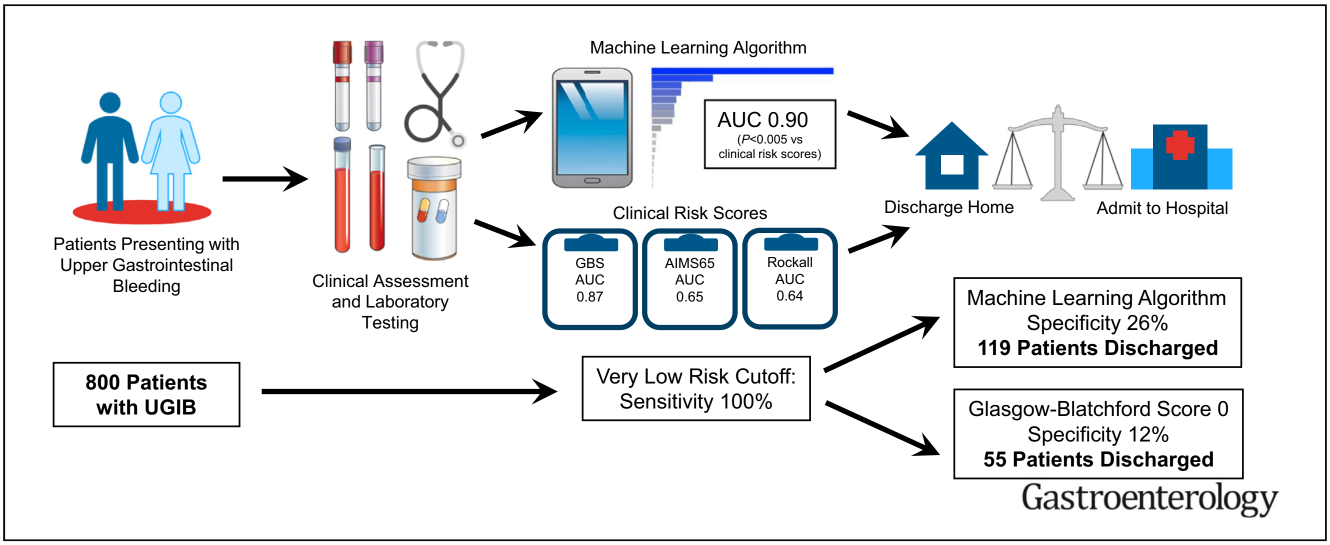
Validation of a Machine Learning Model That Outperforms Clinical Risk Scoring Systems for Upper Gastrointestinal Bleeding

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] and Loren Laine[1,8]

[1]Yale School of Medicine, New Haven, Connecticut; [2]Stanford University, Palo Alto, California; [3]Odense University Hospital, Odense, Denmark; [4]Glasgow Royal Infirmary, Glasgow, United Kingdom; [5]Royal Cornwall Hospital, Cornwall, United Kingdom; [6]Christchurch Hospital, Christchurch, New Zealand; [7]Dunedin Hospital, Dunedin, New Zealand; and [8]Veterans Affairs Connecticut Healthcare System, West Haven, Connecticut


Patients Presenting with Upper Gastrointestinal Bleeding
Clinical Assessment and Laboratory Testing
Machine Learning Algorithm
AUC 0.90
(P<0.005 vs clinical risk scores)
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

BACKGROUND & AIMS: Scoring systems are suboptimal for determining risk in patients with upper gastrointestinal bleeding (UGIB); these might be improved by a machine learning model. We used machine learning to develop a model to calculate the 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 (the United States, Scotland, England, and 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 (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 internal validation set; the clinical scoring systems identified patients who met the composite endpoint with AUC values of 0.88 for the 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 versus 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.

Keywords: Artificial Intelligence; Prognostic Factor; Mortality; Prediction.

Abbreviations used in this paper: **AUC, area under receiver operator characteristic curve; CI, confidence interval; GBS, Glasgow Blatchford Score; ML, machine learning; UGIB, upper gastrointestinal bleeding.**

Most current article

0016-5085/$36.00
https://doi.org/10.1053/j.gastro.2019.09.009

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.

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 approximately 2% to 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 score (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 (ie, area under receiver operator characteristic curve [AUC] $\geq$ 0.90) for the composite outcome, and no clinical risk score performed well (AUC $\geq$ 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 data set 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 health care 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, random forest, and gradient-boosting decision trees).[21–28]

Currently, the one use of clinical risk assessment tools for patients with UGIB generally agreed on 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 radiologic]) 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 (the United States), Glasgow Royal Infirmary (Scotland), Royal Cornwall Hospital Truro (England), Odense University Hospital (Denmark), Singapore General Hospital (Singapore), and Dunedin Hospital (New Zealand).[19]

Participants

Data from 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. The 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 endoscopy, 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 score, GBS, and AIMS65 score. Data for determination of outcome measures were also collected.

Table 1. Clinical Variables Used to Build Machine Learning Models

Category (number of variables)	Variables
Demographic (2)	Age Sex
Comorbidity (6)	American Society of Anesthesiologists score Ischemic heart disease Cardiac failure Renal failure Liver disease Any malignancy
Medications (4)	Aspirin Thienopyridines Anticoagulation Nonsteroidal 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

Feature Selection and Data Transformation

Only nonendoscopic 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, 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 data set was separated into 2 geographic regions: the United States–Europe (United States, Denmark, England, Scotland) and Asia-Pacific (Singapore and New Zealand) (Table 2). The United States–Europe data set was used to train a gradient-boosting model (with the XGBoost package in R) and to perform internal validation using 10-fold cross validation (in which the data set 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 before external validation on a separate data set and maximizes the potential performance of the ML model.[30–34] External validation was performed by using the data set of Asia-Pacific patients. Because clinical risk scores were developed on data from US and European patients, use of the geographically distinct population from the Asia-Pacific group 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.

Before 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 10-fold 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 and ridge regression) appeared to perform best (see Supplementary Materials).

Based on our preliminary findings, the performance was roughly the same across the different models for the composite endpoint, but there was a trend toward 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 uses 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, and AIMS65) were compared with ML models because 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 is most useful well before the time when endoscopy is performed, which is commonly many hours after admission.[17]

Table 2. Comparison of the Training, Internal Validation, and External Validation Groups

Variables	Training set (n = 1958)	External validation set (n = 399)	Difference[a] (95% CI)	*P* value
Demographic				
Age, mean (SD), *y*	62.7 (20.1)	63.6 (17.7)	–0.87 (–2.8 to 1.1)	.38
Men, n (%)	1141 (58)	266 (67)	–0.08 (–0.14 to –0.03)	.002
Comorbidity, n (%)				
ASA score				
1	235 (12)	103 (26)	–0.14 (–0.18 to –0.09)	<.001
2	587 (30)	142 (36)	–0.06 (–0.11 to –0.003)	.03
3	926 (47)	149 (37)	0.10 (0.04 to 0.15)	<.001
4	194 (9.9)	5 (1)	0.08 (0.07 to 0.10)	<.001
5	16 (0.8)	0 (0)	0.01 (0.00 to 0.01)	.14
Ischemic heart disease	384 (20)	78 (20)	0.00 (–0.04 to 0.04)	>.99
Cardiac failure	195 (10)	16 (4)	0.06 (0.03 to 0.08)	<.001
Renal failure	166 (8)	51 (13)	–0.05 (–0.08 to –0.01)	.01
Liver				
None	1581 (81)	339 (85)	–0.04 (–0.08 to –0.001)	.06
Liver disease	94 (5)	13 (3)	0.01 (–0.006 to 0.04)	.22
Liver cirrhosis	113 (6)	44 (11)	–0.06 (–0.09 to –0.02)	<.001
Liver failure	170 (9)	3 (1)	0.07 (0.06 to 0.09)	<.001
Any malignancy	301 (15)	50 (13)	0.03 (–0.01 to 0.07)	.168
Medications, n (%)				
Aspirin	516 (26)	79 (20)	0.07 (0.02 to 0.11)	.007
Adenosine diphosphate inhibitors	148 (7)	28 (7)	0.005 (–0.02 to 0.03)	.78
Anticoagulation	257 (13)	38 (10)	0.04 (0.005 to 0.07)	.04
Nonsteroidal anti-inflammatory drugs	275 (14)	30 (8)	0.07 (0.03 to 0.09)	<.001
Clinical features at presentation				
Pulse, mean (SD), *beats/min*	91.5 (20.3)	91.9 (19.6)	–0.4 (–2.5 to 1.7)	.72
Systolic blood pressure, mean (SD), *mm Hg*	127.2 (24.1)	121.9 (25.5)	6.8 (2.5 to 8.0)	<.001
Syncope, n (%)	190 (10%)	34 (9%)	0.01 (–0.02 to 0.04)	.52
Altered mental status, n (%)	213 (11%)	25 (6%)	0.04 (0.02 to 0.07)	.007
Hematemesis, n (%)	839 (43%)	135 (34%)	0.11 (0.04 to 0.14)	.001
Melena, n (%)	1001 (51%)	277 (69%)	–0.18 (–0.23 to –0.13)	<.001
Hematochezia, n (%)	113 (6%)	23 (6%)	0.0 (–0.02 to 0.02)	>.99
Initial laboratory values, mean (SD)				
Hemoglobin, *g/dL*	112.8 (32.2)	102.7 (30.9)	10.1 (6.8 to 13.5)	<.001
Urea, *mmol/L*	11.0 (9.2)	12.0 (10.1)	–1.0 (–2.1 to 0.04)	.06
Creatinine, *μmol/L*	102.5 (93.1)	116.2 (121.4)	–13.8 (–26.4 to –1.1)	.03
Albumin, *g/L*	35.8 (7.2)	35.4 (6.5)	0.04	.19
INR	1.4 (1.30)	1.40 (1.23)	0.0 (–0.13 to 0.13)	>.99
Transfusion requirement, mean (SD), *RBC units*	1.31 (2.61)	1.54 (2.49)	–0.2 (–0.5 to 0.04)	.09
Outcomes, n (%)				
Mortality (30-day)	154 (8)	20 (5)	0.02 (0.002 to 0.05)	.059
Hemostatic intervention (endoscopic, surgical, or interventional radiology)	396 (20)	90 (23)	–0.02 (–0.07 to 0.02)	.33
Composite outcome (30-day mortality, hemostatic intervention, or transfusion	875 (45)	234 (59)	–0.14 (–0.19 to –0.08)	<.001
Clinical risk scores, mean (SD)				
Glasgow-Blatchford	6.53 (4.56)	7.81 (4.50)	–1.3 (–1.8 to –0.79)	<.001
Admission Rockall	2.82 (1.74)	2.7 (1.8)	0.1 (–0.07 to 0.31)	.22
AIMS65	1.0 (0.94)	0.96 (0.92)	0.04 (–0.05 to 0.14)	.37

ASA, American Society of Anesthesiologists; INR, international normalized ratio; RBC, red blood cell; SD, standard deviation.

[a] Includes mean difference and difference in proportions.

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 2-tailed nonparametric 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, version 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 $\geq$0.80 and $<$0.90 as

Table 3. Performance of XGBoost Machine Learning Model and Clinical Risk Assessment Scores

Composite endpoint[a]	Internal validation AUC (99% CI)	P value	External validation AUC (99% CI)	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)	.001	0.87 (0.84–0.91)	.004
Admission Rockall score	0.69 (0.66–0.71)	<.001	0.65 (0.60–0.71)	<.001
AIMS65	0.72 (0.69–0.74)	<.001	0.64 (0.59–0.69)	<.001

[a] Red blood cell transfusion, hemostatic intervention (endoscopy, surgery, or interventional radiology), or 30-day mortality.

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 3 comparisons, we adjusted the *P* value threshold for significance to $P = .017$ with the Bonferroni correction and present 99% confidence intervals (CIs).

Results

Patient Data

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

The original data set had a total of 3012 patients, with 655 patients (22% of the total data set) having 1 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), international normalized ratio (n = 251), thienopyridine use (n = 105), anticoagulant use (n = 105), or aspirin use (n = 104). All of these patients were excluded from the final data set, 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 Machine Learning Model

Internal Validation. The internal validation group was the 10-fold cross-validation of the final ML model of the training set comprising sites (Denmark, England, Scotland, and the United States) with approximately 390 patients in each fold.

For the composite outcome, ML model (AUC, 0.91; 99% CI, 0.90–0.93) performed better than the GBS (AUC, 0.88; 99% CI, 0.86–0.90; $P = .001$), admission Rockall score (AUC, 0.69; 99% CI, 0.66–0.71; $P < .001$), and AIMS65 (AUC, 0.72; 99% CI, 0.69–0.74; $P < .001$) (Table 3).

External Validation. For the composite endpoint, the ML model performed better than all clinical risk scores: AUC, 0.90; 99% CI, 0.87–0.93 vs GBS: AUC, 0.87; 99% CI, 0.84–0.91; $P = .004$; admission Rockall score: AUC, 0.65; 99% CI, 0.60–0.71; $P < .001$; AIMS65 score; AUC, 0.64; 99% CI, 0.59–0.69; $P < .001$ (Table 3).

Identifying Very-Low-Risk Patients

High-Sensitivity Cutoff for External Validation. Among the clinical risk scores, only GBS of 0 achieved a sensitivity at our prespecified cutoff of 100%; AIMS65 of 0 and pre-endoscopic Rockall of 0 had maximal sensitivities of 74% and 96%, respectively. The ML model performed better than GBS of 0 in correctly classifying patients who did not need a hospital-based intervention or did not die ($P < .001$): the ML model had a specificity of 26% at sensitivity 100% compared with a specificity of 12% at sensitivity of 100% with a GBS of 0. The accuracy for the ML model at this high sensitivity cutoff was 68% (99% CI, 0.64–0.73), whereas for GBS of 0, the accuracy was 63% (99% CI, 0.58–0.68). Because some have suggested a cutoff of GBS ≤ 1,[18,19] given reported sensitivity as high as approximately 99% with this cutoff, we also performed a post hoc comparison of specificities using a threshold for sensitivity of 99%. Specificities for the ML model set at 99% sensitivity and GBS ≤ 1 (which achieved 99% sensitivity threshold) were 35% and 27%, respectively ($P = .02$). To make the tool available for clinicians, we developed an app (US 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 on 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 2-fold as compared with 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, a GBS of 0, which is recommended as a cutoff by US 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 of 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 that 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, compared with 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 (US 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 used 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 this study 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 toward 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 the prediction of mortality on external validation with ML models compared to liver-specific Child-Pugh and Model for End-Stage Liver Disease scores (designed to predict mortality in all patients with cirrhosis rather than in those presenting with gastrointestinal bleeding), and only 1 study found improved performance.[45] All comparisons with clinical risk scores were limited by external validation data sets from the same region.

Strengths

First, this study examines clinically relevant outcome measures. The composite outcome helps triage very-low-risk patients who may be able to be managed as outpatients. Second, unlike prior studies, our patient cohort is large and 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 data set 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 geographic division of US-Europe and Asia-Pacific centers as the criterion to separate the training from the external validation set, and the 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 data set was excluded due to 1 or more missing variables. This may introduce bias into the models because there may be nonrandom 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 before 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 will die are not identified as being at very low risk and are still admitted. However, the improvement in specificity with the ML model compared with GBS potentially should translate into a substantial reduction in health care use. 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 whereas the ML model would identify 119 such patients.

Finally, the data used were 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 the 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 2 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 data fields 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 being at 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, health care use, and patient outcomes.

Supplementary Material

Note: To access the supplementary material accompanying this article, visit the online version of *Gastroenterology* at www.gastrojournal.org, and at https://doi.org/10.1053/j.gastro.2019.09.009.

References

1. Longstreth GF. Epidemiology of hospitalization for acute upper gastrointestinal hemorrhage: a population-based study. Am J Gastroenterol 1995;90:206–210.
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;90:568–573.
3. Blatchford O, Davidson LA, Murray WR, et al. Acute upper gastrointestinal haemorrhage in west of Scotland: case ascertainment study. BMJ 1997;315:510–514.
4. Rockall TA, Logan RF, Devlin HB, et al. Incidence of and mortality from acute upper gastrointestinal haemorrhage in the United Kingdom. BMJ 1995;311:222–226.
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;92:236–243.
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;12:175–181.
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;12:1215–1220.
8. van Leerdam ME. Epidemiology of acute upper gastrointestinal bleeding. Best Pract Res Clin Gastroenterol 2008;22:209–224.
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;60:1327–1335.
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. Gastrointest Endosc 2014;81:882–888.
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.
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. Am J Gastroenterol 2009;104:1633–1641.
13. Wuerth BA, Rockey DC. Changing epidemiology of upper gastrointestinal hemorrhage in the last decade: a nationwide analysis. Dig Dis Sci 2017;63:1286–1293.
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.
15. Laine L, Jensen DM. Management of patients with ulcer bleeding. Am J Gastroenterol 2012;107:345–360.
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;74:1215–1224.
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;152:101–113.
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(10):a1–a46.
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.
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;64:2078–2087.
21. Tibshirani R. Regression shrinkage and selection via the lasso. J R Stat Soc Series B Stat Methodol 1996;58:267–288.
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. J R Stat Soc Series B Stat Methodol 2005;67:301–320.
24. Suykens JA, Vandewalle J. Least squares support vector machine classifiers. Neural Process Lett 1999;9:293–300.
25. Chen T, Guestrin C. XGBoost: a scalable tree boosting system. arXiv Website. https://arxiv.org/pdf/1603.02754.pdf. Accessed March 24, 2019.

26. Breiman L. Random forests. Mach Learn 2001;45:5–32.
27. Hecht-Nielsen R. Theory of the backpropagation neural network. In: Weschler H, ed. *Neural networks for perception*. Amsterdam, The Netherlands: Elsevier, 1992:65–93.
28. Hornik K, Stinchcombe M, White H. Multilayer feedforward networks are universal approximators. Neural Netw 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.
30. Steyerberg EW. Validation in prediction research: the waste by data splitting. J Clin Epidemiol 2018;103:131–133.
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.
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.
33. Kennedy EH, Wiitala WL, Hayward RA, et al. Improved cardiovascular risk prediction using nonparametric regression and electronic health record data. Med Care 2013;51:251–258.
34. Rotondano G, Cipolletta L, Grossi E, et al. Artificial neural networks accurately predict mortality in patients with nonvariceal upper GI bleeding. Gastrointest Endosc 2011;73:218–226.
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;44:837–845.
36. Kuhn M. Building predictive models in R using the caret package. J Stat Softw 2008;28:1–26.
37. Friedman JH, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. J Stat Softw 2010;33:1–22.
38. Sing T, Sander O, Beerenwinkel N, Lengauer T. ROCR: visualizing classifier performance in R. Bioinformatics 2005;21:3940–3941.
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.
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(9657):42–47.
42. Rotondano G, Cipolletta L, Grossi E, et al. Artificial neural networks accurately predict mortality in patients with nonvariceal upper GI bleeding. Gastrointest Endosc 2011;73:218–226.
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. J Clin Gastroenterol 2014;48:712–720.
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.
45. D'Amico G, De Franchis R. Upper digestive bleeding in cirrhosis. Post-therapeutic outcome and prognostic indicators. Hepatology 2003;38:599–612.
46. 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;64:2078–2087.
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(9981):1966–1974.
48. Shimabukuro DW, Barton CW, Feldman MD, 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 Respir Res 2017;4:e000234.
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 Netw Open 2018;1:e181018.

Received March 27, 2019. Accepted September 18, 2019.

Reprint requests

Address requests for reprints to: Dennis Shung, MD, Yale School of Medicine Section of Digestive Diseases, PO Box 208019, New Haven, CT 06520-8019. e-mail: dennis.shung@yale.edu; fax: (203) 737-1755; or Loren Laine, MD, Yale School of Medicine Section of Digestive Diseases, PO Box 208019 New Haven, CT 06520-8019. e-mail: loren.laine@yale.edu; fax: (203) 737-1755.

Acknowledgments

Author contributions: Dennis Shung: study concept and design, data analysis and interpretation, model development, app development, and drafting of the manuscript. Benjamin Au: study concept and design, data analysis, and model development. Richard Andrew Taylor: study design, data analysis and interpretation, app development, and critical revision of manuscript. J. Kenneth Tay: study design, feedback on data interpretation, and critical revision of manuscript. Loren Laine: study concept and design, data analysis and interpretation, drafting of the manuscript, and critical revision of manuscript. Stig Laursen, Adrian Stanley, Harry R. Dalton, Jeffrey Ngu, and Michael Schultz: collaborators for the multicenter data set. All authors had access to the study data and reviewed and approved the final manuscript.

Conflicts of interest

The authors disclose no conflicts.

Funding

This work was supported by the National Institutes of Health (T32 DK007017 to Dennis Shung).