A Boosting Inspired Personalized Threshold Method for Sepsis Screening

Chen Feng[a], Paul Griffin[b], Shravan Kethireddy[c], and Yajun Mei[a]

[a]School of Industrial & Systems Engineering, Georgia Tech, Atlanta, GA; [b]Regenstrief Center for Healthcare Engineering, Purdue University, West Lafayette, IN; [c]Critical Care Medicine, Northeast Georgia Medical Center, Gainesville, GA

ARTICLE HISTORY
Compiled January 15, 2020

ABSTRACT

Sepsis is one of the biggest risks to patient safety, with a natural mortality rate between 25% and 50%. It is difficult to diagnose, and no validated standard for diagnosis currently exists. A commonly used scoring criteria is the quick sequential organ failure assessment (qSOFA). It demonstrates very low specificity in ICU populations, however. We develop a method to personalize thresholds in qSOFA that incorporates easily to measure patient baseline characteristics. We compare the personalized threshold method to qSOFA, five previously published methods that obtain an optimal constant threshold for a single biomarker, and to the machine learning algorithms based on logistic regression and AdaBoosting using patient data in the MIMIC-III database. The personalized threshold method achieves higher accuracy than qSOFA and the five published methods and has comparable performance to machine learning methods. Personalized thresholds, however, are much easier to adopt in real-life monitoring than machine learning methods as they are computed once for a patient and used in the same way as qSOFA, whereas the machine learning methods are hard to implement and interpret.

KEYWORDS
Sepsis; Clinical Risk Score; Personalized Thresholds; Screening; Boosting

1. Introduction

Sepsis is life-threatening organ dysfunction caused by a dysregulated host response to infection, and it is one of the biggest patient safety risks in healthcare settings [10, 20, 35, 49]. Nearly half of patients who die in hospitals are septic, and the natural mortality rate for sepsis is between 25% and 50% [10]. Recently, screening as a decision support mechanism for early detection of sepsis has been widely advocated, since early identification of sepsis and the timely medical intervention could significantly decrease sepsis-related mortality and are cost-effective [3, 14, 21, 27, 38, 42, 48, 53]. In 2016, a task force committee [39] recommended patient screening for sepsis by the scoring criterion termed quick Sequential (sepsis related) Organ Failure Assessment (qSOFA), and conducting laboratory tests to further assess sepsis if needed. The qSOFA score is

Correspondence Author. Email: chelsea.fengchen1993@gmail.com

essentially a constant thresholding technique regardless of patients' baseline information, and it uses three easy-to-measure biomarkers: systolic blood pressure, respiratory rate, and Glasgow Coma Scale score (a score for mental status).

The main statistical approach in qSOFA criteria is to dichotomize each biomarker X, and raise a screening warning if $X \geq c$ (or $X < c$) for some constant threshold c. The constant thresholds in qSOFA are determined based on physiology, clinical experience, and statistical analysis. Indeed, there has been extensive research in the statistical literature to find the optimal threshold. The five most popular methods are: 1) the minP approach [26], 2) the Youden index [51], 3) the closest-to-$(0, 1)$ criteria [29], 4) the concordance probability method [24], and 5) the index of union [47]. All of those five methods are based on Receiver Operating Characteristic (ROC) curve statistical analysis [33, 36].

Unfortunately, there are some drawbacks to the qSOFA score. It demonstrates very low specificity in ICU populations [6, 40]. At emergency department (ED) triage, qSOFA scores poorly in identifying sepsis, and is likewise poor in both pre-hospital and ED triage for predicting intensive care unit (ICU) stays of three or more days [2, 5, 11, 46]. One possible explanation is that the current qSOFA uses thresholds that are constant for all patients, regardless of patients' demographic differences, leading to the low prediction accuracy.

In this paper, we propose to improve qSOFA by self-learning suitable personalized thresholds for different sepsis patients. Our key idea is inspired by boosting [4, 9, 15], a popular machine learning technique, to obtain the personalized thresholds based on patient-related information. As a result, our method can better identify sepsis patients who could benefit from time critical interventions. In addition, our method is similar to original qSOFA, and it is easy to interpret and can be used for real-time monitoring.

Besides screening, another concept in sepsis is assessing, where the patient needs to take more expensive and time-consuming laboratory tests to get more information on various biomarkers so as to further assess the status of sepsis. In such a case, those biomarkers are collected much less frequently, allowing machine learning methods to be applied for sepsis assessing, see [16, 17, 19, 25, 30, 41, 43–45]. For instance, Giannini et al. [12] applied the Random Forest classifier to predict patients at elevated risk of developing severe sepsis and/or septic shock by using Electronic Health Record (EHR) data; Shimabukuro et al. [37] studied a machine learning-based severe sepsis prediction system for reductions in average length of stay and in-hospital mortality rate; Nemati et al. [28] developed and validated an Artificial Intelligence Sepsis Expert algorithm for early prediction of sepsis. However, the context of sepsis screening in ICU is very different from those of sepsis assessing, in that the medical machines or sensors automatically measure those key biomarkers as frequently as per minute. Thus these black-box machine learning methods are not as appropriate a method for screening due to lack of interpretation and implementation difficulties for real-time monitoring. Here, our paper essentially follows the idea of control charts in the field of statistical process control (SPC), where we keep monitoring the frequently observed data, and raise an alarm once any abnormal observation occurs. Here the abnormal observation is defined if a process exceeds the pre-specified cutoff, or equivalently, if the maximum (or minimum) value of the process exceeds the pre-specified cutoff. Instead of using the universal cutoff value, our paper proposes a personalized cutoff value, based on a patient's characteristic data (e.g., age, gender, etc.).

The organization of the remaining sections are as follows. Section 2 introduces the necessary background knowledge, including the details of sepsis and qSOFA score, existing methods to obtain the optimal constant threshold and boosting machine learning method. Section 3 describes the proposed boosting inspired method to estimate the personalized threshold , in which the exponential loss and gradient descent algorithm were applied. Section 4 explores the data we use to demonstrate the method. Section 5 presents the application in sepsis screening and compares the performance of the proposed method to the original qSOFA criteria, other standards used, constant threshold approaches, and machine learning methods. We provide conclusions in Section 6.

2. Background

In this section, we introduce the background information of sepsis and qSOFA score, the existing approaches to obtain constant thresholds, and the boosting method in three subsections, respectively.

2.1. Sepsis and qSOFA

Sepsis is not a specific illness but rather a syndrome producing a similar innate immune response as infection. It is differentiated from infection in a dysregulated host response and the presence of organ dysfunction [39]. Considerable changes have been made on how it is defined based on a better understanding of the underlying pathobiology.

The initial definition of sepsis was developed in 1991, and it was assessed by the Systemic Inflammatory Response Syndromic (SIRS) to infection, which includes a patient's temperature, heart rate, respiratory rate, and white blood cell count [1, 23]. In this definition, sepsis is defined as infection with at least 2 of the 4 SIRS criteria satisfied. However, SIRS criteria do not necessarily indicate organ dysfunction or failure. In 2016, therefore, a task force developed the current definition of sepsis, and recommended two stages of monitoring sepsis: screening and assessing [39]. Instead of the widely used SIRS criteria, the 2016 definitions recommended clinically characterizing a sepsis patient by the Sequential Organ Failure Assessment (SOFA) score, which is used to identify organ dysfuntion. Moreover, the 2016 task force committee recommended a new score criterion termed quickSOFA (qSOFA), a bedside screening measurement identifying patients with suspected infection.

The qSOFA score relies on three important variables for sepsis: respiratory rate, systolic blood pressure, and Glasgow Coma Scale (GCS) score, and the current qSOFA guideline is to check whether these three observed variables are normal or abnormal as compared to their respective constant critical threshold values. A screening alarm is raised if two out of three variables are abnormal. To be specific, in the qSOFA criterion, one is monitoring 1) whether the alteration in mental status occurs (GCS score is less than 15), 2) whether systolic blood pressure is $\leq$100 mm Hg, or 3) whether respiratory rate is $\geq$ 22 breaths/min. These constant critical thresholds are derived from the two group comparisons of sepsis versus non-sepsis patients, and do not take into account the patient's baseline demographic characteristics such as age, sex, admission location, admission type, ethnicity, insurance, and marital status.

2.2. Existing Methods to Obtain the Constant Threshold

The most common approaches to determine the threshold for a biomarker are via Receiver Operating Characteristic (ROC) curve analysis. The ROC curve is a mapping of the sensitivity versus 1-specificity for all possible thresholds. Thresholds dichotomize the test values, and therefore provide the diagnosis whether the subject is diseased or not. A threshold is referred to as optimal when it classifies most of the individuals correctly. Let X be a continuous biomarker that is assumed to be predictive of disease Y ($Y = 1$ for diseased and $Y = -1$ for not diseased). For any possible cut-point c of X, the data can be formed into a 2×2 table,

	$Y = 1$	$Y = -1$
$X \geq c$	$s = s(c)$	$r = r(c)$
$X < c$	$u = u(c)$	$v = v(c)$

Sensitivity(Se) and specificity(Sp) are defined as follows,

$$\mathrm{Se}(c) = P(X \geq c | Y = 1) = \frac{s}{s+u}, \quad \mathrm{Sp}(c) = P(X < c | Y = -1) = \frac{v}{r+v}.$$

Various criteria for the optimal threshold value c have been proposed [24, 26, 29, 47, 51] based on above 2-by-2 table. We briefly describe them here.

minP Approach [26]: The optimal threshold c is selected so as to maximize the standard chi square statistic,

$$\chi_1^2(c) = \frac{(s+r+u+v)(sv-ur)^2}{(s+r)(u+v)(s+u)(r+v)}.$$

Youden Index [51]: Youden index, $\mathrm{Se}(c) + \mathrm{Sp}(c) - 1$, combines sensitivity and specificity into a single measure. Maximizing the Youden index, one is able to find the cut point that has the largest value in the sum of sensitivity and specificity or in the difference between sensitivity and the false positive rate.

Closest-to-(0,1) Criterion [29]: Ideal point (0, 1) on ROC curve represents zero false positives and perfect sensitivity. The "optimal" threshold is defined as the point on the ROC curve closest to (0,1), i.e., find c to minimize $\sqrt{(1 - \mathrm{Se}(c))^2 + (1 - \mathrm{Sp}(c))^2}$.

Concordance Probability [24]: The concordance probability method defines the optimal threshold as the point c maximizing the product of sensitivity and specificity $\mathrm{Se}(c)\mathrm{Sp}(c)$. The concordance probability of dichotomized measure at threshold c can be expressed as the area of a rectangle associated with the ROC curve. The threshold c maximizing $\mathrm{Se}(c)\mathrm{Sp}(c)$ actually maximizes the area of the rectangle.

Index of Union [47]. The method defines the optimal cut-point value c as the point minimizing the summation of absolute values of the differences between AUC and sensitivity and AUC and specificity, $|\mathrm{Se}(c) - \mathrm{AUC}| + |\mathrm{Sp}(c) - \mathrm{AUC}|$.

All these methods are based on the biomarker itself and the outcome variable alone, and do not take into account any individual-specific information. In real clinical practice, patient-related variables are usually available besides the biomarker, such as the easily accessible demographics. How to utilize the extra information to determine a personalized threshold for screening purpose has rarely been considered.

2.3. Boosting Method

Boosting is an ensemble technique that attempts to create a strong classifier from a number of weak classifiers. One well-known boosting algorithm is called "AdaBoost.M1." proposed by Freund and Schapire (1997) [8]. Since our proposed method borrows some of the ideas in boosting, in this section, we will briefly introduce the boosting method.

Given a binary outcome $Y \in \{-1, 1\}$ and a vector of predictor variables $\boldsymbol{X}$, a classifier $f(\boldsymbol{X})$ is a sign function of some statistics, producing a prediction taking one of the two values $\{-1, 1\}$,

$$f(\boldsymbol{X}) = \text{sign}\left(g(\boldsymbol{X})\right), \text{ where } g(\boldsymbol{X}) = \sum_{m=1}^{M} \alpha_m f_m(\boldsymbol{X}), \tag{1}$$

and $f_m(x)$, $m = 1, 2, ..., M$, are weak leaners and $g(\boldsymbol{X})$ combines them together to make a better classification. One of the key ideas in boosting method is to replace the discrete 0-1 classification error by the exponential loss. To be more concrete, in classification with a -1/1 response, the error rate of training data is

$$\overline{\text{err}} = \frac{1}{N}\sum_{i=1}^{N} I(y_i \neq \text{sign}\left(g(\boldsymbol{X})\right)) = \frac{1}{N}\sum_{i=1}^{N} I(y_i \cdot \text{sign}\left(g(\boldsymbol{X})\right) < 0)\,. \tag{2}$$

Observations with $y_i \cdot \text{sign}\left(g(\boldsymbol{X})\right) > 0$ are correctly classified, while those with $y_i \cdot \text{sign}\left(g(\boldsymbol{X})\right) < 0$ are misclassified. The error rate is not a smoothing function, and therefore it is not a favorable loss function for classification. Boosting method replaces the term $I(y_i \cdot \text{sign}\left(g(\boldsymbol{X}) < 0\right))$ by a novel exponential loss function

$$L(y, g(\boldsymbol{X})) = \exp(-y \cdot g(\boldsymbol{X})). \tag{3}$$

It has been proven that the exponential loss is a monotonic continuous approximations to misclassification loss. In the training process, the exponential criterion concentrates much more influence on observations with negative $y \cdot g(\boldsymbol{X})$.

Boosting method essentially is a gradient descent algorithm that finds the parameters $\alpha_m' s$ to minimize the exponential loss function on the training data. It turns out that it can also be thought of as applying different weights to training observations $(\boldsymbol{X}_i, y_i)$, $i = 1, 2, ..., N$. Those observations that were difficult to predict (misclassified by weak leaners) will have larger weights, whereas less weights will be assigned to those easy-to-predict observations that were classified correctly by weak learners.

Our proposed method in Section 3 borrows the following ideas from Boosting: 1) Using a sign function as the classification rule for a binary outcome; 2) Optimizing a novel exponential loss function that approximates the error rate; 3) Assigning different weights on observations based on their influences.

3. Our Proposed Personalized Threshold Method

In this section, we propose a boosting alike method to obtain the personalized threshold. Suppose we have a training data of the form $(Y_i, X_i, u_{i1}, u_{i2}, ..., u_{iq})$, for

$i = 1, \cdots, N$, where $Y_i \in \{-1, 1\}$ is the binary outcome, X_i is the frequently measured biomarker whose threshold needs to be determined for monitoring purpose, $u_{i1}, u_{i2}, ..., u_{iq}$ are q baseline characteristics that will be manually entered to the system when the patient is admitted to ICU. We define $\boldsymbol{u}_i = (1, u_{i1}, u_{i2}, ..., u_{iq})^T \in \mathbb{R}^{q+1}$, where 1 corresponds to the intercept in proposed model. The biomarker X_i could be respiratory rate, systolic blood pressure, etc., and the binary Y_i is predicted by comparing X_i with threshold $c(\boldsymbol{u}_i)$. Here the threshold $c(\boldsymbol{u}_i)$ is a function of baseline characteristics $\boldsymbol{u}_i$, while the existing methods estimate the optimal constant threshold without using the extra information, such as a patient's gender, age, weight, etc. We define the classification rule as:

$$\hat{Y}_i = \begin{cases} 1, & \text{if } X_i \geq c(\boldsymbol{u}_i) \\ -1, & \text{otherwise} \end{cases} = \text{sign}(X_i - c(\boldsymbol{u}_i)), \tag{4}$$

and we assume

$$c(\boldsymbol{u}_i) = \beta_0 + u_{i1}\beta_1 + \cdots + u_{iq}\beta_q = \boldsymbol{u}_i^T \boldsymbol{\beta}, \tag{5}$$

where $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2..., \beta_q)^T \in \mathbb{R}^{q+1}$. Note that for the case where $\hat{Y}_i = 1$ if $X_i \leq c(\boldsymbol{u}_i)$, we could let $X_i' = -X_i$, and then fit X_i' in model (4).

Below we will present our proposed method to obtain the threshold $c(\boldsymbol{u}_i)$ in two subsections. Section 3.1 formulates an optimization problem, and section 3.2 introduces the algorithm to solve the problem.

3.1. Optimization Problem

The personalized threshold can be obtained by solving an optimization problem. To be more concrete, the $(q+1)$-dimensional unknown parameters $\boldsymbol{\beta}$ in (5) can be estimated by minimizing the misclassification rate. The remaining multiple challenges include: First, the function $\text{sign}(X_i - c(\boldsymbol{u}_i))$ in (4) is not continuous; second, the 0-1 loss function $I\{\hat{Y}_i \neq \text{sign}(X_i - c(\boldsymbol{u}_i))\}$ is non-smoothing; third, the consequences of misclassifying sepsis and non-sepsis patients are different, due to the high mortality rate of the sepsis.

In order to address those challenges, we modify the boosting method by minimizing a smooth surrogate weighted exponential loss function,

$$J(\boldsymbol{\beta}) = \frac{1}{N} \sum_{i=1}^{N} \left(w_{Y_i} \cdot e^{-Y_i(X_i - \boldsymbol{\beta}^T \boldsymbol{u}_i)} \right), \tag{6}$$

where the weight w_{Y_i} depends on outcome Y_i. We define $w_{Y_i} = w_+$ when $Y_i = +1$, and $w_{Y_i} = w_-$ when $Y_i = -1$. Equivalently we can write $w_{Y_i} = \big(w_+ \cdot (Y_i + 1) + w_- \cdot (1 - Y_i)\big)/2$. The values of w_+ and w_- are user specified.

Then the unknown coefficients $\boldsymbol{\beta}$ can be estimated as below,

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta} \in \mathbb{R}^{(q+1)}}{\text{argmin}}\, J(\boldsymbol{\beta}). \tag{7}$$

This provides a personalized threshold $\hat{c}_i = \boldsymbol{u}_i^T \hat{\boldsymbol{\beta}}$ in (5) for the i-th subject for $i = 1, \cdots, N$.

The motivation of the proposed weighted exponential loss function (6) is based on the fact that our classification rule in (4) is very similar to that in boosting method (1). To see this, we define $f_i \triangleq X_i - c(\boldsymbol{u}_i)$, then the prediction on the outcome Y_i is

$$\hat{Y}_i = \text{sign}(f_i). \tag{8}$$

Therefore we borrow the idea from boosting and replace the $0-1$ loss function $I(Y_i \neq \text{sign}(f_i))$ by the exponential loss $\exp(-Y_i f_i)$. Besides, boosting method assigns different weights to observations, and inspired by this, we introduce two different weights, w_+ and w_-, depending on whether $Y_i = +1$ or -1, in order to take into account the different consequences of misclassification. These lead us to consider the loss function

$$L(Y_i, f_i) = w_+ \cdot e^{-Y_i f_i} \cdot I(Y_i = 1) + w_- \cdot e^{-Y_i f_i} \cdot I(Y_i = -1).$$

The weighted exponential loss function (6) is then derived from $J(\boldsymbol{\beta}) = \frac{1}{N}\sum_{i=1}^{N} L(Y_i, f_i)$.

3.2. Gradient Descent Algorithm

We apply the gradient descent algorithm to solve the parameter $\hat{\boldsymbol{\beta}}$ in (7). The procedure can be divided into forward propagation and backward propagation steps. The forward propagation step constructs the cost function vector J in (6). If the observed data is $\{Y_i, X_i, u_{i1}, u_{i2}, ..., u_{iq}\}$, we denote $\boldsymbol{X} = (X_1, X_2, ..., X_N)^T$, $\boldsymbol{Y} = (Y_1, Y_2, ..., Y_N)^T$, $\boldsymbol{U} = (\boldsymbol{u}_1, \boldsymbol{u}_2, ..., \boldsymbol{u}_N)$, and $\boldsymbol{u}_i = (1, u_{i1}, u_{i2}, ..., u_{iq})^T$. The unknown parameter is $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, ..., \beta_q)^T$. The weights in (6) are denoted by $\boldsymbol{W} = (w_{Y_1}, w_{Y_2}, ..., w_{Y_N})^T$, where $w_{Y_i} \triangleq \big(w_+ \cdot (Y_i + 1) + w_- \cdot (1 - Y_i)\big)/2$, with w_{Y_i} being w_+ if $Y_i = 1$ and w_- if $Y_i = -1$.

In order to minimize the average cost, we use backwards propagation by finding the derivative of J with respect to $\boldsymbol{\beta}$ using the chain rule, and then moving in the direction to reduce total cost. This is repeated until convergence. If we denote $d\boldsymbol{\beta}$ as the derivatives of J with respect to $\boldsymbol{\beta}$ in current iteration, then the value of $\boldsymbol{\beta}$ will be updated by $\boldsymbol{\beta} \leftarrow \boldsymbol{\beta} - \alpha d\boldsymbol{\beta}$. The learning rate α is a given small number. After T iterations, we obtain $\hat{\boldsymbol{\beta}}$ as the final estimation of unknown $(q+1)$-dim parameter $\boldsymbol{\beta}$, then the estimated thresholds can be calculated as $\hat{\boldsymbol{c}} = \boldsymbol{U}^T\hat{\boldsymbol{\beta}}$ in (5).

In summary, our proposed algorithm for personalized threshold $c(\boldsymbol{u}_i)$ in (5) through solving the optimization problem in (7) can be presented as follows.

The following proposition shows that $\hat{\boldsymbol{c}} = \boldsymbol{U}^T\hat{\boldsymbol{\beta}}$ always exists and is well-defined as a point estimate of the optimal threshold.

Proposition 3.1. *The weighted exponential loss function $J(\boldsymbol{\beta})$ in (6) is a convex function with respect to $\boldsymbol{\beta}$, and thus the gradient descent algorithm converges if we choose a small enough learning rate and long enough optimization steps.*

The proof of Proposition 3.1 is postponed to the Appendix A. By formulating the parameter estimation as a convex optimization problem that has a unique solution, we simplify both the numerical computation and the tuning process for the algorithm parameters. It is now straight forward to evaluate the prediction errors of the classifier in (4) through the testing data and cross-validation.

Remark 1. We propose a knowledgable-based machine learning method that keeps

Algorithm 1 Our Proposed Algorithm for Personalized Threshold

```
Require: $\boldsymbol{Y}$, $\boldsymbol{X}$, $\boldsymbol{U}$, $N$, $w_- > 0$, $w_+ > 0$, $\alpha$, $T$
Initialization: $\beta_i \leftarrow 0 \ \forall i \in \{0, 1, 2, ..., q\}$, $\boldsymbol{W} = \big(w_+(\boldsymbol{Y}+1) + w_-(1-\boldsymbol{Y})\big)/2$
for all $t = 1, 2, \ldots, T$ do
    $\boldsymbol{f} = \boldsymbol{X} - \boldsymbol{U}^T\boldsymbol{\beta}$                                      }
    $\boldsymbol{L} = \exp(-\boldsymbol{Y} * \boldsymbol{f})$ {$*$: Element-wise product}      } Forward Propagation
    $\boldsymbol{J} = \frac{1}{N}\boldsymbol{L}^T\boldsymbol{W}$                                          }
    $d\boldsymbol{f} = -\frac{1}{N}\boldsymbol{W} * \boldsymbol{Y} * \exp(-\boldsymbol{Y} * \boldsymbol{f})$          }
    $d\boldsymbol{\beta} = -\boldsymbol{U}d\boldsymbol{f}$                                        } Backward Propagation
    $\boldsymbol{\beta} \leftarrow \boldsymbol{\beta} - \alpha d\boldsymbol{\beta}$                                  }
end for
$\hat{\boldsymbol{c}} = \boldsymbol{U}^T\boldsymbol{\beta}$
```

the simple thresholding idea, but at the same time, mimics the idea of boosting. Besides, it is easy to notice that our method is close to the logistic regression model with the coefficient of biomarker fixed at 1. Although classical machine learning methods are becoming increasingly popular in improving health care, they are usually a black box for physicians and nurses and are not suitable for screening due to interpretation and implementation difficulties. Therefore, constant thresholding method is still commonly adopted in real life for monitoring frequently measured biomarkers in ICU. As a special case of logistic regression model, our personalized thresholding method is a surrogate that combines the easily implemented thresholding method with machine learning techniques, improving the predictive accuracy without adding too much computational burden or complexity, allowing nurses and physicians to identify and interpret the triggers of the sepsis alert, and guaranteeing timely interventions by early detection of suspicious conditions. In Section 5.3, we will compare our proposed method with logistic regression and AdaBoosting in the context of sepsis screening.

Remark 2. Our proposed method can be extended to multiple biomarkers. When applying our proposed method to sepsis screening, we need to estimate the personalized thresholds for both respiratory rate and systolic blood pressure in qSOFA criteria, but for GCS score, the constant threshold of 15 is kept in that it is a combination of three sub-scores and is scored manually by nurses and physicians based on human judgements and multiple biological indexes, obtaining the personalized threshold for such complicated discrete biomarker is out of scope of this article. We define c_{i1} and c_{i2} to be the personalized thresholds of respiratory rate and systolic blood pressure, respectively, depending on variables $u'_{ij}s$ that are subject's baseline characteristics. We assume $c_{ik} = \beta_{0k} + u_{i1}\beta_{1k} + \cdots + u_{iq}\beta_{qk}$, for some unknown parameters β_{jk}'s, with $k = 1, 2$. Ideally, we want to find suitable choices of the $2(q+1)$ parameters β_{jk}'s from the training data, so that the qSOFA criteria could have good predictive performance for the testing data. However, it is non-trivial to jointly estimate them simultaneously from the training data. We therefore decompose the $2(q+1)$-dimensional estimation problem into 2 different $(q+1)$-dimensional estimation problems, and estimate the $(q+1)$-dimensional vector $(\beta_{0k}, \cdots, \beta_{qk})$ in the personalized thresholds c_{ik} recursively one at a time for each $k = 1, 2$. In general this might lose statistical efficiency since we ignore the intercorrelation between the biomarkers, but it will gain computational efficiency. Moreover, it is a reasonable approach for the sepsis screening context, as the three biomarkers (respiratory rate, systolic blood pressure, and Glasgow Coma Scale (GCS) score) characterize different physical and mental aspects of sepsis patients. In particular, the classifier is considered to have good properties only if each biomarker yields a good prediction of binary outcome, and the constant thresholds of the current qSOFA guideline are also based on the component-to-component optimization.

4. The Data Set

We use the Medical Information Mart for Intensive Care III (MIMIC-III) database (version 1.4) [18, 31], a large and freely-available database comprised of de-identified health-related data associated with over forty thousand patients who stayed in critical care units of the Beth Israel Deaconess Medical Center between 2001 and 2012. There were a total of 46,520 patients in the data set. The International Classification of Diseases, Ninth Revision (ICD-9) coding was used to identify sepsis and non-septic patients. ICD-9 is a list of codes intended for the classification of diseases and a wide variety of signs, symptoms, abnormal findings, complaints, social circumstances, and external causes of injury or disease.

Below we will present the details in four subsections. In section 4.1, we introduce how the study group and control group are selected. Section 4.2 shows the interested variables. We discuss the summary statistics of those variables in Section 4.3 and conduct exploratory analysis on qSOFA biomarkers in Section 4.4.

4.1. Study Population

The details of cohort selection from the MIMIC-III data is provided in Appendix B. There are 36,543 adult patients (aged 18 years or older), 4,233 of which have sepsis-related ICD-9 codes (995.91 for sepsis, 995.92 for severe sepsis, and 785.52 for septic shock). In some cases, a patient is assigned more than one ICD-9 code, and if any of the ICD-9 codes is sepsis related, we consider them as being diagnosed with sepsis. We retrieve comprehensive clinical data, including patient demographic and clinical measurements for qSOFA biomarkers. After excluding those sepsis patients who have no observations in the qSOFA variables within the first 24 hours after admission, we generate a study group of $3,771$ adult patients with sepsis.

We form a control group by randomly sampling $4,000$ adult non-sepsis patients from the MIMIC-III database after excluding those patients with sepsis related ICD-9 codes, having infection plus meeting SIRS criteria, and with missing observations in the qSOFA variables.

4.2. Observed Data

The observed data can be written in the form $\{Y_i, X_{i1}, X_{i2}, X_{i3}, u_{i1}, u_{i2}, ..., u_{i7}\}$ for $i = 1, 2, ..., N$, where $Y_i = -1$ or 1 indicates whether the i-th subject is diagnosed without or with sepsis. The triplet (X_{i1}, X_{i2}, X_{i3}) denotes the qSOFA variables of respiratory rate, systolic blood pressure, and GCS scores observed for the i-th subject within the first 24 hours after ICU admission. The $(u_{i1}, u_{i2}, ..., u_{i7})$ variables represent the demographic variables of age, gender, admission location, admission type, ethnicity, insurance, and marital status. The total number of patients N is $7,771$, including $3,771$ with sepsis and $4,000$ without sepsis. Variable definitions are provided in Table 1.

Table 1. Variables and corresponding definitions.

Variables	Definitions
Y_i	Sepsis indicator for patient i
X_{i1}	Maximum respiratory rate within first 24 hours for patient i
X_{i2}	Minimum systolic blood pressure within first 24 hours for patient i
X_{i3}	Minimum Glasgow Coma Scale (GCS) score within first 24 hours for patient i
u_{i1}	Age for patient i
u_{i2}	Sex for patient i
u_{i3}	Admission location for patient i (1=Emergency room admit; 0=Others)
u_{i4}	Admission type for patient i (1=Emergency and Urgent; 0=Others)
u_{i5}	Ethnicity for patient i (1=White; 2=Black; 3=Hispanic; 4=Others)
u_{i6}	Insurance type for patient i (1=Medicaid; 0= Self pay)
u_{i7}	Marital status for patient i (1=Married; 0=Others)

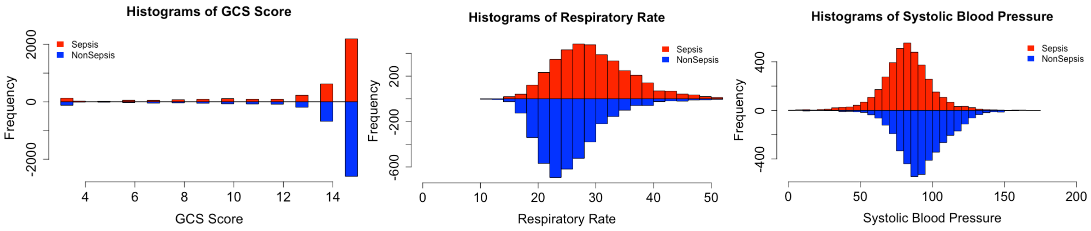

Figure 1. Histograms of qSOFA variables.

4.3. General Characteristics

Descriptive statistics are calculated for all variables of interest, and shown in Appendix C. The mean and standard deviation (SD) are compared across sepsis and non-sepsis groups using two-sample t-tests for continuous data. Categorical data are presented as counts and percentages, and they are compared between two groups by Fisher's exact test or Chi-square test. All of the selected variables are significantly correlated with the sepsis outcome ($p < 0.05$).

4.4. Exploratory Analysis on qSOFA Variables

Histograms for the qSOFA variables are shown in Figure 1. Sepsis patients tend to have lower GCS score, higher respiratory rate, and lower systolic blood pressure as compared to non-sepsis patients. However, the difference of GSC scores among the two cohorts is not strong, and most of the patients have GCS score 14 or 15 (conscious mental status). In our study, we focus on obtaining the personalized thresholds for respiratory rate and systolic blood pressure, and keep the constant cutoff 15 for GCS score, as discussed in Remark 2. The scatter plot of systolic blood pressure against respiratory rate for the two groups is presented in Figure 2. Most of the sepsis cohort lie on the lower right (high respiratory rate and low systolic blood pressure), while most of the non-sepsis cohort lie on the upper left (low respiratory rate and high systolic blood pressure). qSOFA criteria with constant thresholds may classify those patients well, however, the two cohorts overlap in the middle of the plot, and hence constant thresholds may lose power in identifying sepsis patients among them.

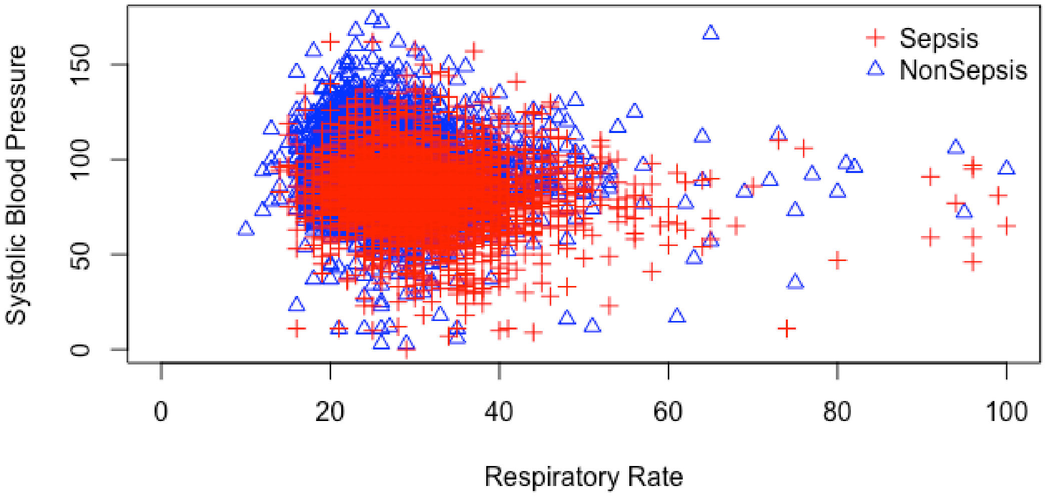

Figure 2. Scatter plot of systolic blood pressure against respiratory rate

5. Application to Sepsis Screening

In this section, we apply our proposed personalized threshold method to MIMIC-III data set for sepsis screening. For the purpose of comparison, we consider 6 baseline methods, including the original qSOFA criteria and five other standard methods, minP, Youden Index, Closest-to-(0,1), Concordance Probability, and Index of Union. Since our focus is on the prediction and classification, we use the random cross-validation to evaluate performances of all methods. Specifically, for each iteration, we randomly divide the real data into training (80%) and testing (20%) sets, and apply our method and the six baseline methods on the training set to obtain the thresholds for respiratory rate and systolic blood pressure, respectively, and then the obtained thresholds are used to classify subjects as septic or non-septic in the testing set to calculate the classification accuracy, sensitivity, and specificity. We then repeat this process 100 times, and report the averaged testing error statistics.

In our proposed personalized threshold method, we replace the constant thresholds for respiratory rate and systolic blood pressure by the personalized thresholds in the qSOFA criteria and keep the constant threshold of 15 for GCS score, i.e., the altered mental status occurs when GCS score is less than 15. The parameters $T = 30000$, $\alpha = 0.001$, $w_+ = 1$, and $w_- = 1$ are selected based on a grid search to maximize the averaged accuracy in a five-fold cross validation on the training data set.

For better presentations, we split this section into four subsection. Section 5.1 discusses the tuning parameters in our proposed method. Section 5.2 and Section 5.3 compare our method with other existing constant threshold methods and machine learning techniques, respectively. In Section 5.4, we provide the interpretation of our personalized threshold method and illustrate how to implement it in practice.

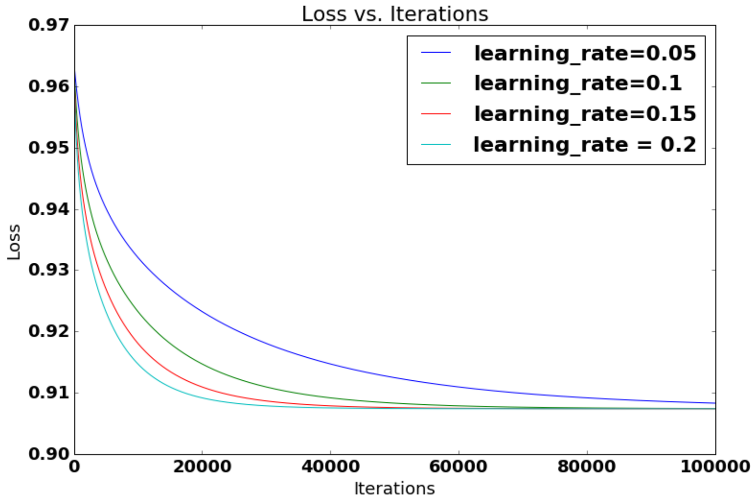

Figure 3. The weighted exponential loss versus the number of iterations with different learning rate α

5.1. Tuning parameters

In this subsection, we illustrate how the learning rate α and the total number of iterations T influence the speed of convergence in our gradient descent algorithm 1. In addition, we discuss the tradeoff between sensitivity and specificity from tuning the weights w_+ and w_- in the proposed cost function (6).

Figure 3 shows how the learning rate influenced the loss function (w_+ and w_- were set to be 1) over all of the training data. The learning rates of 0.05, 0.1, 0.15, and 0.2 were chosen to ensure the algorithm reached the global optimal. Indeed, larger learning rates tend to converge faster.

In our application, the parameters $T = 30000$ and $\alpha = 0.001$ are selected by grid search to maximize the averaged accuracy in a five-fold cross validation on the training data (80% of all data) . Notice that the algorithm with $T = 30000$ and $\alpha = 0.001$ might not converge on the training data, but it led to the highest cross validation classification accuracy. This observation is in line with "early stopping" concept in machine learning, which is to avoid overfitting when training a learner with an iterative method, such as gradient descent [13, 32, 34, 50, 52].

Figure 4 shows the averaged accuracy, sensitivity, and specificity in the five-fold cross validation for fixed $w_- = 1$ and different values for w_+. The learning rate and number of iterations are set as $\alpha = 0.001$ and $T = 30000$. The overall accuracy is essentially constant when we change the values of w_+. There is an obvious increasing trend of sensitivity as we increase w_+, since it penalizes more for the misclassification of patients with sepsis. This demonstrates the flexibility of Algorithm 1.

5.2. Comparison to qSOFA criteria with constant thresholds

We compare the personalized threshold method with the original qSOFA criteria and five other standard methods, minP, Youden Index, Closest-to-(0,1), Concordance Probability, and Index of Union. The averaged classification accuracy, sensitivity, and speci-

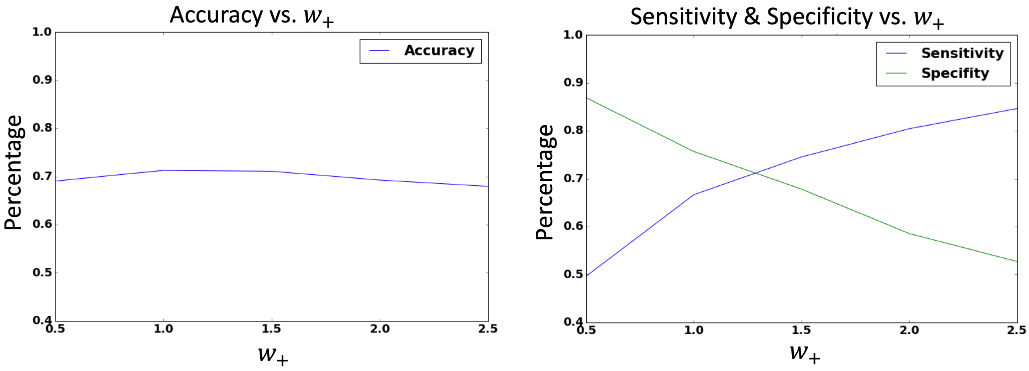

Figure 4. The accuracy (left), sensitivity and specificity (right) with different w_+'s and fixed $w_- = 1$.

Table 2. Overall accuracy

	Individual Biomarker		
Methods	Respiratory Rate	Systolic Blood Pressure	Combined in qSOFA
Personalized threshold	**0.6781**	**0.6889**	**0.6850**
qSOFA threshold	0.5464	0.5899	0.5853
minP	0.6456	0.6359	0.6497
Youden	0.6466	0.6371	0.6524
Closest-to-(0,1)	0.6467	0.6379	0.6536
Concordance Probability	0.6466	0.6380	0.6534
Index of Union	0.6468	0.6377	0.6528

ficity of different methods based on each individual biomarker and their combination in qSOFA criteria are compared as shown in Tables 2, 3, and 4, respectively. When using respiratory rate alone to classify, we identify a patient as having sepsis if the measurement is greater than the obtained threshold, while for systolic blood pressure, a sepsis patient is identified if it is less than the optimal threshold. When combining them in the qSOFA criteria, a patient is identified to be of high risk of developing sepsis if two of the following three criteria are satisfied: respiratory rate is greater than the obtained threshold, systolic blood pressure is less then its corresponding threshold, and GCS score is less than 15.

The personalized threshold method yields the largest overall accuracies. The accuracies for constant thresholds in the qSOFA criteria are the lowest. The five standard methods, minP, Youden Index, Closest-to-(0,1), Concordance Probability, and Index of Union, have similar classification accuracies, which are higher than those for qSOFA with constant cutoffs but lower than our personalized method.

In Table 3, qSOFA with constant thresholds corresponds to the highest sensitivities: 92.94% for respiratory rate, 88.41% for systolic blood pressure, and 89.66% for their combination, while using our personalized threshold method, the sensitivities are 65.09%, 69.24%, and 63.54%, respectively. Other standard methods have sensitivities ranging from 61.79% to 62.41% for respiratory rate, from 58.16% to 63.28% for systolic blood pressure, and from 56.47% to 59.47% for them combined.

The classification specificities are detailed in Table 4. qSOFA with constant thresholds has the lowest specificities (18.51% for respiratory rate, 31.25% for systolic blood pressure, and 29.17% for their combination in qSOFA), while using the personalized thresholds, the specificities are increased to 70.38% for respiratory rate, 68.57% for systolic blood pressure, and 73.18% for their combination in qSOFA. The specificities

Table 3. Sensitivity

Methods	Individual Biomarker		Combined in qSOFA
	Respiratory Rate	Systolic Blood Pressure	
Personalized threshold	0.6509	0.6924	0.6354
qSOFA threshold	**0.9294**	**0.8841**	**0.8966**
minP	0.6179	0.5816	0.5647
Youden	0.6240	0.6156	0.5858
Closest-to-(0,1)	0.6238	0.6328	0.5947
Concordance Probability	0.6241	0.6296	0.5931
Index of Union	0.6240	0.6210	0.5886

Table 4. Specificity

Methods	Individual Biomarker		Combined in qSOFA
	Respiratory Rate	Systolic Blood Pressure	
Personalized threshold	**0.7038**	0.6857	**0.7318**
qSOFA threshold	0.1851	0.3125	0.2917
minP	0.6718	**0.6871**	0.7301
Youden	0.6680	0.6574	0.7152
Closest-to-(0,1)	0.6680	0.6426	0.7092
Concordance Probability	0.6681	0.6458	0.7103
Index of Union	0.6679	0.6534	0.7135

of other standard methods range from 66.79% to 67.18% for respiratory rate, from 64.26% to 68.71% for systolic blood pressure, and from 70.92% to 73.01% for them combined in qSOFA.

In general, the personalized threshold method yields the largest prediction accuracy and the best balance of sensitivity and specificity. Note that in this application, we choose the parameters $w_+ = w_- = 1$ in our cost function, however, we can adjust the balance between sensitivity and specificity by choosing different values of w_+ and w_-.

5.3. Comparison with other machine learning techniques

We also apply logistic regression and AdaBoosting to predict sepsis using the qSOFA variables respiratory rate, systolic blood pressure, and GCS score, together with baseline demographic variables age, sex, admission location, admission type, ethnicity, insurance, and marital status. As described previously, the models are built on the training set (80%) and then applied on the testing set (20%) to classify the subjects with 100 repetitions of randomly splitting. The comparisons between qSOFA (constant and personalized) with logistic regression and AdaBoosting are presented in Table 5.

The averaged classification accuracy of the personalized qSOFA is close to those obtained from logistic regression and AdaBoosting. The averaged sensitivity of the personalized qSOFA is less than those from the machine learning methods, but the specificity of the personalized qSOFA is the largest. In general, the personalized qSOFA is comparable to the more difficult-to-interpret machine learning methods.

Using the personalized qSOFA for sepsis screening is also efficient, since there is only a one time calculation of the personalized thresholds for each patient. Machine learning methods, on the other hand, require an update each time there is a new observation. In addition, there are challenges in implementing machine leaning methods in a practical way for sepsis screening in the intensive care units. First, sepsis screening is essentially to test the null hypothesis of no sepsis repeatedly and frequently whenever there is a new observation, and thus standard machine learning methods

Table 5. Comparison with other machine learning techniques

Methods	Overall Accuracy	Sensitivity	Specificity
Personalized qSOFA	0.6850	0.6354	**0.7318**
Logistic Regression	**0.7182**	**0.7277**	0.7094
AdaBoosting	0.7171	0.7146	0.7196

will yield significantly large false alarm rate due to the multiple tests over the time domain. Second, machine learning methods are often difficult to interpret, making it impossible for physicians, nurses, and patients, etc. to quickly interpret and assess the corresponding results. In other words, in the current healthcare environment, even the most complicated machine learning methods are still incapable of fully substituting for health professionals in the context of Sepsis screening. Meanwhile, the qSOFA is a control-chart based statistical method that is widely used by physicians and nurses. Our approach is to combine machine learning with control chart to develop personalized qSOFA method that can be easily implemented, manipulated and interpreted by physicians and nurses. It does not require deep statistical training or any advanced equipment supports. Once fixed at admission, the personalized thresholds can be used exactly the same way as the currently used constant thresholds in the qSOFA.

5.4. Interpretation and Implementation of Personalized qSOFA

After applying the proposed model to the data, we obtain the estimated personalized thresholds for each individual. Here, we are going to explore how the estimated thresholds are related to the patients' baseline information. Figure 5 plots the predicted personalized threshold for respiratory rate against age. The personalized cutoffs show a decreasing trend as age increases. This suggests that for older patients with lower respiratory rate, the threshold should be set lower than that set for their younger counterparts. Patients with the same age may differ in other baseline characteristics, however, which would lead to different recommended personalized thresholds to improve overall classification accuracy.

An advantage of our approach is that we only use easily accessible demographic variables to estimate the personalized thresholds. Therefore, the threshold can be calculated and fixed as soon as the patient is admitted. The personalized threshold can be treated and manipulated in exactly the same way as the constant threshold in sepsis screening. Figure 6 illustrates this. It shows two selected examples of screening for respiratory rate using the personalized threshold and the constant threshold 22 in qSOFA criteria: one for a non-sepsis patient, and the other for a sepsis patient. As can be seen from the plots, the respiratory rate is frequently measured by medical machine about once every hour. We keep monitoring the measurements and raise an alarm if any one of them reaches the personalized threshold, which is a fixed line for each patient just as the constant qSOFA threshold but are different for different subjects. However, it is not straightforward for any machine learning method to estimate the thresholds for such regularly measured biomarkers. Therefore, to use the machine learning methods for monitoring, one has to update the prediction results each time there is a new observation, which make it hard to implement.

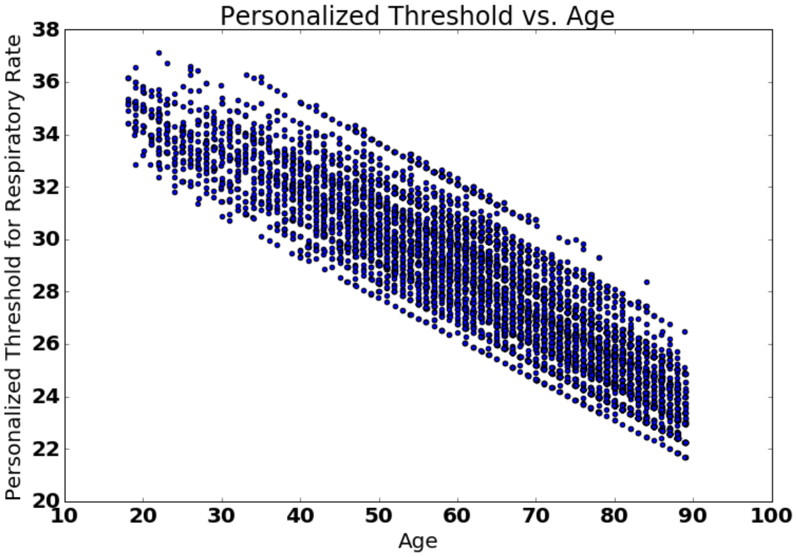

Figure 5. The predicted personalized threshold for respiratory rate against age.

6. Conclusions

Sepsis is difficult to identify and diagnose, and unfortunately, there is not a validated standard diagnostic test for sepsis at present. The idea of "screening" plus " assessing" as recommended by the 2016 Task Force [39] is attractive, but there are some drawbacks in using the qSOFA score. The most important limitation is that it has rather low specificity in identifying sepsis patients in ICU.

We developed a personalized threshold method that is able to adjust the thresholds in the qSOFA criteria based on the subject's baseline characteristics, age, sex, admission location, admission type, ethnicity, insurance, and marital status. We assumed the personalized thresholds were a linear function of those demographic variables and developed a boosting inspired method to obtain the personalized thresholds for efficient screening of sepsis. The gradient descent algorithm was applied to obtain the unknown parameters in the linear function to calculate the personalized thresholds. The method provided an efficient personalized monitoring, enabling the subject-specific intervention in early stages of sepsis, which could significantly reduce the mortality rate in the future.

Our method was applied to the MIMIC-III data (ICU populations) to find the optimal personalized thresholds for the qSOFA variables of respiratory rate and systolic blood pressure. The constant thresholds in qSOFA were replaced by those obtained from our method for classifying patients as septic or non-septic. We compared personalized qSOFA with the original qSOFA criteria and five other standard methods to obtain the optimal constant threshold for a single biomarker (minP, Youden Index, Closest-to-(0,1), Concordance Probability, and Index of Union). Our method yielded the largest overall testing accuracy for identifying sepsis patients. The constant qSOFA had a high sensitivity but a very low specificity in ICU populations, while our personalized qSOFA yielded a better balance. In general, the five standard methods performed better than constant qSOFA but worse than the personalized qSOFA.

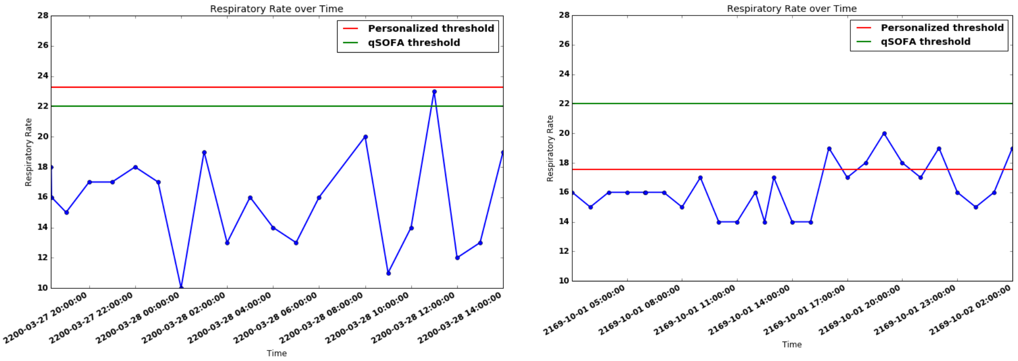

Figure 6. Left: Screening for non-sepsis patient. Right: Screening for sepsis patient.

The personalized qSOFA has comparable performance to logistic regression and AdaBoosintg, and has the benefit of being easily implemented and interpreted by physicians and nurses. Our personalized qSOFA method only requires a one time calculation of the personalized threshold for each patient, and once fixed at admission, the personalized threshold can be used exactly the same way as the constant one. In addition, the balance of sensitivity and specificity can be easily adjusted in our method by tuning the weighting parameters.

There are several limitations that should be mentioned. When obtaining the personalized thresholds for respiratory rate and systolic blood pressure in qSOFA criteria, we did not jointly estimate them, and therefore ignored the possible correlation between them. Although we did preliminary analysis and found that the baseline characteristic variables age, sex, admission location, admission type, ethnicity, insurance, and marital status were significantly correlated with sepsis outcome, we did not consider variable selection when putting them in the model to estimate the thresholds. In future studies, more clinical variables may become available, and therefore performing variable selection will likely be necessary. Finally, we only focused on the scenario where the threshold is a linear function of those baseline characteristics, which might not always be the case.

It is worth noting that although we demonstrated the use of personalized thresholds for sepsis screening, the general approach can be applied to other clinical screening applications. Examples include personalizing the HAVOC score, a clinical score for predicting atrial fibrillation in patients with cyptogenic stroke or transient ischemic attack [22] and the Fong clinical risk score for predicting colorectal cancer recurrence [7].

Acknowledgments

The authors are grateful to the Associate Editor and two anonymous reviewers for their constructive comments that greatly improved the quality and presentation of this article. This research was supported in part by NSF grants CMMI-1362876 and DMS-1830344 through Georgia Institute of Technology, and in part by the National Center for Advancing Translational Sciences of the National Institutes of Health under Award Number UL1TR000454.

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*, Chest 101 (1992), pp. 1644–1655.

[2] M. Brabrand, U. Havshøj, and C.A. Graham, *Validation of the qsofa score for identification of septic patients: a retrospective study*, European Journal of Internal Medicine 36 (2016), pp. e35–e36.

[3] A. Castellanos-Ortega, B. Suberviola, L.A. García-Astudillo, F. Ortiz, J. Llorca, and M. Delgado-Rodríguez, *Late compliance with the sepsis resuscitation bundle: impact on mortality*, Shock 36 (2011), pp. 542–547.

[4] M. Collins, R.E. Schapire, and Y. Singer, *Logistic regression, adaboost and bregman distances*, Machine Learning 48 (2002), pp. 253–285.

[5] M. Dorsett, M. Kroll, C.S. Smith, P. Asaro, S.Y. Liang, and H.P. Moy, *qsofa has poor sensitivity for prehospital identification of severe sepsis and septic shock*, Prehospital Emergency Care 21 (2017), pp. 489–497.

[6] S.M. Fernando, A. Tran, M. Taljaard, W. Cheng, B. Rochwerg, A.J. Seely, and J.J. Perry, *Prognostic accuracy of the quick sequential organ failure assessment for mortality in patients with suspected infection*, Ann Intern Med 168 (2018), pp. 266–75.

[7] Y. Fong, J. Fortner, R.L. Sun, M.F. Brennan, and L.H. Bumgart, *Clinical score for predicting recurrence after hepatic resection for metastatic colorectal cancer*, Annals of Surgery 230 (1999), p. 309.

[8] Y. Freund and R.E. Schapire, *A decision-theoretic generalization of on-line learning and an application to boosting*, Journal of Computer and System Sciences 55 (1997), pp. 119–139.

[9] J. Friedman, T. Hastie, and R. Tibshirani, *The elements of statistical learning*, Vol. 1, Springer Series in Statistics New York, 2001.

[10] D.F. Gaieski, J.M. Edwards, M.J. Kallan, and B.G. Carr, *Benchmarking the incidence and mortality of severe sepsis in the united states*, Critical Care Medicine 41 (2013), pp. 1167–1174.

[11] E.J. Giamarellos-Bourboulis, T. Tsaganos, I. Tsangaris, M. Lada, C. Routsi, D. Sinapidis, M. Koupetori, M. Bristianou, G. Adamis, K. Mandragos, *et al.*, *Validation of the new sepsis-3 definitions: proposal for improvement in early risk identification*, Clinical Microbiology and Infection 23 (2017), pp. 104–109.

[12] H.M. Giannini, C. Chivers, M. Draugelis, A. Hanish, B. Fuchs, P. Donnelly, M. Lynch, L. Meadows, S.J. Parker, W.D. Schweickert, *et al.*, *Development and implementation of a machine-learning algorithm for early identification of sepsis in a multi-hospital academic healthcare system*, in *D15. CriticalL Care: Do We Have a Crystal Ball? Predicting Clinical Deterioration and Outcome in Critically Ill Patients*, American Thoracic Society, 2017, pp. A7015–A7015.

[13] F. Girosi, M. Jones, and T. Poggio, *Regularization theory and neural networks architectures*, Neural Computation 7 (1995), pp. 219–269.

[14] E. Gyang, L. Shieh, L. Forsey, and P. Maggio, *A nurse-driven screening tool for the early identification of sepsis in an intermediate care unit setting*, Journal of Hospital Medicine 10 (2015), pp. 97–103.

[15] T. Hastie, S. Rosset, J. Zhu, and H. Zou, *Multi-class adaboost*, Statistics and its Interface 2 (2009), pp. 349–360.

[16] K.E. Henry, D.N. Hager, P.J. Pronovost, and S. Saria, *A targeted real-time early warning score (trewscore) for septic shock*, Science Translational Medicine 7 (2015), pp. 299ra122–299ra122.

[17] F. Jaimes, J. Farbiarz, D. Alvarez, and C. Martínez, *Comparison between logistic regression and neural networks to predict death in patients with suspected sepsis in the emergency room*, Critical Care 9 (2005), p. R150.

[18] A.E. Johnson, T.J. Pollard, L. Shen, H.L. Li-wei, M. Feng, M. Ghassemi, B. Moody, P.

Szolovits, L.A. Celi, and R.G. Mark, *Mimic-iii, a freely accessible critical care database*, Scientific Data 3 (2016), p. 160035.

[19] D.B. Knox, M.J. Lanspa, K.G. Kuttler, S.C. Brewer, and S.M. Brown, *Phenotypic clusters within sepsis-associated multiple organ dysfunction syndrome*, Intensive Care Medicine 41 (2015), pp. 814–822.

[20] D. Kojic, B.H. Siegler, F. Uhle, C. Lichtenstern, P.P. Nawroth, M.A. Weigand, S. Hofer, and T. Brenner, *Are there new approaches for diagnosis, therapy guidance and outcome prediction of sepsis?*, World Journal of Experimental Medicine 5 (2015), p. 50.

[21] E.D. Krebs, T.E. Hassinger, C.A. Guidry, P.S. Berry, N.R. Elwood, and R.G. Sawyer, *Non-utility of sepsis scores for identifying infection in surgical intensive care unit patients*, The American Journal of Surgery (2018).

[22] C. Kwong, A.Y. Ling, M.H. Crawford, S.X. Zhao, and N.H. Shah, *A clinical score for predicting atrial fibrillation in patients with cryptogenic stroke or transient ischemic attack*, Cardiology 138 (2017), pp. 133–140.

[23] M.M. Levy, M.P. Fink, J.C. Marshall, E. Abraham, D. Angus, D. Cook, J. Cohen, S.M. Opal, J.L. Vincent, G. Ramsay, *et al.*, *2001 sccm/esicm/accp/ats/sis international sepsis definitions conference*, Intensive Care Medicine 29 (2003), pp. 530–538.

[24] X. Liu, *Classification accuracy and cut point selection*, Statistics in Medicine 31 (2012), pp. 2676–2686.

[25] S. Mani, A. Ozdas, C. Aliferis, H.A. Varol, Q. Chen, R. Carnevale, Y. Chen, J. Romano-Keeler, H. Nian, and J.H. Weitkamp, *Medical decision support using machine learning for early detection of late-onset neonatal sepsis*, Journal of the American Medical Informatics Association 21 (2014), pp. 326–336.

[26] R. Miller and D. Siegmund, *Maximally selected chi square statistics*, Biometrics (1982), pp. 1011–1016.

[27] L.J. Moore, S.L. Jones, L.A. Kreiner, B. McKinley, J.F. Sucher, S.R. Todd, K.L. Turner, A. Valdivia, and F.A. Moore, *Validation of a screening tool for the early identification of sepsis*, Journal of Trauma and Acute Care Surgery 66 (2009), pp. 1539–1547.

[28] S. Nemati, A. Holder, F. Razmi, M.D. Stanley, G.D. Clifford, and T.G. Buchman, *An interpretable machine learning model for accurate prediction of sepsis in the icu*, Critical Care Medicine 46 (2018), pp. 547–553.

[29] J. Perkins Neil and F. Schisterman Enrique, *The inconsistency of optimal cut-points using two roc based criteria*, American Journal of Epidemiology 163 (2006), pp. 670–675.

[30] C.C. Polito, A. Isakov, A.H. Yancey, D.K. Wilson, B.A. Anderson, I. Bloom, G.S. Martin, and J.E. Sevransky, *Prehospital recognition of severe sepsis: development and validation of a novel ems screening tool*, American Journal of Emergency Medicine 33 (2015), pp. 1119–1125.

[31] T. Pollard and A. Johnson III, *The mimic-iii clinical database* (2017).

[32] L. Prechelt, *Early stopping-but when?*, in *Neural Networks: Tricks of the trade*, Springer, 1998, pp. 55–69.

[33] E.P. Raith, A.A. Udy, M. Bailey, S. McGloughlin, C. MacIsaac, R. Bellomo, and D.V. Pilcher, *Prognostic accuracy of the sofa score, sirs criteria, and qsofa score for in-hospital mortality among adults with suspected infection admitted to the intensive care unit*, JAMA 317 (2017), pp. 290–300.

[34] G. Raskutti, M.J. Wainwright, and B. Yu, *Early stopping and non-parametric regression: an optimal data-dependent stopping rule*, The Journal of Machine Learning Research 15 (2014), pp. 335–366.

[35] E. Rezende, J.M. Silva Junior, A.M. Isola, E.V. Campos, C.P. Amendola, and S.L. Almeida, *Epidemiology of severe sepsis in the emergency department and difficulties in the initial assistance*, Clinics 63 (2008), pp. 457–464.

[36] C.W. Seymour, V.X. Liu, T.J. Iwashyna, F.M. Brunkhorst, T.D. Rea, A. Scherag, G. Rubenfeld, J.M. Kahn, M. Shankar-Hari, M. Singer, *et al.*, *Assessment of clinical criteria for sepsis: for the third international consensus definitions for sepsis and septic shock (sepsis-3)*, JAMA 315 (2016), pp. 762–774.

[37] D.W. Shimabukuro, C.W. Barton, M.D. Feldman, S.J. Mataraso, and R. Das, *Effect of a machine learning-based severe sepsis prediction algorithm on patient survival and hospital length of stay: a randomised clinical trial*, BMJ Open Respiratory Research 4 (2017), p. e000234.

[38] A.F. Shorr, S.T. Micek, W.L. Jackson, and M.H. Kollef, *Economic implications of an evidence-based sepsis protocol: can we improve outcomes and lower costs?*, Critical Care Medicine 35 (2007), pp. 1257–1262.

[39] M. Singer, C.S. Deutschman, C.W. Seymour, M. Shankar-Hari, D. Annane, M. Bauer, R. Bellomo, G.R. Bernard, J.D. Chiche, C.M. Coopersmith, *et al.*, *The third international consensus definitions for sepsis and septic shock (sepsis-3)*, JAMA 315 (2016), pp. 801–810.

[40] M. Singer and M. Shankar-Hari, *qsofa, cue confusion*, Annals of internal medicine 168 (2018), pp. 293–295.

[41] E.R. Swenson, N.D. Bastian, and H.B. Nembhard, *Data analytics in health promotion: Health market segmentation and classification of total joint replacement surgery patients*, Expert Systems with Applications 60 (2016), pp. 118–129.

[42] D. Talmor, D. Greenberg, M.D. Howell, A. Lisbon, V. Novack, and N. Shapiro, *The costs and cost-effectiveness of an integrated sepsis treatment protocol*, Critical Care Medicine 36 (2008), pp. 1168–1174.

[43] 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 (2010), pp. 19–25.

[44] C. Tucker, Y. Han, H. Black Nembhard, W.C. Lee, M. Lewis, N. Sterling, and X. Huang, *A data mining methodology for predicting early stage parkinson's disease using non-invasive, high-dimensional gait sensor data*, IIE Transactions on Healthcare Systems Engineering 5 (2015), pp. 238–254.

[45] C.S. Tucker, I. Behoora, H.B. Nembhard, M. Lewis, N.W. Sterling, and X. Huang, *Machine learning classification of medication adherence in patients with movement disorders using non-wearable sensors*, Computers in Biology and Medicine 66 (2015), pp. 120–134.

[46] S. Tusgul, P.N. Carron, B. Yersin, T. Calandra, and F. Dami, *Low sensitivity of qsofa, sirs criteria and sepsis definition to identify infected patients at risk of complication in the prehospital setting and at the emergency department triage*, Scandinavian Journal of Trauma, Resuscitation and Emergency Medicine 25 (2017), p. 108.

[47] I. Unal, *Defining an optimal cut-point value in roc analysis: An alternative approach*, Computational and Mathematical Methods in Medicine 2017 (2017).

[48] N. Villegas and L.J. Moore, *Sepsis screening: Current evidence and available tools*, Surgical Infections 19 (2018), pp. 126–130.

[49] J.L. Vincent, S.M. Opal, J.C. Marshall, and K.J. Tracey, *Sepsis definitions: time for change*, Lancet 381 (2013), p. 774.

[50] Y. Yao, L. Rosasco, and A. Caponnetto, *On early stopping in gradient descent learning*, Constructive Approximation 26 (2007), pp. 289–315.

[51] W.J. Youden, *Index for rating diagnostic tests*, Cancer 3 (1950), pp. 32–35.

[52] T. Zhang, B. Yu, *et al.*, *Boosting with early stopping: Convergence and consistency*, The Annals of Statistics 33 (2005), pp. 1538–1579.

[53] X. Zhou, Y. Ye, and G. Tang, *Sepsis screening tools in the era of sepsis 3.0*, Surgical Infections 19 (2018), pp. 553–553.

Appendix A. Proof of Proposition 3.1.

It suffices to show that the weighted exponential loss function $J(\boldsymbol{\beta})$ in (6) is a convex function with respect to $\boldsymbol{\beta}$. Recall that $\boldsymbol{\beta} = [\beta_0, \dots, \beta_q]^T$ and without loss of generality, we can rewrite that $\boldsymbol{u_i} = [u_{i0}, \dots, u_{iq}]^T$ with $u_{i0} = 1$, for any $i \in \{1, \dots, N\}$. It is sufficient to prove that for any $\boldsymbol{z} \in \mathbb{R}^{q+1}$, $\boldsymbol{z}^T[\nabla^2 J(\boldsymbol{\beta})]\boldsymbol{z} \geq 0$, where $\nabla^2 J(\boldsymbol{\beta})$ is the Hessian matrix of $J(\boldsymbol{\beta})$. First, we compute the first order derivative of $J(\boldsymbol{\beta})$ with respect to $\boldsymbol{\beta}$. For any $j \in \{0, \dots, q\}$, we have

$$\frac{\partial J(\boldsymbol{\beta})}{\partial \beta_j} = \frac{1}{N}\sum_{i=1}^{N}\Big(w_{Y_i} \cdot e^{-Y_i f_i} \cdot Y_i u_{ij}\Big)$$

Second, we compute the second order derivatives. For any $j \in \{0, \dots, q\}$ and any $k \in \{0, \dots, q\}$,

$$\frac{\partial^2 J(\boldsymbol{\beta})}{\partial \beta_j^2} = \frac{1}{N}\sum_{i=1}^{N}\Big(w_{Y_i} \cdot e^{-Y_i f_i} \cdot (Y_i u_{ij})^2\Big), \quad \frac{\partial^2 J(\boldsymbol{\beta})}{\partial \beta_j \beta_k} = \frac{1}{N}\sum_{i=1}^{N}\Big(w_{Y_i} \cdot e^{-Y_i f_i} \cdot Y_i^2 u_{ij} u_{ik}\Big).$$

Therefore, for any $\boldsymbol{z} = [z_0, \dots, z_q]^T$, we have

$$\begin{aligned}
\boldsymbol{z}^T[\nabla^2 J(\boldsymbol{\beta})]\boldsymbol{z} &= \frac{1}{N}\sum_{j=0}^{q}\sum_{k=0}^{q}\sum_{i=1}^{N}\Big(w_{Y_i} Y_i^2 \cdot e^{-Y_i f_i} \cdot z_j z_k u_{ij} u_{ik}\Big) \\
&= \frac{1}{N}\sum_{i=1}^{N} w_{Y_i} Y_i^2 \cdot e^{-Y_i f_i} \sum_{j=0}^{q}\sum_{k=0}^{q}(z_j z_k u_{ij} u_{ik}) \qquad \text{(A1)} \\
&= \frac{1}{N}\sum_{i=1}^{N} w_{Y_i} Y_i^2 \cdot e^{-Y_i f_i} (\boldsymbol{z}^T \boldsymbol{u_i})^2.
\end{aligned}$$

Since $w_{Y_i} \geq 0$ for all $i \in \{1, \dots, N\}$, we have that $\boldsymbol{z}^T[\nabla^2 J(\boldsymbol{\beta})]\boldsymbol{z} \geq 0$ for all $\boldsymbol{z}$. □

Appendix B. Flowchart of Cohort Selection from MIMIC-III Data

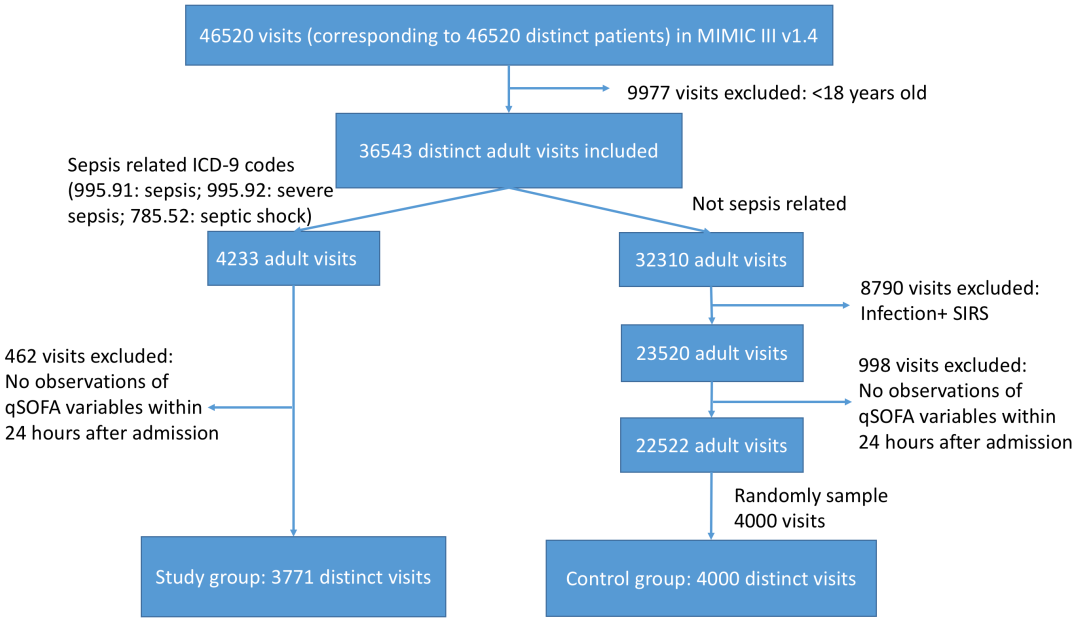
46520 visits (corresponding to 46520 distinct patients) in MIMIC III v1.4
9977 visits excluded: <18 years old
36543 distinct adult visits included
Sepsis related ICD-9 codes (995.91: sepsis; 995.92: severe sepsis; 785.52: septic shock)
Not sepsis related
4233 adult visits
32310 adult visits
8790 visits excluded: Infection+ SIRS
23520 adult visits
462 visits excluded: No observations of qSOFA variables within 24 hours after admission
998 visits excluded: No observations of qSOFA variables within 24 hours after admission
22522 adult visits
Randomly sample 4000 visits
Study group: 3771 distinct visits
Control group: 4000 distinct visits

Figure B1. Flowchart of cohort selection

Appendix C. Descriptive Statistics of Interested Variables

Table C1. Descriptive statistics of interested variables.

Variables		Sepsis	Non-Sepsis	p-Values
		($N = 3771$)	($N = 4000$)	
Max Respiratory Rate, mean (SD)		30.6 (8.9)	26.5 (7.4)	7.6×10^{-106}
Min Systolic Blood Pressure, mean (SD)		82.3 (17.4)	93.5 (18.1)	7.2×10^{-163}
Altered mental status, count (%) (Min GCS $<$ 15)		1580 (41.9)	1407 (35.2)	1.2×10^{-9}
Age, mean (SD)		65.3 (15.6)	61.6 (16.6)	9.9×10^{-25}
Sex, count (%)	Male	2148 (57.0)	2397 (59.9)	8.1×10^{-3}
	Female	1623 (43.0)	1603 (40.1)	
Admission location, count (%)	Emergency Room	1767 (46.9)	1528 (38.2)	1.3×10^{-14}
	Others	2004 (53.1)	2472 (61.8)	
Admission type, count (%)	Emergency and Urgent	3635 (96.4)	3110 (77.8)	0
	Others	136 (3.6)	890 (22.3)	
Ethnicity, count (%)	White	2728 (72.3)	2830 (70.8)	3.6×10^{-6}
	Black	349 (9.3)	280 (7)	
	Hispanic	125 (3.3)	143 (3.6)	
	Others	569 (15.1)	747 (18.7)	
Insurance, count (%)	Medicaid	2703 (71.7)	2302 (57.6)	0
	Self pay	1068 (28.3)	1698 (42.5)	
Marital status, count (%)	Married	1719 (45.6)	2075 (51.9)	3.0×10^{-8}
	Others	2052 (54.4)	1925 (48.1)	