoring system.

All is not lost, however. For assisted monitoring where the assumption is that the system is trained to alert in cases when the caregiver previously missed i.e., identify patients in cohort B and C, model evaluation is feasible. Specifically, we do not need to show generalizability to patients with CMI (category D) where ground truth data are missing. It is sufficient to consider the clear data samples (categories A, B and C) where ground truth labels are observed and evaluate model only on that data. In practice, this implies that the caregiver makes decisions independently from the system but intervenes when either the caregiver suspects that the patient is at risk or if the system identifies as at risk. Thus, with regard to improving outcomes, early prediction systems trained on data from the institutional EMR are just as valuable as they provide decision support to identify patients at risk who were previously missed by the caregivers.

3.2 Developing the Predictive Model

Besides modeling decisions that are generally present in any learning application (e.g. choice of features or learning algorithms), when learning predictive models from EMR the modeler is faced with an additional question: what should be done with records that have confounding medical interventions (CMIs)? Remember that CMIs mask the ground truth label by reducing the risk of observing the adverse condition and making it impossible to distinguish between patients that are truly at risk and patients that have received interventions that were unnecessary or intended to treat a different condition.

In this section we will compare four possible approaches for handling data with CMIs when training a model. To this end, we use SVM-light [23] with a linear kernel and default parameters to train a predictive model for each of the four approaches, and evaluate their performance in the context of an assisted monitoring system, using a random sample of 4000 hospital stays from only cohorts A, B and C as discussed in the previous section. Henceforth we will denote the cohorts A,B and C as *clear data* as the ground truth label for these cohorts is clear. Similarly the cohort D will be denoted as *counfounded data*. For both training and testing, patient stays are broken into windows of length 72 hours. Using a sliding window approach, multiple windows that cover a patient's data through the length of his stay are generated by shifting the window to consecutive clinical events. For septic shock-positive patients, only data from within 3 days of the onset of septic shock is considered. Data prior to that is eliminated as its unclear whether they should receive a positive or negative label. To generate our train and test splits, samples from a given patient can only be included in the training or the test set. This is to mirror the end use more closely where the system is evaluated on patients that are different from those on whom the model was trained.

In this setting, class imbalance is an issue when training our models; there are many more negative samples (from patients who are not at risk) than positive samples (from patients who developed septic shock). Classes are balanced by up-weighting the minority class by the ratio of positive to negative examples. To test how performance depends on the availability of clear data, we used training sets ranging from 50 to 4000 clear stays. In addition, a fixed set of 2000 confounded hospital stays were made available for training. Figure 1 shows the mean AUC and standard deviation over five folds. For each fold, the test set was generated by random sampling (without replacement) from the clear data and within each fold, all four models were evaluated on the same test set so that their performance could be compared. The training set was also generated by random sampling without replacement.

The first approach for dealing with CMIs is to completely ignore the data where the ground truth was masked and only use the clear data for training (i.e. A, C positive, B negative). In the limit, when clear training data is plentiful, this is the best choice as the training and test set distributions match. The test set performance of this approach are depicted in figure 1

[1]To assess the extent of this problem, one can learn a classifier and measure area under the curve (AUC) for distinguishing between samples in categories A, B and C from samples in D. If the AUC is high, clearly the populations are different. Note, that a low AUC does not imply that the populations are similar. It only implies that the populations cannot be distinguished between *given* the current classifier. We performed this test on the septic shock data and obtained and AUC of 0.7

with the curve labeled "Clear Only". As the size of the data increases, performance of the model improves. Note, performance is only evaluated with regard to developing an assisted system and not a stand-alone system. In other words, one cannot claim how this model would perform on patients whose samples are in the confounded data.

The second approach is to use the observed condition as the training label (i.e. A and C as positive, and B and D negative). This would be the natural choice if one were not considering the influence that CMIs have on the ground truth, and indeed, previous work has used this approach both for training and testing the models (e.g., [2, 15]). However, as we discussed in section 3, this approach can introduce unwanted biases in the data and lead to poor results. The results for this approach are shown as the curve labeled "Confounded as Negative" in figure1. We see that indeed, not taking into account the masking effect of the CMIs leads to significantly worse results, performing the worse out of all approaches.

A third approach would be to assume that the caregiver was correct when making the decision to prescribe treatment and therefore, consider all the patients that have received an intervention treatment as positive (i.e. A, C, D as positive, B as negative). This approach is appealing as caregivers are highly trained professionals that usually are correct in their clinical assessment of which patients are at risk. The curve labeled "Confounded as Positive" in figure1 shows the performance of this approach. In our dataset, this approach gives better results than the "Confounded as negative" setting, but performance is still lower than simply ignoring the confounding instances.

Finally, rather than ignoring or choosing an adhoc labeling for the confounded data, one could use a transductive approach for inferring the true labels [24]. We perform a preliminary analysis using a transductive SVM [24]. In this learning framework, the data with masked ground truth is treated as unlabeled data (i.e. A, C as positive, B as negative, D as unlabeled). When only limited labeled training data is available, transductive learning has shown promising results. However, in our vanilla implementation, adding examples as unlabeled data does not seem to help improve performance over "Clear only" (see figure1). It is interesting to note that it outperforms the "Confounding as Negative" approach that is typically used in the literature to train predictive models. There is a rich body of work on transductive and semi-supervised learning. It is likely that models that better exploit our domain characteristics for inferring the labels might improve performance. We think this is a promising line of investigation for the future.

To further understand the utility of each model we performed an error analysis based on the severity of the adverse condition affecting each patient. This analysis revealed that a significant portion of the errors was due to mistakes on patients that develop severe sepsis but do not progress to septic shock. Given that severe sepsis is still a very serious condition with a high mortality rate and that the recommended treatment for severe sepsis is similar to the treatment of septic shock, one could argue that misclassifications of severe sepsis patients should not be considered as an error. Thus, we evaluate the discriminative capacity only on patients with septic shock against patients that do not progress beyond SIRS (i.e. mild sepsis) [8]. We show the results of this analysis in results in figure 2. While relative performance of the different approaches remains the same, all models performed noticably better (AUC increases by approximately 4 points).

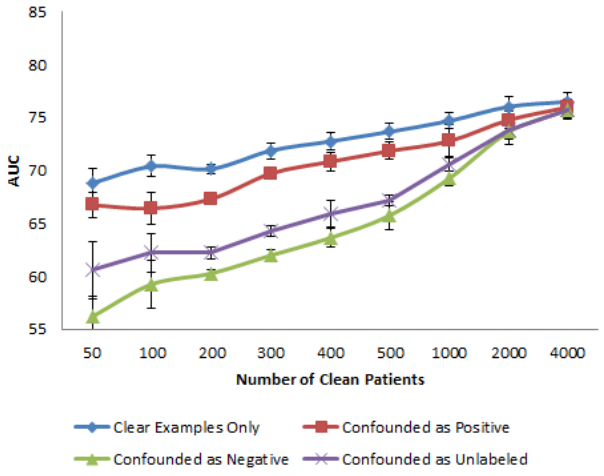

Figure 1: Plot of average AUC across 5 folds for different numbers of clear hospital stays.

4. Conclusions

Electronic Medical Records are becoming an increasingly valuable resource for developing predictive models for improving patient management. We show how systematic biases present in this data make both model development and evaluation challenging. In particular, *confounding medical interventions* (CMIs) systematically mask the ground truth labels needed for training and evaluating a prediction system; we show that not taking this masking into account may lead to models that are useless in practice. We emphasize the importance of considering how this system will be applied in deciding the approach for system development and evaluation. Then, we discuss this in the context of two different clinical applications of early prediction systems: an *assisted monitoring* system that monitors the patients in parallel with the caregiver and is

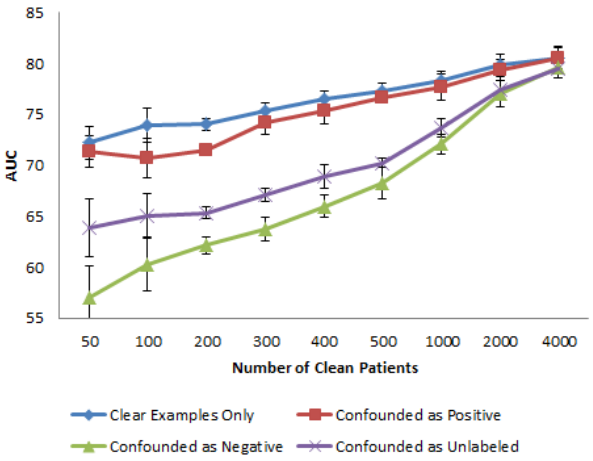

Figure 2: Plot of average AUC across 5 folds for different numbers of clear hospital stays, ignoring stays with severe sepsis.

used to detect at risk patients that would otherwise be missed, versus a *stand-alone monitoring* system that would be used by caregivers to decide whether or not to intervene. We discuss how it would be difficult to train and evaluate stand-alone monitoring systems from observational data, while it is straightforward to evaluate and train assisted monitoring systems using EMRs. Finally, we discuss possible choices available to the modeler in tackling biases such as CMI in the data, and show the effects of these biases within the application of septic shock prediction. In particular, the differences in performance when counfounded data is treated as positive, negative, or ignored entirely suggests that a more nuanced approach to handling this data would be valuable. Future research should examine new approaches for handling counfounded data.

Developing early prediction systems from EMR data is an increasingly important area of research. We hope that ideas presented in this paper will influence design choices made in developing early prediction systems from EMR and through making more principled choices we can improve their generalizability.

References

[1] Saria S, Rajani AK, Gould J, Koller D, Penn AA, et al. Integration of early physiological responses predicts later illness severity in preterm infants. Science translational medicine. 2010;2(48):48ra65–48ra65.

[2] Hug C. Detecting hazardous intensive care patient episodes using real-time mortality models [PhD Thesis]. Massachusetts Institute of Technology. Cambridge, MA; 2009.

[3] Kho AN, Pacheco JA, Peissig PL, Rasmussen L, Newton KM, Weston N, et al. Electronic medical records for genetic research: results of the eMERGE consortium. Sci Transl Med. 2011;3(79):79rel.

[4] Fultz SL, Skanderson M, Mole LA, Gandhi N, Bryant K, Crystal S, et al. Development and verification of a" virtual" cohort using the National VA Health Information System. Medical care. 2006;44(8):S25–S30.

[5] Rivers EP, McIntyre L, Morro DC, Rivers KK. Early and innovative interventions for severe sepsis and septic shock: taking advantage of a window of opportunity. Canadian Medical Association Journal. 2005;173(9):1054–1065.

[6] Lundberg JS, Perl TM, Wiblin T, Costigan MD, Dawson J, Nettleman MD, et al. Septic shock: an analysis of outcomes for patients with onset on hospital wards versus intensive care units. Critical care medicine. 1998;26(6):1020–1024.

[7] Engoren M. The effect of prompt physician visits on intensive care unit mortality and cost. Critical care medicine. 2005;33(4):727–732.

[8] Dellinger RP, Levy MM, Carlet JM, Bion J, Parker MM, Jaeschke R, et al. Surviving Sepsis Campaign: International Guidelines for management of severe sepsis and septic shock. Intensive care medicine. 2008;34(1):17–60.

[9] Rivers E, Nguyen B, Havstad S, Ressler J, Muzzin A, Knoblich B, et al. Early goal-directed therapy in the treatment of severe sepsis and septic shock. New England Journal of Medicine. 2001;345(19):1368–1377.

[10] Jawad I, Lukšić I, Rafnsson SB. Assessing available information on the burden of sepsis: global estimates of incidence, prevalence and mortality. Journal of global health. 2012;2(1).

[11] Martin GS. Sepsis, severe sepsis and septic shock: changes in incidence, pathogens and outcomes. Expert Review of Anti-infective Therapy. 2012;10(6):701–706.

[12] Pilz G, Fraunberger P, Appel R, Kreuzer E, Werdan K, Walli A, et al. Early prediction of outcome in score-identified, postcardiac surgical patients at high risk for sepsis, using soluble tumor necrosis factor receptor-p55 concentrations. Critical care medicine. 1996;24(4):596–600.

[13] Brause R, Hamker F, Paetz J. Septic shock diagnosis by neural networks and rule based systems. Studies in Fuzziness and Soft Computing. 2002;96:323–356.

[14] Seely AJ. Heart rate variability and infection: Diagnosis, prognosis, and prediction. Journal of Critical Care. 2006;21(3):286–289.

[15] Shavdia D. Septic shock: Providing early warnings through multivariate logistic regression models [Master's Thesis]. Massachusetts Institute of Technology. Cambridge, MA; 2007.

[16] Giuliano KK. Physiological monitoring for critically ill patients: testing a predictive model for the early detection of sepsis. American Journal of Critical Care. 2007;16(2):122–130.

[17] Ahmad S, Tejuja A, Newman KD, Zarychanski R, Seely AJ. Clinical review: a review and analysis of heart rate variability and the diagnosis and prognosis of infection. Crit Care Med. 2009;13(6):232.

[18] Tang CH, Middleton PM, Savkin AV, Chan GS, Bishop S, Lovell NH. Non-invasive classification of severe sepsis and systemic inflammatory response syndrome using a nonlinear support vector machine: a preliminary study. Physiological measurement. 2010;31(6):775.

[19] Thiel SW, Rosini JM, Shannon W, Doherty JA, Micek ST, Kollef MH. Early prediction of septic shock in hospitalized patients. Journal of Hospital Medicine. 2010;5(1):19–25.

[20] Wiens J, Guttag J, Horvitz E. Learning Evolving Patient Risk Processes for C. Diff Colonization. ICML Workshop on Health Informatics. 2012;.

[21] Saeed M, Villarroel M, Reisner AT, Clifford G, Lehman LW, Moody G, et al. Multiparameter Intelligent Monitoring in Intensive Care II (MIMIC-II): A public-access intensive care unit database. Critical Care Medicine. 2011 May;39:952–960.

[22] Clifford GD, Long WJ, Moody GB, Szolovits P. Robust parameter extraction for decision support using multimodal intensive care data. Philosophical Transactions of the Royal Society A. 2009;367(1887):411–429.

[23] Joachims T. Making large-Scale SVM Learning Practical. MIT Press; 1999.

[24] Joachims T. Transductive inference for text classification using support vector machines. In: International Conference on Machine Learning; 1999. p. 200–209.