The Trauma Severity Model: An Ensemble Learning Approach to Risk Prediction

Michael T. Gorczyca[1]
Nicole C. Toscano[2]
Julius D. Cheng[2]

[1]Department of Information Science, Cornell University, Ithaca, NY, USA
[2]Department of Surgery, University of Rochester Medical Center, Rochester, NY, USA

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 in the context of mortality risk prediction for trauma patients, as researchers have focused almost exclusively on the use of linear models for risk prediction, and linear models may be too inflexible to capture the potentially complex relationships in trauma data. Due to this, we propose an ensemble machine learning model, the Trauma Severity Model (TSM), for risk prediction. In order to empirically validate TSM's predictive performance, this study compares TSM to three established risk prediction models: the Bayesian Logistic Injury Severity Score (BLISS), the Harborview Assessment for Risk of Mortality (HARM), and the Trauma Mortality Prediction Model (TMPM-ICD9). Our results indicate that TSM has superior predictive performance, and thereby provides improved risk prediction.

1. Introduction

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 [2-14].

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 data sets [2-4, 6, 8, 10-14], which implies that these models may not represent the population appropriately [15]. Another reason is that even if a model was developed from a large data set, nearly every model proposed thus far has been a linear model [2-9], and linear models may be insufficient for capturing the potentially complex relationships that exist within trauma data [16-18].

By taking advantage of recent advances in computational science and statistical

Corresponding Author: Michael T. Gorczyca. Email: mtg62@cornell.edu

theory, which are leading to substantial changes in both the availability of large data sets and the ability to perform intensive analyses on such data sets, our objective is to develop a risk prediction model from machine learning algorithms and compare its predictive performance to the performance of other established risk prediction models. We achieve this by comparing three leading 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-ICD9) [9] — to a new machine learning model for risk prediction. This machine learning model is the Trauma Severity Model (TSM).

2. Materials and Methods

2.1. *Data Summary and Processing*

This study was performed using data from the National Trauma Data Bank on patients hospitalized in 2008, 2009, 2010, and 2012. Our initial data set consisted of 2,865,867 unique patient records, 884 hospitals, and 2,105 ICD-9-CM trauma diagnoses (ICD-9 trauma codes). To facilitate the data cleaning process for this initial data set, we first selected patients from the initial data set (the patient selection process) and then the ICD-9 trauma codes from those remaining (the trauma diagnosis cleaning process).

For our patient selection process, patients were excluded if they had burns or a primary diagnosis unrelated to trauma (e.g., poisoning, drowning, or suffocation) (193,606), were admitted to a hospital that did not maintain complete documentation of relevant trauma diagnoses (655,440), were missing data (for age, comorbidities, gender, Glasgow Coma Scale sub-scores, 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 of age (47,693). Some patients were excluded due to more than one exclusion criteria.

These selection criteria are nearly identical to that used in TMPM-ICD9's study, and are in accordance with the selection criteria for BLISS and HARM (Appendix A). Specifically, there are two differences between the patient selection process in this study and that in TMPM-ICD9's study. One difference between our cleaning procedure and TMPM-ICD9's procedure is in how we selected hospitals from which we selected patients. In TMPM-ICD9's study, the data set 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 all patients that were admitted to any hospital that kept complete records of the ICD-9 trauma codes that we considered relevant. Our reasoning for this is that some trauma centers that would qualify as having substantial trauma experience omitted relevant ICD-9 trauma codes from their registry, and this could harm each model's ability to provide accurate risk predictions [19]. Another difference is that TMPM-ICD9's study only ensured complete documentation for age, gender, and outcome when determining which patients to include. We extended this to also ensure complete documentation for comorbidities, Glasgow Coma Scale sub-scores, injury mechanism, injury type, and intent of trauma. Our 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.

For our trauma diagnosis cleaning process, all trauma diagnoses that were treated as non-injuries in TMPM-ICD9's original study were also treated as non-injuries — these were also in correspondence with the injuries excluded in BLISS and HARM's studies. Following this, the data set was copied such that each model had its own data set, and each model's data set was cleaned to match the trauma code specifications required by that model's study (Appendix B).

For TSM, ICD-9 trauma codes that involved neurologic injuries followed the same re-classification approach as TMPM-ICD9, as these diagnoses codes are differentiated by information that can only be determined at discharge (for instance, several ICD-9 trauma codes for skull fractures are differentiated by the duration of loss of consciousness). ICD-9 trauma codes that appeared fewer than five times in TSM's data set were combined with what was empirically determined to be the closest corresponding trauma code. This 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 one single injury.

Each post-processed data set consisted of the same 1,385,795 unique patient records from 713 hospitals and accounted for 1,920 ICD-9 trauma codes.

2.2. *Model Development*

Injury severity was formalized as a binary classification task in which the independent variables are a patient's trauma diagnoses. Each ICD-9 trauma code is a binary indicator specifying whether or not a patient had that particular trauma diagnosis. The dependent variable is a binary indicator specifying whether or not the patient died prior to discharge. For TSM, we utilized an ensemble machine learning approach for model development — ensemble machine learning is a methodology where several different machine learning models are developed and then combined together to provide a single prediction output [20]. Specifically, we followed an ensemble machine learning framework known as stacked generalization, or "stacking" [21].

Our approach to stacking followed this sequence. First, the data set is divided into four parts: the "base model training set," the "meta-learner data construction set," the "validation set," and the "test set." The base model training set is used to develop a variety of different machine learning models, which are referred to as "base models." In this study, the base models were developed from four machine learning algorithms: penalized linear models, random forests, gradient boosted machines, and neural networks (Appendix C). Each of these base models then predicts each patient from the meta-learner data construction set's risk. These risk predictions are used to create a "meta-learner training set," in which each row is a patient from the meta-learner data construction set, and each column is a base model's risk prediction for that patient. The meta-learner training set is then used to developed a higher-level model (a meta-learner) in which the meta-learner's input variables are each base model's risk prediction and the dependent variable is a binary indicator specifying whether or not the patient died prior to discharge. Qualitatively, the meta-learner is determining how

to combine its base models together in order to output an improved risk prediction.

To avoid over-fitting, we randomly selected 80% of the entire data set to develop each linear model and TSM (the training set), 10% of the entire data set to determine the regression coefficients that optimized each linear model's predictive performance as well as which meta-learner configuration optimized TSM's predictive performance in terms of Akaike's information criterion (the validation set), and 10% of the data set to gather the performance metrics for the resulting models (the test set). Further, for TSM we randomly selected 75% of the training set as the base model training set, and 25% as the meta-learner data construction set (in other words, 60% and 20% of the entire data set was used as the base model training set and the meta-learner data construction set, respectively).

All model development was performed using the h2o [22] and sandwich [23] packages in the R statistical software. Percentile bootstrapped confidence intervals, which provide a range of values that each performance metric for each model lies within, were computed using the R statistical software's boot package [24, 25].

2.3. *Model Assessment*

The predictive performance of each risk prediction model was evaluated using Akaike's information criterion (AIC), the area under the receiver operating characteristic curve (ROC), the Hosmer-Lemeshow statistic (HL), and the reliability statistic (REL). The AIC statistic is a measure of how well a model approximates the true underlying distribution [26]. For the purpose of model selection, the best model for a particular data set is the model with the lowest AIC statistic.

The ROC statistic measures a model's ability to discriminate between outcomes [27]. For example, if a model was given two patients, one who survived treatment and another who succumbed to their injuries, then the ROC statistic would represent how much higher the corresponding prediction output would be for the patient that died compared to that for the patient that survived. A ROC statistic of 1 indicates that the corresponding model is perfect at discriminating between patient outcomes, whereas a ROC statistic of 0.5 indicates that the corresponding model has no discriminatory power. In other words, if a model has a ROC statistic of 0.5, then its prediction outputs would be equivalent to random guesses.

Finally, the HL and REL statistics are measures of a model's probabilistic calibration. In the context of risk prediction, these statistics indicate how close a model's risk prediction is to a patient's true probability of survival. The HL statistic is a standard approach for evaluating a risk prediction model's probabilistic calibration [28]. However, because the HL statistic may be considered inappropriate for large data sets [29-31], we use the REL statistic as an adjunct measure of probabilistic calibration [32]. For both statistics, a value of 0 indicates that the corresponding model is perfectly calibrated to the data; the REL statistic has an upper bound of 1 (implying that the model is completely unreliable), whereas the HL statistic has no upper bound (the higher the HL statistic, the worse the probabilistic calibration).

Table 1: Patient demographics for the processed data set.

Patient Demographics	**Statistic**
Age*	(23, 61)
Female	36.07%
Hospitals†	713
Dead	3.77%
Race	
-White	64.52%
-Black of African American	16.17%
-Hispanic‖	12.74%
-Asian	1.93%
-Native American/Native Hawaiian/Pacific Islander	0.79%
-Other	10.82%
-Not Recorded	5.77%

(*) Interquartile range displayed.
(†) Hospital demographics not displayed as demographic information (such as ACS certification status and number of hospital beds) changed over the course of this study.
(‖) Hispanic is denoted as an ethnicity in NTDB data, not race.

Table 2: ICD-9 trauma code model comparison in terms of AIC, ROC, HL, and REL statistics. Each model's REL statistic was multiplied by 1,000,000.

Model	**AIC (95% CI)**	**ROC (95% CI)**	**HL (95% CI)**	**REL (95% CI)**
TSM	27661.4 [26973, 28349]	0.907 [0.902, 0.911]	110.7 [86.4, 154.4]	29.7 [18.1, 72.9]
BLISS	31414.0 [30706, 32145]	0.900 [0.895, 0.905]	247.1 [210.8, 306.5]	549.7 [459.3, 684.4]
HARM	29561.9 [28869, 30277]	0.871 [0.866, 0.877]	98.4 [72.5, 140.7]	92.9 [66.4, 150.1]
TMPM-ICD9	28868.3 [28168, 29606]	0.897 [0.893, 0.902]	81.6 [58.0, 136.5]	40.9 [24.3, 89.4]

3. Results

The patient demographics are displayed in Table 1. The interquartile range for age was from 23 to 61 years. Females comprised 36.07% of this data set. The mortality rate was 3.77%.

The predictive performance of each model is displayed in Tables 2, 3 as well as Figures 1, 2. TSM demonstrates an improvement over BLISS, HARM, and TMPM-ICD9 in terms of the AIC, ROC, and REL statistics (Table 2). Every model greatly improved in predictive performance when augmented to account for age, comorbidities, gender, Glasgow Coma Scale sub-scores, injury mechanism, injury type, and intent of trauma (Table 3). However, TSM still outperforms each linear model in terms of the AIC and ROC statistics. Although TSM does not uniformly improve over the other models in terms of probabilistic calibration, Figures 1, 2 demonstrates that TSM is still well calibrated to the data.

Table 3: Augmented model comparison in terms of AIC, ROC, HL, and REL statistics. Each model's REL statistic was multiplied by 1,000,000.

Model	AIC (95% CI)	ROC (95% CI)	HL (95% CI)	REL (95% CI)
TSM	19755.8 [19207, 20343]	0.963 [0.961, 0.965]	169.5 [141.9, 217.5]	36.5 [21.8, 87.0]
BLISS	22964.9 [22379, 23568]	0.956 [0.953, 0.959]	48.7 [34.7, 78.1]	114.5 [85.1, 177.3]
HARM	20688.4 [20113, 21282]	0.956 [0.953, 0.959]	64.7 [49.7, 96.0]	93.0 [68.9, 152.8]
TMPM-ICD9	21002.6 [20433, 21609]	0.957 [0.954, 0.960]	109.3 [30.3, 359.7]	18.3 [12.2, 58.3]

Figure 1: Calibration curves for TSM, HARM, BLISS, and TMPM-ICD9 models that only considered ICD-9 trauma codes as input variables.

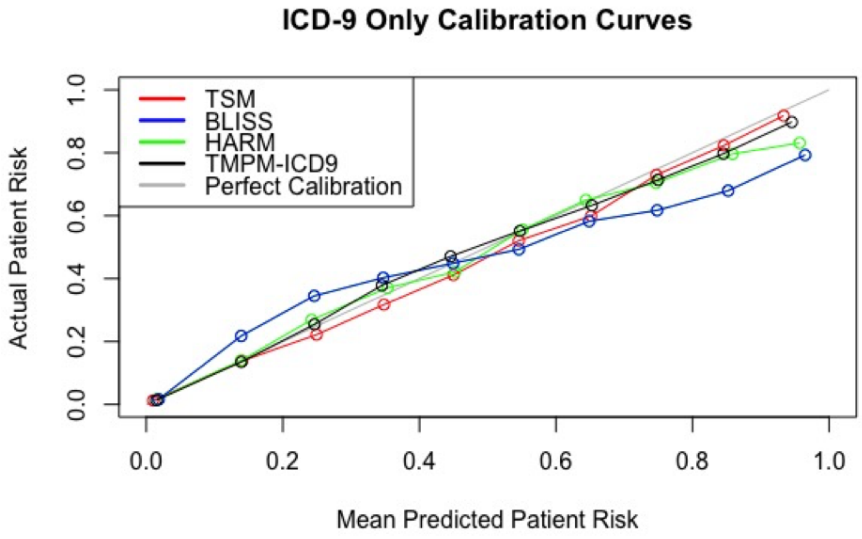
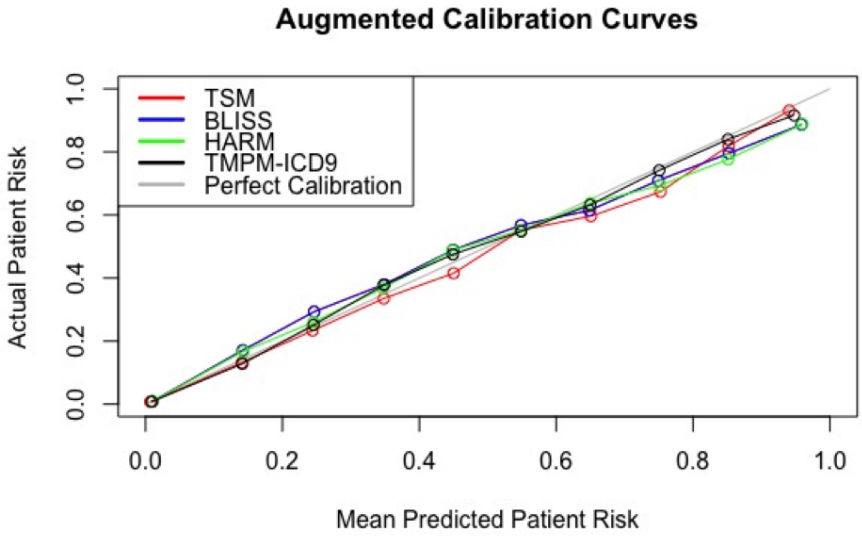


Figure 2: Calibration curves for augmented TSM, HARM, BLISS, and TMPM-ICD9 models.



4. Discussion

The need for quality trauma care is tremendous. In the United States alone, healthcare spending accounts for 17.8% of the Gross Domestic Product [33]. Trauma care specifically comprises a significant amount of this expenditure, as it 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 [34]. But, in order to achieve the goals of improving the quality of trauma care while decreasing the cost of care, we must utilize the best possible risk prediction models in trauma system evaluations. If risk prediction can be improved upon, so too can the quality of trauma care, as better risk prediction models allow for a better evaluation of novel treatments, interventions, and policies. This gives purpose to our study, which was to determine how a machine learning model would perform relative to leading risk prediction models. Our results demonstrate that TSM, a machine learning model, outperforms established risk prediction models in terms of the AIC and ROC statistics. Although it could be argued that TSM does not uniformly improve in terms of probabilistic calibration, there is a tradeoff between discrimination and probabilistic calibration [35], which indicates that TSM improves in discrimination without sacrificing probabilistic calibration.

There have been several previous studies comparing linear models to individual machine learning models for risk prediction, often with contradictory results [10-13]. Pirracchio et al. concluded that this phenomenon "[underscores] the fact that no single algorithm invariably outperforms all others. In any given setting, according to the outcome of interest, the set of explanatory variables available and the underlying population to which it will be applied, the best predictive model might be achieved by a parametric or any of a variety of nonparametric methods" [14]. What separates an ensemble machine learning approach such as stacked generalization from a methodology in which only one model is developed is that, if performed appropriately, stacked generalization will utilize its base models' strengths while compensating for their weaknesses. As a result, it is possible for an ensemble machine learning model developed from stacked generalization to obtain better predictive performance than any base model in its ensemble could possibly obtain alone [20-21, 36-37].

Despite TSM's strong predictive performance, we emphasize that risk prediction can be further improved. One reason for this is that the patients selected during the cleaning process of model development may not be fully representative of the true population of all trauma patients. This indicates that, depending on the degree at which a subpopulation of patients differs from the population of patients used to develop TSM, other mortality prediction models may be better suited than TSM for analyzing that subpopulation. Another reason is that our base model development methodology was rather simple. If we considered more than four machine learning algorithms and performed a more rigorous optimization procedure, we may have been able to further improve predictive performance (Appendix C). A third reason is that there may have been an insufficient amount of data to appropriately capture the effect every ICD-9 trauma code has on outcome. The use of a larger repository of data with more information about each patient's trauma injuries could result in a model that performs substantially better.

We note that given the predictive performance of the other risk prediction models, TSM may not offer a clinically significant advantage. In order to facilitate further

comparison between these models, we have developed an accessible, user-friendly software application that provides each model's risk prediction for a set of trauma injuries (Appendix D).

5. Conclusion

TSM improves over established risk prediction models in terms of the AIC and ROC statistics without sacrificing probabilistic calibration, which gives it prognostic value in trauma system evaluations. To provide researchers access to the risk prediction models from this study, we have developed a user-friendly, freely available software application. The performance of an ensemble machine learning approach on a well-studied problem in epidemiology indicates that ensemble machine learning approaches may be fruitful for other complex problems in healthcare.

6. 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 the Cornell University Data Science Club for their support. These results were presented at the Eastern Association for the Surgery of Trauma Annual Scientific Assembly in 2017.

7. Appendices

7.1. *Appendix A*

Patients were excluded from analysis if they had an ICD-9-CM code that did not pertain to injury (800 to 959.9). We also excluded patients experiencing effects of foreign body entering through body orifice (930 to 939.9), burns (940 to 949.5), and certain late effects of injuries (906.5 to 906.9, 909 to 909.2, 909.4 to 909.9). ICD-9-CM codes that were not considered input variables ("non-injuries") for each model were superficial injuries (910 to 924.9), certain traumatic complications of physical trauma (958.2 to 958.6, 958.8 to 958.99), and late effects of injuries that were not already excluded (905-909.9). These are the same non-injuries defined in TMPM-ICD9's study.

BLISS and HARM had similar selection processes. BLISS excluded patients that had late effects of injury (905 to 909.9), effects of foreign bodies entering through body orifice (930 to 939.9), and burns (940 to 949.5). HARM excluded minor injuries, such as superficial wounds and minor orthopedic injuries (910 to 924.9); burns and burn-related injuries (940 to 949.5); late effects of injuries (905 to 909.9); effects of foreign body entering through body orifice (930 to 939.9); and certain traumatic complications of physical trauma (958 to 958.99) as input variables.

7.2. *Appendix B*

The following describes the procedures for re-estimating BLISS, HARM, and TMPM-ICD9. We first describe the data cleaning process, and then the model development process.

BLISS

For the BLISS data set, ICD-9 trauma codes that involved neurologic injuries followed the same re-classification approach as TMPM-ICD9, as these diagnoses are differentiated by information that can only be determined at discharge. We then developed two Bayesian logistic regression models: one that had a Laplace prior distribution, and another that had a Gaussian prior distribution. These were the prior distributions considered in BLISS's original study. We evaluated each model's predictive performance (in terms of the AIC statistic) on the validation set, and selected the better performing model as BLISS. In this study, both ICD-9 trauma code and augmented BLISS models had a Laplace prior distribution.

The difference between our approach and that of the original study is that the original study used cross-validation rather than a validation set to determine which Bayesian logistic regression model had better predictive performance. The reason we used the validation set instead was to maintain consistency in our study, as every other model in this study used the validation set to optimize predictive performance. We emphasize that second and third-order interaction terms were not included in our BLISS models, as such consideration greatly worsened each BLISS model's AIC statistic.

HARM

When cleaning the HARM data set, we followed HARM's variable combining procedure as closely as possible. However, there are a couple of differences between our cleaning process and that specified in the original study. One difference is that we did not include any diagnoses related to chronic obstructive pulmonary disease and ischemic heart disease (these corresponded to three potential input variables from HARM's original data set), as this information was unavailable in our National Trauma Data Bank data. The other difference is that we allowed the HARM model to utilize variables in our data set that were unavailable in its original data set as potential input variables. We then re-estimated each HARM model using forward selection, as had been performed in its original study — the AIC statistic on the validation set was used for selecting each variable.

Unlike the other models in this study, our re-estimated HARM models used ICD-9 trauma codes that could only be determined at discharge (such as trauma codes for skull fractures that are differentiated by the duration of loss of consciousness). These trauma codes appear to have a significant influence on both the original HARM model and our re-estimated HARM model's predictive performance.

TMPM-ICD9

To clean the TMPM-ICD9 data set, we collapsed ICD-9 trauma codes for neurologic injuries, vertebral column with spinal injury, and open and closed injuries of the larynx and trachea as specified in the original study. We then mapped ICD-9 codes to "region-severity codes" using the voting algorithm described in the original study. Following this, we re-estimated TMPM-ICD9 as specified in its original study, with one difference. We selected the coefficients for TMPM-ICD9 using forward selection, where the AIC statistic on the validation set was used for selecting each variable. The original study only used the five largest "severity measures" (MARC values) [9].

7.3. *Appendix C*

Provided below is a brief overview of the machine learning algorithms used in this study: penalized linear models (a linear algorithm) as well as random forests, gradient boosted machines, and neural networks (non-linear algorithms). Each of these algorithms have hyper-parameters that can be used to tune a corresponding model to the data and improve that model's predictive performance.

In this study, our ensemble of base models consisted of 100 penalized linear models, 50 random forests, 50 gradient boosted machines, and 50 neural networks. Each model had different, randomly selected hyper-parameters [38]. We then trained 100 penalized linear models as meta-learners, and selected the best performing penalized linear model (the model with the lowest AIC statistic for the validation set) as TSM.

Penalized Linear Models

Penalized logistic regression is similar to logistic regression, which utilizes maximum likelihood estimation for its regression coefficients:

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{n} log(1 + exp(-y_i \beta^T x_i))\}$$

The difference between the two is that penalized logistic regression adds a regularization term to its loss function. One popular form of regularization is ridge regression [39], where the penalty applied to a logistic regression model's coefficients is proportional to the sum of the squares of those coefficients. Another popular form of regularization is the LASSO [40], where the penalty applied to a logistic regression model's coefficients is proportional to the sum of the magnitudes of those coefficients. As neither method uniformly outperforms the other for all data sets [41], we developed our penalized linear models using the elastic net penalty [42], which combines both the LASSO and ridge regression penalties:

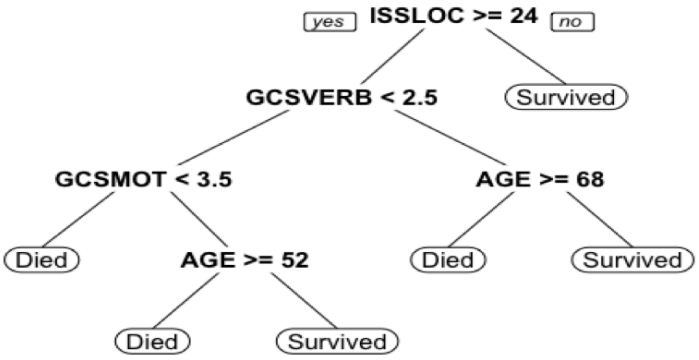
$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{n} log(1 + exp(-y_i \beta^T x_i)) + \lambda \sum_{j=1}^{p} (a|\beta_j| + \frac{(1-a)}{2}|\beta_j|^2)\}$$

Thus, the two hyper-parameters for penalized linear models (with the elastic net penalty) are λ, which is the severity of the penalty applied, and α, which distributes λ between the LASSO and ridge penalties. When $\alpha = 0$, this formulation is equivalent to ridge regression; when $\alpha = 1$, this formulation is equivalent to the LASSO; and when $\lambda = 0$, this formulation is equivalent to ordinary logistic regression. In this study, we varied λ from 10^{-10} to 1 with an order of magnitude of 10 $\{10^{-10}, 10^{-9}, ..., 10^{-1}, 1\}$ and α from 0 to 1 with an increment of 0.001 $\{0, 0.001, ..., 0.999, 1\}$.

Random Forests

A binary classification tree (decision tree) is a simple prediction model that maps observations to a class output using a tree structure (Figure 3). The random forest algorithm is an extension of decision trees, in which several decision trees (an ensemble)

Figure 3: A decision tree that predicts for patient survival using that patient's age ("AGE"); Injury Severity Scores ("ISSLOC"); Glasgow-Coma Scale motor, verbal, and eye response scores ("GCSMOT," "GCSVERB," and "GCSEYE," respectively); and gender ("GENDER"). This tree was developed using a random sample of 100,000 patients from National Trauma Data Bank data in 2008. At each split, the user will move down the left branch if the patient has that condition, and down the right if the patient does not have that condition.



are developed independently of each other and then combined together to provide a single prediction output. What separates random forests from other tree ensemble methods, such as bagging, is that random forests only consider a random subset of all the possible input variables when creating each decision-making "split" in a decision tree. The rationale for this approach is that it enhances diversity [43, 44] — for bagging, it is possible for several highly similar decision trees to be developed and combined together, which can harm predictive performance [45].

We varied three hyper-parameters for our random forests: the number of decision trees, the number of input variables randomly sampled at each split in a tree, and the maximum depth allowed to grow a tree. Specifically, we varied the number of trees from 50 to 200 with an increment 1 $\{50, 51, ..., 199, 200\}$, the number of input variables randomly sampled from 20 to 50 with an increment of 1 $\{20, 21, ..., 49, 50\}$, and the maximum depth allowed to grow a tree from 1 to 20 with an increment of 1 $\{1, 2, ..., 19, 20\}$.

Gradient Boosted Machines

Gradient boosted machines are similar to random forests, yet differ because random forests develop each decision tree independently of the others, while gradient boosted machines develop each decision tree sequentially to the others. This approach allows a gradient boosted machine to "learn" from its prediction errors, and consequently to improve in providing predictions for conditions that other models cannot account for

appropriately. The tradeoff of this approach is that if a user specifies too many trees, a gradient boosted machine may over-fit to the data: a gradient boosted machine may eventually overreact to its prediction errors and try to account for random noise in the data. If a gradient boosted machine over-fits to the data, then it will exhibit poor predictive performance [46, 47].

For our gradient boosted machines, we considered the following hyper-parameters: the number of trees, the learning rate, and the maximum interaction depth. In this study, we varied the number of trees from 40 to 80 with an increment 1 $\{40, 41, ..., 79, 80\}$, the learning rate from 0.05 to 0.30 with an increment of 0.01 $\{0.05, 0.06, ..., 0.29, 0.30\}$, and the maximum interaction depth from 1 to 10 with an increment of 1 $\{1, 2, ..., 9, 10\}$. We emphasize that we scaled the learning rate by a factor of 0.99 after each tree was developed in a gradient boosted machine (in literature, this scaling process is referred to as learning rate annealing).

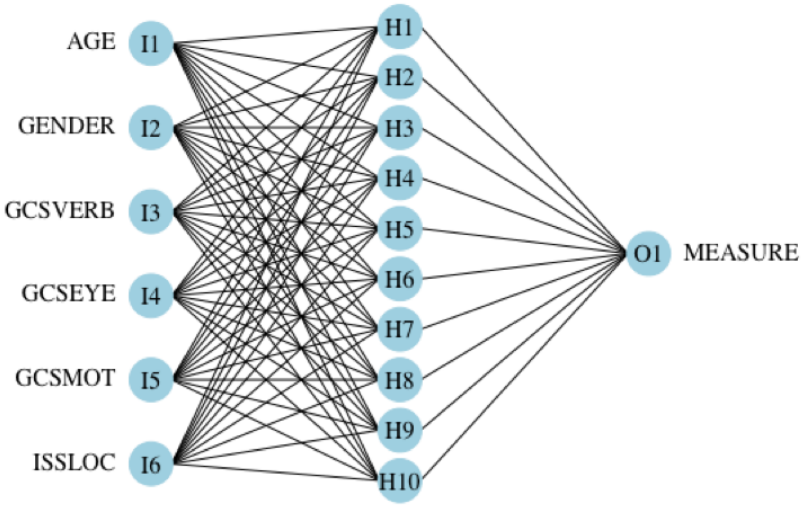
Neural Networks

The term neural network describes a large class of algorithms that are loosely based on the biological brain, and can qualitatively be viewed as a mathematical function with millions of weights that modify and combine information from input data in order to provide a highly accurate response (prediction). In this study, we used feed-forward neural networks, which consist of several smaller functions that are combined together to create a general mathematical function. Specifically, these smaller functions are separated into interconnected "layers." The first layer of functions use the actual data as their inputs, and the output from each function in this layer is used as inputs for the second layer to functions (in literature, these functions are referred to as neurons and these layers are referred to as hidden layers). Each layer of neurons following will use a previous layer of neurons' outputs as inputs until the output layer is reached. The output of the neuron in the output layer is the neural network's prediction for the input data (Figure 4). For a more detailed treatment of neural networks, the authors recommend [48, 49].

We gave consideration to the following hyper-parameters when developing our neural networks: the number of neurons in each layer $\{64, 128, 256, 512, 1024\}$, the number of layers $\{1, 2, 3\}$, and the number of epochs $\{10, 11, ..., 9999, 10000\}$. Further, the activation function (which specifies how each neuron "responds" to its inputs) took the form of a hyperbolic tangent sigmoid, a rectified linear unit [50], or a maxout [51]. The use of dropout, which is a way to prevent neural networks from over-fitting to the data, was also varied [52].

We emphasize that we used the ADADELTA optimization scheme to develop our neural networks, which has two hyper-parameters [53]. One hyper-parameter is ρ, which is similar to momentum and relates to the memory of prior weight updates. The other hyper-parameter is ϵ, which is similar to learning rate annealing during initial training and momentum at later stages of training — in these later stages, ϵ dictates forward progress. We varied ρ from 0.750 to 0.999 with an increment of 0.001 $\{0.750, 0.751, ..., 0.998, 0.999\}$ and ϵ from 10^{-12} to 10^{-3} with a magnitude of $10^{0.5}$ $\{10^{-12}, 10^{-11.5}, ..., 10^{-3.5}, 10^{-3}\}$. We note that we restricted the mini-batch size to 1.

Figure 4: This feed forward neural network, which was developed from the same data as that from Figure 3, consists of one hidden layer with ten neurons (each marked with an "H"). These neurons pass their prediction outputs to the output layer (denoted as "O1"), which outputs a final risk prediction (denoted as "MEASURE").



7.4. *Appendix D*

Our software application allows the user to both manually enter a patient's ICD-9 trauma codes as well as input a data set that consists of a patient's ICD-9 trauma codes (Figures 5, 6). The application will be made freely available by request.

Figure 5: The default state of our software.

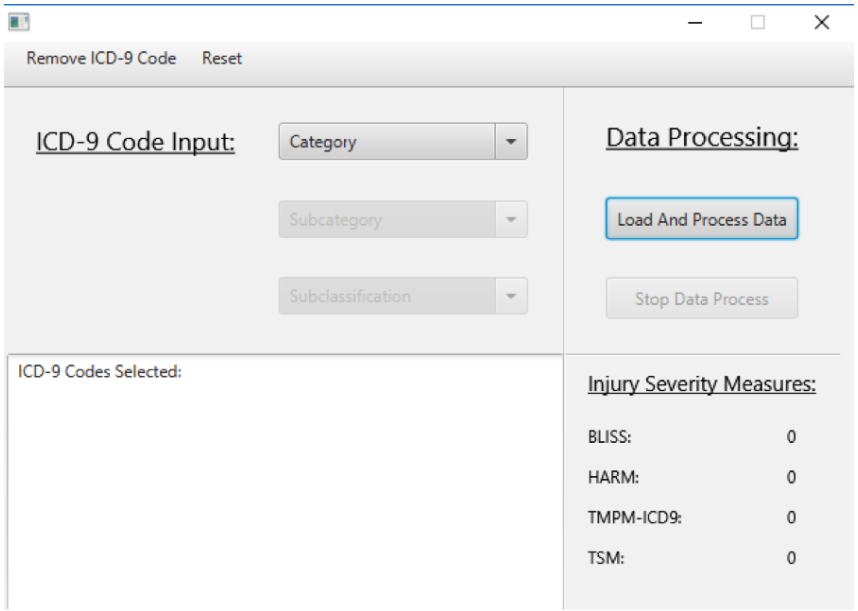
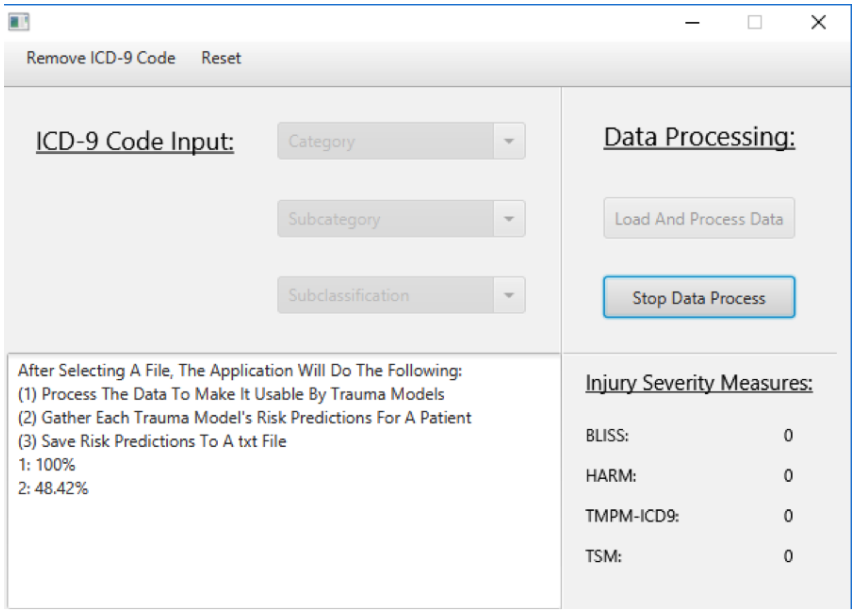


Figure 6: Our software processing a trauma data set.



8. References

[1] Mathes, Tim, et al. Economic Aspects of Trauma Care. Chapter 2. Berlin: Springer. 2016.

[2] Baker SP, O'Neill B, Haddon W Jr, Long WB. The injury severity score: a method for describing patients with multiple injuries and evaluating emergency care. J Trauma. 14: 187-196. 1974.

[3] Champion HR, Sacco WJ, Carnazzo AJ, et al. Trauma score. Crit Care Med. 9: 672-676. 1981.

[4] Boyd CR, Tolson MA, Copes WS. Evaluating trauma care: the TRISS method. J Trauma. 27: 370. 1987.

[5] Champion HR, Copes WS, Sacco WJ, et al. A new characterization of injury severity. J Trauma. 30: 539-545. 1990.

[6] Osler T, Rutledge R, Deis J, Bedrick E. ICISS: an international classification of disease-9 based injury severity score. J Trauma. 41: 380-386. 1996.

[7] Burd RS, Ouyang M, Madigan D: Bayesian logistic injury severity score: a method for predicting mortality using international classification of disease-9 codes. Acad Emerg Med. 15: 466-75. 2008.

[8] T. Al West, F.P. Rivara, P. Cummings, et al. Harborview Assessment for Risk of Mortality: an improved measure of injury severity on the basis of ICD-9CM J. Trauma, 49: 530-540. 2000.

[9] Glance LG, Osler TM, Mukamel DB, Meredith W, Wagner J, Dick AW. TMPM-ICD9: a trauma mortality prediction model based on ICD-9-CM codes. Ann Surg. 249: 1032-1039. 2009.

[10] Dybowski R, Weller P, Chang R, Gant V. Prediction of outcome in critically ill patients using artificial neural network synthesised by genetic algorithm. Lancet. 1996;347:114650.

[11] Clermont G, Angus DC, DiRusso SM, Griffin M, Linde-Zwirble WT. Predicting hospital mortality for patients in the intensive care unit: a comparison of artificial

neural networks with logistic regression models. Crit Care Med. 2001;29:2916.

[12] Ribas VJ1, Lpez JC, Ruiz-Sanmartin A, Ruiz-Rodrguez JC, Rello J, Wojdel A, Vellido A. Severe sepsis mortality prediction with relevance vector machines. Conf Proc IEEE Eng Med Biol Soc. 2011;2011:1003.

[13] Kim S, Kim W, Park RW. A Comparison of Intensive Care Unit Mortality Prediction Models through the Use of Data Mining Techniques. Health Inform Res. 2011;17:23243.

[14] Pirracchio R, Petersen ML, Carone M, Rigon MR, Chevret S, Van der Laan MJ. Mortality prediction in intensive care units with the super ICU learner algorithm (SICULA): A population-based study. The Lancet Respir Med. 2015:3(1), 4252.

[15] Banerjee, A. and Chaudhury, S. 2010. Statistics without tears: Populations and samples. Ind. Psychiatry J, 19, 60-5. 2010.

[16] James G, Witten D, Hastie T, and Tibshirani R. An Introduction to Statistical Learning: with Applications in R. Chapter 2. New York: Springer; 2014.

[17] Hastie T, Tibshirani R, Friedman J. The Elements of Statistical Learning. Chapter 7. New York: Springer; 2003.

[18] Murray IA, Ghahramani Z. 2005 A note on the evidence and Bayesian Occam's razor. Technical Report GCNU-TR 2005-003 London, UK:Gatsby Unit, University College London.

[19] Greene, W.H., 1993. Econometric Analysis, fifth ed. Chapter 21. Macmillan, New York.

[20] Opitz, D.; Maclin, R. (1999). "Popular ensemble methods: An empirical study". Journal of Artificial Intelligence Research. 11: 169198. doi:10.1613/jair.614.

[21] Wolpert DH. Stacked generalization. Neural Networks, 1992; 5(2): 241-259.

[22] Aiello S, Kraljevic T, Maj P and with contributions from the H2O.ai team: h2o: R Interface for H2O. 2016. Available at: https://cran.r-project.org/web/packages/h2o/h2o.pdf, last accessed 30 Nov 2016.

[23] Achim Zeileis (2004). Econometric Computing with HC and HAC Covariance Matrix Estimators. Journal of Statistical Software 11(10), 1-17. URL http://www.jstatsoft.org/v11/i10/.

[24] Angelo Canty and Brian Ripley (2016). boot: Bootstrap R (S-Plus) Functions. R package version 1.3-18.

[25] Davison, A. C. and Hinkley, D. V. (1997) Bootstrap Methods and Their Applications. Cambridge University Press, Cambridge. ISBN 0-521-57391-2

[26] Burnham K, Anderson D. Multimodel inference. Sociological Methods and Research. 2004;33(2):261304.

[27] Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology. 1982;143:29 36.

[28] Hosmer DW, Lemeshow S. Applied Logistic Regression. 2nd ed. Chapter 5. New York, NY: Wiley-Interscience Publication; 2000.

[29] Bertolini G, D'Amico R, Nardi D, Tinazzi A, Apolone G. One model, several results: the paradox of the Hosmer-Lemeshow goodness-of-fit test for the logistic regression model. J Epidemiol Biostat. 2000;5:2513.

[30] Miller ME, Hui SL, Tierney WM. Validation techniques for logistic regression models. Statistics in medicine. 1991;10:121326.

[31] Kramer AA, Zimmerman JE. Assessing the calibration of mortality benchmarks in critical care: The HosmerLemeshow test revisited. Crit Care Med 2007;35:2052e2056.

[32] Siegert, S. (2017), Simplifying and generalising Murphy's Brier score decomposition. Q.J.R. Meteorol. Soc., 143: 11781183. doi:10.1002/qj.2985

[33] Martin AB, Hartman M, Washington B, Catlin A; National Health Expenditure Accounts Team. National health spending: Faster growth in 2015 as coverage expands and utilization increases. Health Aff (Millwood) 2017; 36:166-176.
[34] Ten Leading Causes of Death by Age Group, United States 2014. CDC, editor. CDC. 8. 2-25-2016.
[35] Diamond GA: What price perfection? Calibration and discrimination of clinical prediction models. J Clin Epidemiol 1992; 45: 859
[36] Polikar, R. (2006). "Ensemble based systems in decision making". IEEE Circuits and Systems Magazine. 6 (3): 2145.
[37] Rokach, L. (2010). "Ensemble-based classifiers". Artificial Intelligence Review. 33 (1-2): 139. doi:10.1007/s10462-009-9124-7.
[38] Bergstra, J. and Bengio, Y. (2012). Random search for hyperparameter optimization. J. Machine Learning Res., 13, 281305.
[39] A.E. Hoerl, R.W. Kennard, Ridge regression: biased estimation for nonorthogonal problems, Technometrics 12 (1970) 5567.
[40] R. Tibshirani, Regression shrinkage and selection via the lasso, J. R. Stat. Soc., Ser. B 58 (1994) 267288.
[41] Frank, I. E., and Friedman, J. H. (1993), "A Statistical View of Some Chemometrics Regression Tools," Technometrics, 35, 109148.
[42] H. Zou, T. Hastie, Regularization and variable selection via the elastic net, J. R. Stat. Soc., Ser. B 67 (2005) 301320.
[43] Kuncheva, L. and Whitaker, C., "Measures of diversity in classifier ensembles," Machine Learning, 51, pp. 181-207, 2003.
[44] Brown, G. and Wyatt, J. and Harris, R. and Yao, X., "Diversity creation methods: a survey and categorisation," Information Fusion, 6(1), pp.5-20, 2005.
[45] L. Breiman, "Random forests," Machine Learning, vol. 45, no. 1, pp. 532, 2001.
[46] James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani. 2013. An Introduction to Statistical Learning: With Applications in R. Ch. 8. New York: Springer.
[47] Natekin, A., Knoll, A., 2013. Gradient boosting machines, a tutorial. Front. Neurorobotics 7.
[48] Bengio, Y. (2013). Practical recommendations for gradient-based training of deep architectures. In K-R. Muller, G. Montavon, and G. B. Orr, editors, Neural Networks: Tricks of the Trade. Springer.
[49] Y. Bengio, I. J. Goodfellow, and A. Courville, "Deep learning," 2015, book in preparation for MIT Press. [Online]. Available: http://www.iro.umontreal.ca/bengioy/dlbook
[50] Nair, Vinod, and Geoffrey E. Hinton. "Rectified linear units improve restricted boltzmann machines." Proceedings of the 27th International Conference on Machine Learning (ICML-10). 2010.
[51] Ian J. Goodfellow, David Warde-Farley, Mehdi Mirza, Aaron Courville, Yoshua Bengio (2013). "Maxout Networks."
[52] N. Srivastava, G.E. Hinton, A. Krizhevsky, I. Sutskever, and R. R. Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. The Journal of Machine Learning Research, 15(1):19291958, 2014.
[53] Zeiler, M. D. (2012). ADADELTA: An adaptive learning rate method. arXiv:1212.5701 [cs.LG].