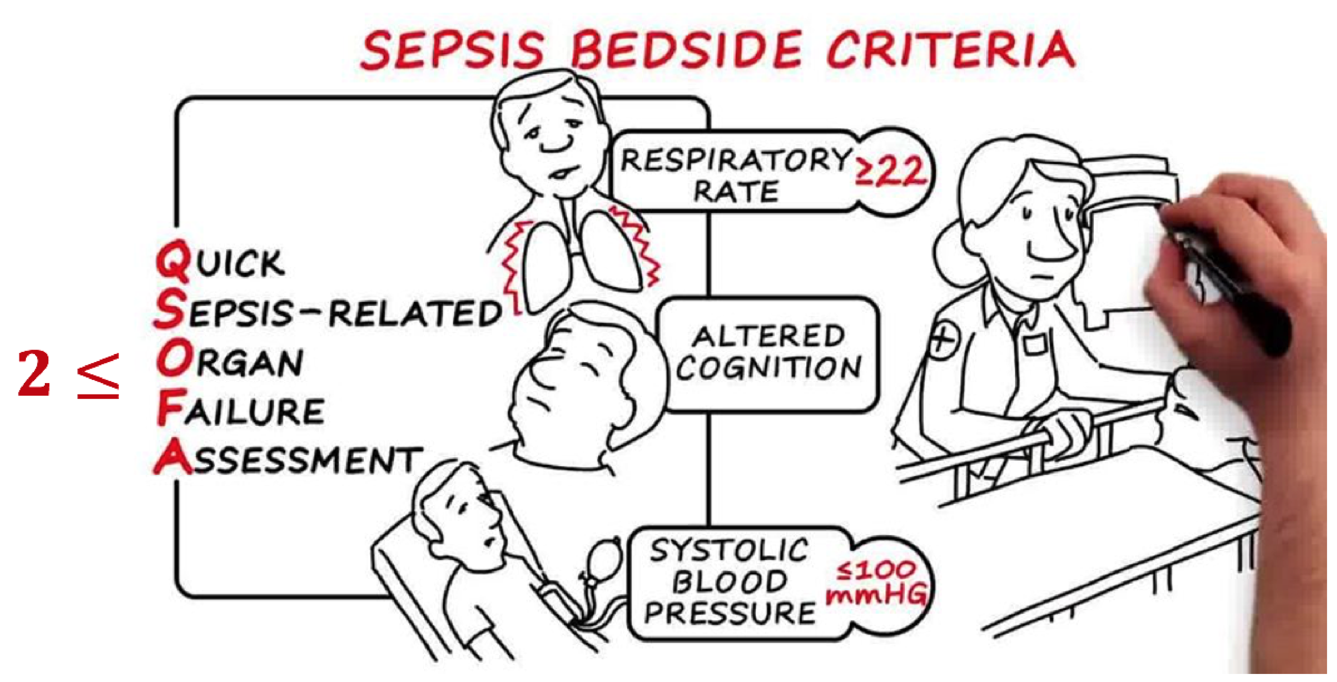
A Personalized Boosting Screening Method and Its Application to Sepsis

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

Motivation and Objective

- The natural mortality rate for **sepsis** is between 25% and 50%. Nearly half of patients who die in hospitals are **septic**
- In 2016, a task force committee recommended **screening for sepsis** by quick Sepsis-related Organ Failure Assessment (**qSOFA**), which uses the constant thresholds in decision making.



- **Our Objective:** Develop a **personalized sepsis screening** method that depends on a patient's baseline characteristics such as age, sex, admission location, etc.

Problem Formulation

- **Data:** $(Y_i, X_i, \boldsymbol{u}_i)$, for $i = 1, \cdots, N$, where $Y_i \in \{-1, 1\}$ is the binary outcome, X_i is the biomarker (e.g., blood pressure, respiratory rate, etc.), and $\boldsymbol{u}_i \in R^q$ are baseline characteristics.
- **Classification Rule:** Predict Y_i by

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

Here we assume

$$c_i(\boldsymbol{u}_i) = \boldsymbol{u}_i^T \boldsymbol{\beta},$$

for some unknown parameters $\boldsymbol{\beta} = (\beta_0, \beta_1, \cdots, \beta_q)^T$.

- **Question**: Estimate $\boldsymbol{\beta}$ so as to minimize misclassification rate.
  - The function $\text{sign}(f)$ is not continuous.
  - The $0-1$ loss function $I(Y \neq \text{sign}(f))$ is non-smoothing.
  - The consequences of misclassifying sepsis and non-sepsis patients are different.

Existing Methods

qSOFA considers the constant threshold, i.e., $c_i(\boldsymbol{u}_i) \equiv c$.

Several approaches to find suitable constant thresholds:

- **minP Approach:** Maximizing the standard chi-square statistic
- **Youden Index:** Maximizing the sum of sensitivity and specificity
- **Closest-to-(0,1) Criterion:** The "optimal" threshold is defined as the point on the ROC curve closest to (0,1)

Our Proposed Method: Personalized Threshold

Key ideas in our proposed method

- We define the threshold $c_i(\boldsymbol{u}_i)$ as a function of the individual subjects' characteristics $\boldsymbol{u}_i$
- We borrow the idea of **boosting** to replace the $0-1$ loss function $I(Y \neq \text{sign}(f))$ by the exponential loss $\exp(-Yf)$
- Introducing 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.

Parameter Estimation as Optimization Problem

We propose to estimate the $(q+1)$-dimensional parameter $\boldsymbol{\beta}$ by minimizing the training error under the weighted exponential loss function:

$$\begin{aligned} J(\boldsymbol{\beta}) &= \frac{1}{N}\sum_{i=1}^{N}\left(w_+ \cdot e^{-Y_i f_i} \cdot I(Y_i = 1) + w_- \cdot e^{-Y_i f_i} \cdot I(Y_i = -1)\right), \\ &= \frac{1}{N}\sum_{i=1}^{N}\left(w_{Y_i} \cdot e^{-Y_i f_i}\right) = \frac{1}{N}\sum_{i=1}^{N}\left(w_{Y_i} \cdot e^{-Y_i\left(X_i - \boldsymbol{\beta}^T \boldsymbol{u}_i\right)}\right) \end{aligned} \quad (1)$$

where $f_i \triangleq f(X_i, \boldsymbol{u}_i) = X_i - c_i(\boldsymbol{u}_i)$ and $w_{Y_i} \triangleq (w_+ \cdot (Y_i + 1) + w_- \cdot (1 - Y_i))/2$.

Computational algorithm: Gradient descent

Algorithm 1 Gradient Descent Using Weighted Exponential Loss

Require: $\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{U}, N, w_- > 0, w_+ > 0, \alpha, T$

Initialization: $\beta_i \leftarrow 0 \ \forall i \in \{0, 1, 2, ..., q\}$, $\boldsymbol{W} = (w_+(\boldsymbol{Y}+1) + w_-(1-\boldsymbol{Y}))/2$

for all $t = 1, 2, \ldots, T$ **do**

Forward Propagation:
- $\boldsymbol{f} = \boldsymbol{X} - \boldsymbol{U}^T\boldsymbol{\beta}$
- $\boldsymbol{L} = \exp(-\boldsymbol{Y} * \boldsymbol{f})$ {$*$: Element-wise product}
- $\boldsymbol{J} = \frac{1}{N}\boldsymbol{L}^T\boldsymbol{W}$

Backward Propagation:
- $d\boldsymbol{f} = -\frac{1}{N}\boldsymbol{W} * \boldsymbol{Y} * \exp(-\boldsymbol{Y} * \boldsymbol{f})$
- $d\boldsymbol{\beta} = -\boldsymbol{U}d\boldsymbol{f}$
- $\boldsymbol{\beta} \leftarrow \boldsymbol{\beta} - \alpha d\boldsymbol{\beta}$

end for

$\hat{c} = \boldsymbol{U}^T\boldsymbol{\beta}$

Proposition: The objective function $J(\boldsymbol{\beta})$ is convex with respect to $\boldsymbol{\beta}$, and thus the gradient descent algorithm converges to the global optimum if the learning rate α is small enough and the optimization steps T is long enough.

Data Set

Medical Information Mart for Intensive Care III (MIMIC-III) database (version 1.4)

Study Population: $3,771$ sepsis patients $(Y = 1)$; $4,000$ non-sepsis patients $(Y = -1)$

qSOFA variables (X): respiratory rate (RR), systolic blood pressure (sysBP), and GCS scores (we keep the constant cutoff 15 for GCS score, since GCS score is a discrete variable)

Baseline characteristics (u's): age, gender, admission location, admission type, ethnicity, insurance, and marital status

Application to Sepsis Screening

Comparison to qSOFA thresholds and existing methods

The parameters $T = 30000$, $\alpha = 0.001$, $w_+ = 1$, and $w_- = 1$ were selected based on a grid search to maximize the prediction accuracy.

Overall Accuracy

	Personalized	qSOFA	minP	Youden	Closest-(0,1)
RR	**0.7045**	0.6637	0.6798	0.6849	0.6817
sysBP	**0.7029**	0.5919	0.6511	0.6514	0.6524
Combined in qSOFA criterion	**0.7093**	0.6204	0.6852	0.6905	0.6889

Sensitivity

	Personalized	qSOFA	minP	Youden	Closest-(0,1)
RR	**0.6731**	0.4203	0.5664	0.6238	0.6386
sysBP	**0.7160**	0.2354	0.6612	0.6621	0.6575
Combined in qSOFA criterion	**0.6668**	0.2903	0.5676	0.6069	0.6128

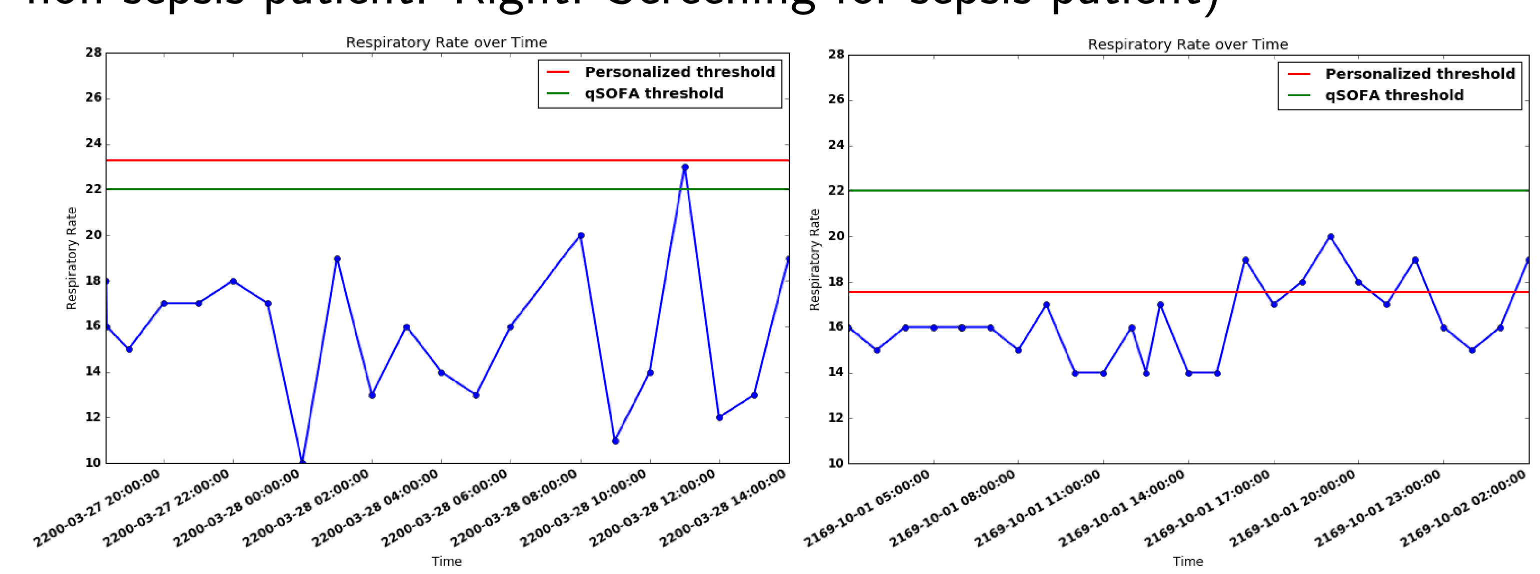
Specificity

	Personalized	qSOFA	minP	Youden	Closest-(0,1)
RR	0.7341	**0.8931**	0.7865	0.7424	0.7222
sysBP	0.6905	**0.9279**	0.6415	0.6413	0.6474
Combined in qSOFA criterion	0.7493	**0.9315**	0.7949	0.7685	0.7601

Interpretation and Implementation of Personalized qSOFA

A) The estimated personalized threshold for RR against age.

B) Compared to qSOFA for two specific patients (Left: Screening for non-sepsis patient. Right: Screening for sepsis patient)



C) Compared to Machine Learning Techniques

Methods	Overall Accuracy	Sensitivity	Specificity
Personalized qSOFA	**0.7093**	**0.6668**	**0.7493**
Logistic Regression	0.7489	0.7428	0.7548
AdaBoosting	0.7456	0.7359	0.7549

- Machine Learning methods are black boxes and difficult for clinicians and nurses to implement for real time monitoring in Intensive Care Unit

Contact information

Chen Feng, email: cfeng@gatech.edu