Journal of Biomedical Informatics 56 (2015) 145–156

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

# Designing optimal mortality risk prediction scores that preserve clinical knowledge

Natalia M. Arzeno [a,1], Karla A. Lawson [b], Sarah V. Duzinski [b], Haris Vikalo [a,*]

[a] Department of Electrical and Computer Engineering, The University of Texas at Austin, 1 University Station C0803, Austin, TX 78712, USA
[b] Trauma Services, Dell Children's Medical Center of Central Texas, 4900 Mueller Blvd., Austin, TX 78723, USA

ARTICLE INFO

*Article history:*
Received 3 September 2014
Revised 26 May 2015
Accepted 28 May 2015
Available online 6 June 2015

*Keywords:*
Prognostic model
Optimizable risk score
Continuous risk score
Nonlinear features
ICU
PRISM III
SOFA

ABSTRACT

Many in-hospital mortality risk prediction scores dichotomize predictive variables to simplify the score calculation. However, hard thresholding in these additive stepwise scores of the form "add $x$ points if variable $v$ is above/below threshold $t$" may lead to critical failures. In this paper, we seek to develop risk prediction scores that preserve clinical knowledge embedded in features and structure of the existing additive stepwise scores while addressing limitations caused by variable dichotomization. To this end, we propose a novel score structure that relies on a transformation of predictive variables by means of nonlinear logistic functions facilitating smooth differentiation between critical and normal values of the variables. We develop an optimization framework for inferring parameters of the logistic functions for a given patient population via cyclic block coordinate descent. The parameters may readily be updated as the patient population and standards of care evolve. We tested the proposed methodology on two populations: (1) brain trauma patients admitted to the intensive care unit of the Dell Children's Medical Center of Central Texas between 2007 and 2012, and (2) adult ICU patient data from the MIMIC II database. The results are compared with those obtained by the widely used PRISM III and SOFA scores. The prediction power of a score is evaluated using area under ROC curve, Youden's index, and precision-recall balance in a cross-validation study. The results demonstrate that the new framework enables significant performance improvements over PRISM III and SOFA in terms of all three criteria.



## 1. Introduction

Technological advancements in medical instrumentation and a growing use of electronic medical records have created an abundance of clinical patient data. Extracting and analyzing useful information from such large and diverse data sets will enable tremendous advancements in clinical decision-making, ultimately leading towards improvements in health and quality of life as well as to reduction of the overall healthcare costs. Availability of data has enabled development of accurate mortality and morbidity risk prediction scores for specific patient populations. Rapid prediction of potentially poor outcomes may provide timely intervention to reduce morbidity and mortality among the patients in the considered group.

Additive stepwise scores of the form "add $x$ points if variable $v$ is above/below threshold $t$" are popular tools for mortality risk prediction in both pediatric and adult intensive care unit (ICU) populations, typically using data acquired at the beginning of an ICU stay. Such scores comprise the Acute Physiology and Chronic Health Evaluation (APACHE) [1,2], the Simplified Acute Physiology Score (SAPS) [3,4], and the Pediatric Risk of Mortality [5,6]. However, while simple enough to allow for fast manual evaluation, these scores dichotomize predictive variables to form prediction scores. Dichotomization of continuous variables results in a loss of information, increased probability of false negatives, and high dependence on cut-off points [7,8]. This, in turn, may lead to critical failures of the prediction process. To remain relevant, risk scores need to be validated and updated periodically so that they reflect innovations and the evolution of the standards of healthcare [9,3], otherwise risking deteriorating accuracy due to the changes in patient populations [10]. Moreover, risk scores might have higher accuracy for certain diseases [1] or may need to be customized for a subpopulation or a location [3].

In this paper, we seek to develop risk prediction scores that preserve the clinical knowledge embedded in the features and structure of the aforementioned additive stepwise risk scores while allowing for soft thresholds in the score calculations. This is accomplished by a novel scoring mechanism that relies on a

* Corresponding author. Tel.: +1 512 232 7922.
*E-mail addresses:* narzeno@utexas.edu (N.M. Arzeno), kalawson@seton.org (K.A. Lawson), svduzinski@seton.org (S.V. Duzinski), hvikalo@ece.utexas.edu (H. Vikalo).
[1] Principle corresponding author.

http://dx.doi.org/10.1016/j.jbi.2015.05.021
1532-0464/

transformation of predictive variables by means of nonlinear logistic functions to facilitate smooth differentiation between critical and normal values of the variables. The parameters in the proposed framework can be readily optimized for specific sub-populations of interest (e.g., particular disease or location) or re-learned to ensure that the score remains relevant as the standards of care evolve. We use the PRISM III score [5] and a pediatric brain trauma population, as well as the Sequential Organ Failure Assessment (SOFA) score [11,12] and an adult ICU population, as motivation and to test the performance of the novel scoring mechanism. The paper is organized as follows. Some of the most common risk scores used in the pediatric and adult ICU are reviewed in Section 2. In Section 3, we present the novel scoring mechanism, describe an algorithm for finding the optimal parameters of the logistic functions used to transform the data, and show quasiconvexity of the corresponding optimization problem. Section 4 presents the results obtained by applying the proposed methodology to predict mortality of pediatric trauma patients admitted to the ICU of the Dell Children's Medical Center of Central Texas between 2007 and 2012 (Section 4.1) and predict mortality of adult ICU patients from the MIMIC database (Section 4.2). Section 5 concludes the paper.

## 2. Existing risk scores

In this section we overview existing prognostic models for both pediatric and adult patient populations, with the emphasis on PRISM III and SOFA. We also briefly mention data-driven techniques that have recently been used in model development and close the section by describing recent publications that incorporate expert knowledge in their models.

### 2.1. Pediatric population risk scores

An example of the pediatric risk prediction scores, the Pediatric Risk of Mortality (PRISM III) [5], is a widely used scoring mechanism in pediatric ICU (PICU) [9] that has been validated in various settings as both an individual predictor and a significant predictor in a multivariate model in the United States and internationally [13–21]. A state-of-the-art scheme, PRISM III has an additive stepwise structure that relies on 17 physiological variables and 26 ranges. The physiological variables it considers are characterized by their maximum or minimum values recorded during the first 12 or 24 h after a patient's admission to the PICU. These variables include the minimum systolic blood pressure, maximum heart rate, the presence of fixed pupils, maximum and minimum body temperature, and a variety of laboratory measures including minimum and maximum $CO_2$ and pH, white blood cell count, glucose and platelet count. Contribution of some variables to the score is determined after evaluating a judiciously chosen logical OR statement, such as in the case of maximum prothrombin time (PT) and partial thromboplastin time (PTT), which can both detect abnormalities in clotting time. Other variables, such as systolic blood pressure and heart rate, have age-dependent ranges.

In PRISM III, the score is incremented when the maximum (minimum) value of a variable in the score is above (below) a predetermined threshold. For example, if a child has a minimum Glasgow Coma Scale (GCS) [22,23] score less than 8, then 5 points are added to her/his PRISM III score. Clearly, calculation of the score is highly dependent on the established cutoff points and thus the prediction may change abruptly due to very small changes in the underlying variables. For instance, interpretation of the variables such as heart rate or blood pressure, which are altered by the simple act of breathing, may widely change due to the strict threshold structure – an adolescent with a maximum heart rate of 144 beats per minute is considered healthy, while another one with a single measure of 145 beats per minute has 3 extra points added in the PRISM III score calculation. Although the PRISM scores have been validated in numerous settings, they have also been shown to overpredict [24,25] and underpredict [26,27] PICU deaths. Poor patient discrimination by PRISM scores, especially in neonates and infants [28,29], and the fact that only a small subset of PRISM variables are significant predictors of the outcome [30], render the PRISM scores sensitive to population characteristics and standards of care and suggest that they may not necessarily be institution independent. Unlike in PRISM III (and the previous versions of PRISM) where the score is computed based on binary indicators of the raw feature values, in this paper each feature is first transformed by a non-linear logistic function whose inflection point and slope we find via an optimization procedure. As a result, we identify a range of values for which the risk changes continuously and monotonically (i.e., increases or decreases with the feature values), which stands in contrast to describing the effect of physiological variables on the risk of mortality by comparing the variables to pre-defined thresholds (the strategy employed by state-of-the-art prediction schemes such as PRISM III).

In addition to PRISM III, widely-used scores in the PICU include the injury severity score (ISS) [31] and the pediatric index of mortality (PIM2) [32], where the former is specific to trauma patients. The ISS is an anatomic score based on the location and severity of the injuries. Limitations of the ISS have led to various modifications as well as risk scores that incorporate the ISS in the calculation [33]. In a pediatric trauma population, PRISM has outperformed the ISS and its variants in identifying in-hospital mortality [17]. Logistic regression has often been employed to learn the weighting coefficients for physiological variables or binary indicators in mortality prediction models [33]. PIM2 is a second generation score, based on recalibrating coefficients of PIM [34] and adding variables for diagnostic groups with poor performance or calibration. PIM2 models risk using logistic regression with 10 variables acquired upon hospital admission or in the first hour after PICU admission, 7 of which are binary indicators. Specifically, the continuous PIM2 variables include first systolic blood pressure, ratio of $FiO_2$ to $PaO_2$, and absolute arterial or capillary base excess. Fixed pupils, mechanical ventilation, elective admission, PICU admission for procedure recovery, and cardiac bypass are included in the model as binary variables. Finally, a selection of high risk or low risk diagnoses, where the model was found to over- or under-estimate mortality, complete the list of logistic regression variables [32].

### 2.2. Adult population risk scores

The Sequential Organ Failure Assessment (SOFA) [11,12] is an additive stepwise score that assigns a score of 0–4 to each of six physiological systems: respiratory, coagulation, hepatic, cardiovascular, renal, and central nervous systems. The total SOFA score is then evaluated as the sum of the individual system scores. SOFA was originally designed to describe the degree of organ dysfunction in patients rather than be a mortality prediction score. However, Vincent et al. acknowledge that there is a correlation between the mortality rate and organ dysfunction [11,12]. Subsequently, SOFA has been successfully implemented as a mortality prediction score, achieving performance comparable to that of other established risk-prediction scores [35,36]. Studies developing SOFA-based models consider the score at either a fixed time or incorporate sequential measurements where the difference in SOFA might be indicative of the risk of mortality [35]. Going a step beyond examining the delta SOFA values, Toma et al. developed risk prediction models that include temporal patterns from SOFA measurements [37]. Moreover, SOFA has been shown to improve

the accuracy of mortality risk-prediction of established scores such as APACHE and SAPS when used in combination with them [38,39].

Commonly used mortality risk prediction scores developed for adult populations [40] include the popular APACHE, [1,2], SAPS [3,4], and Mortality Prediction Model at ICU admission ($MPM_0$) [41,42]. SAPS 3 [3] and $MPM_0$III [42] use data obtained within one hour of ICU admission, while APACHE IV [1] is calculated from data obtained in the first 24 h of the ICU stay. $MPM_0$III is calculated from 15 binary features and age, where the binary features comprise dichotomized physiological variables, CPR before admission, mechanical ventilation, and the existence of certain chronic and acute diagnoses. In addition to these features, $MPM_0$III includes 7 two-way interaction terms. Because feature weights are coefficients from a logistic regression, the weights can be positive or negative. SAPS 3 also has either positive or negative contribution from its variables, but with integer-valued weights. Twenty variables are required for SAPS 3 calculation; these variables can be divided into 3 groups: information before admission such as comorbidities and medications, circumstances of ICU admission such as whether it is a planned admission, and physiological variables. Unlike $MPM_0$III, variable contribution to SAPS 3 follows an additive stepwise structure with multiple hard thresholds for the physiological variables. Mortality risk based on SAPS 3 is additionally differentiated from other scores by the existence of multiple formulas to calculate risk based on the geographic region of the patient. Like SAPS 3, APACHE is an additive stepwise score. The most recent iteration, APACHE IV [1], includes 142 variables, 115 of which are categorical for admission diagnoses. Some of the variables in the model, such as age and the acute physiology score (APS), have linear and nonlinear contribution in the form of restricted cubic splines. The APS is in turn calculated as the APACHE III [2] score in the first 24 h of the ICU stay. The variables in APACHE III are divided into 3 groups: age, physiology, and chronic health. The chronic health variables specify different point contributions based on certain comorbidities, while age and the 17 physiological variables contribute to the score in an additive stepwise fashion. APACHE III includes interactions between variables in the form of AND statements, where the integer-valued contributions to the score are dependent on two physiological variables being within specified ranges. A common customization of prognostic models requires relearning the feature weights for a given population, as in the case of APACHE IV which has been successfully implemented to predict mortality at different time points in a Dutch ICU population [43]. While many have focused on validating and comparing the prognostic models, as described in review articles [44,45], others have sought to understand which component of a score is most predictive of mortality [46] or whether the addition of a particular variable such as age or resuscitation status will increase the discriminatory capability of a model [46,47]. However, these modifications do not address the variable dichotomization limitation of these scores.

### 2.3. Data driven risk scores and the preservation of expert knowledge

Going beyond the traditional statistical methods, data mining and machine learning techniques have recently been applied to the development of prognostic models seeking to aid in determining the time for treatment initiation, therapy choice, and healthcare quality assessment [48]. Recently used methods include decision tree techniques [49,50], neural networks [50], topic models [51], autoregressive implementation of PRISM [52], and techniques that incorporate injury coding schemes in the models [53]. While several scores with additive stepwise structure exist for adult populations (such as APACHE and SAPS), most pediatric mortality risk prediction scores assume that risk depends linearly on the variables and do not consider nonlinear variable transformations. PRISM III is an exception since it uses thresholding to prevent uninformative increase in risk at very high or very low variable values. In neonate morbidity prediction, the PhysiScore [54] achieves greater predictive power than previously established neonatal morbidity scoring systems by relying on a nonlinear transformation of the raw variables in the feature set. In particular, that work employs nonlinear Bayesian models based on log odds ratios of the risk derived from the probability distribution that provides the best fit to the data for each of two patient classes.

Additive stepwise scores, including PRISM III and SOFA, have a strong dependence on expert knowledge during the developmental stage. Paetz [55] proposed a data-driven method for designing additive stepwise scores, where the step weights and the variable ranges are randomly initialized and learned by means of evolutionary strategies. This method does not require experts until the final fine-tuning stage if such a stage is deemed necessary, which presents many advantages in the initial stages of the score design. However, training of the score does not allow for missing data, which is unlikely in the typical practical scenarios where many variables need to be collected and analyzed.

Models that integrate domain expert knowledge with a data driven approach have been reported to result in greater predictive accuracy. In [56], combining knowledge based and data driven risk factors in a prediction model for heart failure greatly improved on the performance of a solely knowledge based classifier while still resulting in a clinically meaningful model. In the task of identifying similar concept pairs in clinical notes, combining context-based similarity and knowledge-based similarity in an algorithm has likewise resulted in a more accurate similarity score [57].

## 3. Methods

Most of the existing scores described in Section 2 require clinical expert input in the score development stage to determine thresholds for physiological variables. Data-driven approaches largely ignore expert knowledge in order to achieve the most discriminative results for a given dataset, which may lead to lack of interpretability. The method for designing an optimal risk prediction algorithm described in this section employs data-driven techniques while preserving the expert knowledge embedded in the existing risk scores. In particular, in this section we propose a novel outcome prediction score that relies on a nonlinear transformation of the features, present an algorithm for optimizing the parameters of the transformation, and discuss optimality of the aforementioned algorithm.

### 3.1. The new score and an algorithm for optimizing parameters of the logistic transformation of predictive features

We describe the risk of mortality using a logistic regression model, where the conditional probability that patient $i$ dies during the hospital stay is given by

$$\mathbf{P}(y_i = 1|\mathbf{w}, \mathbf{z}_i) = \frac{1}{1 + \exp(-\mathbf{w}^T\mathbf{z}_i)}, \quad (1)$$

and the conditional probability of survival is

$$\mathbf{P}(y_i = -1|\mathbf{w}, \mathbf{z}_i) = \frac{1}{1 + \exp(\mathbf{w}^T\mathbf{z}_i)}, \quad (2)$$

where $y_i \in \{-1, 1\}$ is an indicator of mortality, the vector $\mathbf{w} = [w_1\ w_2\ \ldots w_{M_w}]'$ collects weights for the features $\mathbf{z}_i = [z_{i1}\ z_{i2}\ \ldots\ z_{iM_w}]'$, and $M_w$ denotes the total number of features.

In a departure from the commonly used hard thresholding of predictive features and discrete scoring, we introduce a logistic

transformation of the predictive features. The resulting new score is continuous and differentiable which enables computationally efficient search for the optimal parameters of the logistic transformation.[2] In particular, for patient $i$ and feature $j$, the nonlinear transformation $z_{ij}$ of the raw variable $x_{ij}$ is

$$z_{ij} = \begin{cases} \frac{1}{1+\exp(-a_j(x_{ij}-t_{ij}))} & \text{if } x_{ij} \text{ is a maximum} \\ 1 - \frac{1}{1+\exp(-a_j(x_{ij}-t_{ij}))} & \text{if } x_{ij} \text{ is a minimum} \\ 0 & \text{if } x_{ij} \text{ is missing} \end{cases} \quad (3)$$

where $a_j \geqslant 0$ is the slope of the nonlinear transformation and $t_{ij}$ is the inflection point of the logistic function (i.e., a "soft threshold" counterpart to the hard thresholds used by the existing stepwise scoring schemes). It should be noted that if clinical knowledge suggests the use of age-dependent thresholds, the $t_{ij}$'s are not different for every subject $i$ but rather have the same value for all patients within a specific age group and a given feature $j$. In the case where this age-dependence is not evidenced, the soft thresholds $t_{ij}$ in the novel algorithm are shared across the entire patient population and thus the soft threshold for feature $j$ can be written as $t_j$.

The optimal weights and parameters for the nonlinear transformations are determined by minimizing the negative log-likelihood of the logistic regression model,

$$\min \sum_{i=1}^{n} \log(1 + \exp(-y_i(\mathbf{w}^T \mathbf{z}_i))). \quad (4)$$

To preserve and exploit clinical knowledge previously used in the creation of other scores, a lognormal prior is imposed in the optimization for $\mathbf{w}$; this also ensures all features will be associated with a positive weight. In particular, for $\mathbf{w} \in \mathbb{R}^d$ we set

$$\mathbf{P}(\mathbf{w}) = \frac{\exp\left(-\frac{1}{2}(\log \mathbf{w} - \mu)^T \Sigma^{-1} (\log \mathbf{w} - \mu)\right)}{(2\pi)^{d/2} |\Sigma|^{0.5} \prod_{j=1}^{d} w_j}.$$

For a lognormal prior with mean $\mu$ and covariance $\Sigma = \frac{1}{2\lambda} I$, the optimization over $\mathbf{w}$ then becomes

$$\arg \min_{\mathbf{w}} \sum_{i=1}^{n} \log(1 + \exp(-y_i(\mathbf{w}^T \mathbf{z}_i))) + \sum_{j=1}^{d} \log w_j + \lambda \| \log \mathbf{w} - \mu \|_2^2. \quad (5)$$

The joint optimization (5) over $\mathbf{w}$, $\boldsymbol{a} = [a_1\ a_2\ \ldots\ a_{M_a}]$, and/or $\mathbf{t} = [t_1\ t_2\ \ldots\ t_{M_t}]$ is carried out by cyclic block coordinate descent with backtracking line search [59]. Optimization over $\boldsymbol{a}$ and $\mathbf{t}$ includes an additional step of projections onto the constraint set. The blocks for the coordinate descent consist of features derived from the same raw variable. For example, one block contains two features derived from the maximum heart rate, which correspond to two steps with different weights in a simple thresholding-based additive score. The algorithm is formalized as Algorithm 1 given below. Note that the objective function of the optimization (5) is not convex. Nevertheless, even if we use an iterative optimization method that merely ensures the objective function is decreased at each step, in our empirical studies the resulting local minimum leads to an improvement over the existing risk score. In fact, our computational studies show that the proposed scheme is robust with respect to the initial point of the search – for instance, in the application to pediatric population, starting the iterative optimization procedure with a vector $\mathbf{w}$ comprising PRISM III weights and starting with a vector of uniform weights $\mathbf{w} = ([1, \ldots, 1]$ results in almost identical prediction accuracy on the considered dataset.

[2] In machine learning parlance, the new score can broadly be categorized as a generalized additive model [58].

**Algorithm 1.** Optimization over the slopes $\boldsymbol{a}$.

$a_j^{(0)} \leftarrow 0.01,\ j = 1, 2, \ldots, M_a$
$\mathbf{w} \leftarrow \mathbf{w}_{\text{stepwise score}}$
$\mathbf{t} \leftarrow \mathbf{t}_{\text{stepwise score}}$
$k \leftarrow 1$
**repeat**
  $\boldsymbol{a} \leftarrow \boldsymbol{a}^{(k-1)}$
  **for all** $g \in G$ **do**
    $\triangle a_j \leftarrow -(\nabla_a f(\boldsymbol{a}, \mathbf{t}, \mathbf{w}))_j$ if $j \in g$
    $\triangle a_j \leftarrow 0$ if $j \notin g$
    $h \leftarrow 1$
    **while** $f(\boldsymbol{a} + h\triangle\boldsymbol{a}, \mathbf{t}, \mathbf{w}) > f(\boldsymbol{a}, \mathbf{t}, \mathbf{w}) - \alpha h \|\triangle\boldsymbol{a}\|^2$ **do**
      $h \leftarrow \beta h$
    **end while**
    $\boldsymbol{a} \leftarrow \text{Proj}_A(\boldsymbol{a} + h\triangle\boldsymbol{a})$
  **end for**
  $\boldsymbol{a}^{(k)} \leftarrow \boldsymbol{a}$
**until** stopping criterion is met

In Algorithm 1, $G$ is the set of feature groups, $A$ denotes the projection set for $\boldsymbol{a}$, and $\alpha$ and $\beta$ are the backtracking line search parameters. The feature groups $g \in G$ are defined as the groups of nonlinear features related to the same raw variables (e.g., in the application to pediatric population the feature group for maximum pH has two elements corresponding to PRISM III thresholds of 7.55 and 7.48), and make up the blocks for the coordinate descent.

The projection sets are defined so that the clinical knowledge used for the existing score is preserved. In Algorithm 1, $A$ ensures the nonlinear transformations follow the same direction as the steps in the existing score ($A = \{\boldsymbol{a} : a_j \geqslant 0 \text{for } j = 1, \ldots, M_a\}$). The projection set $T$ for optimization over the soft thresholds of the nonlinear transformations in (3), $\mathbf{t}$, preserves the order of soft thresholds for a raw variable with multiple nonlinear transformations. For example, if the feature $i$ in SOFA corresponds to the minimum platelet count between 100 and 150 $\times 10^3/\text{mm}^3$ ($t_i^{(0)} = 150$) and feature $j$ corresponds to the minimum platelet count between 50 and 100 $\times 10^3/\text{mm}^3$ ($t_j^{(0)} = 100$), then the projection ensures that $t_i \geqslant t_j$ at all steps of the optimization procedure. It should be noted that this projection set may lead to a collapse of thresholds if the optimal solution is found at $t_j = t_i$. For example, the minimum platelet count, which has 4 levels in SOFA, may end up having only two different inflection points in the novel score.

The dimensions of the three optimization parameters are not equal due to the existence of binary features (e.g., those associated with pupillary reaction in PRISM III and drug administration in SOFA) that do not have nonlinear transformations ($M_a < M_w$) and the age-dependence of some of the thresholds ($M_w < M_t$). Optimization over the slopes $\boldsymbol{a}$ and soft thresholds $\mathbf{t}$ are implemented without inclusion of a prior in the objective.

### 3.2. Quasiconvexity of the logistic transformation

As stated earlier in this section, the objective function in (5) is not convex. However, we will here show that each block in the block-coordinate descent procedure is both quasiconvex and quasiconcave in the slope parameter $\boldsymbol{a}$, and is thus quasilinear. Since the logistic function is asymptotically flat, the objective in (5) is not strictly quasiconvex. However, since the objectives in the steps of block-coordinate descent procedure are quasilinear, if the initial values of $\boldsymbol{a}$ are such that the gradient is nonzero in every coordinate, the block coordinate descent will reach a global optimum.

For differentiable $f$ and domain $\mathcal{D}$, $f$ is quasiconvex if and only if $\mathcal{D}$ is convex and for all $x, y \in \mathcal{D}$ holds that $f(y) \leqslant f(x) \Rightarrow \nabla f(x)^T(y-x) \leqslant 0$ [59]. Similarly, $f$ is quasiconcave if and only if $\mathcal{D}$ is convex and for all $x, y \in \mathcal{D}$ it holds that $f(y) \geqslant f(x) \Rightarrow \nabla f(x)^T(y-x) \geqslant 0$.

The objective of the optimization is

$$\min\ f = \sum_{i=1}^{n} \log(1+\exp(-y_i(\mathbf{w}^T\mathbf{z}_i))), \tag{6}$$

where

$$\mathbf{w}^T\mathbf{z}_i = \sum_{j\in U} \frac{w_j}{1+\exp(-a_j(x_{ij}-t_{ij}))} + \sum_{k\in D} w_k\left(1-\frac{1}{1+\exp(-a_k(x_{ik}-t_{ik}))}\right) + \sum_{p\in P} w_p\delta(p_i=1), \tag{7}$$

where $U$ denotes the set of indices of features with maximum values whose contribution to the score is in the form of an up-step (i.e., the risk is higher when their values are above a threshold), $D$ is the set of indices of features with minimum values (down-steps), $P$ is the set of indices of the binary features and $\delta(p_i = 1)$ is an indicator function for these features. In PRISM III, the binary pupillary reflex features indicate whether one or both pupils are >3 mm and fixed. In SOFA, the binary features in the cardiovascular component of the score indicate the administered dosage of certain vasoactive drugs and the binary features in the respiratory component indicate respiratory support.

Note that if we want to find when is $f(a_j') \leqslant f(a_j'')$ for a given $a_j'$ and $a_j''$, it is sufficient to find the conditions on $a_j', a_j'', y_i, x_{ij}, t_{ij}$ such that it holds that $f_i(a_j') \leqslant f_i(a_j'')$ for all $i \in 1,\dots n$, where $f = \sum_{i=1}^{n} f_i$. For any $j \in U$, condition $f(a_j') \leqslant f(a_j'')$ is satisfied on the domain where

$$\frac{-y_i w_j}{1+\exp(-a_j'(x_{ij}-t_{ij}))} \leqslant \frac{-y_i w_j}{1+\exp(-a_j''(x_{ij}-t_{ij}))}.$$

This inequality will hold for any of the parameter combinations marked by an X in Table 1. The blank spaces in Table 1 satisfy $f_i(a_j') \geqslant f_i(a_j'')$ for $j \in U$, the fact which we next use to show quasiconcavity. To show quasilinearity, we also need to examine the gradient of $f$ at $a_j''$ which, for $j \in U$, is given by

$$\nabla_{a_j} f_i(a_j'') = -\frac{\exp(-y_i(\mathbf{w}^T\mathbf{z}_i))}{1+\exp(-y_i(\mathbf{w}^T\mathbf{z}_i))} \times \frac{y_i w_j (x_{ij}-t_{ij})\exp(-a_j''(x_{ij}-t_{ij}))}{\left(1+\exp(-a_j''(x_{ij}-t_{ij}))\right)^2}. \tag{8}$$

The sign of the gradient in (8) is determined by $-y_i w_j (x_{ij}-t_{ij})$. Simple arithmetic shows that when $f_i(a_j') \leqslant f_i(a_j')$ (corresponding to the entries marked by X in Table 1), $\nabla_{a_j} f_i(a_j'')(a_j'-a_j'') \leqslant 0$, and hence the condition for quasiconvexity is satisfied. Similarly, when $f_i(a_j') \geqslant f_i(a_j'')$ (corresponding to blanks in Table 1), $\nabla_{a_j} f_i(a_j'')(a_j'-a_j'') \geqslant 0$, which implies quasiconcavity.

The same procedure can be followed to show quasilinearity when $j \in D$, with appropriate sign changes. Note that the blocks of features where all of the features in the block belong to $U$ or all of the features in the block belong to $D$ will also satisfy the quasilinearity condition. Since the coordinate blocks in Algorithm 1 correspond to nonlinear transformations of the same physiological variable, quasilinearity holds. The initial value of $a_j$ in Algorithm 1 is set to 0.01 for all $j$. This corresponds to a small slope in the nonlinear transformations of the variables, and will only result in $\nabla_{a_j} f_i(a_j) = 0$ for subjects $i$ with $x_{ij} = t_{ij}$. Therefore, the initial slope of the cumulative function in (6) will be nonzero and the coordinate descent algorithm will not begin at a stationary point.

**Table 1**
Satisfy $f_i(a_j') \leqslant f_i(a_j')$ for $j \in U$.

| | $a_j' \leqslant a_j''$ | | $a_j' \geqslant a_j''$ | |
|---|---|---|---|---|
| | $y_i = 1$ | $y_i = -1$ | $y_i = 1$ | $y_i = -1$ |
| $x_{ij} - t_{ij} \geqslant 0$ | | X | X | |
| $x_{ij} - t_{ij} \leqslant 0$ | X | | | X |

Since we constrain the slopes $\boldsymbol{a}$ and weights $\mathbf{w}$ to be nonnegative, quasilinearity of the objective function in soft thresholds $\mathbf{t}$ can be shown in fewer steps than that for $\boldsymbol{a}$. We will only show quasiconvexity in $t_{ij}$ when $j \in U$, but quasiconvexity when $j \in D$ and quasiconcavity can be easily shown in a similar manner. To demonstrate quasiconvexity, it is sufficient to show that

$$f_i(t_{ij}') \leqslant f_i(t_{ij}'') \Rightarrow \nabla_{t_{ij}} f_i(t_{ij}'')(t_{ij}'-t_{ij}'') \leqslant 0.$$

Note that for $j \in U, f_i(t_{ij}') \leqslant f_i(t_{ij}'')$ if $y_i = -1$ and $t_{ij}' \geqslant t_{ij}''$, or if $y_i = 1$ and $t_{ij}' \leqslant t_{ij}''$. By examining the gradient of $f_i$, it is clear that since $w_j, a_j \geqslant 0, \nabla_{t_{ij}} f_i(t_{ij}'') \leqslant 0$ only when $y_i = -1$,

$$\nabla_{t_{ij}} f_i(t_{ij}'') = \frac{\exp(-y_i(\mathbf{w}^T\mathbf{z}_i))}{1+\exp(-y_i(\mathbf{w}^T\mathbf{z}_i))} \times \frac{y_i w_j a_j \exp(-a_j'(x_{ij}-t_{ij}''))}{\left(1+\exp(-a_j(x_{ij}-t_{ij}''))\right)^2}. \tag{9}$$

Note that the required conditions for $f_i(t_{ij}') \leqslant f_i(t_{ij}'')$ also guarantee that $\nabla_{t_{ij}} f_i(t_{ij}'')(t_{ij}'-t_{ij}'') \leqslant 0$. However, unlike with the optimization over slopes $\boldsymbol{a}$, we cannot guarantee that the initial $\mathbf{t}$ will be at a point with nonzero slope. Nevertheless, we have empirically observed that although $\nabla_{t_{ij}} f_i = 0$ for some $(i,j), \sum_i \nabla_{t_{ij}} f_i \neq 0$ in the first iteration of the alternating optimization of $\boldsymbol{a}$ and $\mathbf{t}$, indicating the optimization does not start at a stationary point for any of the nonlinear features. As with the weights $\mathbf{w}$, we initialize $\mathbf{t}^{(0)}$ using the hard thresholds of the existing score. Despite the approximations made in order to simplify calculation of the existing score, one expects its thresholds to be close to the optimal values due to the reliance on domain expert knowledge and extensive testing. Therefore, the proposed initialization will likely avoid start of the optimization procedure in the flat part of the quasiconvex curve, implying the global optimality of the block coordinate descent.

### 3.3. Algorithm testing

We validate the algorithm on a pediatric brain injury dataset and an adult ICU dataset. Note that the proposed method allows a straightforward refinement and optimization of the novel risk prediction score for a specific subpopulation and/or location of the hospital. The proposed prediction scheme is tested using leave-one-out cross-validation for the pediatric dataset ($n = 217$) and 10-fold cross-validation for the adult dataset ($n = 3711$) and compared with existing methods in terms of three discrimination criteria: (i) area under the receiver operating characteristic (ROC) curve (AUC), (ii) the Youden index ($J$), which aims to maximize the overall correct classification rate [60,61]: $J = \text{Sensitivity}(Se) + \text{Specificity}(Sp) - 1$, and (iii) the point on the ROC curve that maximizes the minimum of the positive predictivity ($+P$, precision) and sensitivity (recall). The last criterion takes into account class imbalance by balancing the percentage of true positives that are correctly predicted with the percentage of predicted positives that are correct. Thus, correctly predicted true negatives, the majority class, does not affect the performance metric. It should be noted that in evaluating algorithms by the third criterion (PrecRec), a false positive has the same effect as a false negative. These criteria only assess the discriminatory capabilities of the algorithms. In prognostic models,

particularly those that lead to clinical decision-making, the preciseness of the predicted probability, or calibration, should also be examined [48,62]. We assess the model calibration in terms of the Brier score (BS) [63], calculated as

$$BS = \frac{1}{N}\sum_{k=1}^{N}(\pi_k - c_k)^2, \tag{10}$$

where $N$ denotes the number of subjects, $c_k \in \{0,1\}$ is the binary class for subject $k$, and $\pi_k$ is the probability of mortality for subject $k$. The probabilities were calculated by means of Platt scaling [64] during the cross-validation, such that the algorithm outputs were entered in a logistic regression to find the probability of mortality. The Brier score ranges from 0 to 1, where lower values indicate a better calibrated score.

## 4. Results and discussion

In this section, we analyze pediatric and adult datasets to demonstrate the performance of the novel scoring scheme that preserves expert knowledge from PRISM III and SOFA, respectively. Therefore, the variables used by the novel score and the initial parameters for the optimization (feature weights and thresholds) are those from the aforementioned existing scores.

### *4.1. Implementation and results for the novel pediatric population scores*

Here we detail the implementation of the novel algorithm and present the discrimination and calibration results from the leave-one-out cross-validation tests on a population of 217 children with brain trauma and other brain malady or injury. We compare these results with those achieved by the classifiers that use raw non-transformed data as features. Since the latter classifiers do not allow for missing data (as in the case of PRISM III and our novel algorithm), different imputation strategies are examined.

#### *4.1.1. Pediatric patient population*

Data were retrospectively abstracted for 217 children (11.06% mortality rate) admitted to the Dell Children's Medical Center PICU. We included admissions to the PICU between August 2007 and April 2012, age range of 0–14 years, with an ICD-9 code reflecting brain injury, and a PICU stay of at least 24 h. The ICD-9 codes indicate patients with brain trauma (excluding simple concussion) as well as other brain malady or injury such as cerebral palsy, drowning, epilepsy, and asphyxiation. This group of patients was selected in order to emphasize the ease of optimizing the novel score for a specific high mortality population. This particular high-mortality population is of interest since trauma is the leading cause of death in children in the United States [65], with traumatic brain injury being a major contributor [66]. Minimum and maximum variable values from the first 12 h of the PICU stay are used in the calculation of PRISM III and our new score. The use of the data has been approved by the University of Texas at Austin Institutional Review Board and the Seton Clinical Research Steering Committee.

#### *4.1.2. Implementation of the novel prediction scheme in pediatric population*

Our novel score incorporates all of the variables and ranges previously used by PRISM III. Specifically, the 17 variables that our algorithm uses include: systolic blood pressure, heart rate, body temperature, pupillary reflexes, Glasgow Coma Scale, total $CO_2$, pH, $PaO_2$, $PCO_2$, glucose, potassium, creatinine, blood urea nitrogen, white blood cell count, platelet count, PT, and PTT. Detailed information about the ranges and cutoff points for PRISM III can be found in [5]. The features transformed using OR statements in PRISM III (e.g., by adding 6 points to the score if pH $< 7.0$ OR total $CO_2 < 5$) were treated as additive features in order to seamlessly include them in the optimization procedure. The backtracking line search parameters were chosen empirically based on the discrimination and the speed of convergence of the algorithm. The results shown are for $\alpha = 0.2$ and $\beta = 0.5$. In the case of optimization over $\mathbf{w}$, the imposed prior has $\mu = 0, \lambda = 0.25$.

#### *4.1.3. Performance comparison of the novel scores and existing scores*

Fig. 1 illustrates the difference in computing the contribution of a feature to the prediction score between the 12-h PRISM III and the novel scheme. Plots on the left show the logistic transformations of the features after performing optimization over the nonlinear transformation slopes $\boldsymbol{a}$, while the plots on the right show the logistic transformations after performing optimization over both the slopes $\boldsymbol{a}$ and weights $\mathbf{w}$. From the plots, we see that the optimization over $\boldsymbol{a}$ results in mortality risk increasing over a range of 3–4 beats per minute for each step in the maximum heart rate and the risk from minimum systolic blood pressure increasing over a range of 7 mmHg around the 75 mmHg threshold and increasing over a range of 4 mmHg around the 55 mmHg threshold. The risk from the minimum platelet count loses much of the stepwise scoring structure used by PRISM III, and instead increases monotonically for minimum platelet counts between 250,000 and 50,000.

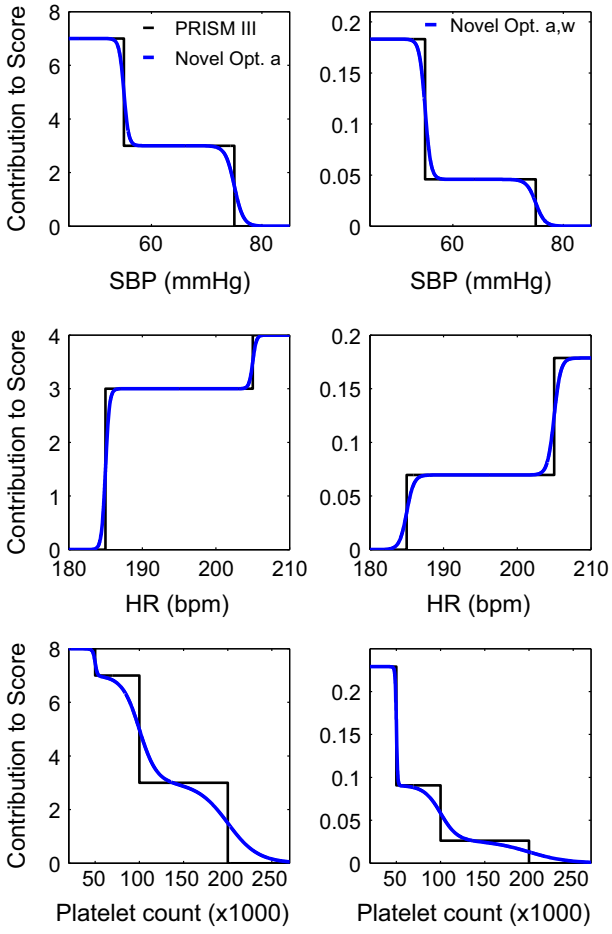


**Fig. 1.** Contribution of systolic blood pressure (SBP), heart rate (HR), and platelet count to risk score for PRISM III (black, left) and novel score optimizing $\boldsymbol{a}$ (blue, left) and optimizing $\boldsymbol{a}, \mathbf{w}$ (blue, right). The black lines on the right panel serve as indicators of a sharp transition between steps. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

This illustrates how our novel score can capture risk that increases continuously throughout a certain range while maintaining sharp thresholds when those are optimal. Optimizing over both $\boldsymbol{a}$ and $\mathbf{w}$ provides further insight into the contribution of variables to the risk of mortality for a given population. For example, the novel score shows that, for this dataset, the second step in the nonlinear transformation of the systolic blood pressure and heart rate should be weighted similarly, if not more heavily, than the lower step. The lower panels of Fig. 1 also indicate that while PRISM III has the mortality risk increasing slightly more when the minimum platelet count falls below 50,000, our inferred scoring function indicates the risk contribution from this variable doubles at approximately 50,000.

The results of the ROC analysis on the cross-validated scores are shown in Table 3. Our novel score optimized over the slopes $\boldsymbol{a}$ of the nonlinear transformations results in a more accurate classifier than PRISM III, in terms of the AUC and $J$, but not in terms of the precision-recall balance (PrecRec). Optimization of the feature weights $\mathbf{w}$ results in a score that performs better than PRISM III in all three evaluation criteria. Alternating optimization over the slopes $\boldsymbol{a}$ and weights $\mathbf{w}$ results in a further improvement of AUC, $J$, and precision-recall balance over individual parameter optimization, as well as the best calibration values (Brier score), and thus provides a significant advancement over PRISM III. The inclusion of soft thresholds $\mathbf{t}$ along with slopes $\boldsymbol{a}$ in the alternating optimization results in some of the soft thresholds falling outside the physiological range of the raw variables – in particular, logistic transformations of these variables result in zero contribution to the risk score for all patients. This optimization results in better discrimination than PRISM III in the upper range of scores (higher $J$ and precision-recall balance) and poorer classification than PRISM in the lower score range (lower AUC). However, we expect that these results would improve with a larger patient population given the age-dependency of some of thresholds and the mortality distribution across age groups (Table 2). The ROC results for PIM2 are also included in Table 3 to compare the novel score to another widely used pediatric risk score. The novel score optimized over any of the parameters outperforms PIM2 in terms of AUC, and optimization over two parameters also yields a higher precision-recall balance than PIM2. Though PIM2 results in a slightly higher $J$ than the novel score optimized over $\boldsymbol{a}, \mathbf{w}$, the large gain in AUC and higher precision-recall balance make the novel score the preferred choice for predicting risk of in-hospital mortality in the studied population.

**Table 2**
Mortality and missing data by age group.

| Age group | No. subjects | No. deaths | Missing variables (mean ± SD) |
|---|---|---|---|
| Neonate (0 mo., 1 mo.) | 2 | 0 | 9 ± 11.31 |
| Infant [1 mo., 12 mo.) | 39 | 5 | 6.95 ± 7.13 |
| Child [12 mo., 144 mo.] | 143 | 15 | 7.02 ± 7.22 |
| Adolescent (144 mo., 180 mo.) | 33 | 4 | 6.03 ± 7.14 |

**Table 3**
Risk score accuracy: Pediatric patient population.

| Score | AUC | $J$ | PrecRec | BS |
|---|---|---|---|---|
| PRISM III | 0.8735 | 0.5840 | 0.5172 | 0.0686 |
| Novel score optimized over $\boldsymbol{a}$ | 0.8897 | 0.6215 | 0.5000 | 0.0680 |
| Novel score optimized over $\mathbf{w}$ | 0.8841 | 0.6369 | 0.5600 | 0.0715 |
| Novel score optimized over $\boldsymbol{a}$ and $\mathbf{t}$ | 0.8358 | 0.6153 | 0.5833 | 0.0708 |
| Novel score optimized over $\boldsymbol{a}$ and $\mathbf{w}$ | 0.8927 | 0.6682 | 0.5833 | 0.0679 |
| PIM 2 | 0.8331 | 0.6729 | 0.5417 | 0.0766 |

The boxplots in Fig. 2 illustrate how the proposed prediction scheme compares with PRISM III. The probability of mortality (right column in Fig. 2) is also included in order to compare the algorithms on the same scale. The probability of mortality is calculated by means of Platt scaling as stated in Section 3.3. Despite additional features used by our scheme (due to splitting of the OR statements into components), the novel scores have similar average values to those of PRISM III. Given the class imbalance, this is likely the result of lower slopes in the nonlinear transformations which decrease the feature contributions of survivors having measurements near the soft thresholds. The movement of soft thresholds outside the physiological range following inclusion of $\mathbf{t}$ as an optimization parameter results in lower average score values compared to PRISM III. Finally, reduction in mean scores when the weights are included as optimization variables is expected due to the prior on the weight distribution. It should be noted that the optimizations which result in overall lower risk scores (i.e., optimizations including $\mathbf{t}$ or $\mathbf{w}$) yield lower risk scores for both the survivor and nonsurvivor groups; the relative difference between the groups remains and the mean probability of mortality is not significantly different from that of PRISM III or the score optimized only over $\boldsymbol{a}$.

The difference in AUC between PRISM III and the novel scores (Fig. 3) is primarily caused by low PRISM III scores for some nonsurvivors. Due to the finite set of PRISM III values, patients are more likely to share the same score which causes a decrease in both the true positive rate and the false positive rate as the cutoff value is lowered in the ROC analysis. The largest differences in the curves can be traced to patients with PRISM III scores between 7 and 13. While 45 out of 193 survivors have PRISM III scores in this range, so do 7 out of 24 nonsurvivors. We further examined the effect of softening the thresholds on patient outcome prediction for the subjects with low PRISM values by selecting the thresholds corresponding to the highest specificity with a sensitivity of at least 0.9. Such a sensitivity restriction promotes accurate identification of the in-hospital mortality class. For the selected cutoff scores, the novel score (optimized over the slopes $\boldsymbol{a}$) correctly identifies three subjects as high risk of in-hospital mortality that are incorrectly classified by PRISM III. One of these subjects is correctly identified by the novel score due to a nonzero contribution to the score from a systolic blood pressure with a value precisely at the PRISM III hard threshold. The other two subjects are correctly identified by the novel score due to nonzero contributions from the Glasgow Coma Scale feature ($\text{GCS} \in \{8, 9\}$), which in the optimized score has an approximately linear relationship to risk whereas PRISM III is affected only if $\text{GCS} < 8$.

#### *4.1.4. Comparison of the novel and existing scores for pediatric age groups*

Analysis of the results for different age groups suggests that the novel algorithm provides the largest gains in discrimination for the *child* age group (12–144 months). The ROC curves for different age groups are shown in Fig. 4, where the neonate and infant groups are combined due to the low number of subjects in the latter group. Note that the presented results display performance of the novel algorithm optimized over the entire population rather than those of three separate optimizations using only subjects from the age groups of interest. For the neonate and infant groups ($n = 41$, 5 deaths), both PRISM III and the novel score optimized over the slopes $\boldsymbol{a}$ and weights $\mathbf{w}$ of the nonlinear transformation achieve the AUC of 0.9833. In the child age group ($n = 143$, 15 deaths), the novel score (AUC = 0.8844) was significantly more discriminative than PRISM III (AUC = 0.8516). Both scores exhibited less accurate performance in the adolescent group ($n = 33$, 4 deaths), with the novel score (AUC = 0.7500) slightly outperforming PRISM III (0.7414). The high AUC in the neonate and infant

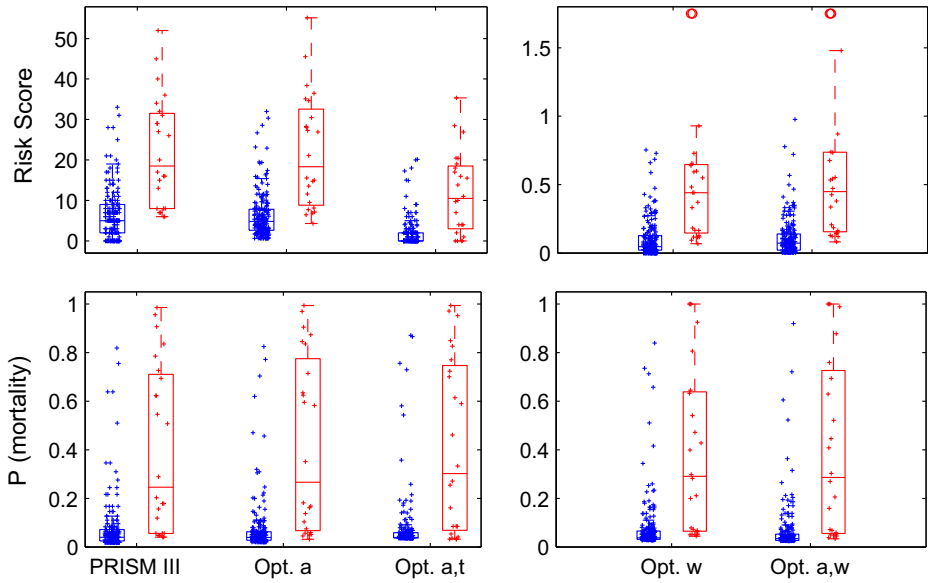


**Fig. 2.** Risk scores and the probability of mortality for survivors (blue) and nonsurvivors (red). The center mark on the box indicates the median while the edges of the box mark the 25th and 75th percentile. Individual scores are plotted as +. In the top right panel, the red circles indicate subjects with scores higher than 2. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

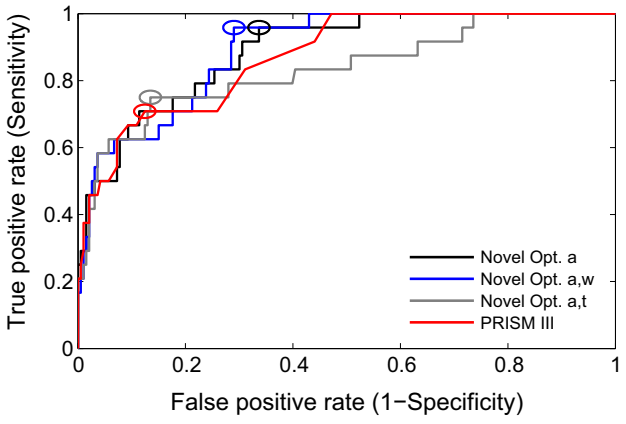


**Fig. 3.** ROC curve for PRISM III (red) and novel scores optimized over the nonlinear transformation slopes $\boldsymbol{a}$ (black), and alternating minimization over the slopes $\boldsymbol{a}$ and weights $\mathbf{w}$ (blue) and the slopes $\boldsymbol{a}$ and soft thresholds $\mathbf{t}$ (gray) of the nonlinear transformations. The circled points correspond to those that maximize Youden's index. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

group can be attributed to the characterization of the subjects in the nonsurvivor group, where most of them were near drowning victims and unresponsive upon ICU admission. The worse discrimination in the adolescent group is likely due to the low number of subjects and deaths, where two of the subjects in the nonsurvivor group had low scores with both the novel algorithm and PRISM III. It should also be noted that the distribution of brain injuries in the adolescent group is different from the other two groups. For example, this age group has a higher incidence of the asphyxiation/strangulation ICD-9 code. Given that the method presented in this paper can be used to design optimal scores for different populations, were more data available, it would have been interesting to design a mortality prediction score specifically for the adolescent population.

#### 4.1.5. Consideration of alternate feature transformation

To illustrate the improvement in discrimination enabled by using nonlinear transformations of the features, we also performed a logistic regression with ridge, lasso [67], and elastic net [68] penalties on the age and raw variables of the datasets. Additionally, to demonstrate the significance of the specific logistic feature transformations used by the novel score, we implemented

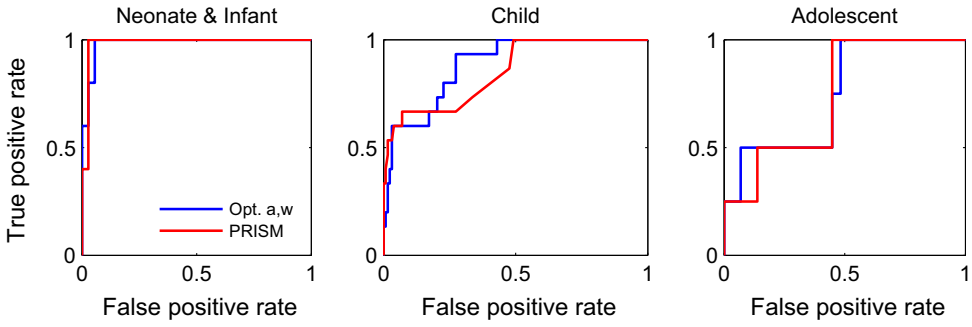


**Fig. 4.** The ROC curve by age group for PRISM III (red) and the novel score optimized over the slopes $\boldsymbol{a}$ and weights $\mathbf{w}$ of the proposed nonlinear transformation (blue). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

and tested logistic regression models with restricted cubic spline transformation of the raw physiological variables. Since missing data is prevalent and complete data sets are required for logistic regression, missing values of the variables are imputed by $k$ nearest-neighbors ($k$NN) [69], the probabilistic principal components analysis (PPCA) method [70], mean values, and normal values [71]. In the $k$NN imputation, Euclidean distances normalized by the number of common features between patients are calculated and the missing values are imputed as the average value of the $k$-nearest neighbors that observe the variable. PPCA aims to reduce the dimensionality of the data by associating a Gaussian latent variable model with the observed data and imputing missing values by an iterative expectation–maximization procedure. In our tests, data were imputed with $k = 5$ and 4 principal components. The best AUC values were achieved using the elastic net for the linear feature set and elastic net or ridge penalty for the nonlinear feature set. The discrimination and calibration measures for the logistic regression with linear (raw) and nonlinear (cubic spline transformation) feature sets are compared to the PRISM III values calculated with the imputed dataset in Table 4.

Though logistic regression with linear features and variables imputed with PPCA leads to a slight improvement in terms of $J$ and the precision-recall balance over the PRISM III scores without imputation, the PRISM III scores calculated with the imputed data outperforms both logistic scores in terms of AUC and $J$. Logistic regression with linear features and variables imputed with mean or normal values results in the same PrecRec value as the novel score. However, the novel score greatly outperforms both of these raw variable models in terms of AUC and $J$. These discrimination results suggest the variables in the scores should be nonlinearly transformed in order to achieve accurate mortality prediction. Note that the novel score without imputation performs better than the scores presented in Table 4 in terms of all of the evaluation criteria except for the Brier score of the logistic model with linear variables and imputation with normal values.

A close inspection of the measures of discrimination of the logistic regression with a nonlinear feature set in Table 4 further emphasizes the importance of performing the nonlinear transformation of physiological variables. Despite slightly higher AUC and $J$ achieved by the logistic models with cubic spline transformations of the features as compared to the novel algorithm, the novel algorithm outperforms the logistic models in terms of precision-recall balance and the Brier score. These two measures are of particular significance when considering the class imbalance and assessing accuracy of identifying subjects at risk for in-hospital mortality. Additionally, expert knowledge is ignored in the logistic regression with cubic spline covariates and the interpretation of the variable contributions to the score is unclear since the range of increasing risk is not a byproduct of the algorithm as in the case of our proposed algorithm.

**Table 4**
Risk score accuracy: Logistic with raw variables vs. PRISM III with imputed data.

| | $k$NN | PPCA | Mean | Normal |
|---|---|---|---|---|
| *AUC* | | | | |
| PRISM III | 0.8790 | 0.8741 | 0.8709 | 0.8735 |
| Raw linear | 0.8437 | 0.8683 | 0.8400 | 0.8141 |
| Raw nonlinear | 0.9011 | 0.8940 | 0.8975 | 0.9223 |
| *J* | | | | |
| PRISM III | 0.6047 | 0.6153 | 0.5788 | 0.5840 |
| Raw linear | 0.5479 | 0.5889 | 0.5637 | 0.5682 |
| Raw nonlinear | 0.6835 | 0.7144 | 0.7198 | 0.7301 |
| *PrecRec* | | | | |
| PRISM III | 0.5000 | 0.5000 | 0.5172 | 0.5172 |
| Raw linear | 0.5000 | 0.5542 | 0.5833 | 0.5833 |
| Raw nonlinear | 0.5172 | 0.5417 | 0.5000 | 0.5152 |
| *BS* | | | | |
| PRISM III | 0.0693 | 0.0685 | 0.0696 | 0.0686 |
| Raw linear | 0.0713 | 0.0689 | 0.0707 | 0.0631 |
| Raw nonlinear | 0.0998 | 0.0783 | 0.0911 | 0.0907 |

### 4.2. Implementation and results for the novel adult population scores

We apply our optimization framework to design scores that preserve the clinical knowledge embedded in SOFA while transforming the features using nonlinear logistic functions (as we did earlier for PRISM III and a pediatric brain injury population).

#### 4.2.1. Adult patient population

We use the data from the MIMIC II clinical database [72,73] to find the parameters of the logistic functions and test the ability of the novel score to predict mortality in the current ICU stay. The examined dataset consists of 3711 adult ICU patients (7.14% mortality).

#### 4.2.2. Implementation of the novel prediction scheme in adult population

The variables included in the SOFA score calculation are $PaO_2/FiO_2$, respiratory support, platelets, bilirubin, mean arterial pressure, Glasgow Coma Scale, creatinine, daily urine output, and certain cardiovascular drugs. The worst value of each physiological variable in the first 24 h is used to compute SOFA and as a raw variable for the calculation of the new score. The features representing the respiratory and cardiovascular systems are treated as indicator variables since the respiratory SOFA calculation includes an AND statement (e.g., the respiratory SOFA = 3 if $PaO_2/FiO_2 < 200$ AND the patient is receiving respiratory support) and because the cardiovascular SOFA greater than 1 is dependent on the administration of certain drugs. The urine output criterion for the renal SOFA is ignored since urine output was not charted consistently and thus the daily value might not be reliable. The parameters of the backtracking line search and the prior on $\mathbf{w}$ are as specified in Section 4.1.2 with the exception of the backtracking line search over $\mathbf{t}$, which was implemented with $\beta = 0.9$ due to a fast convergence of the algorithm.

#### 4.2.3. Performance comparison of novel score and existing score in adult population

The novel score is compared to SOFA in terms of the same discrimination (AUC, J, precision-recall balance) and calibration (Brier score) criteria described in Section 3.3. The average discrimination and calibration results from 10-fold cross-validation for the novel scores with optimization over various parameters are presented in Table 5. In order to visually compare the discriminatory capabilities of the novel score optimized over $\boldsymbol{a}$ and $\mathbf{w}$, we also present the ROC curve calculated from the risk scores obtained from the cross-validation results (in the left panel of Fig. 5) and the ROC curves from the 10 folds (in the right panel of Fig. 5). It should be noted that the novel score has higher discrimination than SOFA for every fold.

The novel scores outperform SOFA in all of the discrimination criteria, while the score optimized over the slopes of the nonlinear

**Table 5**
Risk score accuracy: Adult patient population.

| Score | AUC | $J$ | PrecRec | BS |
|---|---|---|---|---|
| SOFA | 0.6514 | 0.3069 | 0.2258 | 0.0630 |
| Novel score optimized over $\boldsymbol{a}$ | 0.7753 | 0.5213 | 0.3188 | 0.0603 |
| Novel score optimized over $\mathbf{w}$ | 0.7921 | 0.5357 | 0.2886 | 0.0666 |
| Novel score optimized over $\boldsymbol{a}$ and $\mathbf{t}$ | 0.7167 | 0.4498 | 0.2678 | 0.0641 |
| Novel score optimized over $\boldsymbol{a}$ and $\mathbf{w}$ | 0.7885 | 0.5275 | 0.3191 | 0.0617 |

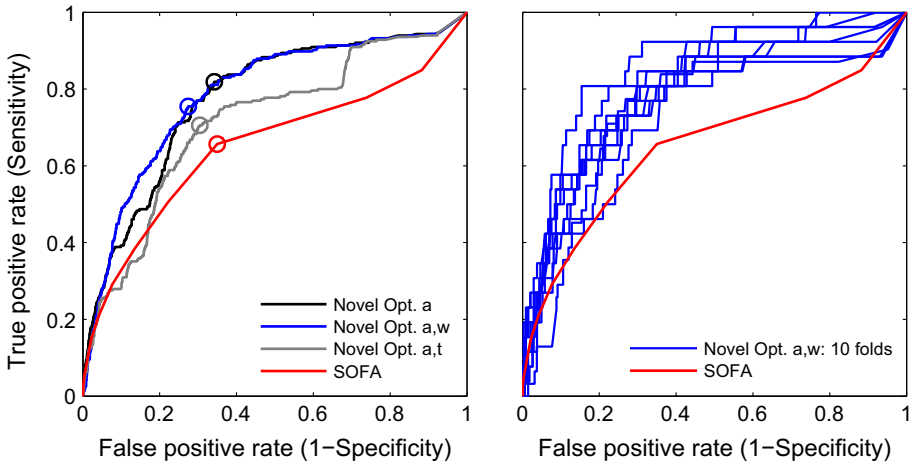


**Fig. 5.** ROC curves for SOFA (red) and the novel score preserving expert knowledge from SOFA. The left panel shows the ROC curves generated using the outputs of the 10-fold cross-validation for the optimization over nonlinear transformation slopes $\boldsymbol{a}$ (black), and alternating minimization over the slopes $\boldsymbol{a}$ and weights $\mathbf{w}$ (blue) and the slopes $\boldsymbol{a}$ and soft thresholds $\mathbf{t}$ (gray). The 10 ROC curves from the cross-validation folds of the novel score optimized over $\boldsymbol{a}$ and $\mathbf{w}$ are presented in the right panel (blue) along with the ROC curve for SOFA (red). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

transformations $\boldsymbol{a}$ and the score optimized over $\boldsymbol{a}$ and the nonlinear feature weights $\mathbf{w}$ also yield better Brier scores. The highest AUC and $J$ are obtained when optimizing over the weights $\mathbf{w}$. Including the slopes $\boldsymbol{a}$ in the optimization along with $\mathbf{w}$ results in a slight decrease in AUC and $J$ but leads to an improvement of the precision-recall balance and the Brier score. This score, optimized over $\boldsymbol{a}$ and $\mathbf{w}$, might therefore be preferred given the class-imbalance of the problem and the importance of identifying true positives. The novel score optimized over the slopes $\boldsymbol{a}$ and soft thresholds $\mathbf{t}$ exhibits worse discrimination than the scores obtained via other optimizations despite achieving the lowest objective value on the training sets. Upon further examination, we found that the final score no longer retained the step-wise structure of SOFA. More specifically, the final soft thresholds either collapsed into a single step or were outside the physiological range for some variables and thus had minor contribution to the final score. This finding stresses the importance of preserving the clinical knowledge embedded in the original score; to perform well on the test set, the novel score structure and optimization methodology should result in a score that resembles the original one.

## 5. Conclusions

We have developed a novel outcome prediction score that exploits advantages of additive stepwise risk scores and addresses the key limitation of hard thresholds typically used by state-of-the-art prediction methods. In particular, by transforming predictive variables using a combination of logistic functions, the developed method allows for a fine differentiation between critical and normal values of the predictive variables. Optimization of the continuous score allows for not only specifying different weights for the variables but, by optimizing over the slope and/or inflection point of the logistic curve in the feature transformation, we can also identify the range of values of each variable where the risk increases. This optimization need only be performed once to determine the optimal parameters and the score is thereafter quickly calculated for each patient. Optimal values of the parameters of logistic functions may be readily re-learned as the patient population and standards of care evolve. The novel scores derived using the proposed optimization framework demonstrate significantly higher predictive power than the widely used PRISM III in a pediatric brain trauma population and the SOFA in an adult ICU population.

The presented method can be broadly applied to devise and optimize risk scores with predictive power superior to schemes that use hard-thresholding of physiological variables, as has been shown in the cases of PRISM III and SOFA. Future applications of the developed scheme include optimization of other in-hospital mortality risk scores designed for adult ICU population, such as the Acute Physiology and Chronic Health Evaluation (APACHE) [1,2] and the Simplified Acute Physiology Score (SAPS) [3]. Moreover, the proposed method may be used to develop risk prediction scores geared towards sub-populations of interest when a hard-thresholding score exists for a larger population (e.g. specific diseases within an ICU population) or in geographical areas with different standards of care than those used in the development of the existing scores.

## Conflict of interest

None declared.

## Acknowledgments

The authors would like to thank Karen Piper for the identification of patients that met the inclusion criteria for the study. This material is based upon work supported by the National Science Foundation Graduate Research Fellowship under Grant No. DGE-1110007 and Jack Kilby/Texas Instruments fellowship.

## References

[1] J.E. Zimmerman, A.A. Kramer, D.S. McNair, F.M. Malila, Acute physiology and chronic health evaluation (APACHE) IV: hospital mortality assessment for today's critically ill patients, Crit. Care Med. 34 (5) (2006) 1297–1310.
[2] W. Knaus, D. Wagner, E. Draper, J. Zimmerman, M. Bergner, P. Bastos, C. Sirio, D. Murphy, T. Lotring, A. Damiano, The APACHE III prognostic system. Risk prediction of hospital mortality for critically ill hospitalized adults, Chest 100 (6) (1991) 1619–1636.
[3] R. Moreno, P. Metnitz, E. Almeida, B. Jordan, P. Bauer, R. Campos, G. Iapichino, D. Edbrooke, M. Capuzzo, J.-R. Le Gall, SAPS 3—From evaluation of the patient to evaluation of the intensive care unit. Part 2: development of a prognostic model for hospital mortality at ICU admission, Inten. Care Med. 31 (10) (2005) 1345–1355.
[4] J. Le Gall, S. Lemeshow, F. Saulnier, A new simplified acute physiology score (SAPS II) based on a European/North American multicenter study, JAMA: J. Am. Med. Assoc. 270 (24) (1993) 2957–2963.
[5] M.M. Pollack, K.M. Patel, U.E. Ruttimann, PRISM III: an updated pediatric risk of mortality score, Crit. Care Med. 24 (5) (1996) 743–752.

[6] M.M. Pollack, U.E. Ruttimann, P.R. Getson, Pediatric risk of mortality (PRISM) score, Crit. Care Med. 16 (11) (1988) 1110–1116.
[7] D.L. Streiner, Breaking up is hard to do: the heartbreak of dichotomizing continuous data, Can. J. Psychiat. 47 (3) (2002) 262.
[8] P. Royston, D.G. Altman, W. Sauerbrei, Dichotomizing continuous predictors in multiple regression: a bad idea, Statist. Med. 25 (1) (2006) 127–141.
[9] J.P. Marcin, M.M. Pollack, Review of the methodologies and applications of scoring systems in neonatal and pediatric intensive care, Pediat. Crit. Care Med. 1 (1) (2000) 20–27.
[10] L. Minne, S. Eslami, N. de Keizer, E. de Jonge, S.E. de Rooij, A. Abu-Hanna, Statistical process control for validating a classification tree model for predicting mortality–a novel approach towards temporal validation, J. Biomed. Inf. 45 (1) (2012) 37–44.
[11] J.L. Vincent, R. Moreno, J. Takala, S. Willatts, A. De Mendonça, H. Bruining, C.K. Reinhart, P.M. Suter, L.G. Thijs, The SOFA (sepsis-related organ failure assessment) score to describe organ dysfunction/failure on behalf of the working group. on sepsis-related problems of the european society of intensive care medicine, Inten. Care Med. 22 (7) (1996) 707–710.
[12] J.L. Vincent, F. Ferreira, R. Moreno, Scoring systems for assessing organ dysfunction and survival, Crit. Care Clin. 16 (2) (2000) 353–366.
[13] D. Scavarda, C. Gabaudan, F. Ughetto, F. Lamy, V. Imada, G. Lena, O. Paut, Initial predictive factors of outcome in severe non-accidental head trauma in children, Child's Nerv. Syst.: ChNS 26 (11) (2010) 1555–1561.
[14] R.J. Gemke, J. van Vught, Scoring systems in pediatric intensive care: PRISM III versus PIM, Inten. Care Med. 28 (2) (2002) 204–207.
[15] A.R. Brady, D. Harrison, S. Black, S. Jones, K. Rowan, G. Pearson, J. Ratcliffe, G.J. Parry, Assessment and optimization of mortality prediction tools for admissions to pediatric intensive care in the united kingdom, Pediatrics 117 (4) (2006) e733–742.
[16] G. Karambelkar, S. Mane, S. Agarkhedkar, R. Karambelkar, S. Singhania, S. Kadam, The relevance of 24 h PRISM III score in predicting mortality in pediatric intensive care unit, Int. J. Pharm. 3 (4) (2012) 214–219.
[17] E. Cantais, O. Paut, R. Giorgi, L. Viard, J. Camboulives, Evaluating the prognosis of multiple, severely traumatized children in the intensive care unit, Inten. Care Med. 27 (9) (2001) 1511–1517.
[18] M. Bahloul, A. Chaari, I. Chabchoub, F. Medhyoub, H. Dammak, H. Kallel, H. Ksibi, S. Haddar, N. Rekik, H. Chelly, M. Bouaziz, Outcome analysis and outcome predictors of traumatic head injury in childhood: analysis of 454 observations, J. Emerg., Trauma, Shock 4 (2) (2011) 198–206.
[19] E.A. Volakli, M. Sdougka, V. Drossou-Agakidou, M. Emporiadou, M. Reizoglou, M. Giala, Short-term and long-term mortality following pediatric intensive care, Pediat. Int. 54 (2) (2012) 248–255.
[20] V.F. Martha, P.C.R. Garcia, J.P. Piva, P.R. Einloft, F. Bruno, V. Rampon, Comparison of two prognostic scores (PRISM and PIM) at a pediatric intensive care unit, J. Pediat. 81 (3) (2005) 259–264.
[21] A.U. Qureshi, A.S. Ali, T.M. Ahmad, Comparison of three prognostic scores (PRISM, PELOD and PIM 2) at pediatric intensive care unit under Pakistani circumstances, Abbottabad: JAMC 19 (2) (2007) 49–53.
[22] H.L. Laurer, D.F. Meaney, S.S. Margulies, T.K. Mcintosh, Modeling brain injury/trauma, in: V. Ramachandran (Ed.), Encyclopedia of the Human Brain, Academic Press, New York, 2002, pp. 93–102.
[23] A.I.R. Maas, C.L. Harrison-Felix, D. Menon, P.D. Adelson, T. Balkin, R. Bullock, D.C. Engel, W. Gordon, J. Langlois-Orman, H.L. Lew, C. Robertson, N. Temkin, A. Valadka, M. Verfaellie, M. Wainwright, D.W. Wright, K. Schwab, Standardizing data collection in traumatic brain injury, J. Neurot. 28 (2) (2011) 177–187.
[24] A. Slater, F. ShannANZICS Paediatric Study Group, The suitability of the pediatric index of mortality (PIM), PIM2, the pediatric risk of mortality (PRISM), and PRISM III for monitoring the quality of pediatric intensive care in australia and new zealand, Pediat. Crit. Care Med. 5 (5) (2004) 447–454.
[25] S. Tibby, D. Taylor, M. Festa, S. Hanna, M. Hatherill, G. Jones, P. Habibi, A. Durward, I. Murdoch, A comparison of three scoring systems for mortality risk among retrieved intensive care patients, Arch. Dis. Childh. 87 (5) (2002) 421–425.
[26] P. Bhadoria, A.G. Bhagwat, Severity scoring systems in paediatric intensive care units, Ind. J. Anaesth. 52 (Suppl. (5)) (2008) 663–675.
[27] A. Thukral, R. Lodha, M. Irshad, N.K. Arora, Performance of pediatric risk of mortality (PRISM), pediatric index of mortality (PIM), and PIM2 in a pediatric intensive care unit in a developing country, Pediat. Crit. Care Med. 7 (4) (2006) 356–361.
[28] M. Wells, J.F. Riera-Fanego, D.K. Luyt, M. Dance, J. Lipman, Poor discriminatory performance of the pediatric risk of mortality (PRISM) score in a south african intensive care unit, Crit. Care Med. 24 (9) (1996) 1507–1513.
[29] J.M. Goddard, Pediatric risk of mortality scoring overestimates severity of illness in infants, Crit. Care Med. 20 (12) (1992) 1662–1665.
[30] A.L. Ponce-Ponce De León, G. Romero-Gutiérrez, C.A. Valenzuela, F.E. González-Bravo, Simplified PRISM III score and outcome in the pediatric intensive care unit, Pediat. Int. 47 (1) (2005) 80–83.
[31] S.P. Baker, B. O'Neill, W. Haddon Jr., W.B. Long, The injury severity score: a method for describing patients with multiple injuries and evaluating emergency care, J. Trauma 14 (3) (1974) 187–196.
[32] A. Slater, F. Shann, G. Pearson, PIM2: a revised version of the paediatric index of mortality, Intens. Care Med. 29 (2) (2003) 278–285.
[33] J.P. Marcin, M.M. Pollack, Triage scoring systems, severity of illness measures, and mortality prediction models in pediatric trauma, Crit. Care Med. 30 (11 Suppl.) (2002) S457–467.
[34] F. Shann, G. Pearson, A. Slater, K. Wilkinson, Paediatric index of mortality (PIM): a mortality prediction model for children in intensive care, Intens. Care Med. 23 (2) (1997) 201–207.
[35] L. Minne, A. Abu-Hanna, E. de Jonge, Evaluation of SOFA-based models for predicting mortality in the ICU: a systematic review, Crit. Care 12 (6) (2008) R161.
[36] S.Y. Hwang, J.H. Lee, Y.H. Lee, C.K. Hong, A.J. Sung, Y.C. Choi, Comparison of the sequential organ failure assessment, acute physiology and chronic health evaluation II scoring system, and trauma and injury severity score method for predicting the outcomes of intensive care unit trauma patients, Am. J. Emerg. Med. 30 (5) (2012) 749–753.
[37] T. Toma, A. Abu-Hanna, R.-J. Bosman, Discovery and inclusion of SOFA score episodes in mortality prediction, J. Biomed. Inf. 40 (6) (2007) 649–660.
[38] P. Fueglistaler, F. Amsler, M. Schüepp, I. Fueglistaler-Montali, C. Attenberger, H. Pargger, A.L. Jacob, T. Gross, Prognostic value of sequential organ failure assessment and simplified acute physiology II score compared with trauma scores in the outcome of multiple-trauma patients, Am. J. Surg. 200 (2) (2010) 204–214.
[39] K.M. Ho, Combining sequential organ failure assessment (SOFA) score with acute physiology and chronic health evaluation (APACHE) II score to predict hospital mortality of critically ill patients, Anaesth. Intens. Care 35 (4) (2007) 515–521.
[40] M.J. Breslow, O. Badawi, Severity scoring in the critically ill: part 1—interpretation and accuracy of outcome prediction scoring systems, Chest 141 (1) (2012) 245–252, http://dx.doi.org/10.1378/chest.11-0330. http://dx.doi.org/10.1378/chest.11-0330.
[41] S. Lemeshow, D. Teres, J. Klar, J.S. Avrunin, S.H. Gehlbach, J. Rapoport, Mortality Probability Models (MPM II) based on an international cohort of intensive care unit patients, JAMA 270 (20) (1993) 2478–2486.
[42] T.L. Higgins, D. Teres, W.S. Copes, B.H. Nathanson, M. Stark, A.A. Kramer, Assessing contemporary intensive care unit outcome: an updated Mortality Probability Admission Model (MPM0-III), Crit. Care Med. 35 (3) (2007) 827–835, http://dx.doi.org/10.1097/01.CCM.0000257337.63529.9F.
[43] S. Brinkman, A. Abu-Hanna, E. de Jonge, N.F. de Keizer, Prediction of long-term mortality in ICU patients: model validation and assessing the effect of using in-hospital versus long-term mortality on benchmarking, Intens. Care Med. 39 (11) (2013) 1925–1931, http://dx.doi.org/10.1007/s00134-013-3042-5.
[44] M.T. Keegan, O. Gajic, B. Afessa, Severity of illness scoring systems in the intensive care unit, Crit. Care Med. 39 (1) (2011) 163–169, http://dx.doi.org/10.1097/CCM.0b013e3181f96f81.
[45] J.-L. Vincent, R. Moreno, Clinical review: scoring systems in the critically ill, Crit. Care (London, England) 14 (2) (2010) 207, http://dx.doi.org/10.1186/cc8204.
[46] D.B. Knox, M.J. Lanspa, C.M. Pratt, K.G. Kuttler, J.P. Jones, S.M. Brown, Glasgow coma scale score dominates the association between admission sequential organ failure assessment score and 30-day mortality in a mixed intensive care unit population, J. Crit. Care 29 (5) (2014) 780–785, http://dx.doi.org/10.1016/j.jcrc.2014.05.009.
[47] M.T. Keegan, O. Gajic, B. Afessa, Comparison of APACHE III, APACHE IV, SAPS 3, and MPM0iii and influence of resuscitation status on model performance, Chest 142 (4) (2012) 851–858, http://dx.doi.org/10.1378/chest.11-2164.
[48] N. Peek, A. Abu-Hanna. Clinical prognostic methods (online).
[49] X.F. Courville, K.J. Koval, B.T. Carney, K.F. Spratt, Early prediction of posttraumatic in-hospital mortality in pediatric patients, J. Pediat. Orthoped. 29 (5) (2009) 439–444.
[50] L.G. Gortzis, F. Sakellaropoulos, I. Ilias, K. Stamoulis, I. Dimopoulou, Predicting ICU survival: a meta-level approach, BMC Health Serv. Res. 8 (1) (2008) 157, http://dx.doi.org/10.1186/1472-6963-8-157.
[51] L.-W. Lehman, M. Saeed, W. Long, J. Lee, R. Mark, Risk stratification of ICU patients using topic models inferred from unstructured progress notes, in: AMIA Annual Symposium Proceedings 2012, 2012, pp. 505–511.
[52] U.E. Ruttimann, M.M. Pollack, A time-series approach to predict outcome from pediatric intensive care, Comp. Biomed. Res., Int. J. 26 (4) (1993) 353–372.
[53] R.S. Burd, D. Madigan, The impact of injury coding schemes on predicting hospital mortality after pediatric injury, Acad. Emerg. Med. 16 (7) (2009) 639–645.
[54] S. Saria, A.K. Rajani, J. Gould, D. Koller, A.A. Penn, Integration of early physiological responses predicts later illness severity in preterm infants, Sci. Transl. Med. 2 (48) (2010). 48ra65-48ra65.
[55] J. Paetz, Finding optimal decision scores by evolutionary strategies, Artif. Intell. Med. 32 (2) (2004) 85–95.
[56] J. Sun, J. Hu, Combining knowledge and data driven insights for identifying risk factors using electronic health records, in: AMIA Annual Symposium Proceedings 2012, 2012, pp. 901–910.
[57] R. Pivovarov, N. Elhadad, A hybrid knowledge-based and data-driven approach to identifying semantically similar concepts, J. Biomed. Inf. 45 (3) (2012) 471–481.
[58] T. Hastie, R. Tibshirani, J. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, second ed., Springer, New York, NY, 2009.
[59] S. Boyd, L. Vandenberghe, Convex Optimization, Cambridge Univ Press, 2009.
[60] J. Hilden, P. Glasziou, Regret graphs, diagnostic uncertainty and Youden's index, Statist. Med. 15 (10) (1996) 969–986.
[61] N.J. Perkins, E.F. Schisterman, The inconsistency of optimal cutpoints obtained using two criteria based on the receiver operating characteristic curve, Am. J. Epidemiol. 163 (7) (2006) 670–675.

[62] S. Medlock, A.C.J. Ravelli, P. Tamminga, B.W.M. Mol, A. Abu-Hanna, Prediction of mortality in very premature infants: a systematic review of prediction models, PloS One 6 (9) (2011) e23441.
[63] G.W. Brier, Verification of forecasts expressed in terms of probability, Month. Weath. Rev. 78 (1) (1950) 1–3.
[64] J.C. Platt, Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods, in: Advances in Large-Margin Classifiers, MIT Press, 1999, pp. 61–74.
[65] S.L. Murphy, J.Q. Xu, K.D. Kochanek, Deaths: Preliminary Data for 2010, National Vital Statistics Reports, vol. 60 (4), National Center for Health Statistics, Hyattsville, MD.
[66] M. Faul, L. Xu, M.M. Wald, V.G. Coronado, Traumatic brain injury in the united states: emergency department visits, hospitalizations and deaths 2002–2006, 2010.
[67] R. Tibshirani, Regression shrinkage and selection via the lasso, J. Roy. Statist. Soc., Ser. B 58 (1994) 267–288.
[68] H. Zou, T. Hastie, Regularization and variable selection via the elastic net, J. Roy. Statist. Soc., Ser. B 67 (2005) 301–320.
[69] T. Hastie, R. Tibshirani, G. Sherlock, M. Eisen, P. Brown, D. Botstein, Imputing Missing Data for Gene Expression Arrays, Tech. rep., Stanford University, 1999.
[70] M.E. Tipping, C.M. Bishop, Probabilistic principal component analysis, J. Roy. Statist. Soc., Ser. B (Statist. Methodol.) 61 (3) (2002) 611–622.
[71] R.E. Behrman, R.M. Kiegman, H.B. Jenson (Eds.), Nelson Textbook of Pediatrics, 17th ed., Elsevier, 2004.
[72] M. Saeed, M. Villarroel, A.T. Reisner, G. Clifford, L.-W. Lehman, G. Moody, T. Heldt, T.H. Kyaw, B. Moody, R.G. Mark, Multiparameter intelligent monitoring in intensive care II (MIMIC-II): a public-access intensive care unit database, Crit. Care Med. 39 (2011) 952–960.
[73] A.L. Goldberger, L.A.N. Amaral, L. Glass, J.M. Hausdorff, P.C. Ivanov, R.G. Mark, J.E. Mietus, G.B. Moody, C.-K. Peng, H.E. Stanley, PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals, Circulation 101 (23) (2000) e215–e220.