Original Paper

Prediction of Blood Transfusion Need and Dose in Patients With Upper Gastrointestinal Bleeding: Retrospective Multicenter Prediction Model Study

Xiaoyu Li[1,2*], PhD; Yuqin He[3*], BMedSci; Mingyang Hou[1], PhD; Zhi Yu[1], PhD; Shuai Miao[1], PhD; Zhiyong Huang[1], PhD, Prof Dr; Min Yang[3], PhD, Prof Dr Med

[1]School of Microelectronics and Communication Engineering, Chongqing University, Chongqing, China
[2]Bioengineering College of Chongqing University, Chongqing University, Chongqing, China
[3]Department of Gastroenterology, Daping Hospital, Army Medical Center of PLA, Army Medical University, Chongqing, China
*these authors contributed equally

Corresponding Author:

Zhiyong Huang, PhD, Prof Dr
School of Microelectronics and Communication Engineering
Chongqing University
No. 174 Shazheng Street, Shapingba District
Chongqing 400044
China
Phone: 86 023-65678860
Email: zyhuang@cqu.edu.cn

Abstract

Background: Transfusion thresholds in upper gastrointestinal bleeding are debated; hemoglobin cutoffs of 70-80 g/L are widely cited yet inconsistently applied. Common risk scores offer limited individualized guidance and rarely provide calibrated, interpretable predictions for transfusion decisions.

Objective: This study aimed to develop and validate a two-stage, clinically constrained gradient-boosting framework (Medically Constrained Gradient Boosting [MCGB]) that predicts transfusion need and estimates transfusion dose with quantified uncertainty and to implement a prototype recommendation system for clinical use.

Methods: We analyzed a retrospective multicenter cohort of 849 adults with endoscopically confirmed upper gastrointestinal bleeding admitted to 3 hospitals in Chongqing, China (January 2019 to August 2025). Predictors available before the transfusion decision included demographics, first recorded vital signs, initial laboratory indices, and clinician-adjudicated etiology. Stage 1 used a calibrated classifier with prespecified monotonic constraints and stability-screened, clinically justified interactions. Stage 2 modeled transfusion dose via quantile predictions with conformal adjustment to generate 95% prediction intervals. Performance was assessed using a cross-site hold-out design. Overall, 2 hospitals were used as the development cohort, within which stratified 5-fold cross-validation was performed for model development, hyperparameter tuning, interaction screening, and calibration. The remaining hospital was held out as an independent test cohort for final evaluation. Hospital-wise alternating external testing was further conducted as a supplementary robustness analysis to assess performance stability across institutions. Classification performance was evaluated using discrimination metrics (area under the receiver operating characteristic curve and area under the precision-recall curve), calibration metrics, and decision-curve analysis; regression performance was evaluated using R^2, mean absolute error, and prediction-interval coverage. A graphical user interface was implemented to enable clinicians to input patient data and obtain calibrated predictions of transfusion probability and corresponding dose recommendations.

Results: MCGB achieved strong discrimination and good calibration across subgroups (area under the receiver operating characteristic curve=0.97 and area under the precision-recall curve=0.91). At a reference probability threshold of .50, sensitivity, specificity, and F_1-scores were 0.99, 0.87, and 0.85, respectively, providing a representative operating point for comparison. For dose prediction among transfused patients, MCGB achieved R^2 of 0.95 and mean absolute error 0.04; 95% prediction-interval coverage was 0.94, indicating accurate point estimates with reliable uncertainty quantification. The software prototype further demonstrated feasibility of real-time decision support at the bedside.

Conclusions: MCGB provides calibrated, interpretable predictions of transfusion need and individualized dose in upper gastrointestinal bleeding and may support bedside decision-making and blood-bank planning, with a prototype interface demonstrating potential for clinical deployment. External validation in additional settings is warranted to confirm generalizability.

JMIR Med Inform 2026;14:e83889; doi: 10.2196/83889

Keywords: blood transfusion; decision support systems, clinical; electronic health records; gastrointestinal hemorrhage; machine learning; regression analysis; ROC curve; receiver operating characteristic

Introduction

Upper gastrointestinal bleeding (UGIB) is a frequent emergency in gastroenterology and acute care [1]. Common etiologies include rupture of esophageal or gastric varices in cirrhosis, peptic ulcer disease, erosive hemorrhagic gastritis, and malignant tumors [2]. Rapid blood loss can precipitate hypotension and coagulopathy with peripheral circulatory failure, progressing to shock and death without timely resuscitation and blood transfusion [3-5]. Although hemoglobin thresholds of 70-80 g/L are widely cited, clinical application varies across centers and patient profiles [6,7], reflecting differences in comorbidity burden, cardiopulmonary reserve, ongoing hemorrhage risk, and the rate and dose of bleeding [8,9]. These realities motivate decision support that is both individualized and reproducible.

Risk scores such as Glasgow-Blatchford, admission Rockall, and AIMS65 are widely used for early triage in UGIB [10-12]. They stratify short-term risk but do not provide calibrated, individualized probabilities of transfusion or quantitative guidance on transfusion dose. Fixed cutoffs may be sensitive to local practice and case mix and can underperform when clinical heterogeneity is high [13,14]. From a systems perspective, uncertainty about transfusion need and dose complicates blood-bank operations, cross-matching, and inventory management during busy endoscopy schedules or resource-constrained periods. A tool that yields reliable risk estimates and dose guidance could improve bedside decision-making while supporting operational planning.

Machine learning has improved discrimination in many clinical prediction tasks and sometimes outperforms traditional scores [15,16]. However, prior work on transfusion decision support has several limitations that hinder clinical uptake. First, models are often difficult to interpret, which impedes bedside discussion and auditability [17]. Second, probability calibration is inconsistent, limiting safe threshold selection and net-benefit optimization [18,19]. Third, performance can degrade across hospitals and etiologies due to distributional shift. Fourth, preprocessing pipelines sometimes use the full dataset rather than being refit within cross-validation folds, increasing the risk of information leakage and optimistic estimates [20,21]. Finally, most studies stop at binary classification and do not estimate transfusion dose or communicate uncertainty for resource allocation. These gaps suggest that methodological choices should encode clinical directionality, preserve interpretability, control information flow during training, and report both discrimination and calibration alongside clinical utility [22,23].

Electronic health records (EHRs) capture the measurements typically available before a transfusion decision, including demographics, initial vital signs, key laboratory indices, and clinician-adjudicated etiology [24]. Leveraging these data for actionable decision support requires models that map to clinical reasoning while remaining robust across centers [25]. Monotonic constraints can encode established relationships such as higher international normalized ratio (INR) or longer prothrombin time (PT) increasing transfusion risk, and higher hemoglobin, hematocrit, or blood pressure decreasing risk. Curating a small, clinically justified set of feature interactions can improve fit without sacrificing transparency. Calibration methods can align predicted probabilities with observed risks, enabling thresholds derived from decision-curve analysis to reflect clinical utility [26].

This study develops and validates a 2-stage, clinically constrained framework—Medically Constrained Gradient Boosting (MCGB)—for transfusion decision support in UGIB using multicenter EHR data. Stage 1 produces calibrated, interpretable probabilities of transfusion need by imposing monotonic clinical directionality and retaining a stability-screened set of clinically justified interactions. Stage 2 estimates transfusion dose using quantile predictions with conformal adjustment to provide 95% prediction intervals, communicating uncertainty for blood-bank planning. The aim is to deliver a reproducible, well-calibrated, and robust tool that supports bedside decisions and operational workflows across heterogeneous hospitals and etiologies.

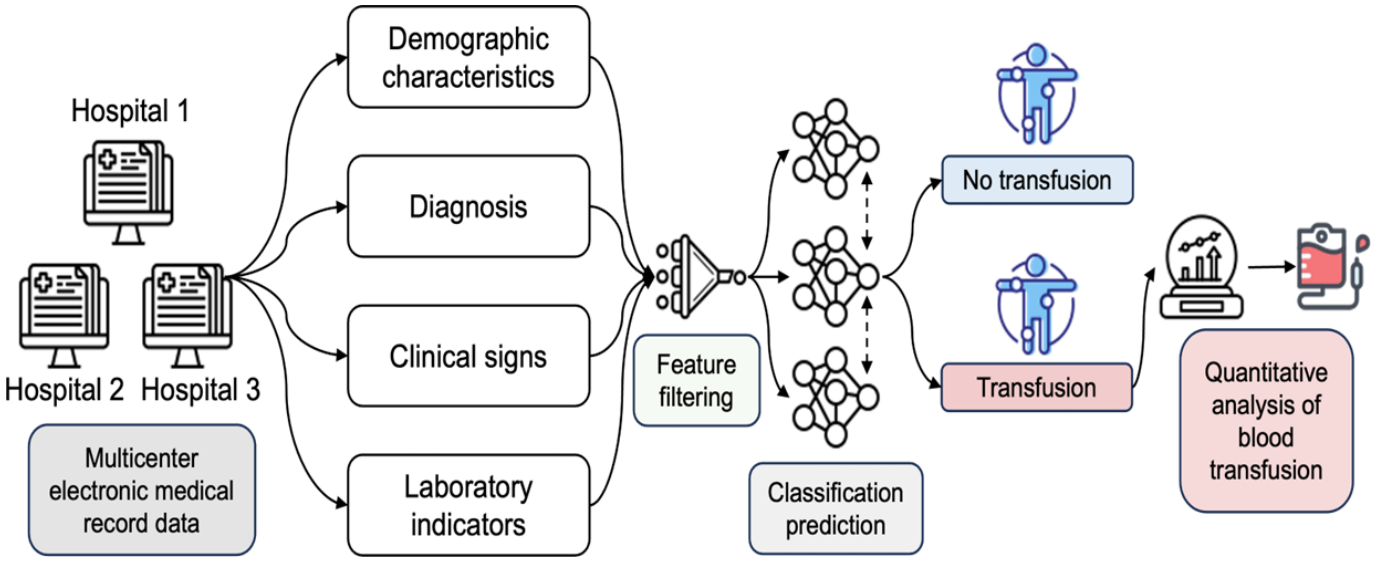
Methods

Overview

The transfusion decision system used a 2-stage framework termed MCGB. A calibrated classifier first estimates the probability of transfusion using monotonic constraints on prespecified laboratory and physiological variables to encode clinically consistent directionality, together with a stability-screened interaction allowlist. To enhance robustness across settings, training optimizes a group distributionally robust objective over hospital, etiology, and admission-period strata; predicted probabilities are calibrated by temperature scaling, and the operating threshold is selected with decision-curve analysis to align classification with clinical utility [27,28]. Patients predicted to require transfusion are then modeled with a regression component that estimates blood units via quantile predictions with conformal adjustment, yielding 95%

prediction intervals to communicate uncertainty for blood-allocation planning [29,30]. All preprocessing—Winsorization at the first and 99th percentiles, median imputation, indicator encoding, and interaction construction—is confined to cross-validation folds and applied only to the corresponding calibration and test splits to prevent information leakage. Candidate interactions arise from continuous-by-etiology and continuous-by-continuous pairs, are filtered for high correlation, ranked by mutual information and distance correlation, and retained when they meet prespecified stability criteria, following a screen-in principle that preserves interpretability, reproducibility, and transparency; the overall workflow is shown in Figure 1.

Figure 1. Overall workflow of the MCGB (Medically Constrained Gradient Boosting) framework for transfusion decision support.



Ethical Considerations

The study protocol was reviewed and approved by the institutional ethics committees of Daping Hospital, Army Medical Center of PLA, Army Medical University; the Third People's Hospital of Chongqing; and the 13th People's Hospital of Chongqing. The requirement for individual informed consent was waived because the analysis used routinely collected EHR data that were deidentified before analysis and posed minimal risk to participants. Data were handled in secure, access-controlled environments under institutional data-use agreements; only coded study identifiers were used, and direct personal identifiers were not accessible to investigators. All procedures complied with the Declaration of Helsinki and applicable local regulations. The project did not involve any prospective intervention and was not registered as a clinical trial.

Data and Dataset

Routinely collected data were retrieved from the clinical data warehouses of 3 hospitals in Chongqing, China. Consecutive adults with endoscopically confirmed UGIB between January 2019 and August 2025 were screened. A total of 849 patients met the inclusion criteria, including 655 nontransfusion cases and 194 transfusion cases. The cohort comprised 264 patients from the Third Affiliated Hospital of the Third Military Medical University, 326 patients from the Third People's Hospital of Chongqing, and 259 patients from the 13th People's Hospital of Chongqing. Exclusion criteria included insufficient clinical data, non-UGIB bleeding, incomplete diagnostic or therapeutic care, self-discharge, transfusion before the index decision, and missing key variables required for model development. The study was approved by the institutional ethics committees of all 3 hospitals, with a waiver of informed consent due to the retrospective design.

Predictors available within 24 hours before the transfusion decision were extracted from the EHR and included demographics, first recorded vital signs at presentation, and initial laboratory results obtained prior to the transfusion decision (covering hematology, coagulation, renal function, albumin, and electrolytes), as well as clinician-adjudicated bleeding etiology encoded as 5 binary indicators. Units were harmonized across centers, and variables not available before the transfusion decision were excluded to prevent data leakage. Patients with missing key variables required for modeling were excluded during cohort construction, resulting in a largely complete analytic dataset. Continuous variables were processed using a median imputation strategy as part of a standardized preprocessing pipeline, although imputation was rarely required in practice due to the near-complete nature of the data. Categorical variables were one-hot encoded, and missingness indicators for selected laboratory tests were constructed to preserve compatibility with potential real-world scenarios in which data may be incomplete. Pairwise correlations were summarized to characterize relationships among predictors while preserving clinically relevant variables. For descriptive comparisons, transfused and nontransfused groups were analyzed using Student *t* tests or nonparametric equivalents for continuous variables and chi-square or Fisher exact tests for categorical variables; 2-sided *P* values are reported in Table 1.

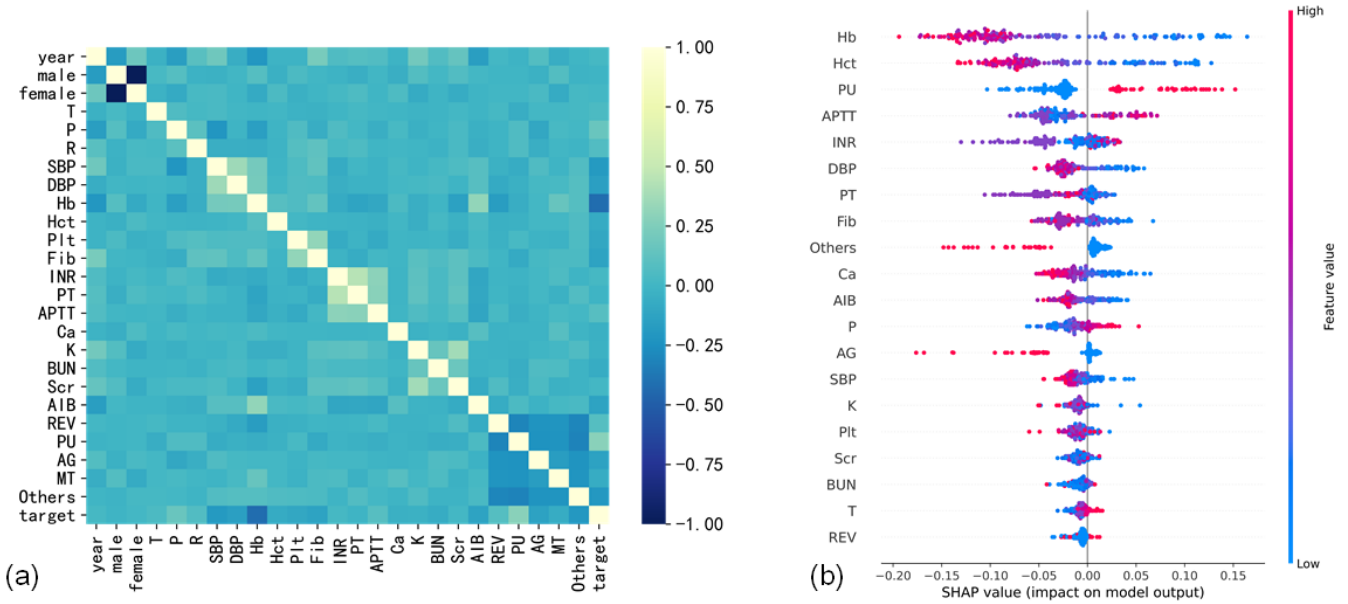
Table 1. Baseline characteristics for the nontransfusion and transfusion groups.

Variable	Nontransfusion group (n=655)	Transfusion group (n=194)	*P* value
Demographics			
Male, n (%)	435 (66.4)	126 (64.9)	.67
Female, n (%)	220 (33.6)	68 (35.1)	.67
Age (y), mean (SD)	66.09 (13.74)	66.77 (13.74)	.55
Vital signs, mean (SD)			
Temperature (°C)	36.53 (0.28)	36.58 (0.28)	.03
Pulse (beats/min)	86.61 (14.72)	93.19 (16.85)	<.001
Respiratory rate (breaths/min)	20.03 (2.62)	20.34 (2.62)	.15
SBP[a] (mm Hg)	123.72 (18.64)	114.37 (20.51)	<.001
DBP[b] (mm Hg)	72.84 (11.28)	65.07 (12.44)	<.001
Laboratory parameters, mean (SD)			
Hematocrit (%)	36.40 (7.21)	22.57 (6.48)	<.001
Platelets ($\times 10^9$/L)	177.00 (32.40)	187.10 (35.20)	<.001
Hemoglobin (g/L)	105.28 (21.46)	71.72 (18.35)	<.001
Fibrinogen (g/L)	3.09 (0.81)	2.76 (0.77)	<.001
INR[c]	1.19 (0.31)	2.72 (1.16)	<.001
PT[d] (s)	12.17 (2.18)	17.64 (6.21)	<.001
APTT[e] (s)	29.39 (5.75)	30.66 (5.75)	.01
BUN[f] (mmol/L)	10.68 (2.46)	11.58 (2.83)	<.001
Albumin (g/L)	37.91 (5.07)	34.51 (5.45)	<.001
Potassium (mmol/L)	4.13 (0.14)	4.08 (0.16)	<.001
Calcium (mmol/L)	2.26 (0.17)	2.07 (0.19)	<.001
Etiology, n (%)			
REV[g]	141 (21.5)	57 (29.4)	.02
PU[h]	116 (17.7)	91 (46.9)	<.001
AG[i]	116 (17.7)	12 (6.2)	<.001
MT[j]	99 (15.1)	19 (9.8)	.01
Others	183 (27.9)	15 (7.7)	<.001

[a]SBP: systolic blood pressure.
[b]DBP: diastolic blood pressure.
[c]INR: international normalized ratio.
[d]PT: prothrombin time.
[e]APTT: activated partial thromboplastin time.
[f]BUN: blood urea nitrogen.
[g]REV: rupture of esophageal varices.
[h]PU: peptic ulcer.
[i]AG: acute gastritis.
[j]MT: malignant tumor.

To further characterize the relationships among predictors and enhance transparency in model interpretation, pairwise correlations were first examined using a heat map Figure 2A, providing an overview of feature dependencies while retaining clinically relevant variables. Building on this analysis, model outputs were subsequently examined using Shapley additive explanations (SHAP), as illustrated in Figure 2B. Each point represents an individual patient from the evaluation set, and the horizontal axis indicates the SHAP value, which quantifies the contribution of each feature to the predicted probability of transfusion [31,32]. Positive SHAP values indicate an increased predicted risk, whereas negative values indicate a decreased risk, and the color represents the original feature value, with red denoting higher values and blue denoting lower values. Key predictors showed clinically coherent patterns: lower hemoglobin and hematocrit levels were associated with a higher predicted probability of transfusion, whereas higher values shifted predictions toward lower risk. Coagulation-related indices and hemodynamic variables also exhibited consistent directional effects that aligned with established clinical understanding of bleeding severity and transfusion requirements. Overall, these findings indicate that the model captures clinically meaningful and plausible relationships among predictors, supporting the interpretability and reliability of the MCGB classifier.

Figure 2. (A) Variable correlation heat map. (B) SHAP summary for transfusion predictors: each dot=patient; x-axis=SHAP value (positive ➞ higher probability, negative ➞ lower); color=feature value (red high, blue low). AG: acute erosive gastritis; Alb: albumin; APTT: activated partial thromboplastin time; BUN: blood urea nitrogen; Ca: calcium; DBP: diastolic blood pressure; Fib: fibrinogen; Hb: hemoglobin; Hct: hematocrit; INR: international normalized ratio; K: potassium; MT: malignant tumor; P: pulse; Plt: platelets; PT: prothrombin time; PU: peptic ulcer; R: respiratory rate; REV: rupture of esophageal varices; SBP: systolic blood pressure; Scr: serum creatinine; SHAP: Shapley additive explanations; T: temperature.



Model Building

In this study, the MCGB framework is proposed to extend standard gradient boosting by integrating 3 clinically motivated components: (1) monotonic constraints to enforce clinically consistent relationships, (2) interaction whitelisting to restrict feature interactions to stable and interpretable pairs, and (3) subgroup-aware reweighting to improve robustness across heterogeneous populations. These extensions are designed to improve both predictive performance and clinical reliability.

The dataset is represented as $D = (x_i, y_i)$, where $x_i \in \mathbb{R}^m$ denotes the input features, $y_i \in \mathbb{R}$ denotes the corresponding outcome, and $i = 1,2,\ldots,n$ indexes the samples. Following the standard gradient boosting formulation, the boosted predictor after T rounds can be expressed as:

$$F_T(x) = \sum_{t=1}^{T} f_t(x), \qquad f_t \in \mathfrak{F}, \tag{1}$$

where $\mathcal{F}$ denotes the space of regression trees mapping $\mathbb{R}^m$ to $\mathbb{R}$. Each tree partitions the data into leaves and assigns a score w_j. The boosting procedure iteratively refines predictions through stage-wise optimization, while the proposed extensions introduce clinically informed constraints and reweighting mechanisms to enhance robustness and consistency. The optimization objective at iteration t is expressed as:

$$Obj_t = \sum_{i=1}^{n} \mathfrak{l}(y_i,\ F_{t-1}(x_i) + f_t(x_i)) + \gamma T_t + \frac{\lambda}{2}\sum_{j=1}^{T_t} w_j^2, \tag{2}$$

where γ controls model complexity by penalizing the number of leaves, and λ is an L2 regularization parameter on leaf weights.

A second-order Taylor expansion is applied around $F_{t-1}(x_i)$. With first- and second-order derivatives defined as $g_i = \partial_{\hat{y}}\ell(y_i,\hat{y})\big|_{\hat{y}=F_{t-1}(x_i)}$ and $h_i = \partial^2_{\hat{y}}\ell(y_i,\hat{y})\big|_{\hat{y}=F_{t-1}(x_i)}$, the approximated loss becomes:

$$l(y_i,\ F_{t-1}(x_i) + f_t(x_i)) \approx l(y_i,\ F_{t-1}(x_i)) + y_i f_t(x_i) + \frac{1}{2} h_i f_t(x_i)^2 \tag{3}$$

To improve robustness across heterogeneous subgroups, MCGB introduces subgroup-aware reweighting of gradients. Specifically, for each predefined subgroup g, a softmax-based weighting scheme is defined as:

$$\pi_g(\eta) = \frac{\exp(\eta L_g)}{\sum_{\acute{g}} \exp(\eta L_{\acute{g}})}, \# \tag{4}$$

where L_g denotes the loss associated with group g, and η §amp;gt; 0 is a temperature parameter that controls the sharpness of the weighting distribution. For each sample i, let $g(i)$ denote the group to which the sample belongs. The reweighted derivatives are then defined as $\widetilde{g}_i = \pi_{g(i)}(\eta)g_i$and $\widetilde{h}_i = \pi_{g(i)}(\eta)h_i$, where g_i and h_i are the original first- and second-order derivatives. This design prevents the model from being dominated by majority subgroups and reduces overfitting to site-specific patterns.

At the tree level, gradients are aggregated within each leaf node. For a leaf j with the instance set I_j, the sufficient statistics are defined as $G_j = \sum_{i \in I_j} \widetilde{g}_i$ and $H_j = \sum_{i \in I_j} \widetilde{h}_i$. The quadratic surrogate objective is then approximated as:

$$\tilde{Obj}^{(t)} = \sum_{j=1}^{T_t}\left(G_j w_j + \frac{1}{2}(H_j+\lambda)w_j^2\right) + \gamma T_t . \# \quad (5)$$

The unconstrained optimal leaf weight is given by $\widehat{w}_j = -\frac{G_j}{H_j+\lambda}$. In MCGB, this solution is further adjusted to satisfy clinical monotonicity constraints of the form $\frac{\partial F_T(x)}{\partial x_k}\sigma_k \geq 0$, where $k \in M$ denotes features subject to monotonic constraints and $\sigma_k \in \{-1, +1\}$ specifies the expected direction. The feasible solution is obtained by projecting $\widehat{w}$onto the isotonic cone:

$$w^* = \underset{Aw \leq b}{\operatorname{argmin}} \frac{1}{2}||w - \widehat{w}||_2^2, \quad (6)$$

ensuring that model predictions follow clinically expected trends and preventing implausible extrapolations, thereby improving generalization and trustworthiness.

The reduction in the surrogate objective when splitting a node into left and right children is quantified by the gain, defined as:

$$\text{Gain} = \frac{1}{2}\left(\frac{G_L^2}{H_L+\lambda} + \frac{G_R^2}{H_R+\lambda} - \frac{(G_L+G_R)^2}{H_L+H_R+\lambda}\right) - \gamma \quad (7)$$

To further enhance interpretability and reproducibility, MCGB restricts candidate splits using an interaction whitelist W. To construct this whitelist, a stability-driven selection procedure is employed. Candidate feature pairs are first evaluated using the Friedman H-statistic to quantify interaction strength. Pairs with a median H-statistic greater than 0.10 across cross-validation folds are retained. To ensure robustness, only interactions that rank among the top five strongest interactions in at least four out of five folds are selected. This process ensures that retained interactions are both statistically stable and clinically plausible. If a node path already includes feature p, then a split on feature q is admissible only if $(p,q) \in W$. The whitelist itself is defined by a reproducibility criterion,

$$1\{(p,\ q) \in W = 1\{median_r\ H_{pq}^{(r)} > 0.10\ \wedge\ Top5_{pq}\ \geq 4\}\}, \# \quad (8)$$

Where $H_{pq}^{(r)}$ is the Friedman H-statistic in fold r and $Top5_{pq}$ is the number of folds in which (p,q) ranks among the 5 strongest interactions. This stability-based selection ensures robustness and reproducibility.

In summary, the MCGB framework extends classical gradient boosting by incorporating 3 clinically motivated components: monotonicity constraints to enforce clinically consistent relationships, interaction whitelisting to restrict feature interactions to stable and interpretable pairs, and group-wise loss reweighting to improve robustness across heterogeneous clinical subgroups. Together, these components enable the model to achieve not only strong predictive performance but also improved interpretability and stability in real-world clinical settings.

To evaluate the effectiveness of the proposed framework, all comparator models were implemented using standard libraries (scikit-learn, extreme gradient boosting [XGBoost], and categorical boosting [CatBoost]) and evaluated under the same cross-site hold-out design as the proposed MCGB model. Within the development cohort, each comparator was trained using the same stratified 5-fold cross-validation splits and fold-specific preprocessing pipeline, and the resulting cross-validation–derived models were subsequently applied to the independent test cohort. This design ensured a fair comparison while maintaining consistent separation between model development and final evaluation. For the classification task, the compared models included logistic regression (LogReg), multilayer perceptron (MLP), random forest (RF), adaptive boosting (AdaBoost), XGBoost [32], and CatBoost [33]; for the regression task, the compared models included linear regression (LR), MLP [34], RF, AdaBoost, XGBoost, and CatBoost. Hyperparameters for all baseline models were tuned via grid search within the training folds only, with model selection based on validation performance.

Model Training

Model development and validation were conducted using a predefined cross-site hold-out design to assess generalizability. Data from the Third People's Hospital of Chongqing and the 13th People's Hospital of Chongqing were combined as the development cohort, whereas data from the Third Affiliated Hospital of the Third Military Medical University were held out as an independent test cohort. The independent test cohort was not used for preprocessing fitting, feature interaction screening, hyperparameter tuning, calibration, threshold selection, or model selection. Within the development cohort, stratified 5-fold cross-validation was used to preserve the transfusion event rate across folds and to perform model development and internal validation. In each fold, preprocessing steps, including median imputation, one-hot encoding, and interaction construction, were fitted exclusively on the training split and then applied to the corresponding validation fold and the independent test cohort, ensuring strict separation and preventing information leakage. Each fold yielded a cross-validation–derived model, and final performance on the independent test cohort was summarized as the mean (SD) across the 5 models. To further examine cross-site robustness, hospital-wise alternating external testing was conducted as a supplementary analysis, in which each hospital was alternately treated as the external test cohort and the remaining 2 hospitals were used for model development. This supplementary analysis was used only to evaluate institutional stability and was not mixed with the primary independent test evaluation.

Given the moderate class imbalance (655 nontransfusion vs 194 transfusion cases), no explicit resampling was performed in order to preserve the original clinical distribution. Model evaluation therefore emphasized metrics robust to class imbalance, including the area under the precision-recall curve (AUPRC) alongside the area under the receiver operating characteristic curve (AUROC). Predicted probabilities were calibrated using temperature scaling within the development cohort, and the learned calibration parameters were then applied unchanged to the independent test cohort.

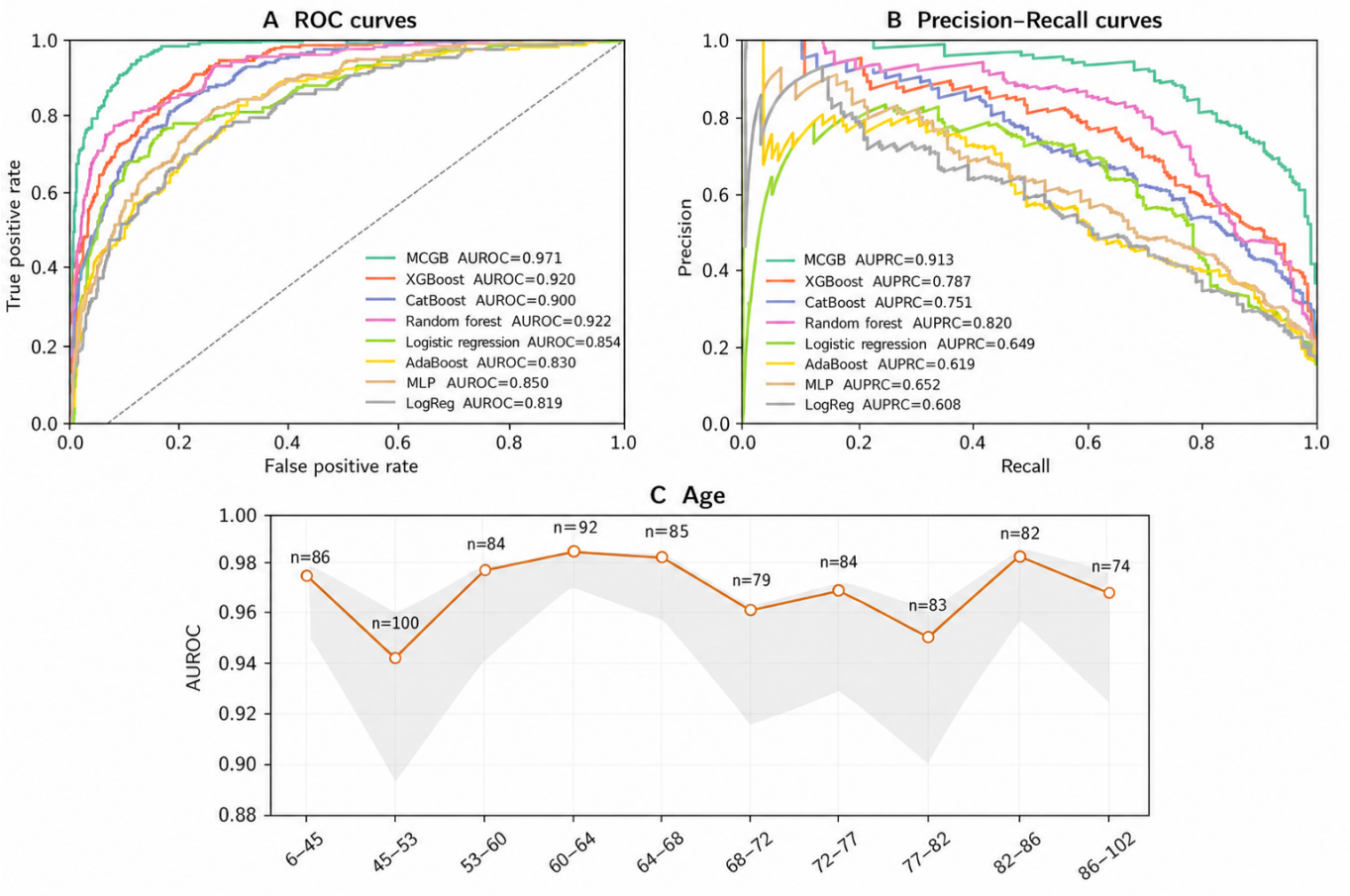
Model performance was primarily assessed across a range of clinically relevant threshold probabilities using decision curve analysis to reflect potential clinical utility. Results at a reference threshold of 0.50 were additionally reported as a conventional operating point for comparison with prior binary prediction studies. For the interaction ablation analysis, net benefit was additionally evaluated at a threshold probability of 0.20 because this value lies within the low-to-moderate risk range in which clinicians may consider early preparation, closer monitoring, or cross-matching rather than immediate transfusion. Therefore, ΔNB@0.20 was used to assess whether the retained interactions improved clinical utility in an early decision-support setting and was not intended to replace the 0.50 reference threshold used for reporting primary classification performance. Primary discrimination metrics included AUROC and AUPRC, whereas secondary metrics included sensitivity, specificity, positive predictive value, negative predictive value, accuracy, F_1-score, and expected calibration error.

For the regression component among transfused patients, predictive accuracy was evaluated using mean absolute error (MAE) and root-mean-square error, explanatory performance using R^2, and uncertainty using empirical coverage and interval width of 95% prediction intervals derived from conformalized quantile prediction. Agreement between predicted and observed transfusion dose was further assessed using Bland–Altman analysis.

Results

As shown in Figures 3A and 3B, MCGB achieved the best discrimination across all models with an AUROC of mean 0.971 (SD 0.008) and an AUPRC of mean 0.913 (SD 0.022). Relative to the strongest baseline, RF, MCGB improved AUROC by 0.049 (0.971 vs 0.922) and AUPRC by 0.093 (0.913 vs 0.820); gains were larger versus XGBoost (ΔAUROC 0.051; ΔAUPRC 0.126) and CatBoost (ΔAUROC 0.071; ΔAUPRC 0.162). The receiver operating characteristic (ROC) and precision–recall curves for MCGB dominate the comparators across clinically relevant thresholds, with the precision–recall separation indicating sustained precision at high recall under class imbalance. For AUROC, the 95% bootstrap intervals do not overlap between MCGB (mean 0.971, SD 0.008) and the leading baselines (RF mean 0.922, SD 0.031; XGBoost mean 0.920, SD 0.020); CatBoost mean 0.900 (SD 0.025), supporting a statistically meaningful improvement. AUPRC intervals are largely separated as well, with only marginal overlap relative to RF.

Figure 3. Discrimination and subgroup robustness of MCGB (Medically Constrained Gradient Boosting). (A) ROC curves on the independent test cohort. (B) Precision–recall curves on the independent test cohort. (C) AUROC stratified by hospital in the supplementary hospital-wise alternating external testing analysis. AdaBoost: adaptive boosting; AUPRC: area under the precision-recall curve; AUROC: area under the receiver operating characteristic curve; CatBoost: categorical boosting; MLP: multilayer perceptron; ROC: receiver operating characteristic; XGBoost: extreme gradient boosting.



As shown in Figures 3A and 3B, MCGB achieved the best discrimination across all models with an AUROC of mean 0.971 (SD 0.008) and an AUPRC of mean 0.913 (SD 0.022). Relative to the strongest baseline, RF, MCGB improved AUROC by 0.049 (0.971 vs 0.922) and AUPRC by 0.093 (0.913 vs 0.820); gains were larger versus XGBoost (ΔAUROC 0.051; ΔAUPRC 0.126) and CatBoost (ΔAUROC 0.071; ΔAUPRC 0.162). The ROC and precision–recall curves for MCGB dominate the comparators across clinically relevant thresholds, with the precision–recall separation indicating sustained precision at high recall under class imbalance. For AUROC, the 95% bootstrap intervals do not overlap between MCGB (mean 0.971, SD 0.008) and the leading baselines (RF mean 0.922, SD 0.031; XGBoost mean 0.920, SD 0.020); CatBoost mean 0.900 (SD 0.025), supporting a statistically meaningful improvement. AUPRC intervals are largely separated as well, with only marginal overlap relative to RF.

Table 2 summarizes classification performance on the independent test cohort. Values are reported as the mean (SD) across the 5 cross-validation–derived models trained within the development cohort. MCGB achieves the best overall performance, with the highest AUROC mean 0.971 (SD 0.008) and AUPRC mean 0.913 (SD 0.022), indicating strong discriminative ability under class imbalance. It also attains the highest accuracy mean 0.922 (SD 0.026) and F_1-score mean 0.847 (SD 0.041), reflecting a favorable balance between precision and recall. Notably, MCGB achieves near-perfect sensitivity mean 0.990 (SD 0.013) while maintaining relatively high specificity mean 0.873 (SD 0.032), suggesting effective identification of patients requiring transfusion with acceptable false-positive rates. Tree-based baselines such as RF and XGBoost show competitive discrimination but consistently lower composite performance, whereas linear and shallow models exhibit reduced recall and F_1-score despite reasonable AUROC. Overall, MCGB provides a more favorable error profile, which is desirable in safety-critical clinical settings.

Table 2. Classification performance.

Model	AUROC[a], mean (SD)	AUPRC[b], mean (SD)	Accuracy, mean (SD)	Sensitivity, mean (SD)	Specificity, mean (SD)	F_1-score, mean (SD)
AdaBoost[c]	0.828 (0.032)	0.617 (0.070)	0.691 (0.028)	0.856 (0.050)	0.642 (0.033)	0.557 (0.044)
CatBoost[d]	0.900 (0.025)	0.751 (0.062)	0.721 (0.033)	0.927 (0.037)	0.661 (0.036)	0.602 (0.051)
LogReg[e]	0.854 (0.031)	0.649 (0.071)	0.855 (0.024)	0.585 (0.068)	0.934 (0.021)	0.647 (0.058)
MLP[f]	0.849 (0.033)	0.652 (0.065)	0.688 (0.030)	0.854 (0.052)	0.639 (0.032)	0.555 (0.049)
RF[g]	0.922 (0.031)	0.820 (0.076)	0.705 (0.031)	0.877 (0.047)	0.654 (0.038)	0.575 (0.047)
XGBoost[h]	0.920 (0.020)	0.787 (0.056)	0.737 (0.029)	0.944 (0.035)	0.676 (0.034)	0.620 (0.047)
MCGB[i]	0.971 (0.008)	0.913 (0.022)	0.922 (0.026)	0.990 (0.013)	0.873 (0.032)	0.847 (0.041)

[a]AUROC: area under the receiver operating characteristic curve.
[b]AUPRC: area under the precision-recall curve.
[c]AdaBoost: adaptive boosting.
[d]CatBoost: categorical boosting.
[e]LogReg: logistic regression.
[f]MLP: multilayer perceptron.
[g]RF: random forest.
[h]XGBoost: extreme gradient boosting.
[i]MCGB: Medically Constrained Gradient Boosting.

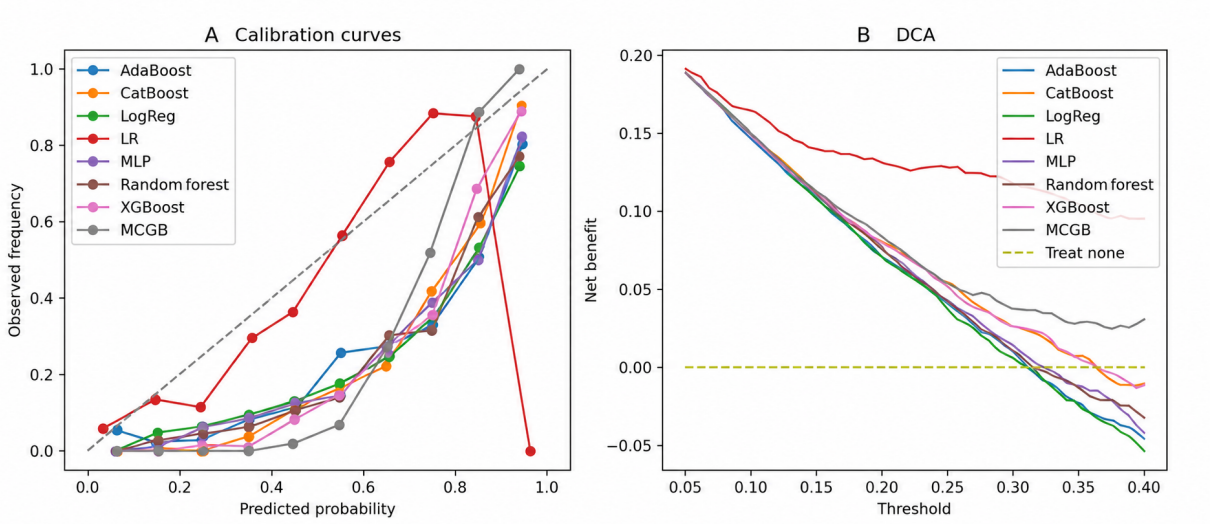
Figure 3C presents model performance stratified by hospital, etiology, and age. Across all strata, MCGB maintained consistently high AUROC with relatively narrow 95% CIs estimated via bootstrapping, indicating stable performance despite variation in subgroup sample sizes (eg, n≈264/326/259 across hospitals). Although some strata contained fewer samples, the corresponding CIs did not show substantial widening, suggesting that performance estimates remain reliable. As a supplementary robustness analysis, hospital-wise alternating external testing was conducted by alternately holding out each institution as the external test cohort and using the remaining 2 institutions for model development. This analysis was separate from the primary independent test evaluation and was intended to examine cross-site stability under institutional heterogeneity. In contrast, comparator models exhibited greater variability across strata, with more pronounced performance fluctuations in smaller subgroups. This stability of MCGB is consistent with its design, including monotonic constraints that encode clinically plausible relationships and a curated interaction structure that reduces overfitting to site-specific patterns. Overall, these results indicate that MCGB achieves not only strong discrimination but also robust and consistent performance across demographic and clinical subgroups, supporting its applicability in multicenter settings.

In addition to discrimination and subgroup robustness, the calibration and clinical usability of the proposed MCGB model were further evaluated. As shown in Figure 2, the reliability curves illustrate the agreement between predicted probabilities and observed outcomes across models. MCGB achieves the lowest Brier score, indicating superior overall probabilistic accuracy, although its calibration slope and intercept suggest a degree of overconfidence, which is commonly observed in high-capacity models. To address this, temperature scaling was applied as a post

hoc calibration method, resulting in improved probability reliability. Furthermore, decision curve analysis demonstrates that MCGB provides competitive or superior net benefit across most clinically relevant threshold ranges. While RF shows slightly higher net benefit in certain threshold intervals, this difference does not persist across clinically relevant ranges, where MCGB demonstrates consistently strong clinical utility.

To move beyond aggregate discrimination, we evaluated whether MCGB captures clinically meaningful interactions as shown in Figure 4. A stability-based screen-in rule combined the Friedman *H*-statistic across folds with mean SHAP interaction values and retained reproducible pairs such as INR × PT and systolic blood pressure × pulse, while interactions including blood urea nitrogen × serum creatinine and age × etiology were discarded. Paired ablation on out-of-fold predictions showed reductions in Brier score and gains in net clinical benefit at a decision threshold of 0.20, and the corresponding 95% CIs did not cross zero. Because the interaction analysis was designed to evaluate whether the retained interaction terms improved clinical utility in the early decision-support range, net benefit was assessed at a threshold probability of 0.20. This threshold corresponds to a low-to-moderate predicted risk level at which clinicians may initiate closer monitoring, prepare blood products, or consider cross-matching, whereas the 0.50 threshold was retained as the conventional reference operating point for reporting primary classification performance. Paired ablation on out-of-fold predictions showed reductions in Brier score and gains in net clinical benefit at this threshold, and the corresponding 95% CIs did not cross zero. These findings indicate that MCGB encodes physiologically plausible structure, with INR × PT reflecting coagulation synergy and systolic blood pressure × pulse capturing hemodynamic coupling, which enhances robustness and interpretability beyond overall discrimination.

Figure 4. Calibration and decision curve analysis of different models. (A) calibration curves comparing predicted probabilities with observed outcomes. The dashed line represents perfect calibration. (B) Decision curve analysis (DCA) showing the net benefit of each model across threshold probabilities. AdaBoost: adaptive boosting; CatBoost: categorical boosting; DCA: decision curve analysis; LR: linear regression; LogReg: logistic regression; MCGB: Medically Constrained Gradient Boosting; MLP: multilayer perceptron; XGBoost: extreme gradient boosting.



Among transfused patients, MCGB achieved the lowest prediction error and the highest explanatory power, with a mean MAE of 0.038 (SD 0.016), mean RMSE of 0.099 (SD 0.042), and a mean R^2 of 0.95 (SD 0.024). Because the transfusion dose was normalized during model training, these error values represent deviations on a relative scale; in practical terms, such small deviations indicate that predicted doses are closely aligned with clinician-determined transfusion decisions, typically differing only by a small margin that is unlikely to alter clinical management or blood allocation planning. Compared with the strongest baseline (RF), MCGB substantially reduced MAE (0.128-0.038) and RMSE (0.245-0.099), while improving R^2 (0.88-0.95), with consistent gains across evaluation metrics, suggesting improved predictive accuracy and stability rather than metric-specific improvements. For uncertainty estimation, conformalized quantile prediction achieved a 95% prediction-interval coverage of mean 0.943 (SD 0.021), which is close to the nominal level while maintaining practically useful interval widths; from a clinical perspective, these prediction intervals provide an interpretable range of likely transfusion requirements, supporting communication between clinicians and blood-bank services and facilitating more informed resource planning. In contrast, baseline models showed either undercoverage or less stable interval performance across folds. Overall, these findings indicate that MCGB not only achieves high statistical accuracy but also produces dose estimates that are clinically meaningful, reliable, and directly applicable to transfusion decision-making and operational planning, as summarized in Table 3.

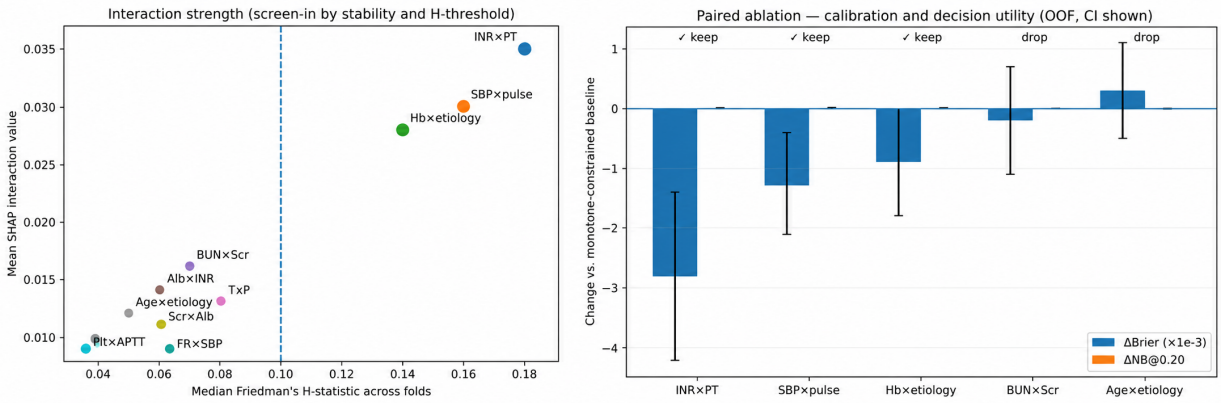
Table 3. Dose regression performance across models among transfused patients. Values are reported as mean (SD) across cross-validation folds. MAE[a] and RMSE[b] measure the deviation between predicted and actual transfusion dose (normalized scale), R^2 reflects explanatory power, and Coverage at 95% denotes the empirical coverage of 95% prediction intervals.

Model	MAE	RMSE	R^2[e]	Coverage at 95%
AdaBoost[c]	0.285 (0.058)	0.373 (0.102)	0.716 (0.041)	0.845 (0.045)
CatBoost[d]	0.262 (0.051)	0.320 (0.094)	0.829 (0.052)	0.867 (0.036)
LR[f]	0.302 (0.049)	0.386 (0.103)	0.670 (0.039)	0.838 (0.048)
MLP[g]	0.247 (0.065)	0.298 (0.098)	0.809 (0.049)	0.918 (0.033)
Random forest	0.128 (0.027)	0.245 (0.086)	0.88 (0.035)	0.912 (0.034)
XGBoost[h]	0.158 (0.036)	0.263 (0.081)	0.85 (0.033)	0.881 (0.032)
MCGB[i]	0.038 (0.016)	0.099 (0.042)	0.95 (0.024)	0.943 (0.021)

[a]MAE: mean absolute error.
[b]RMSE: root-mean-square error.
[c]AdaBoost: adaptive boosting.
[d]CatBoost: categorical boosting.
[e]Regression performance.
[f]LR: linear regression.
[g]MLP: multilayer perceptron.
[h]XGBoost: extreme gradient boosting.
[i]MCGB: Medically Constrained Gradient Boosting.

As illustrated in Figure 5, the MCGB regression model demonstrated a close alignment between predicted and true transfusion dose across all 5 cross-validation folds. The predicted curves (blue) consistently tracked the observed transfusion doses (red) with minimal deviation, yielding low MAE and RMSE, alongside a high coefficient of determination. These results indicate that MCGB was able to capture the fine-grained dose-response relationship while maintaining robustness across heterogeneous patient subgroups. The stability of the predictions across folds further suggests that the model generalizes well and avoids overfitting, making it potentially reliable for practical deployment in transfusion planning and blood inventory management.

Figure 5. Screening and ablation of feature interactions in MCGB (Medically Constrained Gradient Boosting). (A) Stability-based screen-in rule using the Friedman *H*-statistic and Shapley additive explanations.



Comparison with clinical risk scores. To further assess clinical relevance, the proposed MCGB model was compared with established clinical risk scores, including AIMS65, Glasgow-Blatchford Score, and Rockall. As shown in Table 4, these conventional scoring systems demonstrated modest discrimination, with AUROC values ranging from 0.613 to 0.693 and AUPRC from 0.310 to 0.377, whereas MCGB achieved substantially higher performance (AUROC: mean 0.971, SD 0.008; AUPRC: mean 0.913, SD 0.022). Notably, a marked discrepancy between AUROC and AUPRC was observed for the clinical scores. This pattern is consistent with the class imbalance of the cohort and indicates that, although these rule-based scores retain some ability to rank patients by risk, they have limited precision in identifying patients who truly require transfusion, resulting in a higher proportion of false-positive predictions at the individual level. In contrast, MCGB maintained consistently high values across both AUROC and AUPRC, suggesting not only strong global discrimination but also improved precision in detecting clinically relevant cases. This distinction is of particular importance in clinical practice, where AUPRC provides a more informative assessment of a model's ability to correctly identify patients requiring intervention. The observed differences may be attributed to the design of traditional scores, which rely on a small set of predefined variables and are primarily intended for coarse risk stratification, thereby limiting their capacity to capture complex and nonlinear relationships underlying transfusion decisions.

By comparison, MCGB incorporates data-driven modeling with clinically informed constraints, enabling more individualized and reliable predictions. Collectively, these findings suggest that, while established clinical scores remain useful for general risk assessment, they may be insufficient for precise transfusion prediction, and the proposed approach offers potential advantages for supporting clinical decision-making in upper gastrointestinal bleeding.

Table 4. Performance of established clinical risk scores for transfusion prediction under stratified five-fold cross-validation.

Model	AUROC[a]	AUPRC[b]
AIMS65	0.613 (0.050)	0.325 (0.062)
GBS[c]	0.693 (0.053)	0.377 (0.060)
Rockall	0.619 (0.045)	0.310 (0.042)
MCGB[d]	0.971 (0.008)	0.913 (0.022)

[a]AUROC: area under the receiver operating characteristic curve.
[b]AUPRC: area under the precision-recall curve.
[c]GBS: Glasgow-Blatchford Score.
[d]MCGB: Medically Constrained Gradient Boosting.

Through the preceding analyses, the MCGB model demonstrated strong performance in predicting transfusion requirements among patients with UGIB. To facilitate clinical translation, an interactive transfusion recommendation system was developed based on the MCGB model. This system functions as a user-oriented calculator that enables clinicians to estimate the probability of transfusion by inputting routinely available demographic and clinical variables. For patients predicted to require transfusion, the system additionally provides an individualized estimate of the recommended blood dose, thereby supporting both decision-making and treatment planning.

The graphical user interface, implemented using the QT Designer platform (Trolltech), is designed to be intuitive and easy to operate, allowing real-time interaction within clinical workflows. As illustrated in Figure 6, for a representative patient, the system estimated a transfusion probability of 89% and recommended a red blood cell dose of 400 mL. By integrating predictive modeling into a practical interface, the proposed system provides a feasible pathway for incorporating data-driven decision support into routine clinical practice. In this setting, clinicians can obtain rapid, standardized predictions without additional computational burden, which may contribute to more consistent and timely transfusion decision-making (Figure 7).

Figure 6. Cross-validation performance of MCGB (Medically Constrained Gradient Boosting) regression model. Comparison of predicted versus true transfusion dose across five cross-validation folds. Red lines denote true transfusion doses and blue lines denote MCGB predictions. Shaded background regions indicate fold partition. CV: cross-validation; MAE: mean absolute error; RMSE: root-mean-square error.

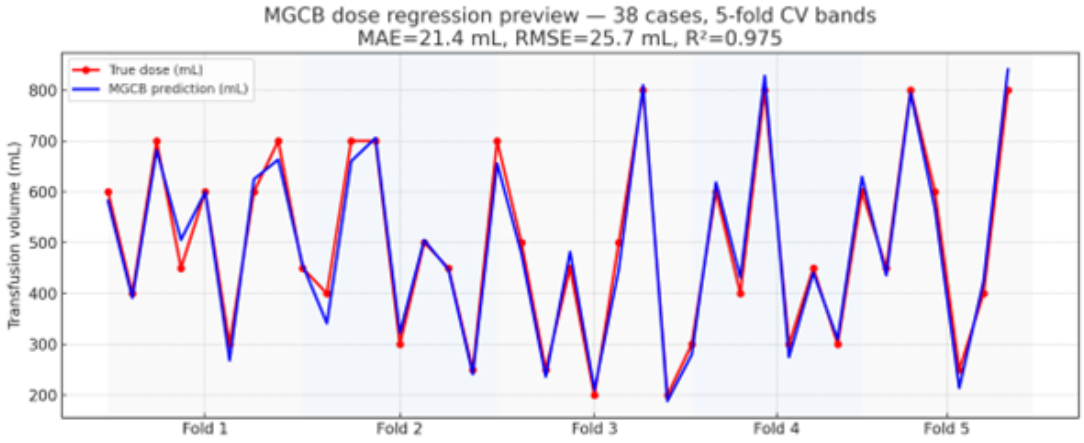
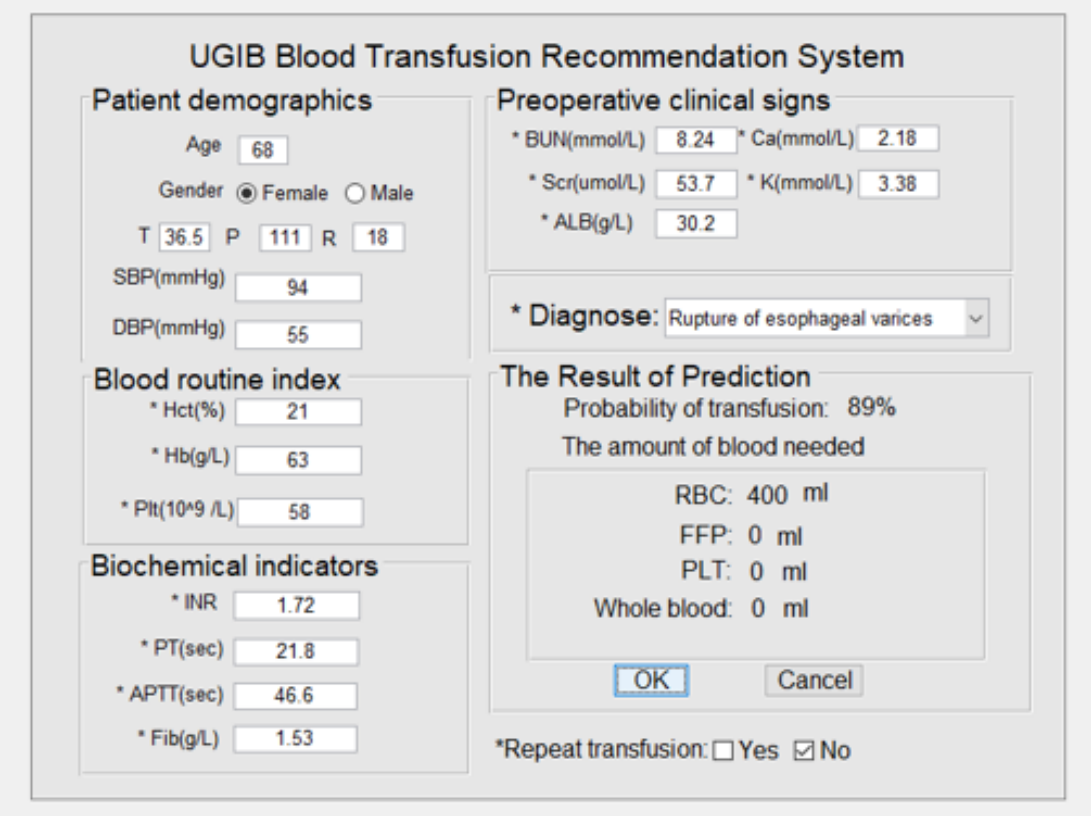


Figure 7. Interface of the UGIB blood transfusion recommendation system. The system integrates patient demographic information, vital signs, laboratory measurements, and clinical diagnosis as inputs, and provides individualized predictions of transfusion probability along with estimated blood component requirements. The interface enables clinicians to obtain real-time decision support for transfusion need assessment and dosage planning, facilitating standardized and interpretable clinical decision-making. ALB: albumin; APPT: activated partial thromboplastin time; BUN: blood urea nitrogen; Ca: calcium; DBP: diastolic blood pressure; FFP: fresh frozen plasma; Fib: fibrinogen; Hb: hemoglobin; Hct: hematocrit; INR: international normalization ratio; K: potassium; Plt: platelet; PT: prothrombin time; RBC: red blood cell; SBP: systolic blood pressure; Scr: serum creatinine.



Discussion

Principal Findings

This study shows that a clinically constrained, 2-stage framework can provide end-to-end decision support for transfusion in UGIB. The classifier delivered very high discrimination together with good calibration, and it maintained performance across hospitals, etiologies, and age strata. These characteristics matter more than a single accuracy number: calibration enables risk thresholds to be interpreted as probabilities, subgroup stability reduces the chance that performance collapses when case mix shifts, and explicit operating thresholds translate model output into consistent actions. In addition, calibration performance is reflected not only in visual agreement but also in quantitative metrics such as expected calibration error and Brier score, which are critical for clinical decision-making [35]. The innovation is not only the use of gradient boosting but the way clinical knowledge is built into the learning process. Monotonic constraints encode clinically expected relationships, and interaction constraints improve interpretability. In the second stage, dose regression provides point estimates with uncertainty, enabling actionable planning rather than binary decisions. In the primary cross-site hold-out evaluation, MCGB maintained strong discrimination and calibration when applied to an independent institutional test cohort.

Comparison With Prior Work

Traditional UGIB scores are useful for triage but do not provide patient-level probabilities of transfusion or quantitative guidance on dose. Many machine-learning studies report high AUROC yet give little attention to calibration, interpretability, or robustness across sites, and most stop at binary classification [36]. The present work addresses those gaps through several design choices. First, clinical constraints are imposed at training time rather than post hoc, which prevents clinically nonsensical relationships from emerging and reduces the need for manual rule fixes later. Second, interaction curation is principled: interactions such as INR × PT and systolic blood pressure × pulse are retained because they are stable across folds and supported by domain knowledge, and ablation confirms that they improve both calibration and net benefit. Third, evaluation extends beyond global curves to threshold performance, decision-curve analysis, and subgroup reporting, which are the quantities clinicians use when deciding whether and how to act. Finally, the dose module elevates the scope of decision support by estimating the quantity of blood required with explicit uncertainty, which is rarely attempted and directly relevant to operations [37,38].

Clinical Implications and Workflow

The proposed MCGB framework is designed to support bedside and blood-bank decision-making using routinely collected clinical data. At presentation, the calibrated classifier provides an individualized probability of transfusion need, and a prespecified threshold derived from decision curve analysis encodes the trade-off between missed transfusion and unnecessary cross-matching, enabling standardized triage and reducing subjective variability [39]. For patients exceeding this threshold, the dose model generates both a unit estimate and a 95% prediction interval, supporting clinical communication, cross-matching, and inventory planning [40]. These outputs can be delivered through the interactive UGIB transfusion recommendation system (Figure 6) and integrated into electronic health record systems to enable automated data extraction and real-time risk assessment without disrupting existing workflows. To support safe use, model outputs are intended as decision support rather than definitive recommendations, with probability estimates and uncertainty information provided to assist clinician interpretation and reduce the risk of automation bias. In practice, implementation also requires careful configuration of thresholds and alerts to align with clinical pathways and avoid alert fatigue [41]. Finally, as a clinical decision support tool, MCGB would require prospective validation, external evaluation, and compliance with relevant regulatory frameworks prior to deployment, with clinician oversight and transparent documentation to ensure accountability and safe integration into routine care.

Limitations and Future Directions

The analysis is retrospective and limited to 3 hospitals within a single regional health care system, which may restrict generalizability to other settings and populations; external and temporal validation is needed to assess transportability and inform recalibration strategies. The dataset is imbalanced with fewer transfusion events, and larger, more balanced cohorts are warranted. Some subgroup strata are relatively small, increasing uncertainty in stratum-specific estimates, although CIs suggest overall stability. In addition, patients with missing key variables were excluded during cohort construction, resulting in model development and validation on a largely complete analytic dataset. While this improves internal consistency, it may not fully reflect real-world deployment conditions, where laboratory measurements may be delayed, unavailable, or selectively ordered at the time of decision-making, requiring explicit strategies for incomplete inputs such as minimum-data requirements or clinician review. Furthermore, thromboelastography and other advanced coagulation indices were not included because they are not routinely available across centers. While such variables may provide additional physiological information, the model was intentionally developed using routinely collected clinical features to enhance generalizability and real-world applicability, and the strong performance observed suggests that commonly available variables capture the majority of clinically relevant signals for transfusion decision-making [42]. Future work should evaluate whether incorporating advanced coagulation markers can further improve performance in settings where such data are available. Although conformal prediction achieved near-nominal coverage, the clinical utility of interval width requires prospective validation. Finally, deployment will require human-in-the-loop safeguards, appropriate alert configuration, and attention to automation bias to ensure safe and effective clinical use.

Conclusions

This study developed a clinically constrained, 2-stage framework MCGB—to support transfusion decisions in UGIB. By incorporating monotonic constraints that encode established clinical directionality and retaining only stability-screened, clinically plausible interactions, MCGB achieved superior discrimination, reliable calibration, and interpretable predictions. Beyond classification of transfusion need, the framework extends to individualized dose estimation with well-calibrated prediction intervals, offering actionable guidance for both bedside decision-making and blood-bank resource planning. These results demonstrate that MCGB can bridge methodological advances in machine learning with clinical requirements for transparency, robustness, and operational relevance, highlighting its potential as a practical decision-support tool for UGIB management.

Acknowledgments

This work was supported by the Special Project to Enhance the Scientific and Technological Innovation Capabilities of the Army Military Medical University (No. 2022XLC07) and the Chongqing Natural Science Foundation (No. CSTB2024NSCQ-MSX0663).

Funding

The authors declared no financial support was received for this work.

Data Availability

The datasets generated and/or analyzed during this study are not publicly available due to institutional privacy regulations and patient confidentiality policies but are available from the corresponding author on reasonable request. Data sharing will follow JMIR Publications' data sharing policy, and deidentified data can be provided upon approval of a data access agreement.

Authors' Contributions

Conceptualization, Methodology, Formal analysis, Visualization, Writing – original draft: XL
Conceptualization, Methodology, Investigation, Writing – original draft: YH

Methodology, Software, Formal analysis, Visualization: MH
Methodology, Software, Formal analysis, Visualization: ZY
Data curation, Investigation, Validation: SM
Supervision, Project administration, Resources, Writing – review & editing: ZH
Supervision, Validation, Resources, Writing – review & editing: MY

Conflicts of Interest

None declared.

References

1. Long B, Gottlieb M. Emergency medicine updates: upper gastrointestinal bleeding. Am J Emerg Med. Jul 2024;81:116-123. [doi: 10.1016/j.ajem.2024.04.052] [Medline: 38723362]
2. Stanley AJ, Laine L. Management of acute upper gastrointestinal bleeding. BMJ. Mar 25, 2019;364:l536. [doi: 10.1136/bmj.l536] [Medline: 30910853]
3. Orpen-Palmer J, Stanley AJ. Update on the management of upper gastrointestinal bleeding. BMJ Med. 2022;1(1):e000202. [doi: 10.1136/bmjmed-2022-000202] [Medline: 36936565]
4. Barkun AN, Almadi M, Kuipers EJ, et al. Management of nonvariceal upper gastrointestinal bleeding: guideline recommendations from the International Consensus Group. Ann Intern Med. Dec 3, 2019;171(11):805-822. [doi: 10.7326/M19-1795] [Medline: 31634917]
5. Gralnek IM, Stanley AJ, Morris AJ, et al. Endoscopic diagnosis and management of nonvariceal upper gastrointestinal hemorrhage (NVUGIH): European Society of Gastrointestinal Endoscopy (ESGE) Guideline - update 2021. Endoscopy. Mar 2021;53(3):300-332. [doi: 10.1055/a-1369-5274] [Medline: 33567467]
6. Laine L, Barkun AN, Saltzman JR, Martel M, Leontiadis GI. ACG clinical guideline: upper gastrointestinal and ulcer bleeding. Am J Gastroenterol. May 1, 2021;116(5):899-917. [doi: 10.14309/ajg.0000000000001245] [Medline: 33929377]
7. de Franchis R, Bosch J, Garcia-Tsao G, Reiberger T, Ripoll C, Baveno VII Faculty. Baveno VII - renewing consensus in portal hypertension. J Hepatol. Apr 2022;76(4):959-974. [doi: 10.1016/j.jhep.2021.12.022] [Medline: 35120736]
8. Carson JL, Stanworth SJ, Dennis JA, et al. Transfusion thresholds for guiding red blood cell transfusion. Cochrane Database Syst Rev. Dec 21, 2021;12(12):CD002042. [doi: 10.1002/14651858.CD002042.pub5] [Medline: 34932836]
9. Vlaar APJ, Dionne JC, de Bruin S, et al. Transfusion strategies in bleeding critically ill adults: a clinical practice guideline from the European Society of Intensive Care Medicine. Intensive Care Med. Dec 2021;47(12):1368-1392. [doi: 10.1007/s00134-021-06531-x] [Medline: 34677620]
10. Blatchford O, Murray WR, Blatchford M. A risk score to predict need for treatment for upper-gastrointestinal haemorrhage. Lancet. Oct 14, 2000;356(9238):1318-1321. [doi: 10.1016/S0140-6736(00)02816-6] [Medline: 11073021]
11. Rockall TA, Logan RFA, Devlin HB, Northfield TC. Risk assessment after acute upper gastrointestinal haemorrhage. Gut. Mar 1996;38(3):316-321. [doi: 10.1136/gut.38.3.316] [Medline: 8675081]
12. 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. Jan 4, 2017;356:i6432. [doi: 10.1136/bmj.i6432] [Medline: 28053181]
13. Laursen SB, Oakland K, Laine L, et al. ABC score: a new risk score that accurately predicts mortality in acute upper and lower gastrointestinal bleeding: an international multicentre study. Gut. Apr 2021;70(4):707-716. [doi: 10.1136/gutjnl-2019-320002] [Medline: 32723845]
14. Saltzman JR, Tabak YP, Hyett BH, Sun X, Travis AC, Johannes RS. A simple risk score accurately predicts in-hospital mortality, length of stay, and cost in acute upper GI bleeding. Gastrointest Endosc. Dec 2011;74(6):1215-1224. [doi: 10.1016/j.gie.2011.06.024] [Medline: 21907980]
15. Redondo-Cerezo E, Vadillo-Calles F, Stanley AJ, et al. MAP(ASH): a new scoring system for the prediction of intervention and mortality in upper gastrointestinal bleeding. J Gastroenterol Hepatol. Jan 2020;35(1):82-89. [doi: 10.1111/jgh.14811] [Medline: 31359521]
16. Shung DL, Au B, Taylor RA, et al. Validation of a machine learning model that outperforms clinical risk scoring systems for upper gastrointestinal bleeding. Gastroenterology. Jan 2020;158(1):160-167. [doi: 10.1053/j.gastro.2019.09.009] [Medline: 31562847]
17. Shung DL, Simonov M, Gentry M, Au B, Laine L. Machine learning to predict outcomes in patients with acute gastrointestinal bleeding: a systematic review. Dig Dis Sci. Aug 2019;64(8):2078-2087. [doi: 10.1007/s10620-019-05645-z] [Medline: 31055722]
18. Shung DL, Chan CE, You K, et al. Validation of an electronic health record-based machine learning model compared with clinical risk scores for gastrointestinal bleeding. Gastroenterology. Nov 2024;167(6):1198-1212. [doi: 10.1053/j.gastro.2024.06.030] [Medline: 38971198]

19. Shung DL, Lin JK, Laine L. Achieving value by risk stratification with machine learning model or clinical risk score in acute upper gastrointestinal bleeding: a cost minimization analysis. Am J Gastroenterol. Feb 1, 2024;119(2):371-373. [doi: 10.14309/ajg.0000000000002520] [Medline: 37753930]
20. Collins GS, Reitsma JB, Altman DG, Moons KGM, Group T. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. Ann Intern Med. Jan 6, 2015;162(1):55-63. [doi: 10.7326/M14-0697] [Medline: 25560714]
21. Wolff RF, Moons KGM, Riley RD, et al. PROBAST: a tool to assess the risk of bias and applicability of prediction model studies. Ann Intern Med. Jan 1, 2019;170(1):51-58. [doi: 10.7326/M18-1376] [Medline: 30596875]
22. Dhiman P, Ma J, Andaur Navarro CL, et al. Methodological conduct of prognostic prediction models developed using machine learning in oncology: a systematic review. BMC Med Res Methodol. Apr 8, 2022;22(1):101. [doi: 10.1186/s12874-022-01577-x] [Medline: 35395724]
23. Andaur Navarro CL, Damen JAA, Takada T, et al. Risk of bias in studies on prediction models developed using supervised machine learning techniques: systematic review. BMJ. Oct 20, 2021;375:n2281. [doi: 10.1136/bmj.n2281] [Medline: 34670780]
24. Steyerberg EW, Vergouwe Y. Towards better clinical prediction models: seven steps for development and an ABCD for validation. Eur Heart J. Aug 1, 2014;35(29):1925-1931. [doi: 10.1093/eurheartj/ehu207] [Medline: 24898551]
25. Steyerberg EW, Vickers AJ, Cook NR, et al. Assessing the performance of prediction models: a framework for traditional and novel measures. Epidemiology. Jan 2010;21(1):128-138. [doi: 10.1097/EDE.0b013e3181c30fb2] [Medline: 20010215]
26. Van Calster B, Nieboer D, Vergouwe Y, De Cock B, Pencina MJ, Steyerberg EW. A calibration hierarchy for risk models was defined: from utopia to empirical data. J Clin Epidemiol. Jun 2016;74:167-176. [doi: 10.1016/j.jclinepi.2015.12.005] [Medline: 26772608]
27. Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. Med Decis Making. 2006;26(6):565-574. [doi: 10.1177/0272989X06295361] [Medline: 17099194]
28. Vickers AJ, van Calster B, Steyerberg EW. A simple, step-by-step guide to interpreting decision curve analysis. Diagn Progn Res. 2019;3:18. [doi: 10.1186/s41512-019-0064-7] [Medline: 31592444]
29. Brier GW. Verification of forecasts expressed in terms of probability. Mon Wea Rev. Jan 1950;78(1):1-3. [doi: 10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2]
30. Romano Y, Patterson E, Candès EJ. Conformalized quantile regression. In: Advances in Neural Information Processing Systems. Vol 2019. :3543-3553.
31. Ponce-Bobadilla AV, Schmitt V, Maier CS, Mensing S, Stodtmann S. Practical guide to SHAP analysis: explaining supervised machine learning model predictions in drug development. Clin Transl Sci. Nov 2024;17(11):e70056. [doi: 10.1111/cts.70056] [Medline: 39463176]
32. ChenT, GuestrinC. XGBoost: a scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Association for Computing Machinery; 785-794. [doi: 10.1145/2939672.2939785]
33. Prokhorenkova L, Gusev G, Vorobev A, Dorogush AV, Gulin A. CatBoost: unbiased boosting with categorical features. In: Advances in Neural Information Processing Systems. Vol 2018. :6638-6648.
34. Guo C, Pleiss G, Sun Y, Weinberger KQ. On calibration of modern neural networks. Presented at: Proceedings of the 34th international conference on machine learning; Aug 6-11, 2017; Sydney, Australia.
35. Meier JM, Tschoellitsch T. Artificial intelligence and machine learning in patient blood management: a scoping review. Anesth Analg. Sep 1, 2022;135(3):524-531. [doi: 10.1213/ANE.0000000000006047] [Medline: 35977362]
36. Maynard S, Farrington J, Alimam S, et al. Machine learning in transfusion medicine: a scoping review. Transfusion. Jan 2024;64(1):162-184. [doi: 10.1111/trf.17582] [Medline: 37950535]
37. Goodnough LT, Hollenhorst MA. Clinical decision support and improved blood use in patient blood management. Hematology Am Soc Hematol Educ Program. Dec 6, 2019;2019(1):577-582. [doi: 10.1182/hematology.2019000062] [Medline: 31808902]
38. Goodnough LT, Shieh L, Hadhazy E, Cheng N, Khari P, Maggio P. Improved blood utilization using real-time clinical decision support. Transfusion. May 2014;54(5):1358-1365. [doi: 10.1111/trf.12445] [Medline: 24117533]
39. Strickland M, Nguyen A, Wu S, et al. Assessment of machine learning methods to predict massive blood transfusion in trauma. World J Surg. Oct 2023;47(10):2340-2346. [doi: 10.1007/s00268-023-07098-y] [Medline: 37389644]
40. Goodnough LT, Shah N. Is there a “magic” hemoglobin number? clinical decision support promoting restrictive blood transfusion practices. Am J Hematol. Oct 2015;90(10):927-933. [doi: 10.1002/ajh.24101] [Medline: 26113442]
41. Crispin P, Akers C, Brown K, et al. A review of electronic medical records and safe transfusion practice for guideline development. Vox Sang. Jun 2022;117(6):761-768. [doi: 10.1111/vox.13254] [Medline: 35089600]

42. Zapf MAC, Fabbri DV, Andrews J, et al. Development of a machine learning model to predict intraoperative transfusion and guide type and screen ordering. J Clin Anesth. Dec 2023;91:111272. [doi: 10.1016/j.jclinane.2023.111272] [Medline: 37774648]

Abbreviations

AdaBoost: adaptive boosting
AUPRC: area under the precision-recall curve
AUROC: area under the receiver operating characteristic curve
CatBoost: categorical boosting
EHR: electronic health record
INR: international normalized ratio
LogReg: logistic regression
LR: linear regression
MAE: mean absolute error
MCGB: Medically Constrained Gradient Boosting
MLP: multilayer perceptron
PT: prothrombin time
RF: random forest
RMSE: root-mean-square error
ROC: receiver operating characteristic
SHAP: Shapley additive explanations
UGIB: upper gastrointestinal bleeding
XGBoost: extreme gradient boosting

Edited by Arriel Benis; peer-reviewed by Chih-Yuan Yang, Rinku Sharma Dixit, Valentina Palama; submitted 10.Sep.2025; final revised version received 01.Jun.2026; accepted 12.Jun.2026; published 04.Aug.2026

Please cite as:
Li X, He Y, Hou M, Yu Z, Miao S, Huang Z, Yang M
Prediction of Blood Transfusion Need and Dose in Patients With Upper Gastrointestinal Bleeding: Retrospective Multicenter Prediction Model Study
JMIR Med Inform 2026;14:e83889
URL: https://medinform.jmir.org/2026/1/e83889
doi: 10.2196/83889