Published in final edited form as:

Gastroenterology. 2020 January ; 158(1): 160–167. doi:10.1053/j.gastro.2019.09.009.

Validation of a Machine Learning Model That Outperforms Clinical Risk Scoring Systems for Upper Gastrointestinal Bleeding

Dennis L. Shung[1], **Benjamin Au**[1], **Richard Andrew Taylor**[1], **J. Kenneth Tay**[2], **Stig B Laursen**[3], **Adrian J Stanley**[4], **Harry R. Dalton**[5], **Jeffrey Ngu**[6], **Michael Schultz**[7], **Loren Laine**[1,8]

[1]Yale School of Medicine, New Haven, USA [2]Stanford University, Palo Alto, USA [3]Odense University Hospital, Odense, Denmark [4]Glasgow Royal Infirmary, Glasgow, UK [5]Royal Cornwall Hospital, Cornwall, UK [6]Christchurch Hospital, Christchurch, New Zealand [7]Dunedin Hospital, Dunedin, New Zealand [8]VA Connecticut Healthcare System, West Haven, USA

Abstract

Background & Aims: Scoring systems are suboptimal for determining risk in patients with gastrointestinal bleeding (UGIB); these might be improved by a machine learning model. We used machine learning to develop a model to calculate risk of hospital-based intervention or death in patients with UGIB and compared its performance with other scoring systems.

Methods: We analyzed data collected from consecutive unselected patients with UGIB from medical centers in 4 countries (United States, Scotland, England, Denmark; n=1958) from March 2014 through March 2015. We used the data to derive and internally validate a gradient-boosting machine learning model to identify patients who met a composite endpoint of hospital-based intervention (transfusion or hemostatic intervention) or death within 30 days. We compared the performance of the machine learning prediction model with validated pre-endoscopic clinical risk scoring systems (the Glasgow-Blatchford score [GBS], admission-Rockall score, and AIMS65). We externally validated the machine learning model using data from 2 Asia-Pacific sites

Corresponding Authors: Dennis Shung M.D. and Loren Laine, M.D., Yale School of Medicine Section of Digestive Diseases P.O. Box 208019 New Haven, CT 06520-8019, Phone: 203 785-7312, dennis.shung@yale.edu, loren.laine@yale.edu.

Author contributions:

Dennis Shung M.D.: study concept and design, data analysis and interpretation, model development, app development, drafting of manuscript.

Benjamin Au, M.S.: study concept and design, data analysis and model development.

Richard Andrew Taylor, M.D., M.H.S.: study design, data analysis and interpretation, app development, critical revision of manuscript

J. Kenneth Tay, M.S., PhD candidate: study design, feedback on data interpretation, critical revision of manuscript.

Loren Laine M.D.: study concept and design, data analysis and interpretation, drafting of manuscript, critical revision of manuscript.

Stig Laursen, M.D. Adrian Stanley, M.D. Harry R. Dalton, M.D. Jeffrey Ngu, M.D. Michael Schultz, M.D.: collaborators for the multicenter dataset

All authors had access to study data, reviewed and approved the final manuscript.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

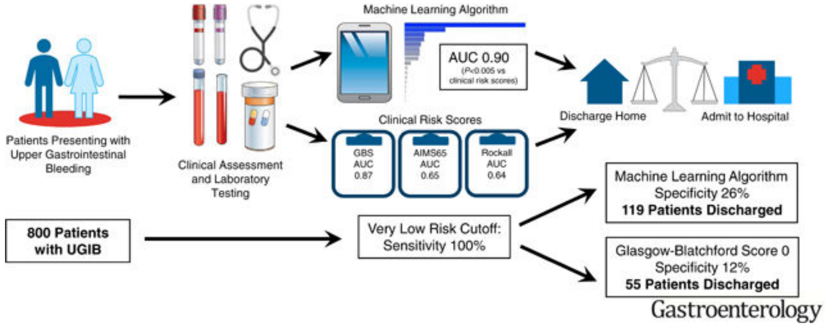
Disclosures: The authors disclose no conflicts.

(Singapore and New Zealand; n=399). Performance was measured by area under receiver operating characteristic curve (AUC) analysis.

Results: The machine learning model identified patients who met the composite endpoint with an AUC of 0.91 in the in the internal validation set; the clinical scoring systems identified patients who met the composite endpoint with AUC values of 0.88 for GBS (P=.001), 0.73 for Rockall score (P<.001), and 0.78 for AIMS65 score (P<.001). In the external validation cohort, the machine learning model identified patients who met the composite endpoint with an AUC of 0.90, the GBS with an AUC of 0.87 (P=.004), the Rockall score with an AUC of 0.66 (P<.001), and the AIMS65 with an AUC of 0.64 (P<.001). At cutoff scores at which the machine learning model and GBS identified patients who met the composite endpoint with 100% sensitivity, the specificity values were 26% with the machine learning model vs 12% with GBS (P<.001).

Conclusions: We developed a machine learning model that identifies patients with UGIB who met a composite endpoint of hospital-based intervention or death within 30 days with a greater AUC and higher levels of specificity, at 100% sensitivity, than validated clinical risk scoring systems. This model could increase identification of low-risk patients who can be safely discharged from the emergency department for outpatient management.

Graphical Abstract


Machine Learning Algorithm
AUC 0.90
(P<0.005 vs clinical risk scores)
Patients Presenting with Upper Gastrointestinal Bleeding
Clinical Assessment and Laboratory Testing
Clinical Risk Scores
GBS AUC 0.87
AIMS65 AUC 0.65
Rockall AUC 0.64
Discharge Home
Admit to Hospital
800 Patients with UGIB
Very Low Risk Cutoff: Sensitivity 100%
Machine Learning Algorithm Specificity 26% 119 Patients Discharged
Glasgow-Blatchford Score 0 Specificity 12% 55 Patients Discharged
Gastroenterology

Keywords

artificial intelligence; prognostic factor; mortality; prediction

BACKGROUND

Acute upper gastrointestinal bleeding (UGIB) is a common gastrointestinal diagnosis requiring hospital admission, with reported annual incidences in the range of 48 to 172 per 100,000 [1–8] and mortality of ~2–10%.[9–16] Multiple guidelines recommend stratification of patients into low and high risk groups, and some recommend using risk assessment scores.[15, 17, 18]

Pre-endoscopic risk scores such as the Glasgow-Blatchford (GBS), admission-Rockall score, and AIMS65 synthesize clinical, hemodynamic, and initial laboratory variables to help guide patient triage. Recently a large prospective multicenter study comparing current clinical risk scores in UGIB suggested that only the GBS provided good results for a composite outcome of transfusion, hemostatic intervention, or death. None of the scores had excellent performance (i.e. area under receiver operator characteristic curve (AUC) ≥0.90) for the composite outcome, and no clinical risk score performed well (AUC ≥0.80) for the outcomes of mortality alone or hemostatic intervention alone.[19]

All current clinical models use standard statistical analyses to identify predictors and most assign fixed weights based on the original dataset used to derive a score. Machine learning (ML) is a discipline that uses computational modeling to learn from data, meaning that performance at executing a specific task improves with experience (i.e., more data). Thus, ML models may improve upon the risk stratification provided by existing clinical risk scores. However, studies of ML models in gastrointestinal bleeding have been limited by small sample sizes, absence of internal and external validation, and/or absence of head-to-head comparisons with existing clinical risk assessment scores.[20]

Electronic health records are increasingly becoming not only repositories of healthcare data, but platforms that can be used to deploy ML models as tools to help guide clinical decision-making. Currently the models deployed on electronic health records include support vector machines, regression models, and decision trees (classification and regression trees (CART), random forest, and gradient boosting decision trees).[21–28]

Currently, the one use of clinical risk assessment tools for patients with UGIB generally agreed upon by guidelines and experts is to identify very low-risk patients who may be safely discharged from emergency departments with outpatient management[15, 18, 29]. A composite endpoint is typically used in assessments to identify very low-risk patients, most commonly a combination of hospital-based intervention (transfusion or hemostatic intervention (endoscopic, surgical, or interventional radiological)) and mortality.[19] The aim of this study was to develop and validate a pre-endoscopic ML model to identify very-low risk patients presenting with UGIB, and compare its performance to existing pre-endoscopic clinical risk scores in predicting the need for hospital-based intervention or mortality in patients with acute UGIB.

METHODS

The data were taken from a study that involved Yale-New Haven Hospital (USA), Glasgow Royal Infirmary (Scotland), Royal Cornwall Hospital Truro (England), Odense University Hospital (Denmark), Singapore General Hospital (Singapore), and Dunedin Hospital (New Zealand).[19]

Subjects

Consecutive, unselected patients presenting with UGIB were collected between March 2014 and March 2015. Inclusion required overt bleeding, defined as hematemesis or melena. Exclusion criteria were patients who were already inpatients when UGIB occurred. Initial assessment of patients was performed in the emergency department or acute assessment unit. Details of the care of these patients has been previously published.[19]

Outcome and Data Collection

The endpoint selected to develop ML models was a composite endpoint of need for hospital-based intervention or death (transfusion of red blood cells; hemostatic intervention with endoscopic, interventional radiology, or surgery; and 30-day all-cause mortality). This is generally considered the most useful outcome in identifying patients at very low risk of poor outcomes.

Data entry was performed by a dedicated research nurse, physician, or medical student at each site, and data collection included patient characteristics, clinical variables, and laboratory results at presentation required to calculate the admission-Rockall, Glasgow-Blatchford, and AIMS65 scores. Data for determination of outcome measures were also collected.

Feature Selection and Data Transformation

Only non-endoscopic variables were included for model development (Table 1). For complete case analysis, continuous clinical variables (age, pulse, systolic blood pressure) and selected laboratory variables (albumin, international normalized ratio (INR), urea and creatinine) were transformed and centered to ensure that all variables were on the same scale. Related categorical variables with increased correlation (defined as correlation >0.55) were decorrelated by consolidating them into a single variable: “any malignancy” and “disseminated malignancy” were consolidated into “malignancy”. Liver disease variables were transformed into an ordinal variable as follows: 0=no liver disease, 1=liver disease, 2=liver cirrhosis, 3=liver failure.

Study Design

The dataset was separated into two geographical regions: U.S.-Europe (United States, Denmark, England, Scotland) and Asia-Pacific (Singapore and New Zealand) (Table 2). The U.S.-Europe dataset was used to train a gradient boosting model (with the XGBoost package in R) and to perform internal validation using tenfold cross validation (in which the dataset is divided into 10 folds and each of the folds is used for internal validation with the remaining 90% used for training to develop the model. Use of cross-validation and hyperparameter tuning for internal validation is considered a robust method for model evaluation prior to external validation on a separate dataset and maximizes the potential performance of the ML model.[30–34] External validation was performed using the dataset of Asia-Pacific patients. Because clinical risk scores were developed on data from U.S. and European patients, use of the geographically distinct population from Asia-Pacific should provide an appropriate assessment for external validation.

The primary use of risk scores in clinical practice that has been recommended by guidelines is identification of very low-risk patients for outpatient management.[15, 18, 29] Achieving a high sensitivity is important for this group: false negatives need to be very rare so that patients who require hospital-based intervention or will die are not sent home. Therefore, to assess the clinical utility of the ML model, we planned to use the clinical risk tool (or tools) at a cutoff score that achieved a sensitivity of 100% (or closest to 100% if none reached 100%) as a comparator and choose the low-risk cutoff for the ML model by setting sensitivity at 100% in the external validation population. We then compared specificities at this cutoff given that the predictive tool with the higher specificity would indicate that tool would identify a greater proportion of patients presenting with UGIB who could be safely discharged.

Prior to choosing and optimizing the gradient boosting (XGBoost) model, we performed rigorous exploratory analyses of logistic regression (with and without regularization), support vector machines, decision trees, and neural networks. For regularized logistic regression, the lasso, ridge and elastic net penalties were studied. A linear support vector machine algorithm, decision-tree models, random forest, gradient boosting (XGBoost), and a multilayered feed-forward perceptron neural network were studied. Separate models were generated with hyperparameter tuning to optimize their performance for each of the outcomes, and all models underwent tenfold cross-validation for internal validation and external validation on Asia-Pacific patients. The preliminary findings suggested that decision tree models (gradient boosting and random forest) and regression models (elastic net, ridge regression) appeared to perform best. (Appendix 1)

Based on our preliminary findings, the performance was roughly the same across the different models for the composite endpoint but there was a trend towards improved performance for decision tree models (random forest and XGBoost) on internal validation and random forest for external validation. We tested both and found that the XGBoost algorithm, which utilizes gradient tree boosting, performed best and was the algorithm of choice for our final model. This is a regression tree-based ML algorithm that combines the output of other decision trees to improve classification. XGBoost is a recently developed gradient tree boosting algorithm that is scalable and allows for faster computation.[25]

Clinical Risk Scores

Only pre-endoscopic clinical risk scores (GBS, admission-Rockall, AIMS65) were compared with ML models, since in practice only pre-endoscopic variables would be available to clinicians when they decide to triage patients to outpatient or inpatient management—and decide on the level of inpatient care and timing of endoscopy. Furthermore, the use of risk stratification to inform management decisions is recommended and most useful well before the time endoscopy is performed, which is commonly many hours after admission.[17]

Statistical Analysis

Two-tailed t-tests and Chi-squared tests were used to compare baseline characteristics between the training set and external validation set. For internal validation, the Wilcoxon signed-rank test, a nonparametric test for matched samples, was used for pairwise comparisons of AUC. For external validation, the AUC was calculated and then compared using a two-tailed non-parametric method.[35] McNemar's matched pairs test was used to compare specificities. For ML models, the caret[36] and glmnet[37] packages were used to create models and tune hyperparameters in R 3.5.1. (R Foundation for Statistical Computing, Vienna, Austria). The ROCR[38] and ggplot2[39] packages were used for visualizing data and generating AUC statistics. We predefined AUC ≥0.80 and <0.90 as good performance, and

AUC ≥0.90 as excellent performance. Our primary analysis was comparison of AUCs for ML models vs. the 3 clinical risk scores; because we performed three comparisons, we adjusted the p-value threshold for significance to p=0.017 with the Bonferroni correction and present 99% confidence intervals.

RESULTS

Patient Data

The study included 2357 patients divided into 1958 for the training and internal validation group and 399 for the external validation group with complete case analysis, all with 30-day follow-up (Table 2).

The original dataset had a total of 3012 patients, with 655 patients (22% of the total dataset) having one or a combination of missing variables. The variables with the greatest missingness (number of patients with either the variable alone or in combination with others) include albumin (N = 302), INR (N = 251), thienopyridine use (N = 105), anticoagulant use (N = 105), or aspirin use (N = 104). All of these patients were excluded from the final dataset, which was a complete case analysis.

Comparisons of the training and external validation groups are shown in Table 2. The mean age for the training group was 62.7 years, and 58% were men. For the external validation group, the mean age was 63.6 years and 67% were men. The mortality rate was 7% in the training group, and 5% in the external validation group. Hemostatic intervention was performed in 19% of patients in the training group, and 21% of the external validation group. The composite endpoint occurred in 43% of the training group, and in 58% of the external validation group.

Performance of the ML Model

Internal Validation—The internal validation group was the tenfold cross-validation of the final ML model of the training set comprised of four sites (Denmark, England, Scotland and U.S.A) with approximately 390 patients in each fold.

For the composite outcome, ML model (AUC=0.91, 0.90–0.93) performed better than GBS (AUC=0.88, 0.86–0.90; p=0.001), admission-Rockall score (AUC=0.69, 0.66–0.71; p<0.001), and AIMS65 (AUC=0.72, 0.69–0.74; p<0.001) (Table 3).

External Validation—For the composite endpoint, the ML model performed better than all clinical risk scores: AUC=0.90, 0.87–0.93 versus GBS AUC=0.87, 0.84–0.91; P=0.004; admission-Rockall AUC=0.65, 0.60–0.71; p<0.001; AIMS65 AUC=0.64, 0.59–0.69; p<0.001 (Table 3).

Identifying Very Low-Risk Patients

High Sensitivity Cutoff for External Validation—Among the clinical risk scores, only GBS=0 achieved a sensitivity at our pre-specified cutoff of 100%; AIMS65=0 and pre-endoscopic Rockall=0 had maximal sensitivities of 74% and 96%, respectively. The ML model performed better than GBS=0 in correctly classifying patients who did not need a hospital-based intervention or die ($p<0.001$): the ML model had a specificity of 26% at sensitivity 100% compared to a specificity of 12% at sensitivity of 100% with GBS=0. The accuracy for the ML model at this high sensitivity cutoff is 68% (0.64–0.73), whereas for GBS=0 accuracy is 63% (0.58–0.68). Because some have suggested a cutoff of GBS≤1[18, 19] given reported sensitivity as high as ~99% with this cutoff, we also performed a post hoc comparison of specificities using a threshold for sensitivity of 99%. Specificities for ML model set at 99% sensitivity and GBS≤1 (which achieved 99% sensitivity threshold) were 35% and 27%, respectively (p=0.02). In order to make the tool available for clinicians, we developed an app (U.S. version: https://dshung.shinyapps.io/UGIB_App_USA/; International version: https://dshung.shinyapps.io/UGIB_App_INTL/) that allows for point of care entry of the variables.

DISCUSSION

In acute UGIB, a gradient-boosting ML model derived from a large international multicenter cohort predicts the composite outcome of transfusion, hemostatic intervention, or death better than the current commonly used clinical risk scores, GBS, admission-Rockall, and AIMS65, on internal and external validation. Thus, this ML model improves upon the ability to identify very-low risk patients who can be safely discharged from the emergency department. Importantly, this ML model increases the number of very-low risk patients who can be identified by more than two-fold as compared to the best performing clinical risk tool currently available.

Risk stratification scores are used in clinical practice by choosing threshold scores to guide care, with the goal of choosing thresholds that maximize sensitivity (minimize false negatives). Guidelines suggest that patients with low GBS scores may be discharged from the emergency department with outpatient management arranged because very few of these patients die or require transfusion or hemostatic intervention. In our study, GBS of 0, which is recommended as a cut-off by U.S. and Asia-Pacific guidelines, had a sensitivity of 100% and a specificity of 12% for the composite outcome of transfusion, hemostatic intervention or death. A meta-analysis reported similar results for GBS=0, with sensitivity of 99% and specificity of 8% for a composite outcome of recurrent UGIB, intervention, or death.[40] At the matched sensitivity of 100% our ML model had specificity of 26%. Sensitivity of 100% means that no patients who will die or require transfusion or hemostatic intervention have a score above the cutoff—and suggests these patients generally can be sent home with outpatient management.[41] The significant increase in specificity from 12% with GBS to 26% with the ML model, with the same 100% sensitivity in both, suggests that, as compared to GBS, the ML model can increase the number of patients who can be safely discharged from the emergency department by more than 2-fold. We provide an app (U.S. version: https://dshung.shinyapps.io/UGIB_App_USA/; International version: https://dshung.shinyapps.io/UGIB_App_INTL/) that allows for point of care entry of the input variables and an immediate feedback if the patient meets the threshold for very low risk.

Previous studies of ML models in UGIB have been limited by sample size, homogeneous patient cohorts, and lack of external validation. For example, the largest study of ML in UGIB utilized a total of 2380 patients and found that a neural network model had improved performance over the full Rockall score (which included endoscopic findings) for 30-day mortality on internal validation only--but had no external validation of the model and no assessment of more clinically relevant pre-endoscopic clinical risk scores or the clinically important composite outcome.[42] For mortality, other ML models have a trend towards better performance than clinical risk scores, although only 3 studies compared ML models to clinical risk scores designed to assess risk in gastrointestinal bleeding and only 1 of them compared to a pre-endoscopic clinical risk score which is the appropriate comparator for risk stratification of gastrointestinal bleeding.[42–44] Two studies examined prediction of mortality on external validation with ML models compared to liver-specific Child-Pugh and MELD scores (designed to predict mortality in all patients with cirrhosis rather than in those presenting with gastrointestinal bleeding), and only one study found improved performance.[45] All comparisons with clinical risk scores were limited by external validation datasets from the same region.

Strengths

First, this study examines clinically relevant outcome measures. The composite outcome helps to triage very low-risk patients who may be able to be managed as outpatients. Second, unlike prior studies our patient cohort is large, prospective and spans multiple centers throughout the world. Third, we initially assessed a variety of different types of ML models to assess their performance in modeling the same dataset to inform our choice of a final ML model for use in clinical practice. Fourth, this study provides direct comparison to multiple pre-endoscopic clinical risk scores developed for prognostication in UGIB. Finally, this study includes both internal and external validation, which allows a more rigorous evaluation of ML model performance. Most prior studies of ML models in UGIB, including the largest ML study published to date, did not have an external validation group.[46]

Limitations

We used the geographical division of U.S.-Europe and Asia-Pacific centers as the criterion to separate training from external validation set and use of other external validation sets might provide different results. Also, selection bias is present due to complete case analysis without integrating missingness: 22% of the dataset was excluded due to one or more missing variables. This may introduce bias into the models, since there may be non-random differences between those who have data elements missing and complete cases. However, all previous studies examining ML and gastrointestinal bleeding have been conducted with complete case analysis, and this approach provides a necessary baseline prior to exploring missingness and the impact of integrating missingness into ML models.

Despite the improvement in performance at the high sensitivity cutoff, the specificity of 26% is less than optimal. The low specificity means that most patients who will not require hospital-based intervention or die are not identified as very low risk and are still admitted. However, the improvement in specificity with the ML model compared to GBS potentially should translate into a substantial reduction in healthcare utilization. For example, based on this increase in specificity, among the 800 patients seen at our emergency department with hematemesis or melena in 1 year, a GBS of 0 would identify 55 very low-risk patients while the ML model would identify 119 such patients.

Finally, the data used was prospectively collected and entered into a registry, which is different from electronic health record data, which is usually more heterogenous with a higher rate of missingness.

Future Directions

In summary, our findings suggest that an ML model trained on predictors derived from existing clinical risk scores provides excellent performance that is better than existing pre-endoscopic clinical risk scores for a composite outcome commonly considered most clinically appropriate in identification of very low-risk patients presenting with UGIB: transfusion, hemostatic intervention, and mortality. For very low-risk patients, there is an improvement in the specificity, meaning that more patients may be safely discharged from the emergency department with outpatient management with the ML model than with use of GBS.

ML models have two key advantages over clinical risk scores: inclusion of a larger number of variables and the potential to improve over time. Electronic health records are becoming platforms for deploying prognostic ML models, which have already been used in clinical care for sepsis, acute kidney injury, and delirium.[47–49] The next steps would be an electronic health record based study that would reliably identify patients presenting with acute UGIB, use structured datafields as predictive variables to develop models based on local patterns of disease and outcomes, and then prospectively validate the models in patients presenting in the emergency department with acute UGIB.

Implementation of the ML model would automatically identify patients with UGIB and generate risk profiles for decision support. For example, the results of the ML model could provide recommendations for outpatient management in patients who are at thresholds accepted as very low risk for mortality, needing transfusion, and requiring hemostatic intervention. Finally, a randomized controlled trial should be conducted to evaluate the effect of ML models as clinical decision support on clinician behavior, healthcare utilization and patient outcomes.

Grant support:

National Institutes of Health T32 DK007017 (D.S.)

Appendix 1:: Exploratory Methodological Study

METHODS

Machine Learning (ML) Models

Figure 1 shows the different categories of ML models. We chose to study the following models in our initial exploratory analyses: logistic regression (with and without regularization), support vector machines, decision trees, and neural networks. For regularized logistic regression, the lasso, ridge and elastic net penalties were studied. A linear support vector machine algorithm, decision-tree models, random forest, gradient boosting (XGBoost), and a multilayered feed-forward perceptron neural network were studied. Separate models were generated for each of the three outcomes

Study Design: Model Development, Internal and External Validation

The dataset was separated into two geographical regions: U.S.-Europe (United States, Denmark, England, Scotland) and Asia-Pacific (Singapore and New Zealand) (Table 2). For internal validation, the models were trained on U.S.-Europe patients with a randomized site-stratified training group (80% of patients randomly sampled from each site) that underwent tenfold cross-validation and hyperparameter tuning. These models were then tested on U.S.-Europe patients with a randomized site-stratified test set (i.e., the remaining 20% of patients). The performance of clinical risk scores recorded from the test set were then used as a comparison to the performance of the ML models. This process was performed for ten iterations to generate a dataset of internal validation test AUCs for ML models and the clinical risk scores. For external validation, ML models were trained on the entire U.S.-Europe patient dataset, underwent tenfold cross-validation and hyperparameter tuning, and then were tested on Asia-Pacific patients.

RESULTS

Internal Validation

The internal validation group was a random stratified sample of 20% of patients from each of the four sites (Denmark, England, Scotland and U.S.A) with 390 patients in each iteration for ten iterations. In Table 3, the characteristics of the first iteration are provided.

For the composite outcome, the random forest and XGBoost models performed best, with AUCs that were not significantly higher than GBS AUC (0.91 versus 0.88, p=0.02) (Table 3). All ML models had higher AUCs than the admission-Rockall score (AUC=0.69, p<0.001) and AIMS65 (AUC=0.72, p<0.001).

For 30-day mortality, the random forest model performed best (AUC=0.85), and all ML models except support vector performed better (p<0.001) than all three clinical risk scores (GBS AUC=0.69, admission-Rockall AUC=0.73, AIMS65 AUC=0.78) (Table 3).

For hemostatic intervention, the elastic net and XGBoost models performed best, although AUCs (0.78) were relatively similar to GBS (AUC=0.76, p=0.03) though better than admission-Rockall (AUC=0.60, p<0.002) and AIMS65 (AUC=0.63, p<0.001).

External Validation

For the composite endpoint, the random forest model performed best (AUC=0.90) on external validation. All ML models (AUC range 0.87–0.90) performed better than the admission-Rockall (AUC=0.65, p<0.001) and AIMS65 (AUC=0.64, p<0.001), but similar to GBS (AUC=0.87).

For 30-day mortality, the random forest and elastic net models performed better than GBS (AUC=0.86–0.87 versus 0.67, p<0.002), although not significantly different from the best performing clinical risk score (AIMS65 AUC=0.81; p=0.10 vs. random forest).

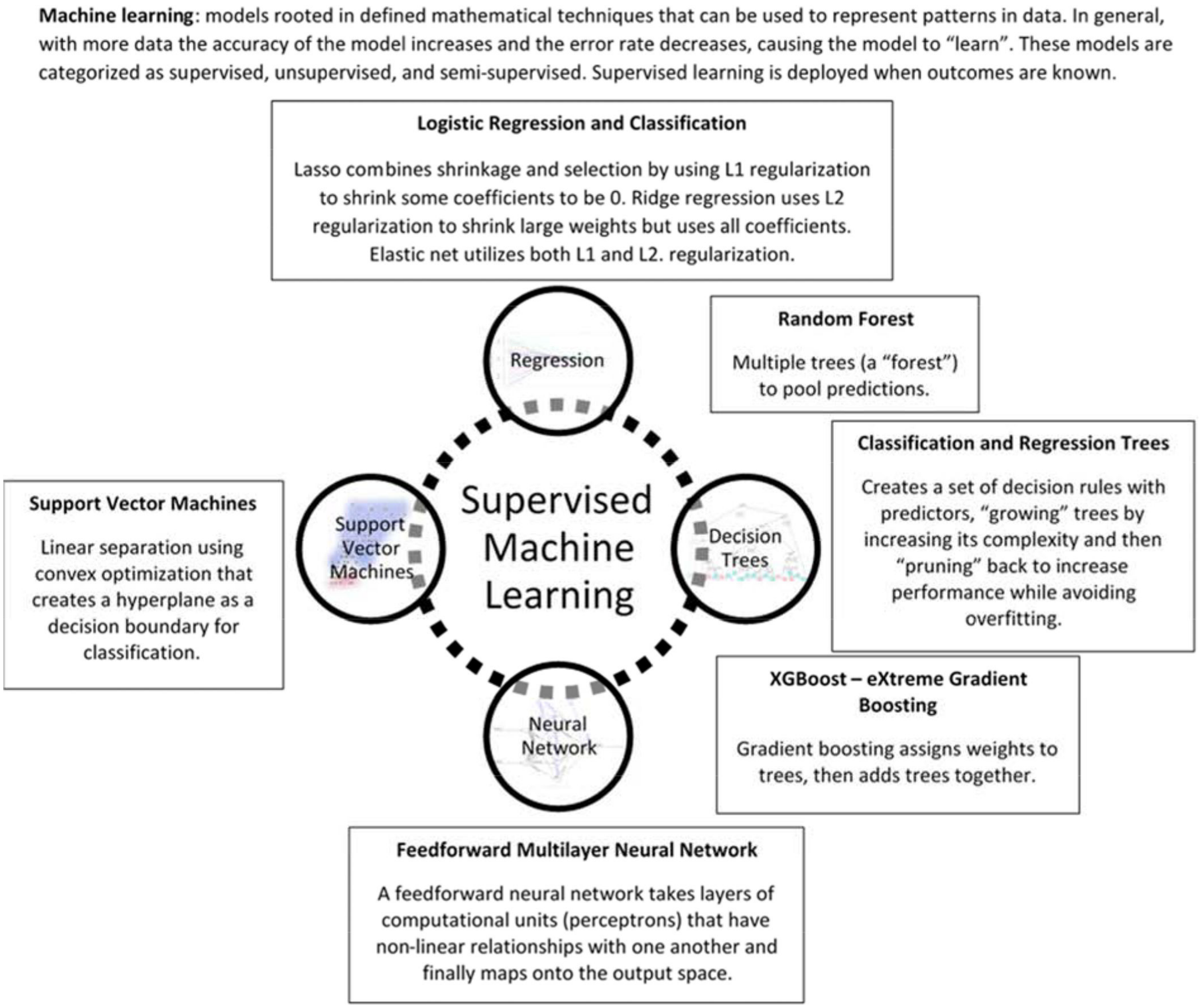
For hemostatic intervention, the support vector model performed best (AUC=0.75) on external validation. The support vector model performed better than admission-Rockall (AUC=0.75 versus 0.61, p<0.002) and AIMS65 (AUC=0.75 versus 0.52, p<0.001) but similar to GBS (AUC=0.69). All other ML models performed better than AIMS65 (AUC range 0.68–0.72 versus 0.52 p<0.001) and had similar results to GBS (AUC range 0.68–0.75 versus 0.69).

Specificities at High Sensitivity Cutoffs on External Validation

For the composite outcome, GBS 0 had a sensitivity of 100% and specificity of 12%. The random forest model was the ML model with the highest AUC, and performed with a similar specificity (15%) at sensitivity 100%. Among the other ML models, the ridge regression model had the highest specificity (32%) at sensitivity 100%.

For mortality, the best performing clinical score was AIMS65 and at the cutoff of 0 had sensitivity 95% and specificity 36%. The random forest model had the highest AUC and performed with a higher specificity 51% at sensitivity 95%. Among the other ML models, the lasso, ridge regression, and neural network models had the highest specificity (63%) at sensitivity 95%.

For hemostatic intervention, the best performing clinical score was GBS, and at a cutoff of ≤3 had sensitivity 97% and specificity 27%. The support vector model had the highest AUC and performed with similar sensitivity and specificity to GBS at 97% and 28% respectively.



Machine learning: models rooted in defined mathematical techniques that can be used to represent patterns in data. In general, with more data the accuracy of the model increases and the error rate decreases, causing the model to "learn". These models are categorized as supervised, unsupervised, and semi-supervised. Supervised learning is deployed when outcomes are known.

Logistic Regression and Classification

Lasso combines shrinkage and selection by using L1 regularization to shrink some coefficients to be 0. Ridge regression uses L2 regularization to shrink large weights but uses all coefficients. Elastic net utilizes both L1 and L2. regularization.

Random Forest

Multiple trees (a "forest") to pool predictions.

Classification and Regression Trees

Creates a set of decision rules with predictors, "growing" trees by increasing its complexity and then "pruning" back to increase performance while avoiding overfitting.

Support Vector Machines

Linear separation using convex optimization that creates a hyperplane as a decision boundary for classification.

XGBoost – eXtreme Gradient Boosting

Gradient boosting assigns weights to trees, then adds trees together.

Feedforward Multilayer Neural Network

A feedforward neural network takes layers of computational units (perceptrons) that have non-linear relationships with one another and finally maps onto the output space.

Figure 1:

Overview of Machine Learning and Models Studied

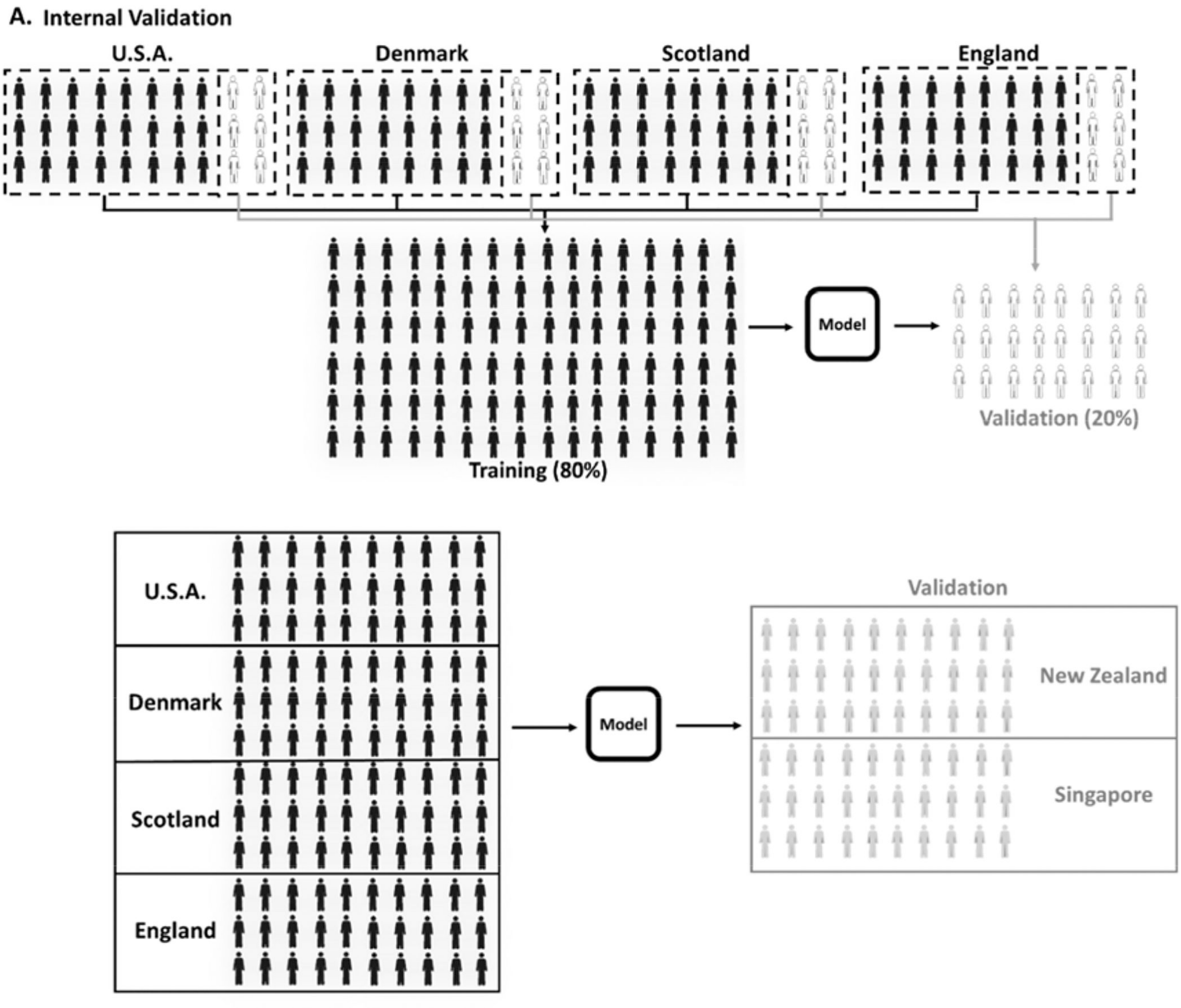


Figure 2.

Distribution of patients used for development and internal or external validation of machine learning models

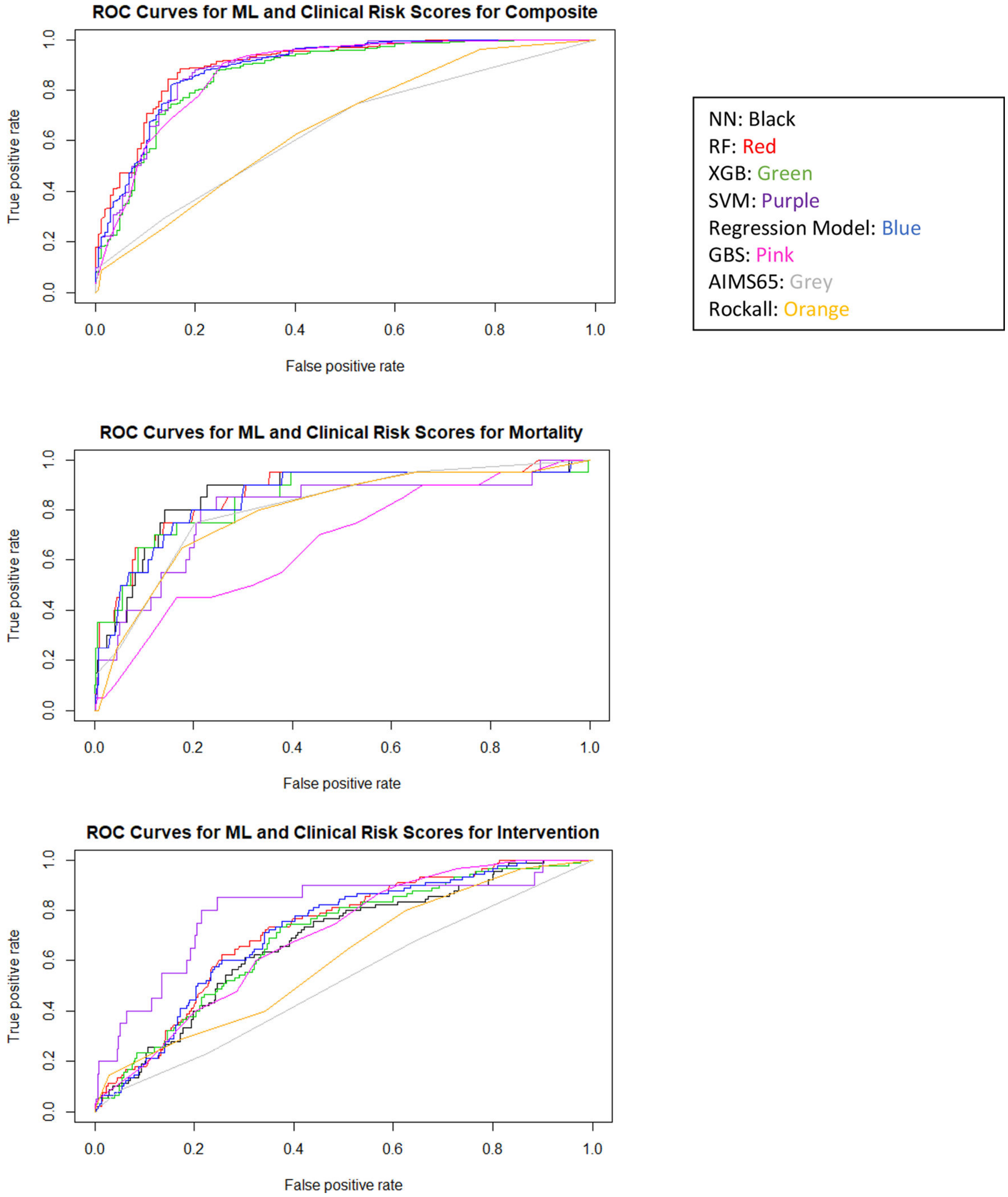


Figure 3:

ROC curves for Composite, Mortality, and Intervention

Table 1.

Performance of Machine Learning Models (Internal and External Validation) and Clinical Risk Assessment Scores

Outcome	Composite Endpoint[a]	Mortality	Hemostatic Intervention
Internal validation dataset (AUC with 99.8% CIs)			
Logistic Regression	0.90[R,A] (0.87 – 0.92)	0.85[G,R,A] (0.79 – 0.90)	0.78[R,A] (0.76–0.81)
Lasso	0.89[R,A] (0.87 – 0.92)	0.85[G,R,A] (0.80 – 0.90)	0.78[R,A] (0.75 – 0.80)
Elastic Net	0.90[R,A] (0.87 – 0.92)	0.85[G,R,A] (0.80 – 0.90)	0.78[R,A] (0.76–0.81)
Ridge Regression	0.89[R,A] (0.87 – 0.92)	0.85[G,R,A] (0.80 – 0.90)	0.78[R,A] (0.74–0.81)
Support Vector Machines	0.90[R,A] (0.87–0.92)	0.77[G] (0.72 – 0.83)	0.75[R,A] (0.72 – 0.78)
Random Forest	0.91[R,A] (0.88 – 0.93)	0.85[G,R,A] (0.80–0.89)	0.77[R,A] (0.75 – 0.79)
XG Boost	0.91[R,A] (0.88 – 0.93)	0.84[G,R,A] (0.80–0.89)	0.78[R,A] (0.76 – 0.80)
Feed-Forward Neural Network	0.88[R,A] (0.85 – 0.92)	0.83[G,R] (0.78 – 0.87)	0.73[R,A] (0.69 – 0.76)
Glasgow-Blatchford score	0.88 (0.85–0.91)	0.69 (0.64 – 0.74)	0.76 (0.72 – 0.79)
Admission-Rockall score	0.69 (0.65 – 0.72)	0.73 (0.68 – 0.77)	0.60 (0.57 – 0.64)
AIMS65	0.72 (0.69 – 0.75)	0.78 (0.72 – 0.83)	0.63 (0.59 – 0.68)
External validation dataset (AUC)			
Logistic Regression	0.88[R,A] (0.83 – 0.94)	0.85 (0.70 – 1.0)	0.72[R] (0.63 – 0.80)
Lasso	0.89[R,A] (0.83 – 0.94)	0.85 (0.70 – 1.0)	0.71[R] (0.63 – 0.80)
Elastic Net	0.89[R,A] (0.84 – 0.94)	0.86[G] (0.71 – 1.0)	0.71[R] (0.62 – 0.79)
Ridge Regression	0.88[R,A] (0.83 – 0.94)	0.85 (0.70 – 1.0)	0.70[R] (0.62 – 0.80)
Support Vector Machine	0.89[R,A] (0.83 – 0.94)	0.80 (0.62 – 0.98)	0.75[R,A] (0.67 – 0.84)
Random Forest	0.90[R,A] (0.85 – 0.95)	0.87[G] (0.69 – 1.0)	0.72[R] (0.63–0.81)
XGBoost	0.88[R,A] (0.83 – 0.94)	0.85 (0.68 – 1.0)	0.70[R] (0.61 – 0.80)
Feed-forward Neural Network	0.87[R,A] (0.81 – 0.93)	0.86 (0.71 – 1.0)	0.68[R] (0.58 – 0.77)
Glasgow-Blatchford score	0.87 (0.82 – 0.93)	0.67 (0.48 – 0.86)	0.69 (0.60 – 0.78)
Admission-Rockall score	0.66 (0.56 – 0.72)	0.79 (0.63 – 0.96)	0.61 (0.51 – 0.71)
AIMS65	0.64 (0.57 – 0.74)	0.81 (0.66 – 0.95)	0.52 (0.42 – 0.62)

[a] Blood transfusion, hemostatic intervention, or 30-day mortality

[G] p < 0.002 compared to GBS

[R] p < 0.002 compared to admission-Rockall

[A] p < 0.001 compared to AIMS65

Table 2:

Performance Characteristics for Clinical Risk Scores and ML Models Matched to Highest Clinical Risk Score Sensitivity on External Validation

Composite Outcome	Model	Cutoff	Sensitivity	Specificity	PPV	NPV	Prevalence
Clinical Risk Scores	GBS	0	1.00	0.12	0.61	1.00	0.58
	AIMS65	0	0.74	0.48	0.67	0.57	
	Rockall	0	0.96	0.23	0.63	0.81	

Composite Outcome	Model	Cutoff	Sensitivity	Specificity	PPV	NPV	Prevalence
Regression Models	Logistic Regression		1.00	0.29	0.66	1.00	
	Lasso		1.00	0.30	0.66	1.00	
	Elastic Net		1.00	0.29	0.66	1.00	
	Ridge Regression		1.00	0.32	0.67	1.00	
Linear Support Vector Machines	Support Vector Machines		1.00	0.26	0.65	1.00	
Decision Tree Models	Random Forest		1.00	0.15	0.62	1.00	
	XG Boost		1.00	0.15	0.62	1.00	
Neural Network	Neural Network		1.00	0.16	0.62	1.00	

Mortality	Model	Cutoff	Sensitivity	Specificity	PPV	NPV	Prevalence
Clinical Risk Scores	AIMS65	0	0.95	0.36	0.07	0.99	0.05
	GBS	1	0.95	0.12	0.05	0.98	
	Rockall	0	0.95	0.12	0.05	0.98	
Regression Models	Logistic Regression		0.95	0.62	0.12	1.00	
	Lasso		0.95	0.63	0.12	1.00	
	Elastic Net		0.95	0.61	0.11	1.00	
	Ridge Regression		0.95	0.63	0.12	1.00	
Linear Support Vector Machines	Support Vector Machines		0.95	0.12	0.05	0.98	
Decision Tree Models	Random Forest		0.95	0.51	0.09	1.00	
	XG Boost Neural		0.95	0.51	0.09	1.00	
Neural Network	Network		0.95	0.63	0.12	1.00	

Hemostatic Intervention	Model	Cutoff	Sensitivity	Specificity	PPV	NPV	Prevalence
Clinical Risk Scores	GBS	3	0.97	0.27	0.27	0.97	0.21
	AIMS65	0	0.68	0.36	0.22	0.80	
	Rockall	0	0.97	0.14	0.24	0.94	
Regression Models	Logistic Regression		0.97	0.20	0.25	0.94	
	Lasso		0.97	0.24	0.26	0.95	
	Elastic Net		0.97	0.27	0.27	0.96	
	Ridge Regression		0.97	0.22	0.26	0.95	
Linear Support Vector Machines	Support Vector Machines		0.97	0.28	0.29	0.97	
Decision Tree Models	Random Forest		0.97	0.22	0.26	0.95	
	XG Boost		0.97	0.21	0.25	0.95	
Neural Network	Neural Network		0.97	0.18	0.24	0.94	

Abbreviations:

UGIB	(Upper Gastrointestinal Bleeding)

ML	(Machine Learning)
GBS	(Glasgow Blatchford Score)
CART	(Classification and Regression Trees)
INR	(International Normalized Ratio)
XGBoost	(Extreme Gradient Boosting)
MELD	(Model for End-Stage Liver Disease)

References

1. Longstreth GF. Epidemiology of hospitalization for acute upper gastrointestinal hemorrhage: a population-based study. Am J Gastroenterol 1995;90:206–10. [PubMed: 7847286]
2. Yavorski RT, Wong RK, Maydonovitch C, et al. Analysis of 3,294 cases of upper gastrointestinal bleeding in military medical facilities. Am J Gastroenterol 1995;90:568–73. [PubMed: 7717312]
3. Blatchford O, Davidson LA, Murray WR, et al. Acute upper gastrointestinal haemorrhage in west of Scotland: case ascertainment study. Bmj 1997;315:510–4. [PubMed: 9329304]
4. Rockall TA, Logan RF, Devlin HB, et al. Incidence of and mortality from acute upper gastrointestinal haemorrhage in the United Kingdom. Steering Committee and members of the National Audit of Acute Upper Gastrointestinal Haemorrhage. Bmj 1995;311:222–6. [PubMed: 7627034]
5. Vreeburg EM, Snel P, de Bruijne JW, et al. Acute upper gastrointestinal bleeding in the Amsterdam area: incidence, diagnosis, and clinical outcome. Am J Gastroenterol 1997;92:236–43. [PubMed: 9040198]
6. Czernichow P, Hochain P, Nousbaum JB, et al. Epidemiology and course of acute upper gastro-intestinal haemorrhage in four French geographical areas. Eur J Gastroenterol Hepatol 2000;12:175–81. [PubMed: 10741931]
7. Paspatis GA, Matrella E, Kapsoritakis A, et al. An epidemiological study of acute upper gastrointestinal bleeding in Crete, Greece. Eur J Gastroenterol Hepatol 2000;12:1215–20. [PubMed: 11111778]
8. van Leerdam ME. Epidemiology of acute upper gastrointestinal bleeding. Best Pract Res Clin Gastroenterol 2008;22:209–24. [PubMed: 18346679]
9. Hearnshaw SA, Logan RF, Lowe D, et al. Acute upper gastrointestinal bleeding in the UK: patient characteristics, diagnoses and outcomes in the 2007 UK audit. Gut 2011;60:1327–35. [PubMed: 21490373]
10. Abougergi MS, Travis AC, Saltzman JR. The in-hospital mortality rate for upper GI hemorrhage has decreased over 2 decades in the United States: a nationwide analysis. Gastrointestinal Endoscopy 2014;81:882–8.e1. [PubMed: 25484324]
11. Nahon S, Hagege H, Latrive JP, et al. Epidemiological and prognostic factors involved in upper gastrointestinal bleeding: results of a French prospective multicenter study. Endoscopy 2012;44:998–1008. [PubMed: 23108771]
12. Lanas A, Garcia-Rodriguez LA, Polo-Tomas M, et al. Time trends and impact of upper and lower gastrointestinal bleeding and perforation in clinical practice. American Journal of Gastroenterology 2009;104:1633–41. [PubMed: 19574968]
13. Wuerth BA, Rockey DC. Changing Epidemiology of Upper Gastrointestinal Hemorrhage in the Last Decade: A Nationwide Analysis. Digestive Diseases and Sciences 2017;63:1286–1293. [PubMed: 29282637]
14. Peery AF, Crockett SD, Barritt AS, et al. Burden of Gastrointestinal, Liver, and Pancreatic Diseases in the United States. Gastroenterology 2015;149:1731–1741 e3. [PubMed: 26327134]
15. Laine L, Jensen DM. Management of patients with ulcer bleeding. Am J Gastroenterol 2012;107:345–60; quiz 361. [PubMed: 22310222]

16. Saltzman JR, Tabak YP, Hyett BH, et al. A simple risk score accurately predicts in-hospital mortality, length of stay, and cost in acute upper GI bleeding. Gastrointest Endosc 2011;74:1215–24. [PubMed: 21907980]
17. Barkun AN, Bardou M, Kuipers EJ, et al. International consensus recommendations on the management of patients with nonvariceal upper gastrointestinal bleeding. Ann Intern Med 2010;152:101–13. [PubMed: 20083829]
18. Gralnek IM, Dumonceau JM, Kuipers EJ, et al. Diagnosis and management of nonvariceal upper gastrointestinal hemorrhage: European Society of Gastrointestinal Endoscopy (ESGE) Guideline. Endoscopy 2015;47:a1–46. [PubMed: 26417980]
19. Stanley AJ, Laine L, Dalton HR, et al. Comparison of risk scoring systems for patients presenting with upper gastrointestinal bleeding: international multicentre prospective study. BMJ 2017;356:i6432. [PubMed: 28053181]
20. Shung D, Simonov M, Gentry M, et al. Machine Learning to Predict Outcomes in Patients with Acute Gastrointestinal Bleeding: A Systematic Review. Dig Dis Sci 2019.
21. Tibshirani R. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society. Series B (Methodological) 1996:267–288.
22. Hoerl AE, Kennard RW. Ridge regression: Biased estimation for nonorthogonal problems. Technometrics 1970;12:55–67.
23. Zou H, Hastie T. Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 2005;67:301–320.
24. Suykens JA, Vandewalle J. Least squares support vector machine classifiers. Neural processing letters 1999;9:293–300.
25. Chen T, Guestrin C. Xgboost: A scalable tree boosting system, In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, ACM, 2016.
26. Breiman L. Random forests. Machine learning 2001;45:5–32.
27. Hecht-Nielsen R. Theory of the backpropagation neural network. Neural networks for perception: Elsevier, 1992:65–93.
28. Hornik K, Stinchcombe M, White H. Multilayer feedforward networks are universal approximators. Neural networks 1989;2:359–366.
29. Sung JJ, Chiu PW, Chan FKL, et al. Asia-Pacific working group consensus on non-variceal upper gastrointestinal bleeding: an update 2018. Gut 2018;67:1757–1768. [PubMed: 29691276]
30. Steyerberg EW. Validation in prediction research: the waste by data splitting. J Clin Epidemiol 2018;103:131–133. [PubMed: 30063954]
31. Motwani M, Dey D, Berman DS, et al. Machine learning for prediction of all-cause mortality in patients with suspected coronary artery disease: a 5-year multicentre prospective registry analysis. Eur Heart J 2017;38:500–507. [PubMed: 27252451]
32. Samad MD, Ulloa A, Wehner GJ, et al. Predicting Survival From Large Echocardiography and Electronic Health Record Datasets: Optimization With Machine Learning. JACC Cardiovasc Imaging 2019;12:681–689. [PubMed: 29909114]
33. Kennedy EH, Wiitala WL, Hayward RA, et al. Improved cardiovascular risk prediction using nonparametric regression and electronic health record data. Med Care 2013;51:251–8. [PubMed: 23269109]
34. Rotondano G, Cipolletta L, Grossi E, et al. Artificial neural networks accurately predict mortality in patients with nonvariceal upper GI bleeding. Gastrointestinal Endoscopy 2011;73:218–226.e2. [PubMed: 21295635]
35. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics 1988;44:837–45. [PubMed: 3203132]
36. Kuhn M. Building Predictive Models in R Using the caret Package. 2008 2008;28:26.
37. Friedman JH, Hastie T, Tibshirani R. Regularization Paths for Generalized Linear Models via Coordinate Descent. 2010 2010;33:22.
38. Lengauer TSaOSaNBaT. ROCR: visualizing classifier performance in R. Bioinformatics 2005;21:7881.

39. Wickham H. ggplot2: Elegant Graphics for Data Analysis. New York: Springer-Verlag 2016.

40. Ramaekers R, Mukarram M, Smith CA, et al. The Predictive Value of Preendoscopic Risk Scores to Predict Adverse Outcomes in Emergency Department Patients With Upper Gastrointestinal Bleeding: A Systematic Review. Acad Emerg Med 2016;23:1218–1227. [PubMed: 27640399]

41. Stanley AJ, Ashley D, Dalton HR, et al. Outpatient management of patients with low-risk upper-gastrointestinal haemorrhage: multicentre validation and prospective evaluation. Lancet 2009;373:42–7. [PubMed: 19091393]

42. Rotondano G, Cipolletta L, Grossi E, et al. Artificial neural networks accurately predict mortality in patients with nonvariceal upper GI bleeding. Gastrointestinal Endoscopy 2011;73:218–26, 226.e1. [PubMed: 21295635]

43. Lyles T, Elliott A, Rockey DC. A risk scoring system to predict in-Hospital mortality in patients with cirrhosis presenting with upper gastrointestinal bleeding. Journal of Clinical Gastroenterology 2014;48:712–720. [PubMed: 24172184]

44. Lee HH, Park JM, Han S, et al. A simplified prognostic model to predict mortality in patients with acute variceal bleeding. Dig Liver Dis 2018;50:247–253. [PubMed: 29208551]

45. D’Amico G, De Franchis R. Upper digestive bleeding in cirrhosis. Post-therapeutic outcome and prognostic indicators. Hepatology 2003;38:599–612. [PubMed: 12939586]

46. Shung D, Simonov M, Au B, et al. Machine Learning to Predict Outcomes in Patients with Acute Gastrointestinal Bleeding: Systematic Review and Meta-Analysis. Gastroenterology 2018;154.

47. Wilson FP, Shashaty M, Testani J, et al. Automated, electronic alerts for acute kidney injury: a single-blind, parallel-group, randomised controlled trial. Lancet 2015;385:1966–74. [PubMed: 25726515]

48. Shimabukuro DW, Christopher WB, Mitchell DF, et al. Effect of a machine learning-based severe sepsis prediction algorithm on patient survival and hospital length of stay: a randomised clinical trial. BMJ Open Respiratory Research 2017;4.

49. Wong A, Young AT, Liang AS, et al. Development and Validation of an Electronic Health Record–Based Machine Learning Model to Estimate Delirium Risk in Newly Hospitalized Patients Without Known Cognitive Impairment. JAMA Network Open 2018;1:e181018–e181018. [PubMed: 30646095]

What you need to know:

BACKGROUND AND CONTEXT:

We used machine learning to develop a model to calculate risk hospital-based intervention or death in patients with upper gastrointestinal bleeding (UGIB) and compared its accuracy with current scoring systems.

NEW FINDINGS:

We developed a machine learning model that identifies patients with UGIB who met a composite endpoint of hospital-based intervention or death within 30 days with a greater AUC and higher levels of specificity (at 100% sensitivity) than validated clinical risk scoring systems. This model could increase identification of low-risk patients who can be safely discharged from the emergency department for outpatient management.

LIMITATIONS:

This model requires validation in other populations.

IMPACT:

This model could increase identification of low-risk patients who can be safely discharged from the emergency department for outpatient management.

Lay Summary:

We used machine learning to analyze data from patients with upper gastrointestinal bleeding and identify those at risk for hospital-based intervention or death within 30 days.

Table 1:

Clinical Variables Used to Build Machine Learning Models

Demographic (2)	
	Age
	Sex
Comorbidity (6)	
	American Society of Anesthesiologists (ASA) Score
	Ischemic Heart Disease
	Cardiac Failure
	Renal Failure
	Liver Disease
	Any Malignancy
Medications (4)	
	Aspirin
	Thienopyridines
	Anticoagulation
	Non-steroidal anti-inflammatory drugs
Clinical Features at Presentation (7)	
	Pulse
	Systolic Blood Pressure
	Syncope
	Altered Mental Status
	Hematemesis
	Melena
	Hematochezia
Initial laboratory values (5)	
	Hemoglobin
	Urea
	Creatinine
	Albumin
	International Normalized Ratio (INR)

Table 2:

Comparison of the Training, Internal Validation, and External Validation Groups

		Training Set	External Validation Set		
Variables				Difference* (95% Confidence Interval)	P-value
		N = 1958	N = 399		
Demographic					
	Age	62.7 (20.1)	63.6 (17.7)	−0.87 (−2.8 to 1.1)	0.38
	Men	1141 (58%)	266 (67%)	−0.08 (−0.14 to(−0.03)	0.002
Comorbidity					
	ASA Score				
	1	235 (12%)	103(26%)	−0.14 (−0.18 to −0.09)	<0.001
	2	587 (30%)	142(36%)	−0.06 (−0.11 to −0.003)	0.03
	3	926 (47%)	149(37%)	0.10 (0.04 to 0.15)	<0.001
	4	194(9.9%)	5(1%)	0.08 (0.07 to 0.10)	<0.001
	5	16(0.8%)	0 (0%)	0.01 (0.00 to 0.01)	0.14
	Ischemic Heart Disease	384 (20%)	78 (20%)	0.00 (-0.04 to 0.04)	1
	Cardiac Failure	195(10%)	16(4%)	0.06 (0.03 to 0.08)	<0.001
	Renal Failure	166(8%)	51 (13%)	−0.05 (−0.08 to −0.01)	0.01
	Liver				
	None	1581 (81%)	339 (85%)	−0.04 (−0.08 to −0.001)	0.06
	Liver Disease	94 (5%)	13(3%)	0.01 (-0.006 to 0.04)	0.22
	Liver Cirrhosis	113(6%)	44(11%)	−0.06 (−0.09 to −0.02)	<0.001
	Liver Failure	170(9%)	3(1%)	0.07 (0.06 to 0.09)	<0.001
	Any Malignancy	301 (15%)	50(13%)	0.03 (−0.01 to 0.07)	0.168
Medications					
	Aspirin	516(26%)	79 (20%)	0.07 (0.02 to 0.11)	0.007
	Adenosine diphosphate (ADP) Inhibitors	148(7%)	28 (7%)	0.005 (−0.02 to 0.03)	0.78
	Anticoagulation	257(13%)	38(10%)	0.04 (0.005 to 0.07)	0.04
	Non-steroidal anti-inflammatory drugs (NSAIDs)	275(14%)	30 (8%)	0.07 (0.03 to 0.09)	<0.001
Clinical Features at Presentation					
	Pulse	91.5(20.3)	91.9(19.6)	−0.4 (−2.5 to 1.7)	0.72
	Systolic Blood Pressure	127.2 (24.1)	121.9(25.5)	6.8 (2.5 to 8.0)	<0.001
	Syncope	190(10%)	34 (9%)	0.01 (−0.02 to 0.04)	0.52
	Altered Mental Status	213(11%)	25 (6%)	0.04 (0.02 to 0.07)	0.007
	Hematemesis	839 (43%)	135(34%)	0.11 (0.04 to 0.14)	0.001
	Melena	1001 (51%)	277 (69%)	−0.18 (−0.23 to −0.13)	<0.001
	Hematochezia	113(6%)	23 (6%)	0.0 (−0.02 to 0.02)	1

		Training Set	External Validation Set		
Initial Laboratory Values					
	Hemoglobin	112.8 (32.2)	102.7(30.9)	10.1 (6.8 to 13.5)	<0.001
	Urea	11.0(9.2)	12.0(10.1)	−1.0 (−2.1 to 0.04)	0.06
	Creatinine	102.5 (93.1)	116.2(121.4)	−13.8 (−26.4 to −1.1)	0.03
	Albumin	35.8 (7.2)	35.4 (6.5)	0.04	0.19
	INR	1.4(1.30)	1.40(1.23)	0.0 (−0.13 to 0.13)	1.0
Transfusion Requirement (number of RBC units) Outcomes		1.31 (2.61)	1.54 (2.49)	−0.2 (−0.5 to 0.04)	0.09
	Mortality (30-day)	154(8%)	20 (5%)	0.02 (0.002 to 0.05)	0.059
	Hemostatic Intervention (endoscopic, surgical, or interventional radiology)	396 (20%)	90 (23%)	−0.02 (−0.07 to 0.02)	0.33
	Composite outcome (30-day mortality, hemostatic intervention, or transfusion	875 (45%)	234 (59%)	−0.14 (−0.19 to −0.08)	<0.001
Clinical Risk Scores	Glasgow-Blatchford	6.53 (4.56)	7.81 (4.50)	−1.3 (−1.8 to −0.79)	<0.001
	Admission-Rockall	2.82(1.74)	2.7(1.8)	0.1 (−0.07 to 0.31)	0.22
	AIMS65	1.0 (0.94)	0.96 (0.92)	0.04 (−0.05 to 0.14)	0.37

[*] Includes mean difference and difference in proportions

Table 3:

Performance of XGBoost Machine Learning Model and Clinical Risk Assessment Scores

Composite Endpoint[a]	Internal Validation AUC with 99% Confidence Interval	p-value	External Validation AUC with 99% Confidence Interval	p-value
XGBoost Machine Learning Model	0.91 (0.90–0.93)		0.90 (0.87–0.93)	
Glasgow-Blatchford score	0.88 (0.86–0.90)	0.001	0.87 (0.84–0.91)	0.004
Admission-Rockall score	0.69 (0.66–0.71)	<0.001	0.65 (0.60–0.71)	<0.001
AIMS65	0.72 (0.69–0.74)	<0.001	0.64 (0.59–0.69)	<0.001