Computers in Biology and Medicine 108 (2019) 9–19

Contents lists available at ScienceDirect

Computers in Biology and Medicine

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

The trauma severity model: An ensemble machine learning approach to risk prediction

Michael T. Gorczyca[a,*], Nicole C. Toscano[b], Julius D. Cheng[b]

[a] Department of Biological and Environmental Engineering, Cornell University, Ithaca, NY, 14853, USA
[b] Department of Surgery, University of Rochester Medical Center, Rochester, NY, 14642, USA

ARTICLE INFO

Keywords:
Risk prediction
Trauma quality improvement
Machine learning in medicine
National trauma data bank
Nationwide readmission database

ABSTRACT

Statistical theory indicates that a flexible model can attain a lower generalization error than an inflexible model, provided that the setting is appropriate. This is highly relevant for mortality risk prediction with trauma patients, as researchers have focused exclusively on the use of generalized linear models for trauma risk prediction, and generalized linear models may be too inflexible to capture the potentially complex relationships in trauma data. To improve trauma risk prediction, we propose a machine learning model, the Trauma Severity Model (TSM). In order to validate TSM's performance, this study compares TSM to three established risk prediction models: the Bayesian Logistic Injury Severity Score, the Harborview Assessment for Risk of Mortality, and the Trauma Mortality Prediction Model. Our results indicate that TSM has superior predictive performance on National Trauma Data Bank data and on Nationwide Readmission Database data.

1. Background

Trauma is a global healthcare epidemic, accounting for 9.2% of all deaths and 10.9% of disability-adjusted life-years [1]. The potential impact of trauma injuries on one's quality of life has inspired several studies on how we can improve the quality of trauma care, and consequently improve trauma patient outcomes. However, several of these studies require that we take a trauma patient's injury severity (risk of mortality) into account, and there is no consensus as to which risk prediction model is most appropriate for use [2–9].

Interestingly, careful consideration of the methodologies used to develop these risk prediction models indicates that regardless of which model is most appropriate, there may be room for substantial improvement in the quality of risk prediction. One reason for this is that several risk prediction models have been developed from small datasets [2–5], which implies that these models may not represent the population appropriately [10]. Another reason is that every model developed from a large dataset thus far has been a generalized linear model [6–9], and generalized linear models may be insufficient for capturing the potentially complex relationships in trauma data.

For these reasons, our objective is to develop a risk prediction model from machine learning algorithms with data from the National Trauma Data Bank (NTDB) and to compare this model to established risk prediction models on NTDB data as well as Nationwide Readmission Database (NRD) data. This is achieved by comparing three established risk prediction models – the Bayesian Logistic Injury Severity Score (BLISS) [7], the Harborview Assessment for Risk of Mortality (HARM) [8], and the Trauma Mortality Prediction Model (TMPM) [9] – to a new machine learning model for risk prediction. This machine learning model is the Trauma Severity Model (TSM).

2. Methods

2.1. Dataset overview

We utilized two datasets in this study. One dataset is the National Trauma Data Bank (NTDB) for patients hospitalized in 2008, 2009, 2010, and 2012 (2011 NTDB data was not entirely available to the authors during this study). The other dataset is the Nationwide Readmission Database (NRD) for patients hospitalized in 2013. Risk prediction with these datasets is formalized as a binary classification task, where the output variable is a binary indicator specifying whether or not a patient died before discharge from a hospital.

The NTDB is currently the largest aggregation of trauma data in the United States and provides patient demographics, hospital demographics, ICD-9-CM diagnoses codes (ICD-9 codes), general trauma assessments, hospital identifiers, physiology values, and in-hospital mortality [11]. The NTDB dataset initially consisted of 2,865,867 patient records from 884 hospitals with 7283 ICD-9 codes.

The 2013 NRD dataset is unique in that it possesses a large sample

* Corresponding author. 321 Riley-Robb Hall, Cornell University, Ithaca, NY, 14853, USA.
E-mail address: mtg62@cornell.edu (M.T. Gorczyca).

https://doi.org/10.1016/j.compbiomed.2019.02.025
Received 6 October 2018; Received in revised form 28 February 2019; Accepted 28 February 2019
0010-4825/

size (the discharge data accounts for 49.1 percent of all United States hospitalizations in 2013) and can support several different types of analyses [12]. The NRD dataset initially consisted of 14,325,172 patient records from 2006 hospitals with 12,403 ICD-9 codes.

We developed two sets of models from the NTDB dataset. One set of models considered ICD-9 codes as input variables (ICD-9 models). The other set of models considered ICD-9 codes, patient demographics (age and gender), and general trauma assessments (comorbidities, Glasgow Coma Scale response scores recorded by a physician, injury mechanism, injury type, and intent of trauma) as input variables (augmented models). Injury mechanism, injury type, and intent of trauma are based on a qualitative mapping of each external cause of injury code (E-Code) to a category defined by the American College of Surgeons [13]. To provide an example, E-Code E966, "assault by cutting and piercing instrument," would have an injury mechanism category of "cut/pierce," injury type category of "penetrating," and intent of trauma category of "assault."

For the augmented models, all input variables except age are treated as binary indicators, which specify whether or not a patient had a particular condition. For Glasgow Coma Scale response scores, the eye response score is represented by five binary indicators (four for response scores and one for not provided in the dataset), the verbal score is represented by six binary indicator variables (five for response scores and one for not provided in the dataset), and the motor score is represented by seven binary indicators (six for response scores and one for not provided in the dataset). These response scores are treated as binary indicators, as this improved the performance of the established risk prediction models relative to treating each response score as an ordinal variable or the summation of all response scores as a single numeric variable.

The NRD dataset is utilized for external validation of the ICD-9 models. We did not consider validating the augmented models on the NRD dataset because (1) risk prediction is commonly performed with models developed from ICD-9 codes only, and (2) the NRD dataset does not possess all the information utilized by the augmented models (such as Glasgow Coma Scale response scores) [14].

2.2. Data processing

2.2.1. NTDB data processing

We followed the data cleaning procedure described by TMPM's study when processing the NTDB dataset (BLISS and HARM follow similar data cleaning procedures). These data cleaning procedures are also advocated in Refs. [15–17]. Patient selection involves excluding patients that had burns or an ICD-9 code unrelated to trauma (e.g., poisoning, drowning, or suffocation) (193,606), were admitted to a hospital that did not submit ICD-9 codes considered relevant in TMPM's study to NTDB (655,440), were missing data (for age, comorbidities, gender, injury mechanism, injury type, intent of trauma, and outcome) (335,980), had pre-hospital mortality (60,234), were transferred to another hospital (848,885), were discharged to hospice care or another acute care hospital (16,429), withdrew care (18,395), or were less than one year old (47,693). Fig. 1 provides an overview of the NTDB patient selection process.

There are two differences between the patient selection process in this study and that in TMPM's study. One difference is how we selected hospitals from which we selected patients. In TMPM's study, the dataset consisted of patients from hospitals that admitted at least 500 patients during at least one year of the study (hospitals with "substantial trauma experience") [9]. We instead used every patient that was admitted to a hospital that submitted every ICD-9 code considered relevant in TMPM's study to the NTDB. The reasoning for this is that each hospital has different criteria for which ICD-9 codes are recorded for billing; these records are submitted to NTDB [13]. Excluding patients from hospitals that do not use relevant ICD-9 codes removes heterogeneity from the dataset, which can compromise a model's ability to provide

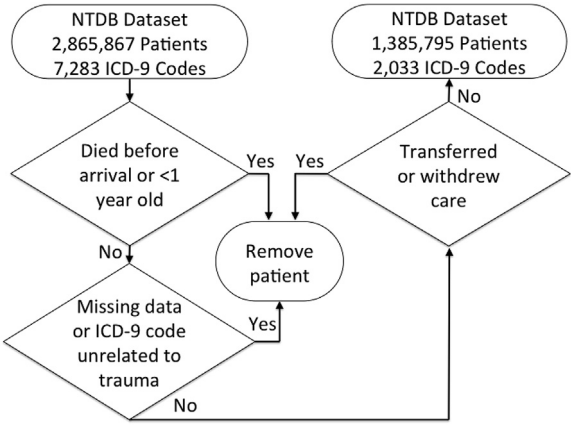

Fig. 1. Patient selection process for NTDB data.

accurate risk predictions that correspond to informed medical assessments [18,19]. Another difference is that TMPM's study ensured complete documentation only for age, gender, and outcome when determining which patients to include. We extended this to also ensure complete documentation for comorbidities, injury mechanism, injury type, and intent of trauma. The reasoning for this is that (1) no additional patients were excluded because of these criteria, (2) this information is typically known at the time of admission and is relevant in determining patient outcome, and (3) no risk prediction model has ever given consideration to such a combination of variables.

Once patient selection was completed, we performed an ICD-9 code combining procedure, which involved (1) combining ICD-9 codes together as recommended in TMPM's study, and then (2) combining each ICD-9 code that appeared fewer than five times with its closest corresponding ICD-9 code (based on expert consensus). The second part of this ICD-9 code combining procedure consisted of combining a specific injury with a more general injury; an open injury with a closed injury; or a group of highly similar injuries that were poorly represented to a single injury. We found that combining each ICD-9 code that appeared fewer than five times with its closest corresponding ICD-9 code improved the performance of every model in this study. Fig. 2 provides an overview of the ICD-9 code combining procedure for NTDB data.

The patient selection process kept 1,385,795 patient records out of 2,865,867 patient records and 2033 ICD-9 codes out of 7283 ICD-9 codes in the NTDB dataset. The ICD-9 code cleaning procedure from TMPM's study collapsed these 2033 ICD-9 codes into 1272 binary indicators representing ICD-9 codes. Combining ICD-9 codes that appeared fewer than five times with the closest corresponding ICD-9 code collapsed these 1272 binary indicators into 1234 binary indicators. There are 74 other variables that represent patient demographics, general trauma assessments, and patient outcome, which leaves us with a sparse 1,385,795 by 1308 matrix (all variables are binary indicators

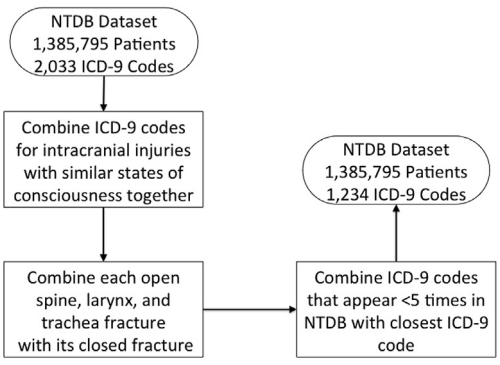

Fig. 2. ICD-9 code combining process for NTDB data.

Table 1
Demographics for the processed NTDB dataset.

Demographic Characteristics	
Age (Interquartile Range)	(23, 61)
Male (%)	63.93
Number of Hospitals[a]	713
Mortality before Discharge (%)	3.77
Racial Characteristics	
White (%)	64.52
Black or African American (%)	16.17
Hispanic[b] (%)	12.74
Asian (%)	1.93
Native American/Native Hawaiian/Pacific Islander (%)	0.79
Other (%)	10.82
Not Recorded (%)	5.77

[a] Specific hospital demographics not displayed due to this information changing each year in NTDB.
[b] Hispanic is denoted as an ethnicity in NTDB data, not race.

except age, which is numeric). Table 1 provides a brief summary of the demographics for the processed NTDB dataset.

2.2.2. NRD data processing

The established risk prediction models in this study have been applied to populations of patients that differ from what these models assumed for model development [20,21]. Due to this, we followed similar data cleaning procedures to these studies when processing the NRD dataset. For patient selection, we included all patients that satisfied the following three criteria. (1) The patient must have at least one ICD-9 code between 800 and 959.9 recorded during a hospital visit (excluding ICD-9 codes for burns, late effects of trauma, foreign bodies or complications of traumatic injury). (2) The patient's admission must be an emergency admission and the patient must have a valid E-Code (excluding drowning or submersion, bites or stings, overexertion, poisoning, suffocation, and adverse effects of medical and surgical interventions or medications). (3) The patient must not have a discharge status concerning transfer to another hospital. This cleaning procedure kept 820,694 patient records out of 14,325,172 patients records. Fig. 3 provides an overview of the patient selection procedure for NRD data.

The ICD-9 code combining procedure for NRD data is nearly identical to the ICD-9 code combining procedure for NTDB data. One difference is that we kept patients with ICD-9 codes unrelated to trauma, but omitted these ICD-9 codes from analysis. The other difference is that there was an ICD-9 trauma code that was present in the NRD dataset, but not present in the NTDB dataset. This ICD-9 code was combined with the closest corresponding ICD-9 code in NTDB data. This

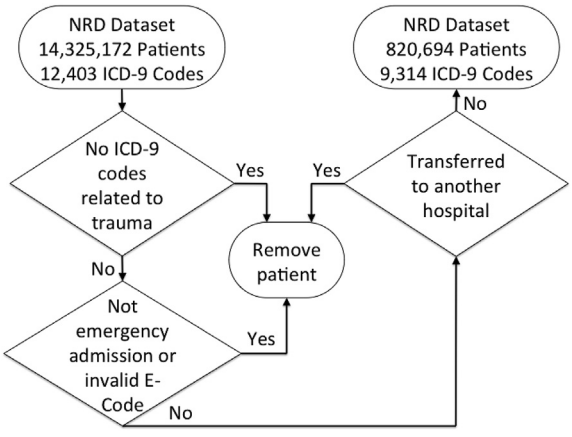

Fig. 3. Patient selection process for NRD data.

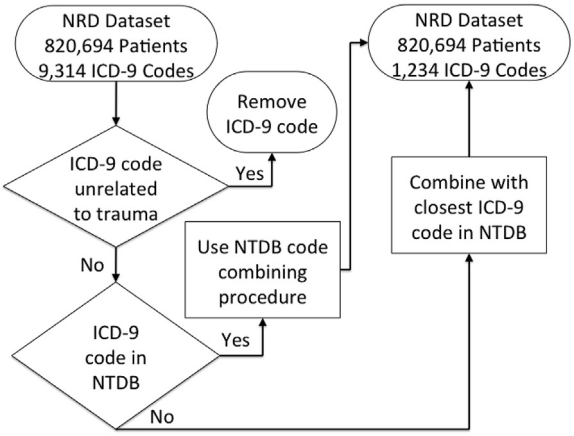

Fig. 4. ICD-9 code combining process for NRD data.

Table 2
Demographics for the processed NRD dataset.

Demographic Characteristics	
Age (Interquartile Range)	(47, 82)
Male (%)	49.86
Number of Hospitals	1844
Mortality before Discharge (%)	3.22
Economic Characteristics[a]	
$1-$37,999 (%)	26.45
$38,000-$47,999 (%)	25.30
$48,000-$63,999 (%)	24.25
$64,000 or more (%)	22.32
Not Recorded (%)	1.68

[a] Median household income quartiles for a patient's ZIP Code.

reduced the number of ICD-9 codes from 12,403 ICD-9 codes to 1234 binary indicators representing ICD-9 codes. Fig. 4 provides an overview of the ICD-9 code combining procedure for NRD data, and Table 2 provides a brief summary of the demographics for the processed NRD dataset.

2.3. Experimental setup

The ICD-9 models (BLISS, HARM, TMPM, and TSM models that considered ICD-9 codes only as input variables) and the augmented models (BLISS, HARM, TMPM, and TSM models that considered ICD-9 codes, patient demographics, and general trauma assessments as input variables) were developed from a random sample of the NTDB dataset (60% of the NTDB dataset). We then used an out-of-sample validation set (20% of the NTDB dataset) to assess each model's predictive performance and modify each model to further improve its predictive performance. These models were modified until their log-loss (LL) was minimized on the validation set [22]. Once a model's LL was minimized on the validation set, we used an out-of-sample test set as a final assessment of model performance on the NTDB dataset (20% of the NTDB dataset). Model development was performed using the h2o [23] and sandwich [24] packages in the R statistical software (Version 3.3.1) [25]. Each model in this study was developed from the same training set, validation set, and test set.

Two experiments were performed to fully assess the ICD-9 models on the NRD dataset. For the first experiment, the models were assessed by their predictive performance on the entire NRD dataset. For the second experiment, a random sample of the NRD dataset (50% of the NRD dataset) was used to calibrate each model with Platt scaling (a logistic regression model where the input variable is the prediction output of a risk prediction model, and the output variable is whether or

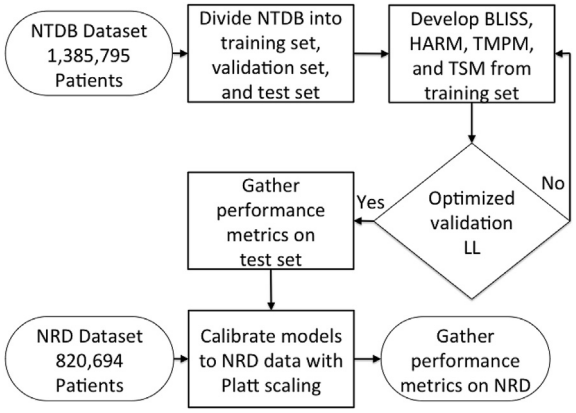

Fig. 5. Model development and assessment on NTDB and NRD data (experimental setup). Validation LL denotes log-loss on the validation set.

not a patient survived care) [26]. We used Platt scaling instead of a more sophisticated calibration procedure because the prediction outputs of risk prediction models are commonly used as the input variables for logistic regression models [14]. The remaining 50% of the NRD dataset was used to assess the performance of these calibrated models. Fig. 5 provides an overview of the experimental setup.

2.4. Established risk prediction model development

2.4.1. BLISS model development

BLISS utilizes Bayesian logistic regression for risk prediction. We considered two different prior distributions for developing BLISS in this study: a Gaussian distribution and a Laplace distribution. The Laplace distribution was chosen for the ICD-9 BLISS model as well as the augmented BLISS model because it produced BLISS models with a lower LL on the validation set than BLISS models with a Gaussian distribution. Fig. 6 provides an overview of BLISS model development.

2.4.2. HARM model development

HARM is a logistic regression model that takes advantage of the hierarchical structure of ICD-9 codes in order to reduce dimensionality. Specifically, ICD-9 codes and patient demographics are combined together to create new variables, or "collapsed variables" (based on expert consensus). These collapsed variables are selected as the input variables for HARM using forward selection [22].

To develop HARM in this study, we followed HARM's variable combining procedure as closely as possible. One difference between our model development procedure for HARM and that specified in HARM's study is that HARM originally accounted for diagnoses related to chronic obstructive pulmonary disease and ischemic heart disease (these corresponded to three input variables in the original HARM model). The NTDB does not account for this information, and none of the other models use diagnoses unrelated to trauma, so we did not attempt to include this information as input variables. Another difference is that the NTDB dataset possesses several ICD-9 codes absent from HARM's original dataset. We still considered ICD-9 codes present in the NTDB dataset but absent from HARM's original dataset as input variables, but these ICD-9 codes were not combined with any of the collapsed variables. Forward selection was performed until LL was minimized on the validation set. Fig. 6 provides an overview of HARM model development.

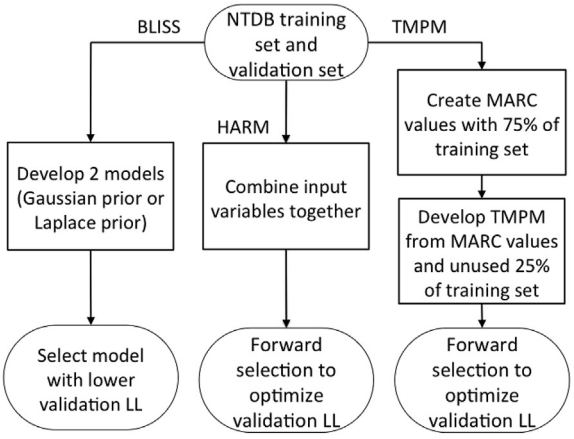

Fig. 6. BLISS, HARM, and TMPM model development. Validation LL denotes log-loss on the validation set.

2.4.3. TMPM model development

TMPM is a probit regression model that maps ICD-9 codes to numeric severity values ("MARC values") in order to reduce dimensionality. The only difference between our model development procedure for TMPM and that specified in its original study is that TMPM was originally developed using information pertaining to a patient's five largest MARC values as input variables. We instead used forward selection to select the input variables for TMPM, which was performed until LL was minimized on the validation set. We found that this forward selection procedure improved LL of TMPM relative to following TMPM's model development procedure exactly. Fig. 6 provides an overview of TMPM model development.

2.5. TSM model development overview

2.5.1. TSM model development

TSM was developed using stacked generalization [27,28]. Our approach to stacked generalization followed this sequence. First, several machine learning models, or base models, are created from four machine learning algorithms: logistic regression with the elastic net penalty [29], random forests [30], gradient boosted machines [31], and feed-forward neural networks [32]. The feed-forward neural networks were developed with the AdaDelta optimizer [33] and the Hogwild stochastic gradient update scheme [34]. During the training process, five-fold cross-validation was used to gather approximate out-of-sample risk predictions (cross-validated risk predictions) from each base model [22]. Each base model's cross-validated risk predictions are then combined to create a "meta-learner training set," which is used to develop a higher-level model (a meta-learner). For clarity, the meta-learner training set consists of each base model's cross-validated risk predictions as the input variables, and a binary indicator specifying whether or not the corresponding patient died prior to discharge as the output variable. Fig. 7 provides an overview of TSM model development.

2.5.2. TSM base model hyper-parameter search procedure

A benefit of using stacked generalization with cross-validation is that the meta-learner for TSM is developed from the same patients used to develop its base models. This allows comparison between TSM's base models, TSM's meta-learner, BLISS, HARM, and TMPM. However, the relationships that machine learning models learn from data are influenced by hyper-parameters, or user-specified values that are input into a machine learning algorithm before model development begins. In order to ensure appropriate model comparison, strong performing base models must be developed, which depends on the range of values a user considers for each hyper-parameter (the range of values considered for every hyper-parameter in a machine learning algorithm is referred to as a hyper-parameter space). This can be problematic in practice, as the user may configure a hyper-parameter space that does not contain the optimal hyper-parameters, which inhibits the development of strong performing machine learning models [36]. To avoid this potential issue, we propose a simple, heuristic search procedure for hyper-parameter optimization.

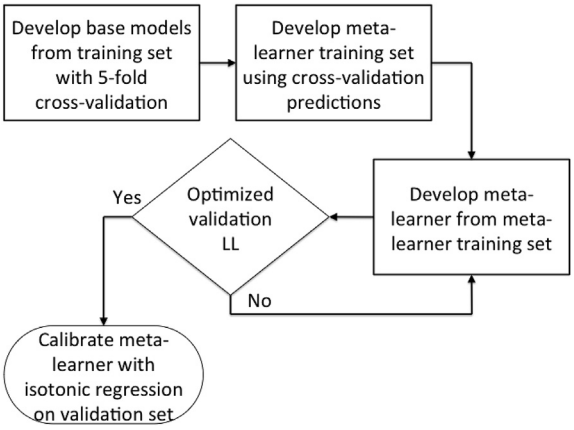

Fig. 7. TSM model development. Validation LL denotes log-loss on the validation set.

First, a manual search is performed to determine an initial hyper-parameter space configuration for a machine learning algorithm. Then, machine learning models are developed using a random search for hyper-parameters within this initial hyper-parameter space configuration [37]. After five models are developed from this initial hyper-parameter space (ten for neural networks due to its larger number of hyper-parameters), a checking procedure is performed. For clarity when describing the sequence of this checking procedure, input values denote the hyper-parameters used to develop a machine learning model, and hyper-parameter interval denotes the range of values considered for a hyper-parameter. (1) The top two performing models are selected based on their LL on the validation set (three for neural networks). (2) The input values of these selected models are examined to determine where the input values lie on their corresponding hyper-parameter intervals. (3) Each time every selected model has an input value for a hyper-parameter in the top (or bottom) quarter of its hyper-parameter interval, that hyper-parameter interval shifts such that the top (or bottom) quarter of the original hyper-parameter interval now represents the bottom (or top) quarter of a new hyper-parameter interval.

If none of the hyper-parameter intervals shift after five (ten) models are developed, then this checking procedure is performed after each subsequent model is developed and the checking procedure will give consideration to all models developed in the current hyper-parameter space. If a hyper-parameter interval does shift, then the checking procedure is not performed until five (ten) new models are developed, and the checking procedure will only give consideration to the models developed in this new hyper-parameter space. In this study, this hyper-parameter search procedure was performed until 40 models were developed from each machine learning algorithm except neural networks, from which 80 models were developed. Table 3 provides the initial hyper-parameter space configuration for each algorithm. Fig. 8 provides an overview of this hyper-parameter search procedure.

2.5.3. TSM meta-learner hyper-parameter search procedure

The meta-learner for TSM is a gradient boosted machine, which was developed using an exhaustive grid search where the only hyper-parameter varied was the maximum depth the trees in a gradient boosted machine were allowed to grow (from one to sixteen with an increment of one) [37]. The remaining hyper-parameters were set to their default values in h2o except the learning rate (which was set to 0.05), the annealing parameter for the learning rate (which was set to 0.99), and the number of trees developed (the number of trees was determined using the default early stopping protocol with a validation set in h2o) [35]. The gradient boosted machine with the lowest LL on the validation set was selected as the meta-learner for TSM.

Table 3
Initial hyper-parameter space configured for each machine learning algorithm. The number of trees in a random forest was not allowed to shift during our hyper-parameter search procedure.

Hyper-Parameter	ICD-9 Models	Augmented Models
Logistic Regression with Elastic Net Penalty		
α	$\mathscr{U}(0, 1)$[a]	$\mathscr{U}(0, 1)$
λ	$10^{\mathscr{U}^{d}(-10, -1)}$	$10^{\mathscr{U}^{d}(-10, -1)}$
Random Forest		
#Trees	$\mathscr{U}_d(50, 150)$[a]	$\mathscr{U}_d(50, 150)$
MNL[b]	$\mathscr{U}_d(1, 75)$	$\mathscr{U}_d(1, 75)$
NVS[c]	$\mathscr{U}_d(40, 150)$	$\mathscr{U}_d(40, 150)$
Max. Tree Depth	$\mathscr{U}_d(1, 50)$	$\mathscr{U}_d(60, 180)$
Gradient Boosted Machine		
#Trees	$\mathscr{U}_d(10, 80)$	$\mathscr{U}_d(10, 80)$
Max. Tree Depth	$\mathscr{U}_d(1, 15)$	$\mathscr{U}_d(1, 15)$
Learning rate	$\mathscr{U}(0.05, 0.50)$	$\mathscr{U}(0.05, 0.50)$
Annealing	$\mathscr{U}(0.850, 0.999)$	$\mathscr{U}(0.850, 0.999)$
Neural Networks		
#Hidden layers	$\mathscr{U}_d(1, 4)$	$\mathscr{U}_d(1, 4)$
#Neurons	$\mathscr{U}_d(1, 2^{11-\#\text{Hidden Layers}})$	$\mathscr{U}_d(1, 2^{11-\#\text{Hidden Layers}})$
Activation function	ReLU or Hyperbolic Tangent	ReLU or Hyperbolic Tangent
Dropout rates[d]	$\mathscr{U}(0, 0.33)$	$\mathscr{U}(0, 0.33)$
Epochs	$\mathscr{U}_d(10, 10{,}000)$	$\mathscr{U}_d(10, 10{,}000)$
ρ[e]	$\mathscr{U}(0.75, 0.999)$	$\mathscr{U}(0.75, 0.999)$
ε[e]	$10^{\mathscr{U}(-12, -3)}$	$10^{\mathscr{U}(-12, -3)}$
Meta-Learner (Gradient Boosted Machine)		
Max. Tree Depth	$\mathscr{U}_d(1, 15)$	$\mathscr{U}_d(1, 15)$

[a] $\mathscr{U}(a, b)$ denotes uniform continuous distribution from a to b, $\mathscr{U}_d(a, b)$ denotes uniform discrete distribution from a to b.
[b] MNL: minimum number of observations in a leaf.
[c] NVS: number of variables used in each split.
[d] The dropout rate was allowed to differ for each hidden layer, as this improved predictive performance.
[e] Hyper-parameters from the AdaDelta optimizer.

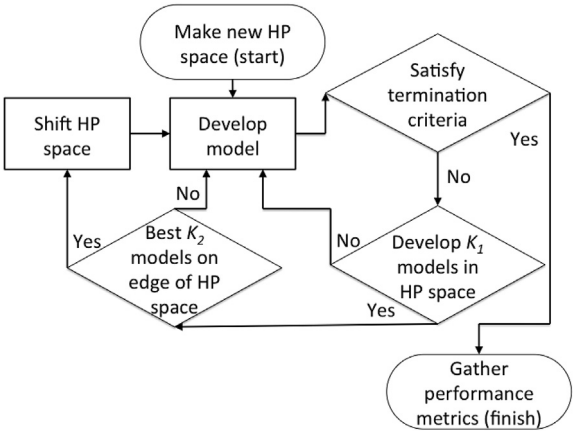

Fig. 8. Hyper-parameter search procedure. HP denotes hyper-parameter. K_1 was set to five for every machine learning algorithm except neural networks, which was set to ten. K_2 was set to two for every machine learning algorithm except neural networks, which was set to three.

2.5.4. Machine learning model calibration

Naively assessing the probabilistic calibration of each model in this study may be problematic, as non-linear machine learning models (random forests, gradient boosted machines, and neural networks) can have poor probabilistic calibration when the outcome event is rare, which Tables 1 and 2 indicate [26,38]. To avoid this potential issue, the training set (the portion of the processed NTDB dataset used for model development) was "balanced" before developing the non-linear machine learning models (all non-linear base models and the meta-learner for TSM). This data balancing procedure involved randomly over-sampling patients in the training set until the number of patients who

survived care was approximately the same as the number of patients who did not survive care and the balanced training set was approximately five times the size of the original training set [39]. This oversampling procedure was not performed with the NTDB validation set, the NTDB test set, and the NRD dataset.

The base model from each non-linear algorithm that had the lowest LL on the NTDB validation set as well as the meta-learner for TSM are calibrated using isotonic regression before assessing their performance on the processed NTDB test set and NRD dataset [37]. These isotonic regression models are developed from the prediction outputs of these models on the NTDB validation set.

2.6. Model assessment

The performance of the established risk prediction models; TSM; and the logistic regression model developed with the elastic net penalty, random forest, gradient boosted machine, and neural network models in TSM's ensemble that has the lowest LL on the validation set (selected base models) are evaluated with six performance metrics, which may be divided into three groups: threshold metrics, rank metrics, and probabilistic calibration metrics (calibration metrics). The threshold metrics are classification accuracy (ACC) and F-score (FSC). These metrics are computed based on whether or not a risk prediction is above a user-specified threshold value. ACC and FSC range from zero to one, where larger values indicate better performance. A threshold of 0.5 was used when computing these metrics [40].

The rank metrics used in this study are the area under the receiver operating characteristic curve (ROC) [41] and the area under the precision-recall curve (APR) [42]. Rank metrics depend on the ordering of outcomes, and not the actual risk predictions. Provided that this ordering is preserved, the range of a model's risk predictions does not affect its rank metric. These metrics measure how well positive cases (survival) are ordered before negative cases (mortality) and can be viewed as a summary of model performance across all possible thresholds. The ROC statistic may range from approximately 0.5 to one, and APR may range from zero to one. Larger values indicate better performance.

Calibration metrics assess how well a risk prediction corresponds to a patient's true risk of mortality. The calibration metrics considered in this study are LL [22] and the Hosmer-Lemeshow statistic with ten subgroups (HL) [43]. For these metrics, smaller values indicate better performance, where zero represents perfect probabilistic calibration.

In addition to assessing the models with these performance metrics, calibration curves [43] and precision-recall curves [44] were developed for model assessment. We also compared the ten largest variable importance measures from the selected base models [22,45]. Model assessment was performed using the boot [46,47], Metrics [48], and ResourceSelection [49] packages in the R statistical software. Variable importance measures were gathered using the h2o package [23].

3. Results

3.1. Model performance on NTDB data

The performance metrics of the ICD-9 models on the NTDB test set are displayed in Table 4. When ICD-9 codes are considered as input variables, TSM demonstrates an improvement over BLISS, HARM, and TMPM on every performance metric. TSM also demonstrates an improvement over its base models on most performance metrics (the selected random forest model has better HL). The performance of the augmented models on the NTDB test set is displayed in Table 5. Every model greatly improved in performance when augmented to account for patient demographics (age and gender) as well as general trauma assessments (comorbidities, Glasgow Coma Scale response scores, injury mechanism, injury type, and intent of trauma). But, TSM outperforms BLISS, HARM, TMPM, and its base models on every performance metric. BLISS generally outperforms TSM's base models on most performance metrics for the NTDB test set.

The calibration curves of the ICD-9 models are displayed in Fig. 9; the calibration curves of the augmented models are displayed in Fig. 10. TSM and TMPM consistently provide well-calibrated prediction outputs, while BLISS and HARM do not provide well-calibrated prediction outputs. The precision-recall curves of the ICD-9 models are displayed in Fig. 11; the precision-recall curves of the augmented models are displayed in Fig. 12. TSM generally displays higher precision and recall than BLISS, HARM, and TMPM.

3.2. Model performance on NRD data

The performance metrics of the un-calibrated ICD-9 models on the NRD dataset are displayed in Table 6. Every model had worse performance on the processed NRD dataset than the processed NTDB dataset, but TSM has better threshold and rank metrics than BLISS, HARM, and TMPM. TSM, BLISS, and HARM have similar LL, and HARM has better HL than BLISS, TMPM, and TSM (the selected gradient boosted machine base model has the best LL and HL overall). The performance metrics of the calibrated ICD-9 models on the processed NRD dataset are displayed in Table 7. When the ICD-9 models are calibrated with Platt scaling using half the NRD dataset, and these calibrated models are assessed on the other half of the NRD dataset, TSM maintains better performance than BLISS, HARM, and TMPM on every threshold and rank metric. TSM also attains the best LL when calibrated, and similar HL to HARM, which maintains better HL than BLISS, TMPM, and TSM.

3.3. TSM base model variable importance

The ten largest variable importance measures from the selected ICD-9 base models are displayed in Fig. 13, and the augmented base models in Fig. 14. Each model ranks the significance of their input variables differently. But, intracranial hemorrhages, subdural hemorrhages, and concussions were amongst the ten largest variable importance measures for every selected ICD-9 base model. Further, a Glasgow Coma Scale eye response score of one and a patient's age were amongst the ten largest variable important measures for every selected augmented base model. This indicates that these base models rely on information pertaining to head trauma when predicting patient outcomes.

3.4. Hyper-parameter search procedure results

Fig. 15 shows the performance of our hyper-parameter search scheme for developing the ICD-9 base models, and Fig. 16 shows the performance of our hyper-parameter search scheme for developing the augmented base models. Fig. 15 demonstrates that our hyper-parameter search scheme was particularly successful when developing random forest models, as hyper-parameter space shifts correspond to decreasing LL on the validation set. Fig. 16 shows that no hyper-parameter space shifting occurred during augmented base model development. Table 8 displays the hyper-parameters of the selected base models.

4. Discussion

4.1. Importance of machine learning in trauma care

Trauma is the leading cause of death for people younger than 44, and the fourth leading cause of death for all age groups in the United States [50]. As healthcare spending has grown to 17.8% of the Gross Domestic Product, it has become increasingly important to take the cost of care into consideration when improving the quality of trauma care [51]. But, in order to achieve the goals of improving the quality of trauma care while controlling the cost of care, we must utilize the best risk prediction models in trauma system evaluations. If risk prediction can be improved, so too can the quality of trauma care, as better risk

Table 4
Model comparison on the NTDB test set for the ICD-9 models. Standard error of the metric is denoted in the parenthesis, and the best metric attained is denoted in bold. PLM, RF, GBM, and NN denote logistic regression with the elastic net penalty, random forest, gradient boosted machine, and neural network, respectively.

Model	ACC	FSC	ROC	APR	HL	LL
TSM	**0.968 ($3.368\cdot10^{-4}$)**	**0.404 ($5.172\cdot10^{-3}$)**	**0.912 ($1.556\cdot10^{-3}$)**	**0.489 ($5.387\cdot10^{-3}$)**	**84.400 (17.183)**	**0.098 ($8.597\cdot10^{-4}$)**
BLISS	0.967 ($3.420\cdot10^{-4}$)	0.369 ($5.097\cdot10^{-3}$)	0.900 ($1.747\cdot10^{-3}$)	0.448 ($5.518\cdot10^{-3}$)	556.357 (38.350)	0.108 ($9.439\cdot10^{-4}$)
HARM	0.965 ($3.426\cdot10^{-4}$)	0.299 ($4.798\cdot10^{-3}$)	0.866 ($2.027\cdot10^{-3}$)	0.378 ($5.021\cdot10^{-3}$)	140.257 (22.619)	0.114 ($9.230\cdot10^{-4}$)
TMPM	0.966 ($3.421\cdot10^{-4}$)	0.336 ($5.135\cdot10^{-3}$)	0.898 ($1.753\cdot10^{-3}$)	0.435 ($5.450\cdot10^{-3}$)	154.692 (40.537)	0.105 ($9.221\cdot10^{-4}$)
PLM	0.966 ($3.419\cdot10^{-4}$)	**0.371 ($5.076\cdot10^{-3}$)**	0.899 ($1.749\cdot10^{-3}$)	0.448 ($5.511\cdot10^{-3}$)	529.964 (37.479)	0.108 ($9.462\cdot10^{-4}$)
RF	**0.967 ($3.369\cdot10^{-4}$)**	0.364 ($5.132\cdot10^{-3}$)	0.899 ($1.727\cdot10^{-3}$)	**0.452 ($5.224\cdot10^{-3}$)**	**77.420 (16.719)**	**0.104 ($8.880\cdot10^{-4}$)**
GBM	0.966 ($3.475\cdot10^{-4}$)	0.313 ($5.254\cdot10^{-3}$)	0.887 ($1.896\cdot10^{-3}$)	0.420 ($5.439\cdot10^{-3}$)	2510.427 (60.227)	0.114 ($7.794\cdot10^{-4}$)
NN	0.966 ($3.452\cdot10^{-4}$)	0.291 ($5.111\cdot10^{-3}$)	**0.902 ($1.678\cdot10^{-3}$)**	0.434 ($5.419\cdot10^{-3}$)	305.524 (29.035)	0.104 ($8.563\cdot10^{-4}$)

Table 5
Model performance comparison on the NTDB test set for the augmented models. Standard error of the metric is denoted in the parenthesis, and the best metric attained is denoted in bold. PLM, RF, GBM, and NN denote logistic regression with the elastic net penalty, random forest, gradient boosted machine, and neural network, respectively.

Model	ACC	FSC	ROC	APR	HL	LL
TSM	**0.976 ($2.964\cdot10^{-4}$)**	**0.621 ($4.277\cdot10^{-3}$)**	**0.965 ($7.936\cdot10^{-4}$)**	**0.696 ($4.485\cdot10^{-3}$)**	**23.341 (14.147)**	**$6.889\cdot10^{-2}$ ($7.165\cdot10^{-4}$)**
BLISS	0.975 ($2.925\cdot10^{-4}$)	0.601 ($4.235\cdot10^{-3}$)	0.957 ($9.828\cdot10^{-4}$)	0.665 ($4.795\cdot10^{-3}$)	95.063 (17.530)	$7.498\cdot10^{-2}$ ($7.676\cdot10^{-4}$)
HARM	0.973 ($2.972\cdot10^{-4}$)	0.564 ($4.353\cdot10^{-3}$)	0.955 ($9.914\cdot10^{-4}$)	0.631 ($5.086\cdot10^{-3}$)	115.840 (16.765)	$7.810\cdot10^{-2}$ ($7.416\cdot10^{-4}$)
TMPM	0.973 ($2.992\cdot10^{-4}$)	0.573 ($4.378\cdot10^{-3}$)	0.958 ($9.118\cdot10^{-4}$)	0.643 ($4.771\cdot10^{-3}$)	135.461 (84.870)	$7.577\cdot10^{-2}$ ($7.472\cdot10^{-4}$)
PLM	**0.974 ($2.949\cdot10^{-4}$)**	**0.593 ($4.291\cdot10^{-3}$)**	0.955 ($1.007\cdot10^{-4}$)	0.653 ($4.862\cdot10^{-3}$)	96.256 (16.654)	**$7.599\cdot10^{-2}$ ($7.719\cdot10^{-4}$)**
RF	0.974 ($2.996\cdot10^{-4}$)	0.577 ($4.456\cdot10^{-3}$)	**0.957 ($8.913\cdot10^{-4}$)**	**0.653 ($4.641\cdot10^{-3}$)**	40.772 (21.683)	$7.608\cdot10^{-2}$ ($7.449\cdot10^{-4}$)
GBM	0.974 ($3.063\cdot10^{-4}$)	0.561 ($4.569\cdot10^{-3}$)	0.957 ($9.171\cdot10^{-4}$)	0.641 ($5.060\cdot10^{-3}$)	135.050 (18.694)	$7.644\cdot10^{-2}$ ($7.136\cdot10^{-4}$)
NN	0.971 ($3.193\cdot10^{-4}$)	0.540 ($4.528\cdot10^{-3}$)	0.955 ($1.052\cdot10^{-4}$)	0.598 ($5.322\cdot10^{-3}$)	**39.630 (12.818)**	$7.780\cdot10^{-2}$ ($7.608\cdot10^{-4}$)

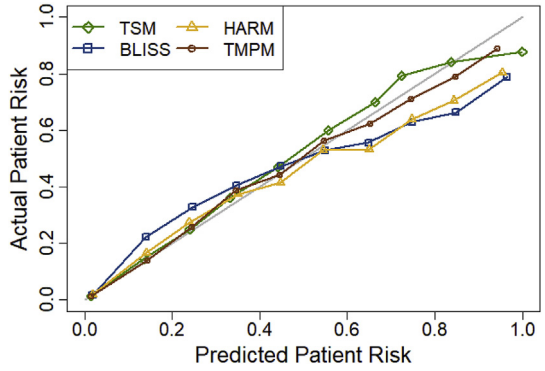

Fig. 9. Calibration curves for TSM, HARM, BLISS, and TMPM models that considered ICD-9 codes only as input variables. The grey line represents perfect probabilistic calibration. TSM and TMPM provide well-calibrated prediction outputs.

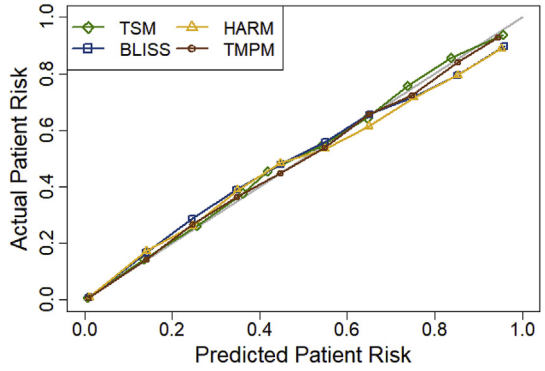

Fig. 10. Calibration curves for TSM, HARM, BLISS, and TMPM models that considered ICD-9 codes, patient demographics, and general trauma assessments as input variables. The grey line represents perfect probabilistic calibration. Every model provides well-calibrated prediction outputs.

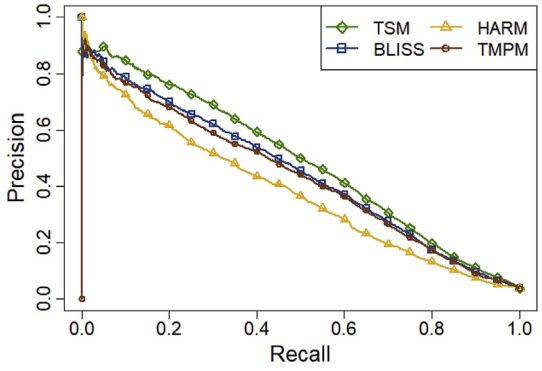

Fig. 11. Precision-recall curves for TSM, BLISS, HARM, and TMPM models that considered ICD-9 codes only as input variables. TSM generally had higher precision and recall at every threshold.

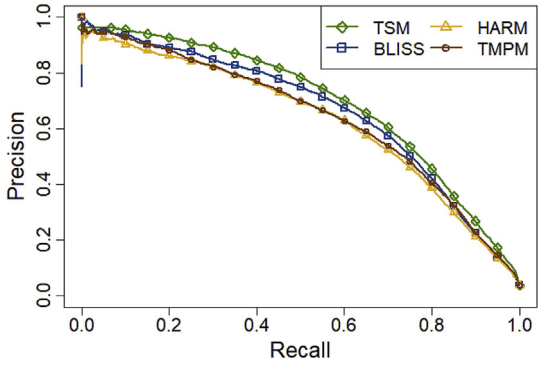

Fig. 12. Precision-recall curves for TSM, BLISS, HARM, and TMPM models that considered ICD-9 codes, patient demographics, and general trauma assessments as input variables. TSM generally had higher precision and recall at every threshold.

Table 6
Model comparison on the NRD dataset before model calibration. Standard error of the metric is denoted in the parenthesis, and the best metric attained is denoted in bold. PLM, RF, GBM, and NN denote logistic regression with the elastic net penalty, random forest, gradient boosted machine, and neural network, respectively.

Model	ACC	FSC	ROC	APR	HL	LL
TSM	**0.968 ($1.923\cdot10^{-4}$)**	**0.064 ($2.008\cdot10^{-3}$)**	**0.656 ($1.929\cdot10^{-3}$)**	**0.126 ($2.069\cdot10^{-3}$)**	58696 (2077)	0.140 ($7.675\cdot10^{-4}$)
BLISS	0.968 ($1.937\cdot10^{-4}$)	0.049 ($1.797\cdot10^{-3}$)	0.641 ($1.908\cdot10^{-3}$)	0.110 ($1.854\cdot10^{-3}$)	23382 (830)	**0.139 ($7.366\cdot10^{-4}$)**
HARM	0.968 ($1.936\cdot10^{-4}$)	0.052 ($1.865\cdot10^{-3}$)	0.624 ($1.799\cdot10^{-3}$)	0.101 ($1.698\cdot10^{-3}$)	**11843 (331)**	0.140 ($7.473\cdot10^{-4}$)
TMPM	0.965 ($2.009\cdot10^{-4}$)	0.038 ($1.579\cdot10^{-3}$)	0.620 ($1.984\cdot10^{-3}$)	0.083 ($1.183\cdot10^{-3}$)	180213 (6013)	0.152 ($9.140\cdot10^{-4}$)
PLM	**0.968 ($1.936\cdot10^{-4}$)**	0.050 ($1.810\cdot10^{-3}$)	0.641 ($1.905\cdot10^{-3}$)	0.110 ($1.853\cdot10^{-3}$)	25268 (898)	0.139 ($7.394\cdot10^{-4}$)
RF	0.967 ($1.946\cdot10^{-4}$)	**0.074 ($2.149\cdot10^{-3}$)**	0.623 ($1.917\cdot10^{-3}$)	**0.117 ($1.983\cdot10^{-3}$)**	42224 (1043)	0.145 ($8.011\cdot10^{-4}$)
GBM	0.968 ($1.934\cdot10^{-4}$)	0.035 ($1.569\cdot10^{-3}$)	0.635 ($1.896\cdot10^{-3}$)	0.109 ($1.775\cdot10^{-3}$)	**7058 (224)**	**0.136 ($6.043\cdot10^{-4}$)**
NN	0.968 ($1.925\cdot10^{-4}$)	0.031 ($1.487\cdot10^{-3}$)	**0.653 ($1.917\cdot10^{-3}$)**	0.113 ($1.899\cdot10^{-3}$)	34321 (1275)	0.138 ($7.228\cdot10^{-4}$)

Table 7
Model comparison on the out-of-sample NRD dataset after model calibration. Standard error of the metric is denoted in the parenthesis, and the best metric attained is denoted in bold. PLM, RF, GBM, and NN denote logistic regression with the elastic net penalty, random forest, gradient boosted machine, and neural network, respectively.

Model	ACC	FSC	ROC	APR	HL	LL
TSM	**0.968 ($2.758\cdot10^{-4}$)**	**0.064 ($2.881\cdot10^{-3}$)**	**0.656 ($2.789\cdot10^{-3}$)**	**0.125 ($2.898\cdot10^{-3}$)**	1409 (60)	**0.133 ($9.148\cdot10^{-4}$)**
BLISS	0.968 ($2.783\cdot10^{-4}$)	0.046 ($2.547\cdot10^{-3}$)	0.641 ($2.770\cdot10^{-3}$)	0.111 ($2.635\cdot10^{-3}$)	2465 (99)	0.136 ($9.167\cdot10^{-4}$)
HARM	0.968 ($2.774\cdot10^{-4}$)	0.046 ($2.464\cdot10^{-3}$)	0.622 ($2.636\cdot10^{-3}$)	0.101 ($2.479\cdot10^{-3}$)	**1399 (63)**	0.136 ($9.168\cdot10^{-4}$)
TMPM	0.966 ($2.796\cdot10^{-4}$)	0.023 ($1.734\cdot10^{-3}$)	0.621 ($2.805\cdot10^{-3}$)	0.083 ($1.687\cdot10^{-3}$)	2392 (102)	0.139 ($9.211\cdot10^{-4}$)
PLM	0.968 ($2.781\cdot10^{-4}$)	0.047 ($2.550\cdot10^{-3}$)	0.640 ($2.765\cdot10^{-3}$)	0.111 ($2.631\cdot10^{-3}$)	2445 (98)	0.136 ($9.167\cdot10^{-4}$)
RF	**0.968 ($2.759\cdot10^{-4}$)**	0.047 ($2.509\cdot10^{-3}$)	0.623 ($2.807\cdot10^{-3}$)	**0.116 ($2.736\cdot10^{-3}$)**	761[a] (49)	**0.134 ($9.180\cdot10^{-4}$)**
GBM	0.968 ($2.768\cdot10^{-4}$)	0.036 ($2.276\cdot10^{-3}$)	0.635 ($2.794\cdot10^{-3}$)	0.109 ($2.545\cdot10^{-3}$)	**1176 (67)**	0.134 ($9.132\cdot10^{-4}$)
NN	0.968 ($2.737\cdot10^{-4}$)	**0.048 ($2.483\cdot10^{-3}$)**	**0.654 ($2.748\cdot10^{-3}$)**	0.112 ($2.610\cdot10^{-3}$)	1463 (66)	0.134 ($9.103\cdot10^{-4}$)

[a] HL for RF could not be computed with 10 subgroups, and was computed with 11 instead. Due to this, HL for RF is not directly comparable.

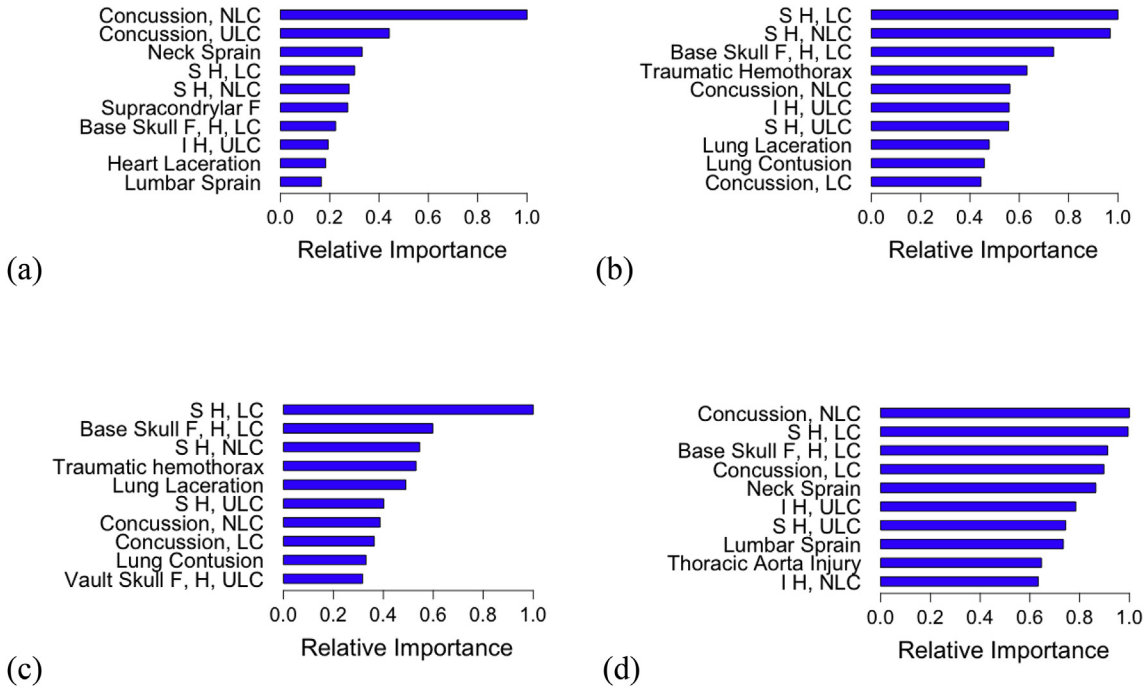

Fig. 13. Ten largest variable importance measures of the selected machine learning models that considered ICD-9 codes only as input variables. The selected logistic regression model developed with the elastic net penalty is represented by (a), random forest (b), gradient boosted machine (c), and neural network (d). Further, F denotes fracture, H hemorrhage, I intracranial, LC loss of consciousness, NLC no loss of consciousness, S subdural, and ULC unspecified loss of consciousness.

prediction models allow for a better evaluation of novel treatments, interventions, protocols, and policies.

There is controversy regarding the utility of machine learning in healthcare [52]. This is in part motivated by several studies that compare generalized linear models to individual machine learning models for risk prediction, often with contradictory results [53–56]. This phenomenon is due, in part, to the fact that no single algorithm is inherently better than all others – depending on the performance metric and the complexity of the data, the best predictive model could be developed from any algorithm [57,58]. This claim is evidenced by the results of this study, as it is unclear which model is best when comparing the established risk prediction models to TSM's machine learning

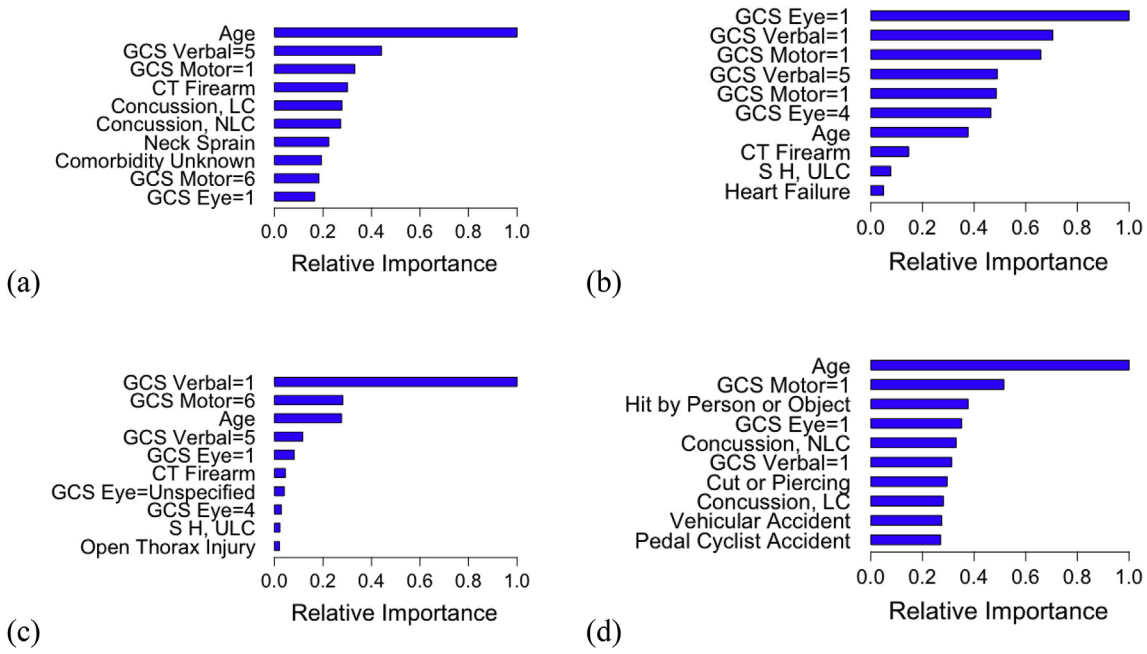

Fig. 14. Ten largest variable importance measures of the selected machine learning models that considered ICD-9 codes, patient demographics, and general trauma assessments as input variables. The selected logistic regression model developed with the elastic net penalty is denoted by (a), random forest (b), gradient boosted machine (c), and neural network (d). Further, GCS denotes Glasgow Coma Scale, CT cause of trauma, H hemorrhage, LC loss of consciousness, NLC no loss of consciousness, S subdural, and ULC unspecified loss of consciousness.

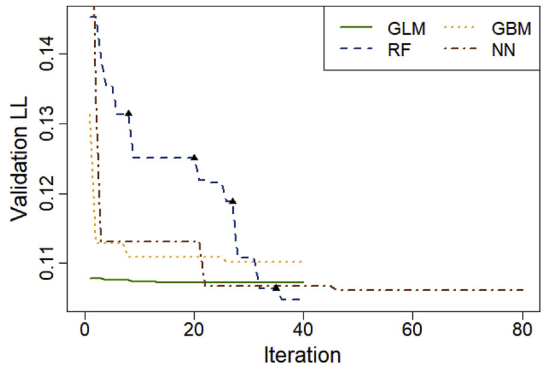

Fig. 15. Lowest recorded log-loss on the validation set (validation LL) after each iteration of our hyper-parameter search procedure when developing machine learning models that considered ICD-9 codes only as input variables. The hyper-parameter space shifted 4 times for the random forest algorithm, and each shift is associated with a decrease in validation LL. Shifts are denoted by a black triangle (▲).

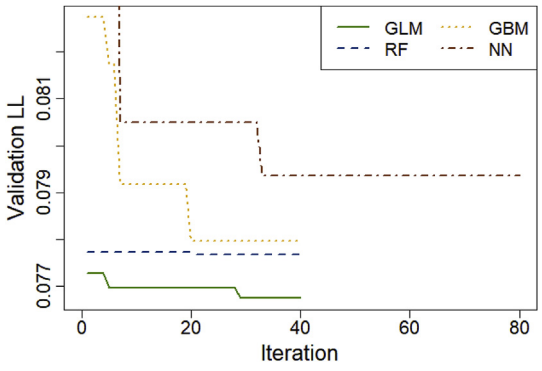

Fig. 16. Lowest recorded log-loss on the validation set (validation LL) after each iteration of our hyper-parameter search procedure when developing machine learning models that considered ICD-9 codes, patient demographics, and general trauma assessments as input variables. The hyper-parameter space did not shift for any machine learning algorithm.

base models.

What separates an ensemble machine learning approach, such as stacked generalization, from a methodology where a single model is selected and assessed is that, if performed appropriately, stacked generalization will utilize the strengths of its base models while compensating for their weaknesses. As a result, it is likely that a well-designed ensemble machine learning model developed from stacked generalization will obtain better predictive performance than any base model in its ensemble [27,28,59–61]. This is empirically demonstrated in this study, as TSM, an ensemble machine learning model, outperforms its base models as well as established risk prediction models on NTDB and NRD data for most performance metrics. While TSM does not have the best calibration metrics on the NRD dataset, we demonstrate that this issue may be resolved by calibrating TSM to NRD data, which gives it better calibration metrics than the calibrated BLISS, HARM, and TMPM models (HARM maintains better HL when calibrated).

While this study highlights the benefits of using an ensemble machine learning model, most medical studies do not require anything more sophisticated than a generalized linear model. This is due to their low computational cost, their simple functional form (which captures the underlying relationships in most medical datasets), and the interpretability as well as consistency of their weights. Arguably, generalized linear models could still be considered most appropriate for mortality risk prediction with trauma patients, as the established risk prediction models have strong predictive performance on NTDB data.

However, as modeling problems with healthcare data become increasingly complex, non-linear machine learning algorithms should be considered, as they are data-driven approaches that automatically find non-linear relationships in data [22]. Generalized linear models, on the other hand, would require extensive feature engineering for such data, and depending on the setting such measures will not result in the development of a model that performs as well as a model developed from machine learning algorithms. This claim is partly validated by our results, as HARM, which is developed using extensive feature engineering based on clinical intuition and expert consensus, generally had the

Table 8
Hyper-parameters of the selected base models.

Hyper-Parameter	ICD-9 Models	Augmented Models
Logistic Regression with Elastic Net Penalty		
λ	10^{-4}	10^{-4}
α	0.845	0.720
Random Forest		
#Trees	83	138
MNL[a]	29	70
NVS[b]	70	107
Max. Tree Depth	157	71
Gradient Boosted Machine		
#Trees	80	66
Max. Tree Depth	15	8
Learning rate	0.670	0.250
Annealing	0.962	0.984
Neural Network		
#Neurons	(41, 42, 12, 49)	(13, 60, 35, 62)
Activation function	ReLU	ReLU
Dropout rates	(0.15, 0.03, 0.28, 0.31)	(0.02, 0.01, 0.23, 0.32)
Epochs	163	91
ρ	0.978	0.991
ε	$3.162{\cdot}10^{-10}$	10^{-10}
Meta-Learner (Gradient Boosted Machine)		
Max. Tree Depth	1	1

[a] MNL: minimum number of observations in a leaf.
[b] NVS: number of variables used in each split.

worst performance metrics on NTDB data. Further, the ICD-9 base models outperform the established risk prediction models on every performance metric with NTDB data.

4.2. Concerning the decline in model performance on NRD data

This study highlights a problem with the assumptions made during model development, as every model experienced a decline in performance when assessed on the NRD dataset (a different population of trauma patients). This decline in performance does not indicate that these assumptions are inherently poor, but rather that they reflect the quality of data that was available at the time these assumptions were proposed.

Limitations due to data quality are obvious when we look at previous iterations of trauma risk prediction models, such as the Injury Severity Score (ISS), which was introduced in 1974. Specifically, the population of patients used to develop ISS was limited to “2128 motor vehicle occupants, pedestrians, and other road users”, and risk prediction from ISS was entirely based on an expert's evaluation of a patient's trauma injuries [2]. While ISS was suitable for several studies at the time of its introduction, it is apparent that ISS is losing its utility in trauma research [62–64].

In order to improve trauma risk prediction, fewer assumptions need to be made about a population of trauma patients when developing a model from a high quality dataset. One assumption that is particularly limiting in this study is the exclusion of patients with incomplete data from analysis, as this is known to introduce error due to bias. This issue may be resolved by using additional binary indicators to represent missing data or by applying imputation methods on a dataset [65]. Another assumption that is particularly limiting is how the established risk prediction models define trauma injury – for example, patients with burns are excluded from model development. This issue may be resolved by including these patients in the dataset for model development.

4.3. Limitations and strengths of the hyper-parameter search procedure

The challenge with developing a well-designed ensemble machine learning model is that the ensemble must consist of base models that have strong predictive performance (ensemble strength) as well as base models that provide different prediction outputs for the same conditions (ensemble diversity) [66,67]. Our hyper-parameter search scheme attempts to address both of these, as a random search can provide both ensemble strength and ensemble diversity with regards to a hyper-parameter space (as randomly selected hyper-parameters may result in the development of a strong or a weak performing model), and hyper-parameter space shifting attempts to improve hyper-parameter space configuration if the optimal hyper-parameters lie beyond the initial hyper-parameter space configured (stronger performing models may be developed after shifting the hyper-parameter space).

While our hyper-parameter search scheme worked in this study, our search scheme is not guaranteed to work for all settings, as it is dependent on the sensitivity of the initial hyper-parameter space configured as well as the state of the random number generator. Issues pertaining to sensitivity may be addressed by taking careful measures to configure an appropriate initial hyper-parameter space. Issues pertaining to random number generation may be addressed by developing a large number of models, examining the hyper-parameters of several models before shifting the hyper-parameter space, and specifying a small distance to shift the hyper-parameter space.

5. Conclusions

The Trauma Severity Model performs better than established risk prediction models on National Trauma Data Bank data as well as Nationwide Readmission Database data, which give it prognostic value in trauma system evaluations. The hyper-parameter search scheme proposed in this study performed well and developed strong performing machine learning models. The performance of an ensemble machine learning model on a well-studied problem in epidemiology indicates that ensemble machine learning approaches may be fruitful for other complex problems in healthcare, and could play a major role in evaluating the quality of care delivered by health care systems.

Conflicts of interest

The authors declare that they have no conflicts of interest in regards to the content in this article.

Acknowledgements

The authors would like to thank Dr. Laurent Glance and Dr. Turner Osler for their invaluable assistance on this project. The authors would also like to thank Ph.D. student Hugo Milan for conversations concerning this manuscript. These results were presented at the Eastern Association for the Surgery of Trauma Annual Scientific Assembly in 2017.

References

[1] T. Mathes, C. Mosch, M. Eikermann, Economic Aspects of Trauma Care, Springer, 2016.
[2] S.P. Baker, B. O'Neill, W. Haddon, W.B. Long, The injury severity score: a method for describing patients with multiple injuries and evaluating emergency care, Journal of Trauma and Acute Care Surgery 14 (3) (1974) 187–196.
[3] H.R. Champion, W.J. Sacco, A.J. Carnazzo, W. Copes, W.J. Fouty, Trauma Score, Critical Care Medicine, 1981.
[4] C.R. Boyd, M.A. Tolson, W.S. Copes, Evaluating trauma care: the triss method, Journal of Trauma and Acute Care Surgery 27 (4) (1987) 370–378.
[5] H.R. Champion, W.S. Copes, W.J. Sacco, M.M. Lawnick, L.W. Bain, D.S. Gann, T. Gennarelli, E. Mackenzie, S. Schwaitzberg, A new characterization of injury severity, Journal of Trauma and Acute Care Surgery 30 (5) (1990) 539–545.
[6] T. Osler, R. Rutledge, J. Deis, E. Bedrick, ICISS: an international classification of disease-9 based injury severity score, Journal of Trauma and Acute Care Surgery 41 (3) (1996) 380–386.
[7] R.S. Burd, M. Ouyang, D. Madigan, Bayesian logistic injury severity score: a method for predicting mortality using International Classification of Disease-9 Codes, Acad. Emerg. Med. 15 (5) (2008) 466–475.
[8] T.A. West, F.P. Rivara, P. Cummings, G.J. Jurkovich, R.V. Maier, Harborview assessment for risk of mortality: an improved measure of injury severity on the basis

of icd-9cm, Journal of Trauma and Acute Care Surgery 49 (3) (2000) 530–540.
[9] L.G. Glance, T.M. Osler, D.B. Mukamel, W. Meredith, J. Wagner, A.W. Dick, Tmpm-icd9: a trauma mortality prediction model based on icd-9-cm codes, Ann. Surg. 249 (6) (2009) 1032–1039.
[10] A. Banerjee, S. Chaudhury, Statistics without tears: populations and samples, Ind. Psychiatry J. 19 (1) (2010) 60–65.
[11] American college of surgeons, Ntdb annual report, https://www.facs.org/~/media/files/quality%20programs/trauma/ntdb/ntdb%20annual%20report%202016.ashx, (2016).
[12] Healthcare cost and utilization project, Overview of the Nationwide Readmissions Database (NRD), https://www.hcup-us.ahrq.gov/nrdoverview.jsp.
[13] American college of surgeons, National trauma data bank, User manual, https://www.facs.org/~/media/files/quality%20programs/trauma/ntdb/ntdbmanual2010.ashx, (2010).
[14] L. de Munter, S. Polinder, K.W.W. Lansink, M.C. Cnossen, E.W. Steyerberg, M.A. C de Jongh, Mortality prediction models in the general trauma population: a systematic review, Injury 48 (2) (2017) 221–229.
[15] J. Langley, R. Brenner, What is an injury? Inj. Prev. 10 (2) (2004) 69–71.
[16] Injury Surveillance Workgroup, Consensus Recommendations for Using Hospital Discharge Data for Injury Surveillance, (2003).
[17] D.E. Clark, T.M. Osler, D.R. Hahn, ICDPIC: Stata Module to Provide Methods for Translating International Classification of Diseases (Ninth Revision) Diagnosis Codes into Standard Injury Categories And/or Scores, Statistical Software Components S457028, (2009).
[18] R. Caruana, Y. Lou, J. Gehrke, P. Koch, M. Sturm, N. Elhadad, Intelligible Models for Healthcare: Predicting Pneumonia Risk and Hospital 30-day Readmission, Knowledge Discovery and Data Mining, 2015.
[19] W.H. Greene, Econometric Analysis, fifth ed., (1993) Ch. 21.
[20] T. Osler, D. Yuan, J. Holden, Z. Huang, A. Cook, L.G. Glance, J.S. Buzas, D.W. Hosmer, Variation in Readmission Rates Among Hospitals Following Admission for Traumatic Injury, Injury, 2018.
[21] R.Y. Calvo, V. Bansal, C.E. Dunne, J. Badiee, C.B. Sise, M.J. Sise, A population-based analysis of outcomes after repair of thoracic aortic emergencies in trauma, J. Surg. Res. 231 (2018) 352–360.
[22] T. Hastie, R. Tibshirani, J. Friedman, The Elements of Statistical Learning, Springer, 2003.
[23] The H2O.ai team, h2o: R Interface for H2O, Version 3.16.0.2, (2017).
[24] A. Zeileis, Econometric computing with hc and hac covariance matrix estimators, J. Stat. Softw. 11 (2004) 1–17.
[25] R Core Team, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, 2017.
[26] A. Niculescu-Mizil, R. Caruana, Predicting Good Probabilities with Supervised Learning, ICML, 2005.
[27] D.H. Wolpert, Stacked Generalization, Neural Networks, 1992.
[28] M.J. van der Laan, E.C. Polley, A.E. Hubbard, Super Learner, Statistical Applications in Genetics and Molecular Biology, (2007).
[29] H. Zou, T. Hastie, Regularization and variable selection via the elastic net, J. R. Stat. Soc. Ser. B 67 (2) (2005) 301–320.
[30] L. Breiman, Random Forests, Machine Learning, 2001.
[31] J.H. Friedman, Greedy function approximation: a gradient boosting machine, Ann. Stat. 29 (5) (2001) 1189–1232.
[32] Y. Bengio, I.J. Goodfellow, A. Courville, Deep Learning, MIT Press, 2015.
[33] M.D. Zeiler, Adadelta: an Adaptive Learning Rate Method, (2012) arXiv:1212.5701.
[34] B. Recht, C. Re, S. Wright, F. Niu, Hogwild: A Lock-free Approach to Parallelizing Stochastic Gradient Descent, NIPS, 2011.
[35] R. Caruana, S. Lawrence, C.L. Giles, Overfitting in Neural Networks: Backpropagation, Conjugate Gradient, and Early Stopping, NIPS, 2001.
[36] Y. Bengio, Practical Recommendations for Gradient-Based Training of Deep Architectures, Springer, 2012.
[37] J. Bergstra, Y. Bengio, Random search for hyper-parameter optimization, J. Mach. Learn. Res. 13 (1) (2012) 281–305.
[38] C. Guo, G. Pleiss, Y. Sun, K.Q. Weinberger, On Calibration of Modern Neural Networks, (2017) 1706.04599.
[39] G.E.A.P.A. Batista, R.C. Prati, M.C. Monard, A Study of the Behavior of Several Methods for Balancing Machine Learning Training Data, SIGKDD Explorations, 2004.
[40] R. Caruana, A. Niculescu-Mizil, G. Crew, A. Ksikes, Ensemble Selection from Libraries of Models, ICML, 2004.
[41] J.A. Hanley, B.J. McNeil, The meaning and use of the area under a receiver operating characteristic (roc) curve, Radiology 143 (1) (1982) 29–36.
[42] J. Davis, M. Goadrich, The Relationship between Precision-Recall and Roc Curves, ICML, 2006.
[43] D.W. Hosmer, S. Lemeshow, Applied Logistic Regression, second ed., Wiley, 2000 Ch. 5.
[44] T. Saito, M. Rehmsmeier, The precision-recall plot is more informative than the roc plot when evaluating binary classifiers on imbalanced datasets, PLoS One 10 (3) (2015) e0118432, , https://doi.org/10.1371/journal.pone.0118432.
[45] T.D. Gedeon, Data mining of inputs: analysing magnitude and functional measures, Int. J. Neural Syst. 8 (2) (1997) 209–218.
[46] A. Canty, B. Ripley, boot, Bootstrap R (S-plus) functions, https://cran.r-project.org/web/packages/boot/boot.pdf.
[47] A.C. Davison, D.V. Hinkley, Bootstrap Methods and Their Applications, Cambridge University Press, 1997.
[48] B. Hamner, M. Frasco, Metrics: evaluation metrics for machine learning, https://CRAN.R-project.org/package=Metrics, (2018).
[49] S.R. Lele, J.L. Keim, P. Solymos, ResourceSelection: resource selection (probability) functions for use-availability data, https://CRAN.R-project.org/package=ResourceSelection, (2017).
[50] Center for disease control, Ten leading causes of death by age group, United States, https://www.cdc.gov/injury/wisqars/pdf/leading_causes_of_death_by_age_group_2014-a.pdf, (2014).
[51] A.B. Martin, M. Hartman, B. Washington, A. Catlin, National Health Expenditure Accounts Team. National health spending: faster growth in 2015 as coverage expands and utilization increases, Health Aff. (2017).
[52] A. Verghese, N.H. Shah, R.A. Harrington, What this computer needs is a physician - humanism and artificial intelligence, J. Am. Med. Assoc. 319 (1) (2017) 19–20.
[53] R. Dybowski, P. Weller, R. Chang, V. Gant, Prediction of outcome in critically ill patients using artificial neural network synthesised by genetic algorithm, The Lancet Respiratory Medicine 347 (9009) (1996) 1146–1150.
[54] G. Clermont, D.C. Angus, S.M. DiRusso, M. Griffin, W.T. Linde-Zwirble, Predicting hospital mortality for patients in the intensive care unit: a comparison of artificial neural networks with logistic regression models, Crit. Care Med. 29 (2) (2001) 291–296.
[55] V.J. Ribas, J.C. Lopez, A. Ruiz-Sanmartin, J.C. Ruiz-Rodriguez, J. Rello, A. Wojdel, A. Vellido, Severe Sepsis Mortality Prediction with Relevance Vector Machines, EMBC, 2011.
[56] S. Kim, W. Kim, R.W. Park, A comparison of intensive care unit mortality prediction models through the use of data mining techniques, Health Informatics Research 17 (4) (2011) 232–243.
[57] R. Pirracchio, M.L. Petersen, M. Carone, M.R. Rigon, S. Chevret, M.J. van der Laan, Mortality prediction in intensive care units with the super icu learner algorithm (sicula): a population-based study, The Lancet Respiratory Medicine 3 (1) (2015) 42–52.
[58] D.H. Wolpert, W.G. Macready, No free lunch theorems for optimization, IEEE Transactions of Evolutionary Compution 1 (1) (1997) 67–82.
[59] D. Opitz, R. Maclin, Popular ensemble methods: an empirical study, J. Artif. Intell. Res. 11 (1) (1999) 169–198.
[60] R. Polikar, Ensemble based systems in decision making, IEEE Circuit Syst. Mag. 6 (3) (2006) 21–45.
[61] L. Rokach, Ensemble-based classifiers, Artif. Intell. Rev. 30 (2010) 1–39.
[62] T. Osler, L. Glance, J.S. Buzas, D. Mukamel, J. Wagner, A. Dick, A trauma mortality prediction model based on the anatomic injury scale, Ann. Surg. 247 (6) (2008) 1041–1048.
[63] P.D. Kilgo, J.W. Meredith, R. Hensberry, T.M. Osler, A note on the disjointed nature of the injury severity score, J. Trauma 57 (3) (2004) 479–485.
[64] R. Rutledge, J Trauma, the injury severity score is unable to differentiate between poor care and severe injury, J. Trauma 40 (6) (1996) 944–950.
[65] R.J.A. Little, D.B. Rubin, Statistical Analysis with Missing Data, Wiley, 1987.
[66] M. Gashler, C. Giraud-Carrier, T. Martinez, Decision Tree Ensemble: Small Heterogeneous Is Better than Large Homogeneous, ICML, 2008.
[67] L. Kuncheva, C. Whitaker, Measures of diversity in classifier ensembles, Mach. Learn. 51 (2) (2003) 181–207.