

On-scene machine learning prediction model for massive transfusion in trauma and its association with in-hospital mortality

Byungchul Yu^{1,2,†}, Jaehyeong Cho^{3,4,†}, Hyunjee Kim^{3,5,†}, Seung Ha Hwang^{3,6,7}, Jiyoung Hwang³ , Soeun Kim^{3,5}, Jiyeon Oh^{3,8}, Sooji Lee^{3,8}, Do Wan Kim⁹, Junepill Seok^{10,11} , Kyoungwan Kim¹², Jinseok Lee^{3,6,7}, Dong Keon Yon^{3,5,13,*}  and Wu Seong Kang^{2,*} 

¹Traumatology, Gachon University College of Medicine, Incheon, Korea

²Department of Trauma Surgery, Gachon University Gil Medical Center, Incheon, Korea

³Center for Digital Health, Medical Science Research Institute, Kyung Hee University College of Medicine, Seoul, Korea

⁴Department of Medicine, CHA University School of Medicine, Seongnam, Korea

⁵Department of Precision Medicine, Kyung Hee University College of Medicine, Seoul, Korea

⁶Department of Electronics and Information Convergence Engineering, Kyung Hee University, Yongin, Korea

⁷Department of Biomedical Engineering, Kyung Hee University, Yongin, Korea

⁸Department of Medicine, Kyung Hee University College of Medicine, Seoul, Korea

⁹Department of Thoracic and Cardiovascular Surgery, Chonnam National University Hospital and Chonnam National University Medical School, Gwangju, Korea

¹⁰Department of Thoracic and Cardiovascular Surgery, Chungbuk National University Hospital, Cheongju, Korea

¹¹Department of Thoracic and Cardiovascular Surgery, Chungbuk National University, College of Medicine, Cheongju, Korea

¹²Department of Trauma Surgery, Jeju Regional Trauma Center, Cheju Halla General Hospital, Jeju, Korea

¹³Department of Pediatrics, Kyung Hee University Medical Center, Kyung Hee University College of Medicine, Seoul, Korea

*Correspondence to: Dong Keon Yon, Department of Pediatrics, Kyung Hee University College of Medicine, 23 Kyungheedaero-ro, Dongdaemun-gu, Seoul 02447, South Korea (e-mail: yonkang@gmail.com); Wu Seong Kang, Department of Trauma Surgery, Gachon University Gil Medical Center, 65, 21, Namdong-daero 774 beon-gil, Namdong-gu, Incheon, South Korea (e-mail: wuseongkang@naver.com).

†Contributed equally to this work

‡joint last authors

Abstract

Background: Early triage for massive transfusion (MT) is essential in trauma care but most existing scoring systems rely on in-hospital data. To address this limitation, a machine learning model using only prehospital variables to predict MT and stratify mortality risk was developed and externally validated.

Methods: Data from the Korean Trauma Data Bank from 19 trauma centres (2017–22) was used for model development and internal validation, with 2023 data for patients from four additional centres used for external validation. Trauma cases were identified using S or T codes from the Korean Classification of Diseases, 7th edition. MT was defined as ≥ 5 units packed red blood cells within 4 hours or ≥ 10 units within 24 hours. Machine learning models were trained using 21 prehospital variables, with a final ensemble model constructed from the top-performing algorithms. Model interpretability was assessed using Shapley additive explanations (SHAP), and the association between predicted probability tertiles (T1–T3) and in-hospital mortality was evaluated using logistic regression.

Results: In all, 227 567 patients were included in the development cohort and internal validation cohort, with 8867 patients in the external validation cohort. The soft-voting ensemble model, combining random forest and AdaBoost, showed high predictive performance, with area under the receiver operating characteristic curve values of 0.837 (internal validation) and 0.837 (external validation). SHAP analysis identified accident type as the most influential predictor, followed by consciousness level, and circulatory assistance. Higher model probability was associated with increased in-hospital mortality (adjusted odds ratios (95% confidence intervals) 2.34 (2.16 to 2.55), 2.70 (2.49 to 2.92), and 3.53 (3.25 to 3.83) for T1, T2, and T3, respectively).

Conclusion: A prehospital ensemble learning model to predict MT was developed and validated, and its predictions were significantly associated with in-hospital mortality. However, this study is limited by the inclusion of a single ethnicity, and future research needs to integrate data from multiple populations to enhance generalizability.

Keyword: artificial intelligence

Received: October 13, 2025. **Revised:** October 17, 2025. **Accepted:** December 01, 2025

© The Author(s) 2026. Published by Oxford University Press on behalf of BJS Foundation Ltd.

This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial License (<https://creativecommons.org/licenses/by-nc/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact reprints@oup.com for reprints and translation rights for reprints. All other permissions can be obtained through our RightsLink service via the Permissions link on the article page on our site—for further information please contact journals.permissions@oup.com.

Introduction

Among patients with trauma, uncontrolled haemorrhage remains one of the leading causes of preventable death worldwide¹. For patients with severe traumatic bleeding, massive transfusion (MT) is a life-saving intervention, with numerous studies showing that timely activation of MT protocols can reduce mortality^{2,3}. To support early recognition of patients requiring MT, several trauma scoring tools, such as the shock index (SI)⁴, Trauma-Associated Severe Haemorrhage (TASH) score⁵, and Assessment of Blood Consumption (ABC) score, have been proposed. However, these scores largely depend on imaging or laboratory parameters that are only available after hospital arrival, making them unusable for prehospital triage and decision-making^{6,7}. More recently, the Bleeding Audit and Triage Trauma (BATT) score, which uses prehospital variables, has been developed and validated⁸. However, the BATT score was designed to predict mortality and the use of tranexamic acid, with limited applicability for predicting the need for MT⁸.

Given these limitations, various machine learning (ML) approaches have emerged as promising alternatives for risk prediction in acute care, enabling the integration of complex and multidimensional data to identify non-linear patterns and improve predictive accuracy⁹. However, most existing ML models for MT prediction have focused on perioperative or in-hospital data. Few studies have developed explainable ML models using only prehospital information, a setting in which early and accurate decision support may be most valuable¹⁰.

To address these gaps, an explainable ML model to predict the likelihood of MT was developed and validated using prehospital clinical metadata derived from a large-scale nationwide trauma registry, namely the Korean Trauma Data Bank (KTDB). Furthermore, the association between model-predicted probabilities and in-hospital mortality was examined, and Shapley additive explanations (SHAP) was used to identify the most influential variables driving model predictions. These findings may facilitate the integration of artificial intelligence (AI)-assisted decision tools into trauma care workflows, particularly in time-sensitive prehospital environments where conventional scores underperform.

Methods

Data source and study population

This retrospective cohort study was conducted following the TRIPOD + Artificial Intelligence checklist¹¹. Data from the KTDB, a nationwide trauma registry established to improve trauma care by integrating trauma registries from 19 major trauma centres across Korea, was used¹². The study population comprised patients with trauma recorded between 1 January 2017 and 31 December 2022 (data set identification number K2024-09-2-02-29). Trauma cases were identified using S and T codes from the Korean Standard Classification of Diseases, 7th version, which is based on the International Classification of Diseases, Tenth Revision¹³. MT was identified from electronic medical records at each centre and was defined as meeting either of two established thresholds: the transfusion of ≥ 5 units of packed red blood cells within 4 hours; or transfusion of ≥ 10 units within 24 hours¹⁴. This composite definition preserves continuity with the conventional 24-hour criterion while capturing early resuscitation haemorrhage and reducing survivor bias. Following the application of exclusion criteria, the remaining data were used as a discovery cohort for model development, training, and internal validation.

For external validation, data from four major trauma centres in Korea (Cheju Halla General Hospital, Gachon University Gil Medical Center, Chonnam National University Hospital, and Chungbuk National University Hospital) were used. The external validation cohort included patients with traumatic injuries enrolled between 1 January 2023 and 31 December 2023. Trauma cases were identified using S and T codes from the Korean Standard Classification of Diseases, 7th version, and the same exclusion criteria and MT definition used for the discovery cohort were used for the validation cohort.

The research protocol was approved by the institutional review boards of the Gachon University Gil Medical Center (GCIRB 2025-127), Chonnam National University Hospital (CNUH-2025-070), Chungbuk National University Hospital (CBNUH-2025-03-002-003), and Cheju Halla General Hospital (CHH-2401-L02). The requirement for written informed consent was waived in accordance with relevant ethical guidelines for the secondary use of administrative data. This study followed the ethical guidelines outlined in the Declaration of Helsinki.

Data preprocessing

Patients younger than 15 years of age were excluded, because paediatric physiology and transfusion guidelines differ substantially from those of adults, with weight-adjusted definitions of MT and distinct haemodynamic responses to haemorrhage¹⁵. Missing values in key prehospital variables were handled using multiple imputation by chained equations. To develop a model for predicting MT among patients with trauma before hospital arrival, the following 21 metadata features were used: sex, age group, prehospital cardiac arrest, suspected alcohol consumption, accident type, type of injury, the use of protective equipment, time of injury, initial consciousness, abnormal SI, abnormal pulse, abnormal respiration, abnormal body temperature, abnormal oxygen saturation, prehospital airway management, prehospital oxygen administration, prehospital cardiopulmonary resuscitation, prehospital automated external defibrillator use, prehospital circulatory assistance, prehospital bleeding control, and prehospital management for pregnant women¹⁶. A detailed description of these variables is provided in [Table S1](#).

Given the class imbalance between MT (3.10%) and non-MT (96.90%) patients, the synthetic minority over-sampling technique was used to address this issue¹⁷. The synthetic minority over-sampling technique generated synthetic MT samples by interpolating between existing data points, thereby improving class balance and reducing bias during model training¹⁸. To standardize the data set, min-max normalization was applied by scaling all feature values to a range between 0 and 1¹⁹. The normalization parameters were derived from the training set and subsequently applied to the internal and external validation sets to ensure consistency and comparability across data sets.

Model development and training

Six ML models were used to predict the likelihood of MT: adaptive boosting (AdaBoost), extreme gradient boosting (XGBoost), gradient boosting machine (GBM), LightGBM, random forest (RF), and logistic regression (LR)^{17–19}. Each model was optimized through a grid search to identify the optimal hyperparameters for the prediction task²⁰. By fine-tuning these parameters, both the predictive accuracy and stability of the individual models were improved, thereby enhancing the overall performance of

the ensemble. The finalized hyperparameters for each model are summarized in [Table S2](#).

Subsequently, an ensemble approach was used by evaluating all possible combinations of the top three models among the six listed above. To enhance predictive performance and generalizability, a voting-based combination of RF and AdaBoost was implemented²¹. The voting method effectively leveraged the complementary strengths of RF and AdaBoost, resulting in superior predictive capability and robustness. Furthermore, five-fold cross-validation was used to ensure robust generalization and reduce the risk of overfitting²². The training data set was randomly shuffled and stratified to maintain the original class distribution of MT and non-MT patients across all folds. In each repeat, the model was trained on four folds and validated on the remaining fold, with the process repeated five times¹⁷. This stratified cross-validation approach enabled effective learning from the imbalanced data set while preserving representative validation subsets.

Model evaluation and external validation

Each model developed to predict the likelihood of MT was evaluated using an independent testing data set, which was held out from the discovery cohort exclusively for performance assessment. This data set was not used during model training or cross-validation, ensuring an unbiased evaluation of generalization performance. Model performance was assessed using sensitivity, specificity, accuracy, balanced accuracy, and the area under the receiver operating characteristic curve (AUROC)²³. These metrics were collectively used to provide a comprehensive evaluation and comparison of model performance.

Given the substantial class imbalance in the internal validation cohort (MT : non-MT = 1 : 31.2), balanced accuracy and AUROC were selected as the primary evaluation metrics. To enhance performance, likelihood thresholds for certain models were fine-tuned based on AUROC analysis. To improve performance, the models were tuned to use optimal likelihood thresholds based on AUROC analysis²⁴. The optimized thresholds were as follows: 0.46 for the AdaBoost and GBM model; 0.55 for the RF and GBM model; 0.53 for the AdaBoost and RF model; and 0.54 for the ensemble model combining AdaBoost, GBM, and RF. These thresholds were consistently applied throughout all evaluation procedures^{19,25}.

To assess the ML model's comparative reliability and performance, a single-variable model using SI, a widely recognized and easily obtainable predictor of MT in prehospital settings²⁶, was developed through LR analysis. The area under the curve was calculated to evaluate its predictive capacity, and Youden's index was used to determine the optimal SI threshold, sensitivity, and specificity²⁷. In addition, SHAP values were calculated to quantify the contribution of each feature to the model's predictions²⁸, ranking features based on their relative importance in predicting the likelihood of MT with on-site clinical data among patients with trauma.

To enhance the clinical applicability of the predicted probabilities, the model outputs were calibrated using Platt scaling, which was trained on the training/validation data through cross-validation²⁹. Model calibration was evaluated in the test set of the discovery cohort using calibration-in-the-large (intercept) and calibration slope³⁰. Calibration-in-the-large quantifies the difference between the average predicted probability and the observed event rate, whereas calibration slope reflects the agreement between predicted and observed risks across the probability range³⁰. A calibration-in-the-large value closer to 0.000

indicates minimal average prediction bias, whereas a calibration slope closer to 1.000 reflects strong concordance between predicted and observed risks. Finally, decision curve analysis (DCA) was conducted in the test set to assess the net clinical benefit of the model across a range of clinically relevant thresholds, compared with default strategies such as treating all patients or none³¹. Calibration plots and DCA curves were generated for both pre- and post-calibration models.

The generalizability of the model was further assessed using external data sets from four independent regional trauma centres. Given the substantial class imbalance in the external validation cohort, the same primary evaluation metrics, balanced accuracy and AUROC, were applied, consistent with the internal validation³². Based on performance across all evaluation stages, the model representing consistently high balanced accuracy and AUROC was selected as the final proposed ML model. All models were implemented using Python (version 3.11.11; Python Software Foundation, Wilmington, DE, USA) with the following libraries: TensorFlow (2.19.0; Google LLC, Mountain View, CA, USA), Keras (3.10.0; Google LLC), NumPy (2.1.0; NumFOCUS, Austin, TX, USA), Pandas (2.2.3; NumFOCUS), Matplotlib (3.10.1; NumFOCUS), and Scikit-learn (1.6.1; NumFOCUS).

Risk stratification and in-hospital mortality analysis

Based on the probabilities generated by the proposed ML model in the discovery cohort, individuals with predicted probabilities exceeding the optimized threshold (0.46) from the final ensemble model were classified as the exposed group, whereas those with probabilities below the threshold served as the reference group. The exposed group was further divided into tertiles (T1, T2, and T3) based on the distribution of predicted probabilities, corresponding to the lower (T1), middle (T2), and upper (T3) thirds of the exposed group¹⁷. The outcome was defined as any form of in-hospital mortality following transfer to a trauma centre, with all relevant conditions confirmed using data from the discovery cohort. In the external validation set, this analysis was not performed due to the insufficient sample size of in-hospital mortality.

To assess the association between the predicted risk of MT and subsequent in-hospital mortality, logistic regression analysis was used to estimate adjusted odds ratios (ORs) and the association 95% confidence interval (c.i.). Two models were constructed: Model 1 was adjusted for age and sex only, whereas Model 2 was adjusted for all covariates listed in [Table S3](#). Two-sided $P < 0.050$ was considered indicative of statistical significance. All statistical analyses were conducted using SAS® version 9.4 (SAS Institute, Cary, NC, USA).

Patient and public involvement

No patients were involved in setting the research question or the outcome measures, nor were they involved in developing plans for the design or implementation of the study. No patients were asked to advise on the interpretation or writing up of results. However, the results of this study will be disseminated to any of the study participants or wider relevant communities on request.

Results

Study population

Among the 227 567 patients admitted to major trauma centres in Korea included in the discovery cohort, 12 795 patients younger than 15 years of age were excluded ([Fig. S1](#)). Similarly, of the 8867

Table 1 Comparison of prediction model performance on the training and test data sets

	Sensitivity	Specificity	Accuracy	Balanced accuracy	AUROC
Five-fold cross-validation*					
AdaBoost	0.679 (0.663, 0.695)	0.809 (0.803, 0.814)	0.805 (0.800, 0.810)	0.744 (0.739, 0.749)	0.834 (0.829, 0.838)
RF	0.676 (0.644, 0.708)	0.823 (0.807, 0.838)	0.818 (0.804, 0.832)	0.749 (0.740, 0.758)	0.840 (0.836, 0.844)
GBM	0.668 (0.660, 0.676)	0.812 (0.807, 0.817)	0.808 (0.803, 0.812)	0.740 (0.737, 0.743)	0.827 (0.824, 0.830)
LightGBM	0.489 (0.464, 0.515)	0.867 (0.862, 0.872)	0.855 (0.850, 0.860)	0.678 (0.666, 0.690)	0.803 (0.797, 0.809)
XGBoost	0.525 (0.504, 0.546)	0.858 (0.854, 0.861)	0.847 (0.844, 0.851)	0.691 (0.681, 0.701)	0.808 (0.802, 0.813)
LR	0.691 (0.671, 0.711)	0.773 (0.762, 0.784)	0.770 (0.760, 0.780)	0.732 (0.726, 0.737)	0.807 (0.800, 0.814)
AdaBoost and RF†	0.576 (0.539, 0.613)	0.869 (0.858, 0.880)	0.860 (0.851, 0.869)	0.722 (0.709, 0.736)	0.837 (0.832, 0.841)
RF and GBM†	0.563 (0.525, 0.600)	0.871 (0.858, 0.883)	0.861 (0.850, 0.872)	0.717 (0.704, 0.730)	0.834 (0.831, 0.838)
GBM and AdaBoost†	0.667 (0.655, 0.680)	0.813 (0.808, 0.818)	0.808 (0.803, 0.813)	0.740 (0.735, 0.745)	0.828 (0.824, 0.831)
AdaBoost, GBM, and XGBoost†	0.560 (0.521, 0.599)	0.872 (0.859, 0.885)	0.862 (0.851, 0.873)	0.716 (0.702, 0.730)	0.834 (0.831, 0.838)
Test set					
AdaBoost	0.700	0.803	0.800	0.751	0.834
RF	0.698	0.813	0.809	0.755	0.836
GBM	0.672	0.809	0.805	0.741	0.825
LightGBM	0.512	0.860	0.849	0.686	0.804
XGBoost	0.542	0.852	0.842	0.697	0.808
LR	0.702	0.768	0.766	0.735	0.800
AdaBoost and RF†	0.600	0.857	0.849	0.729	0.837
RF and GBM†	0.583	0.857	0.849	0.720	0.833
GBM and AdaBoost†	0.671	0.814	0.810	0.742	0.826
AdaBoost, GBM, and XGBoost†	0.586	0.860	0.851	0.723	0.832
Extra validation set					
AdaBoost	0.750	0.722	0.723	0.736	0.812
RF	0.770	0.720	0.721	0.745	0.820
GBM	0.690	0.781	0.779	0.736	0.822
LightGBM	0.390	0.910	0.897	0.650	0.795
XGBoost	0.445	0.897	0.885	0.671	0.801
LR	0.800	0.615	0.620	0.707	0.783
AdaBoost and RF†	0.550	0.845	0.837	0.698	0.823
RF and GBM†	0.525	0.856	0.847	0.691	0.827
GBM and AdaBoost†	0.685	0.789	0.786	0.737	0.826
AdaBoost, GBM, and XGBoost†	0.560	0.851	0.843	0.706	0.826

Bolding indicates the model with the best performance. *Data show mean values with 95% confidence intervals in parentheses. †All ensemble models used a soft-voting method. AUROC, area under the receiver operating characteristic curve; AdaBoost, adaptive boosting; RF, random forest; GBM, gradient boosting machine; XGBoost, extreme gradient boosting; LR, logistic regression.

patients in the external validation cohort, 1299 patients younger than 15 years of age were excluded. Basic characteristics of development and validation cohort are shown in [Table S3](#).

Five-fold cross-validation and internal validation results

The performance of the ML models was evaluated using five-fold cross-validation and multiple evaluation metrics, namely sensitivity, specificity, accuracy, balanced accuracy, and AUROC, as summarized in [Table 1](#). Among the six ML models, AdaBoost, RF, and GBM exhibited superior performance as individual models, with relatively higher balanced accuracy and AUROC. Based on these results, the top three models were selected to develop ensemble models to improve predictive performance. Two ensemble approaches, including voting and stacking, were applied to all possible combinations of the three models. Among these, the ensemble learning model with RF and AdaBoost using a soft-voting approach achieved the best performance, with a sensitivity of 0.576 (95% c.i. 0.539 to 0.613), a specificity of 0.869 (95% c.i. 0.858 to 0.880), accuracy of 0.860 (95% c.i. 0.851 to 0.869), balanced accuracy of 0.722 (95% c.i. 0.709 to 0.736), and an AUROC of 0.837 (95% c.i. 0.832 to 0.841). Given the class imbalance, balanced accuracy and AUROC were used as the primary evaluation metrics, and the ensemble model based on RF and AdaBoost with a voting strategy showed the highest performance across these metrics and was identified as the optimal model for predicting the likelihood of MT in patients with trauma. Simplified model architecture is illustrated in [Fig. 1a](#).

Similarly, the proposed ensemble learning model with RF and AdaBoost achieved superior performance in internal validation using the test set from the discovery cohort. The model achieved a sensitivity of 0.600, a specificity of 0.857, an accuracy of 0.849, a balanced accuracy of 0.729, and an AUROC of 0.837 ([Table 1](#)). Given its consistently high performance across both cross-validation and internal validation, the model showed no evidence of overfitting or underfitting. Moreover, the single-variable model using SI in the test set yielded an area under the curve of 0.67, which was lower than that of the ensemble learning model ([Fig. S2](#)). At an optimal threshold of 0.51, the SI model showed a sensitivity of 0.42 and a specificity of 0.87.

External validation results

External validation was performed using data from four major trauma centres, as presented in [Table 1](#). The proposed ensemble model exhibited robust performance, achieving a sensitivity of 0.550, a specificity of 0.845, an accuracy of 0.837, a balanced accuracy of 0.698, and an AUROC of 0.823. These results indicate that the model maintained consistent predictive performance in identifying the likelihood of MT among patients with trauma, thereby validating its generalizability.

Model interpretation

To analyse the contribution of each feature in predicting the likelihood of MT among patients with trauma, SHAP value analysis was conducted for the proposed ensemble model ([Fig. 2](#)). Among the features included in the model, accident type

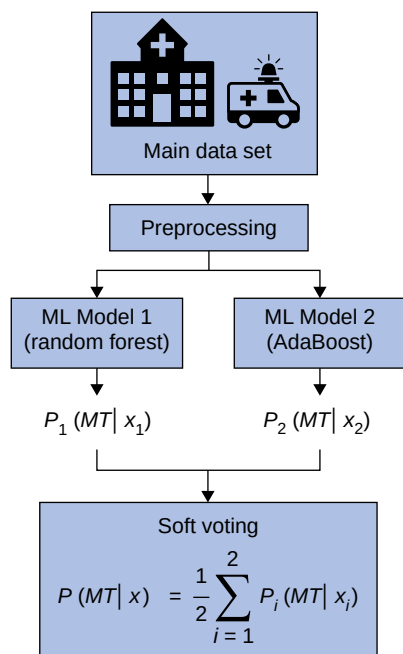
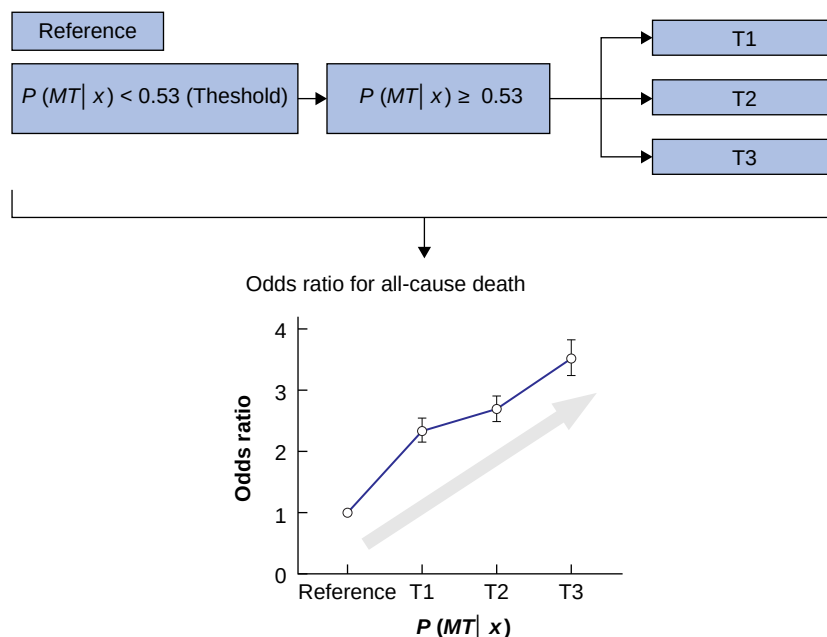
a Proposed ensemble learning model**b Mortality risk based on model probability**

Fig. 1 Model structure of proposed ensemble learning model to predict massive transfusion and mortality risk based on model probability

a Proposed ensemble learning model to predict massive transfusion. **b** Mortality risk based on model probability. The graph shows odds ratios for all-cause death relative to the reference for predicted likelihood tertiles. Tertiles (T1–T3) were defined based on the distribution of predicted probabilities among the exposed group, corresponding to the lower (T1), middle (T2), and upper (T3) thirds of predicted risk. Error bars indicate 95% confidence intervals. ML, machine learning; MT, massive transfusion; T1, first tertile; T2, second tertile; T3, third tertile.

was identified as the most influential predictor, followed by initial level of consciousness, circulatory assistance, initial on-scene oxygen saturation, and initial on-scene SI.

In addition to discrimination metrics, model calibration was assessed using mean calibration, calibration slope, and calibration plots, whereas clinical utility was evaluated through DCA. [Figure S2](#) shows that calibration markedly improved predictive reliability and clinical usefulness, with predicted risks more closely aligned to the ideal calibration line and higher net benefit maintained across a range of threshold probabilities. The model achieved a calibration-in-the-large (intercept) of -0.000 , indicating minimal average bias, and a calibration slope of 1.008 , suggesting strong concordance between predicted and observed risks with a slight tendency towards risk overestimation.

Risk of in-hospital mortality among patients after MT

[Table 2](#) presents the risk of in-hospital mortality according to the predicted likelihood of MT in the discovery cohort. The overall analytical process is illustrated in [Fig. 1b](#). In the adjusted model, patients in T1, T2, and T3, stratified by increasing predicted probabilities, exhibited progressively higher association with mortality, with ORs of 2.34 (95% c.i. 2.16 to 2.55), 2.70 (95% c.i. 2.49 to 2.92), and 3.53 (95% c.i. 3.25 to 3.83), respectively. Overall, patients predicted to require MT had a significantly higher risk of death than those not predicted to require MT, with the strength of this association increasing proportionally with the predicted likelihood of MT.

Discussion

This study used the KTDB, a nationwide trauma registry comprising patients treated at 21 major trauma centres in

Korea, to develop an ML model for predicting the likelihood of MT based on prehospital clinical information. The soft-voting ensemble learning model with RF and AdaBoost exhibited superior predictive performance compared with individual models and other ensemble combinations, achieving an AUROC of 0.837 (discovery) and 0.837 (external validation). Model calibration showed adequate agreement with observed outcomes (calibration-in-the-large (intercept), -0.000 ; calibration slope, 1.008). Calibrated predictions yielded higher net benefit across thresholds in DCA. Model interpretation using SHAP values showed that accident type was the most influential predictor, followed by level of consciousness and circulatory assistance. Furthermore, a higher predicted likelihood of MT was significantly associated with an increased risk of in-hospital mortality. These findings suggest that the proposed ensemble model, leveraging prehospital clinical metadata, may serve as an effective tool for the early identification of patients at high risk of MT and for stratifying in-hospital mortality risk in trauma care settings.

Consistent with the proposed ensemble model, which achieved an AUROC of 0.837 , previous studies^{6,7,10,33,34} have reported mean AUROC values around 0.856 for ML models predicting MT among patients with trauma. However, this comparison may be biased, given that only one study reported substantially higher AUROC values (Qatar; RF, AUROC = 0.950)¹⁰, whereas others reported comparable or lower performance, including studies from Korea (Gradient Boosting, AUROC = 0.839)⁶, the US (LR, AUROC = 0.830 ³³; XGBoost, AUROC = 0.760 ⁷), and China (XGBoost, AUROC = 0.828)³⁴. Moreover, these existing models have several limitations, including focus on the perioperative period⁶, limited sample sizes ($17\,986$ ⁶, 2438 ³³, 805 ³⁴, and $17\,390$ ¹⁰ patients versus $214\,772$ patients in the present study), restriction to specific trauma types (for example, hip fractures)³⁴, inclusion

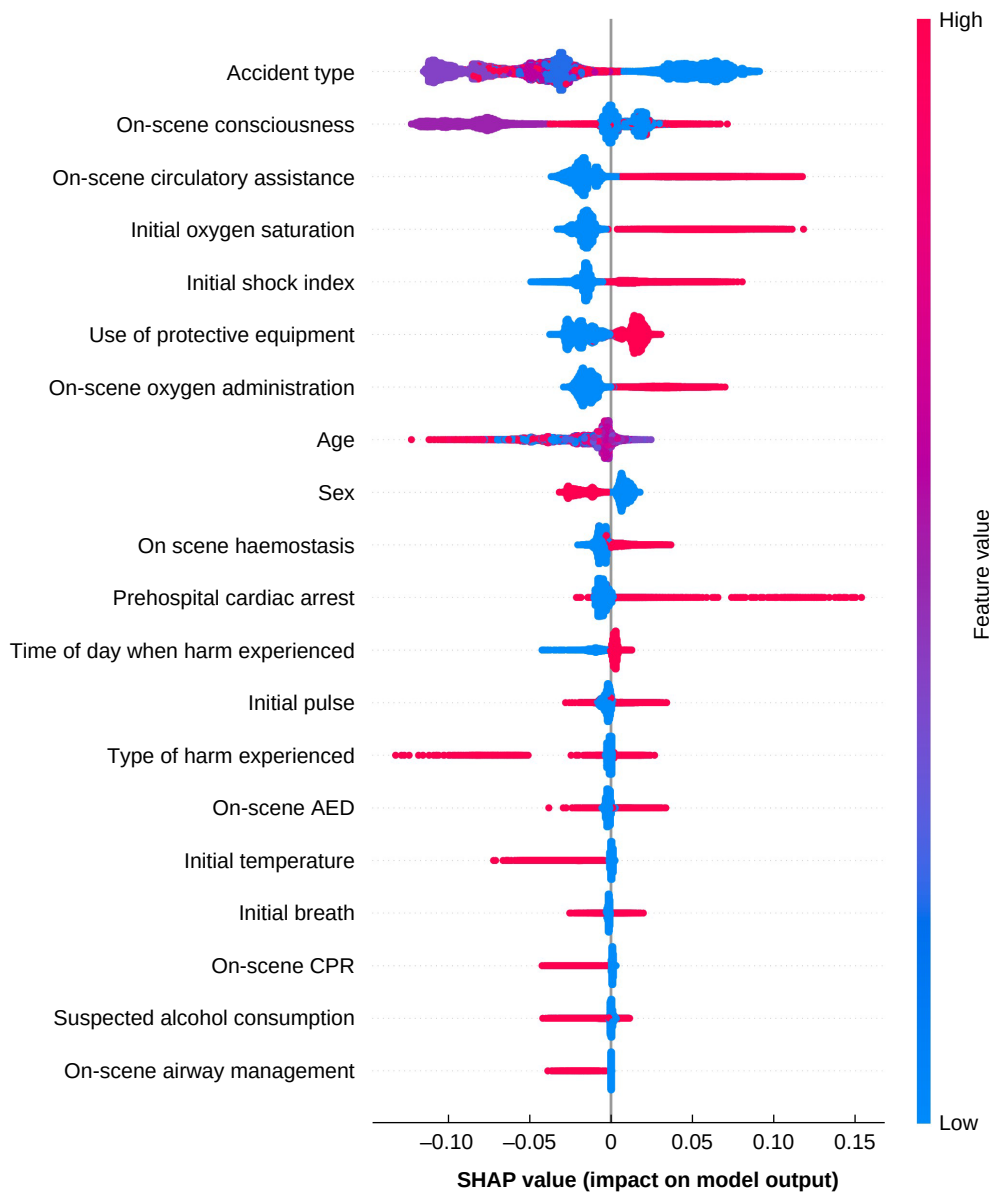


Fig. 2 SHAP values for the proposed ensemble learning model
AED, automated external defibrillator; CPR, cardiopulmonary resuscitation; SHAP, Shapley additive explanations.

Table 2 Risk of in-hospital mortality according to the likelihood of massive transfusion < 0.5 and according to predicted likelihood tertiles in the discovery cohort

Likelihood	Model 1		Model 2	
	OR	P	OR	P
<0.5	1.00 (reference)		1.00 (reference)	
Predicted likelihood tertiles				
T1	2.76 (2.61, 2.92)	<0.001	2.34 (2.16, 2.55)	<0.001
T2	2.28 (2.14, 2.42)	<0.001	2.70 (2.49, 2.92)	<0.001
T3	9.16 (8.77, 9.57)	<0.001	3.53 (3.25, 3.83)	<0.001

Values in parentheses are 95% confidence intervals. Model 1 was adjusted for age and sex. Model 2 was adjusted for age, sex, prehospital cardiac arrest, suspected alcohol consumption, type of accident, type of harm experienced, the use of protective equipment, time of day when harm experienced, initial on-scene consciousness status, initial on-scene vital signs, and initial on-scene emergency treatment. Tertiles (T1–T3) were defined based on the distribution of predicted probabilities among the exposed group, corresponding to the lower (T1), middle (T2), and upper (T3) thirds of predicted risk. OR, odds ratio.

of features obtained after hospital arrival^{33,35}, and reliance on data from a single centre^{6,10,33,34}. However, the present study used KTDB, a large-scale nationwide and multicentre trauma

registry without limitations on trauma type, enabling MT prediction based solely on prehospital clinical information before hospital arrival.

Moreover, SHAP values were calculated to identify the features that most strongly contributed to the prediction of MT among patients with trauma. In contrast, few previous studies have used explainable AI approaches^{10,36}. Notably, the present proposed model identified accident type as the most influential feature, whereas previous studies have reported triage level¹⁰ or a positive focused assessment with sonography for trauma³⁶ as the most important predictors of MT. In addition, whereas previous studies focused solely on predicting MT without incorporating associated outcomes such as mortality^{6,7,33–35}, the present study extends this approach by showing that mortality risk increases progressively with higher model-predicted probabilities of MT.

In clinical settings, massive bleeding represents a critical condition that can rapidly lead to death, making timely preparation and transfusion through MT essential for patient survival³⁷.

However, clinical gestalt, even among experienced clinicians, has limited effectiveness in detecting massive internal haemorrhage³⁸. Notably, delayed transfusion, often due to unrecognized bleeding, is one of the most common causes of preventable death³⁹. From this perspective, the present study offers valuable clinical insights by introducing an ML model that leverages prehospital clinical metadata to predict MT among patients with trauma. Furthermore, although the SI²⁶, TASH score⁵, and ABC score⁴⁰ are widely used global indicators for predicting MT, the present study validated that the ensemble learning model, which incorporates prehospital clinical metadata, achieves superior predictive performance. Among the predictors included in the model, accident type emerged as the most influential feature, consistent with the clinical observation that high-energy mechanisms (for example, traffic accidents, falls from heights, machinery-related injuries) are frequently associated with severe haemorrhage requiring MT, whereas low-energy mechanisms (for example, slips or minor lacerations) are generally associated with a substantially lower risk. It is also important to acknowledge the consequences of false predictions: false positives may waste limited trauma resources, whereas false negatives could delay transfusion and increase preventable deaths. Thus, future use of this model should balance sensitivity and specificity to optimize its clinical utility. With further validation and refinement in real-world environments, this model could be integrated into prehospital clinical workflows and support paramedics in making timely decisions at the scene of trauma.

This study has several limitations that should be considered to avoid misinterpretation of the results. First, the data set used in this study was limited to a single ethnic group (Korean), which may limit the generalizability of the findings to other populations. Furthermore, accident type was classified according to the KTDB system; however, generalizability to trauma systems outside Korea may be limited because categories for mechanism of injury and coding practices vary. In addition, although external validation was undertaken, all validation centres were in Korea, which further limits generalizability to settings with different clinical practices and healthcare infrastructures. Second, statistical analysis of mortality was not conducted in the external validation cohort due to the limited number of events. Nevertheless, findings from the main data set, the KTDB, exhibited a significant association between a higher predicted likelihood of MT and an increased risk of in-hospital mortality. Third, a substantial proportion of patients had missing prehospital vital signs, raising concerns

about potential selection bias and generalizability. To address this issue, multiple imputation by chained equations was used and the entire cohort was reanalysed. By incorporating all available patients, this approach mitigates the impact of missing data and supports the robustness of the reported findings. Fourth, because this was a retrospective study, there is potential for selection and analysis bias. Therefore, future prospective studies are warranted to validate and extend these findings. Fifth, although SI is a key component of widely used tools such as the Advanced Trauma Life Support programme for predicting MT²⁶, the present model identified SI as a less influential predictor. This discrepancy may stem from the retrospective nature of the data or variations in the accuracy of prehospital measurements. Future prospective studies with standardized collection of vital signs are warranted to validate this finding. Sixth, by using both the conventional ≥ 10 units/24 hours and the early-phase ≥ 5 units/4 hours definitions of MT, the present study provides a more comprehensive assessment that reduces survivor bias and better captures clinically relevant early outcomes, although this broadened approach may limit direct comparability with studies using only the 24-hour criterion. Finally, several MT prediction tools (ABC, BATT, and TASH) have been proposed, but they depend on variables often not available prehospital (that is, mechanism and ultrasound for ABC, Glasgow Coma Scale for BATT, haemoglobin or fracture data for TASH). Given these constraints, the SI was used as the most feasible comparator. Although head-to-head comparisons were not possible, the present model complements these tools by integrating prehospital variables via ML and highlights the need for future data sets enabling direct benchmarking.

Using a national data set from multiple trauma centres in Korea, this study highlights the potential of an AI model to predict the likelihood of MT in patients with trauma based on real-time, on-scene prehospital clinical data. By applying an explainable AI approach, key features that contributed to the model's predictions were identified. In addition, a potential association between model probability and an increased risk of in-hospital mortality was observed. With further validation and refinement in real-world settings, this model could be integrated into prehospital clinical workflows to support emergency medical services personnel in making timely and informed decisions at the scene of trauma.

Funding

This research was supported by the Ministry of Science and ICT (RS-2024-00509257 and IITP-2024-RS-2024-00438239) and the Ministry of Health and Welfare (RS-2025-02220492), Korea. The funders had no role in the study design, data collection, analysis, interpretation, or manuscript preparation.

Author contributions

Byungchul Yu (Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Resources, Software, Validation, Visualization, Writing—original draft, Writing—review & editing), Jaehyeong Cho (Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Resources, Software, Validation, Visualization, Writing—original draft, Writing—review & editing), Hyunjee Kim (Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Resources,

Software, Validation, Visualization, Writing—original draft, Writing—review & editing), Seung Ha Hwang (Writing—review & editing), Jiyoung Hwang (Writing—review & editing), Soeun Kim (Writing—review & editing), Jiyeon Oh (Writing—review & editing), Sooji Lee (Writing—review & editing), Do Wan Kim (Writing—review & editing), Junepill Seok (Writing—review & editing), Kyoungwan Kim (Writing—review & editing), Jinseok Lee (Writing—review & editing), Dong Keon Yon (Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Software, Supervision, Validation, Visualization, Writing—original draft, Writing—review & editing), and Wu Seong Kang (Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Resources, Software, Validation, Visualization, Writing—original draft, Writing—review & editing)

Disclosure

The authors declare no conflict of interest.

Supplementary material

Supplementary material is available at [BJS Open](https://academic.oup.com/bjsopen) online.

Data availability

The study protocol and statistical code are available from the corresponding author. The data set is available from the KTDB (discovery data set) through data use agreements. However, the validation data set cannot be shared due to privacy concerns because it originates from internal hospital data. The underlying code for this study (and the training/validation data set) are not publicly available but may be made available to qualified researchers upon reasonable request to the corresponding author.

References

- Mizobata Y. Damage control resuscitation: a practical approach for severely hemorrhagic patients and its effects on trauma surgery. *J Intensive Care* 2017;**5**:4
- Meneses E, Boneva D, McKenney M, Elkbuli A. Massive transfusion protocol in adult trauma population. *Am J Emerg Med* 2020;**38**:2661–2666
- Seo D, Heo I, Park J, Kwon J, Sohn HM, Jung K. Predictors of massive transfusion protocols activation in patients with trauma in Korea: a systematic review. *J Trauma Inj* 2024;**37**:97–105
- El-Menyar A, Goyal P, Tilley BC, Latifi R. The clinical utility of shock index to predict the need for blood transfusion and outcomes in trauma. *J Surg Res* 2018;**227**:52–59
- Mackenzie D, Briatico D, Livingston MH, Beshay T, Robinson T, Farrokhhyar F et al. Elevated international normalized ratio is correlated with large volume transfusion in pediatric trauma patients. *J Pediatr Surg* 2022;**57**:903–907
- Lee SM, Lee G, Kim TK, Le T, Hao J, Jung YM et al. Development and validation of a prediction model for need for massive transfusion during surgery using intraoperative hemodynamic monitoring data. *JAMA Netw Open* 2022;**5**:e2246637
- Nikouline A, Feng J, Rudzicz F, Nathens A, Nolan B. Machine learning in the prediction of massive transfusion in trauma: a retrospective analysis as a proof-of-concept. *Eur J Trauma Emerg Surg* 2024;**50**:1073–1081
- Ageron FX, Coats TJ, Darioli V, Roberts I. Validation of the BATT score for prehospital risk stratification of traumatic haemorrhagic death: usefulness for tranexamic acid treatment criteria. *Scand J Trauma Resusc Emerg Med* 2021;**29**:6
- Rajkomar A, Dean J, Kohane I. Machine learning in medicine. *N Engl J Med* 2019;**380**:1347–1358
- El-Menyar A, Naduvilekandy M, Asim M, Rizoli S, Al-Thani H. Machine learning models predict triage levels, massive transfusion protocol activation, and mortality in trauma utilizing patients hemodynamics on admission. *Comput Biol Med* 2024;**179**:108880
- Collins GS, Moons KGM, Dhiman P, Riley RD, Beam AL, Van Calster B et al. TRIPOD + AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ* 2024;**385**:e078378
- Kim MH, Cho HM, Kim SH, Kim Y, Shin YK, Kim KH. Prevalence of and factors associated with trauma surgeons' referral and patients' willingness to acupuncture treatment after traumatic rib fractures: a single-center cross-sectional study. *Integr Med Res* 2024;**13**:101096
- Im J, Seo EW, Jung K, Kwon J. Understanding regional trauma centers and managing a trauma care system in South Korea: a systematic review. *Ann Surg Treat Res* 2023;**104**:61–70
- Kang WS, Shin IS, Pyo JS, Ahn S, Chung S, Ki YJ et al. Prognostic accuracy of massive transfusion, critical administration threshold, and resuscitation intensity in assessing mortality in traumatic patients with severe hemorrhage: a meta-analysis. *J Korean Med Sci* 2019;**34**:e318
- Holcomb JB, Tilley BC, Baraniuk S, Fox EE, Wade CE, Podbielski JM et al. Transfusion of plasma, platelets, and red blood cells in a 1 : 1 : 1 vs a 1 : 1 : 2 ratio and mortality in patients with severe trauma: the PROPPR randomized clinical trial. *JAMA* 2015;**313**:471–482
- Lee S, Kim DW, Oh NE, Lee H, Park S, Yon DK et al. External validation of an artificial intelligence model using clinical variables, including ICD-10 codes, for predicting in-hospital mortality among trauma patients: a multicenter retrospective cohort study. *Sci Rep* 2025;**15**:1100
- Lee H, Hwang SH, Park S, Choi Y, Lee S, Park J et al. Prediction model for type 2 diabetes mellitus and its association with mortality using machine learning in three independent cohorts from South Korea, Japan, and the UK: a model development and validation study. *EclinicalMedicine* 2025;**80**:103069
- Kim J, Lee H, Lee J, Rhee SY, Shin JI, Lee SW et al. Quantification of identifying cognitive impairment using olfactory-stimulated functional near-infrared spectroscopy with machine learning: a post hoc analysis of a diagnostic trial and validation of an external additional trial. *Alzheimers Res Ther* 2023;**15**:127
- Hwang SH, Lee H, Lee JH, Lee M, Koyanagi A, Smith L et al. Machine learning-based prediction for incident hypertension based on regular health checkup data: derivation and validation in 2 independent nationwide cohorts in South Korea and Japan. *J Med Internet Res* 2024;**26**:e52794
- Yang L, Shami A. On hyperparameter optimization of machine learning algorithms: theory and practice. *Neurocomputing* 2020;**415**:295–316
- Mohammed A, Kora R. A comprehensive review on ensemble deep learning: opportunities and challenges. *J King Saud Univ Comput Inf Sci* 2023;**35**:757–774
- Rodríguez JD, Pérez A, Lozano JA. Sensitivity analysis of kappa-fold cross validation in prediction error estimation. *IEEE Trans Pattern Anal Mach Intell* 2010;**32**:569–575

23. Sang H, Lee H, Park J, Kim S, Woo HG, Koyanagi A et al. Machine learning-based prediction of neurodegenerative disease in patients with type 2 diabetes by derivation and validation in 2 independent Korean cohorts: model development and validation study. *J Med Internet Res* 2024;**26**:e56922
24. Kim S, Kim H, Kim S, Lee H, Hammoodi A, Choi Y et al. Machine learning-based prediction of substance use in adolescents in three independent worldwide cohorts: algorithm development and validation study. *J Med Internet Res* 2025;**27**:e62805
25. Yang J, Li Y, Li X, Tao S, Zhang Y, Chen T et al. A machine learning model for predicting in-hospital mortality in Chinese patients with ST-segment elevation myocardial infarction: findings from the China myocardial infarction registry. *J Med Internet Res* 2024;**26**:e50067
26. Carsetti A, Antolini R, Casarotta E, Damiani E, Gasparri F, Marini B et al. Shock index as predictor of massive transfusion and mortality in patients with trauma: a systematic review and meta-analysis. *Crit Care* 2023;**27**:85
27. Sau A, Pastika L, Sieliwonzcyk E, Patlatzoglou K, Ribeiro AH, McGurk KA et al. Artificial intelligence-enabled electrocardiogram for mortality and cardiovascular risk estimation: a model development and validation study. *Lancet Digit Health* 2024;**6**:e791–e802
28. Lundberg SM, Erion G, Chen H, DeGrave A, Prutkin JM, Nair B et al. From local explanations to global understanding with explainable AI for trees. *Nat Mach Intell* 2020;**2**:56–67
29. Huang Y, Li W, Macheret F, Gabriel RA, Ohno-Machado L. A tutorial on calibration measurements and calibration models for clinical prediction models. *J Am Med Inform Assoc* 2020;**27**:621–633
30. Van Calster B, Nieboer D, Vergouwe Y, Cock D, Pencina B, Steyerberg MJ et al. A calibration hierarchy for risk models was defined: from utopia to empirical data. *J Clin Epidemiol* 2016;**74**:167–176
31. Vickers AJ, van Calster B, Steyerberg EW. A simple, step-by-step guide to interpreting decision curve analysis. *Diagn Progn Res* 2019;**3**:1–8
32. Kim H, Son Y, Lee H, Kang J, Hammoodi A, Choi Y et al. Machine learning-based prediction of suicidal thinking in adolescents by derivation and validation in 3 independent worldwide cohorts: algorithm development and validation study. *J Med Internet Res* 2024;**26**:e55913
33. Strickland M, Nguyen A, Wu S, Suen SC, Mu Y, Del Rio Cuervo J et al. Assessment of machine learning methods to predict massive blood transfusion in trauma. *World J Surg* 2023;**47**:2340–2346
34. Guo J, He Q, Li Y. Development and validation of machine learning models to predict perioperative transfusion risk for hip fractures in the elderly. *Ann Med* 2024;**56**:2357225
35. Akaraborworn O, Siribumrungwong B, Sangthong B, Thongkhao K. Massive blood transfusion for trauma score to predict massive blood transfusion in trauma. *Crit Care Res Pract* 2021;**2021**:3165390
36. Benjamin AJ, Young AJ, Holcomb JB, Fox EE, Wade CE, Meador C et al. Early prediction of massive transfusion for patients with traumatic hemorrhage: development of a multivariable machine learning model. *Ann Surg Open* 2023;**4**:e314
37. Rossaint R, Afshari A, Bouillon B, Cerny V, Cimpoesu D, Curry N et al. The European guideline on management of major bleeding and coagulopathy following trauma: sixth edition. *Crit Care* 2023;**27**:80
38. Pommerening MJ, Goodman MD, Holcomb JB, Wade CE, Fox EE, Del Junco DJ et al. Clinical gestalt and the prediction of massive transfusion after trauma. *Injury* 2015;**46**:807–813
39. Kwon J, Lee M, Kim Y, Moon J, Huh Y, Song S et al. Trauma system establishment and outcome improvement: a retrospective national cohort study in South Korea. *Int J Surg* 2023;**109**:2293–2302
40. Hanlin E, Meyer DE, Heft N, Stilgenbauer H, Cotton B, Bourgeois M et al. Prehospital validation of the Assessment of Blood Consumption (ABC) score. *Prehosp Emerg Care* 2024;**28**:495–500