

Accurate, fair, and generalisable scaling of injury severity score-based AI with demographics in terms of mortality in patients with trauma: multi-centre, multi-national retrospective cohort study



Yunjeong Choi,^{a,n} Junepill Seok,^{b,c,n} Thomas Young-Chul Oh,^d Jeremy Hsu,^d Do Wan Kim,^e Byungchul Yu,^{f,g} Jayun Cho,^g Woocheol Jang,^a Jina Kim,^a Na-eun Oh,^a Jehyeuk Ahn,^a Robert J. Femia,^h Paul A. Testa,^h Dong Keon Yon,^{i,j} Daniel K. Sodickson,^{k,l} Wu Seong Kang,^{f,g,m,o,*} and Jinseok Lee^{a,**,o}



^aDepartment of Biomedical Engineering, Kyung Hee University, Yongin, Republic of Korea

^bDepartment of Thoracic and Cardiovascular Surgery, Chungbuk National University Hospital, Cheongju, Republic of Korea

^cDepartment of Thoracic and Cardiovascular Surgery, Chungbuk National University, College of Medicine, Cheongju, Republic of Korea

^dTrauma Service, Westmead Hospital, University of Sydney, NSW, Australia

^eDepartment of Thoracic and Cardiovascular Surgery, Chonnam National University Hospital and Chonnam National University Medical School, Gwangju, Republic of Korea

^fDepartment of Traumatology, Gachon University College of Medicine, Incheon, Republic of Korea

^gDepartment of Trauma Surgery, Gachon University Gil Medical Center, Incheon, Republic of Korea

^hRonald O. Perelman Department of Emergency Medicine, New York University Grossman School of Medicine, New York, NY, USA

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

^jDepartment of Precision Medicine, Kyung Hee University College of Medicine, Seoul, South Korea

^kDepartment of Radiology, Bernard and Irene Schwartz Centre for Biomedical Imaging, New York University Grossman School of Medicine, New York, NY, USA

^lCentre for Advanced Imaging Innovation and Research (CAI2R), Department of Radiology, New York University Grossman School of Medicine, New York, NY, USA

^mDepartment of Trauma Surgery, Jeju Regional Trauma Centre, Cheju Halla General Hospital, Jeju, Republic of Korea

Summary

Background Accurate and equitable prediction of trauma-related in-hospital mortality is critical for guiding clinical decisions and optimising trauma care resources. Traditional severity scoring systems like the Injury Severity Score (ISS) do not account for demographic factors, potentially limiting their fairness and generalisability across diverse populations.

Methods We developed and externally validated an artificial intelligence (AI) model based on ISS and integrated demographic features (age and sex) to predict in-hospital mortality after trauma. Data from the Korean Trauma Data Bank were used for model development and internal validation, comprising 121,418 patients with trauma aged ≥ 15 years treated at 19 trauma centres in South Korea (2017–2022). External validation was performed on an independent cohort of 7458 patients from five trauma centres (four in South Korea and one in Australia, 2022–2024). The primary outcome was trauma-related in-hospital mortality. Predictive performance was assessed using area under the receiver operating characteristic curve (AUROC), sensitivity, specificity, accuracy, and balanced accuracy. Fairness was evaluated by comparing AUROC differences across age (<65 vs ≥ 65 years) and sex (female vs male) subgroups.

Findings The ISS-based AI model incorporating age and sex achieved high predictive performance (internal validation AUROC, 0.934; external validation AUROC range, 0.901–0.920), outperforming conventional ISS-based methods. The model also demonstrated improved fairness, showing reduced AUROC differences across subgroups (age: 0.068 vs 0.091; sex: 0.021 vs 0.046 for AI model vs ISS, respectively).

eBioMedicine
2026;126: 106206

Published Online xxx
<https://doi.org/10.1016/j.ebiom.2026.106206>

*Corresponding author. Department of Traumatology, Gachon University College of Medicine; Department of Trauma Surgery, Gachon University Gil Medical Center, Incheon, Republic of Korea.

**Corresponding author. Department of Biomedical Engineering, Kyung Hee University, 446-701 Electronic Information College building, Kyunghee Univ, Global Campus, Seocheon-dong, Giheung-gu, Yongin-si, Gyeonggi, Republic of Korea.

E-mail addresses: wuseongkang@gachon.ac.kr (W.S. Kang), gonasago@khu.ac.kr (J. Lee).

[†]These authors contributed equally to this work and should be considered co-first authors.

[‡]WSK and JL contributed equally and should be considered as corresponding authors.

Interpretation Scaling an ISS-based AI model through demographic integration yielded accurate, fair, and generalisable predictions of trauma-related in-hospital mortality. This approach may enhance trauma care decision-making and enable more equitable resource allocation across diverse clinical settings.

Funding This research was supported by the MSIT (Ministry of Science and ICT), Korea, under the ITRC (Information Technology Research Center) support program (IITP-2025-RS-2024-00438239) and the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (RS-2024-00509257, Global AI Frontier Lab). In addition, this research was supported by the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health & Welfare, Republic of Korea (RS-2025-02220492).

Copyright © 2026 The Author(s). Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

Keywords: Artificial intelligence; Injury severity score; Trauma; Mortality; Demographic factors; Prediction model; Fairness; Generalisability

Research in context

Evidence before this study

Accurate prediction of in-hospital mortality in patients with trauma is essential for guiding clinical decision-making and allocating limited healthcare resources. The Injury Severity Score (ISS) has long been used as the standard for anatomical injury severity quantification, but it is limited by its overly simplified structure and lack of demographic considerations. Prior artificial intelligence (AI) models using raw AIS codes (Region-1999) or conventional groupings (Region-6) offered either high complexity or low resolution. Although several machine learning models have been developed to improve trauma prognostication, many lacked scalability, fairness evaluation, or external validation across different populations. We searched PubMed and Embase up to May 2025 using terms such as “trauma”, “mortality prediction”, “machine learning”, “artificial intelligence”, “deep learning”, “injury severity score”, and “demographic fairness”. Few studies have integrated patient demographic factors into anatomical injury-based AI models and validated them using large-scale, real-world, multinational datasets.

Added value of this study

This is the large-scale, multinational study to demonstrate that integrating demographic variables (age and sex) into a Region-46-based AI model significantly improves predictive accuracy and fairness for trauma-related in-hospital mortality. The model was trained on over 120,000 patients

with trauma from South Korea and externally validated on 7458 patients across five independent centres in South Korea and Australia. The proposed deep neural network (DNN) model achieved superior performance (internal AUROC: 0.934; external AUROC: 0.901–0.920) compared to conventional ISS-based models and alternative anatomical encodings. It also demonstrated lower prediction bias across age and sex subgroups and showed generalisability across healthcare systems. The study provides rigorous validation of a demographically scaled, anatomically detailed AI model for trauma prognosis.

Implications of all the available evidence

This study supports a paradigm shift from traditional ISS-based stratification to AI-driven mortality prediction models that incorporate both anatomical detail and patient-specific demographic features. By using a balanced anatomical representation (Region-46) and adding age and sex, the proposed model achieves high accuracy, fairness, and external validity. These findings suggest that such models can be incorporated into trauma workflows and healthcare data integration platforms to support early prognostication and optimise clinical decision-making across diverse healthcare environments. This approach may enhance equitable trauma care and facilitate data-driven triage across local, national, and global trauma systems.

Introduction

Trauma is a leading cause of death and disability worldwide, particularly among young and middle-aged adults.¹ In-hospital mortality represents a critical outcome reflecting both injury severity and the quality of acute clinical care.² Accurate and timely prognosis prediction enables optimal resource allocation and informed clinical decisions, potentially improving patient outcomes.²

To date, the Injury Severity Score (ISS), derived from the Abbreviated Injury Scale (AIS), has been the standard method for quantifying anatomical injury burden.^{3,4} However, ISS aggregates detailed AIS injury codes into only six broad anatomical regions (Region-6),⁵ equally weighting injuries that have substantially different prognostic implications. For example, ISS assigns equivalent weight to a pancreatic injury and a rib fracture if they share the same AIS severity score, despite

their markedly different clinical consequences.^{6,7} At the other extreme, directly using the original AIS injury descriptors (Region-1999)—which individually encode each injury—preserves extensive detail but introduces significant computational complexity and risks of model overfitting.⁸

To overcome these limitations, our group previously developed Region-46, a balanced anatomical encoding scheme transforming the original AIS injury descriptors into 46 clinically meaningful anatomical categories.⁸ By providing a middle ground between the excessive simplicity of Region-6 and the computational complexity of Region-1999, Region-46 has shown potential as a superior set of input features for AI-based mortality prediction models.⁸ Although initial AI models using Region-46 demonstrated improved predictive accuracy, our previous work was limited by a relatively small dataset ($n = 37,762$) and lacked comprehensive external validation.⁸

The current study, therefore, builds upon this foundation by systematically addressing five key research questions. First, does the predictive advantage of the Region-46–based AI model over Region-6– and Region-1999–based models, and conventional ISS-based approaches persist when trained on a substantially larger nationwide dataset ($>120,000$ patients)? Second, does this substantial increase in training data significantly enhance the predictive accuracy of the Region-46–based AI model compared with our previous smaller-scale dataset?⁸ Third, does incorporating demographic variables (patient age and sex) further improve predictive accuracy beyond anatomical injury features alone? Fourth, does the Region-46–based AI model demonstrate fair (unbiased) performance across demographic subgroups?^{9,10} Fifth, does the predictive performance of the Region-46–based AI model generalise effectively across diverse clinical settings?

By explicitly addressing these critical questions, this study aims to establish an accurate, robust, generalisable, and fair AI-driven framework for trauma prognosis prediction, ultimately contributing to improved clinical decision-making and trauma care outcomes.

Methods

Data source and ethical considerations

This cohort study adhered to the TRIPOD + AI (Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis extended for Artificial Intelligence) guidelines¹¹ for model development and validation, with the completed checklist provided in [Table S1](#).

For model development and internal validation, data were obtained from the Korean Trauma Data Bank (KTDB), a nationwide registry that includes all 17 regional trauma centres—representing the highest level of trauma care in South Korea—and two affiliated educational trauma centres. The KTDB dataset included

patients with trauma registered between 2017 and 2022 (dataset identification number: K2024-03-2-02-29).

To assess generalisability, external validation was performed using five independent datasets collected in 2023 from four regional trauma centres in South Korea—Chonnam National University Hospital (CNUH), Cheju Halla General Hospital (CHGH), Chungbuk National University Hospital (CNUH), and Gachon University Gil Medical centre (GUGC)—and from one Level I trauma centre in Australia, Westmead Hospital (WH), with data collected between 2022 and 2024.

Ethical approval for the use of KTDB data was granted by the Institutional Review Board of the National Emergency Medical Centre. The requirement for informed consent was waived due to the retrospective nature of the study and the use of deidentified data. External validation using independent datasets was approved by the IRBs of each participating institution: CHGH (CHH-2024-L03), CNUH (CBNUH 2024-10-017), GUGC (GFIRB2024-357), CNUH (CNUH-2024-016), and WH (2024/PID0326). Notably, the approval from CHGH encompassed both the analysis of the national KTDB dataset and the clinical validation data. All study procedures were conducted in accordance with applicable national regulations and institutional ethical standards. The requirement for informed consent was waived in accordance with relevant ethical guidelines for secondary use of administrative data.

Patient selection

Patients from both the KTDB dataset and the external validation cohorts were screened using identical inclusion and exclusion criteria. Eligible patients were those aged 15 years or older who were admitted to the hospital following a traumatic injury. Patients were excluded if they died in the emergency department prior to admission, were discharged home without hospitalisation, were transferred to another facility without hospital admission, had duplicate episodes, or had cardiopulmonary resuscitation (CPR) or do-not-resuscitate (DNR) orders denied at the initial site of care.

Additionally, patients with missing key variables (age, sex, or AIS-based injury codes) were excluded to ensure completeness of the model inputs. Cases lacking either injury or discharge timestamps were also removed because these variables were essential to confirm that each record represented a valid hospitalisation episode following trauma. These exclusions were applied to maintain data integrity and ensure reliable outcome labelling. In the development and internal validation cohort ($n = 228,116$ before exclusion), 540 (0.2%) missing age, 3262 (1.4%) missing AIS codes, 1376 (0.6%) missing injury time, and 1156 (0.5%) missing discharge time were excluded. In the external validation cohort ($n = 12,473$ before exclusion), 5 ($<0.1\%$) missing age, 6 ($<0.1\%$) missing sex, 359 (2.9%) missing AIS codes, 422 (3.4%) missing injury

time, and 797 (6.4%) missing discharge time were excluded. No imputation techniques were applied, and all analyses were performed using complete cases only. Biological sex was recorded at the time of admission based on official electronic health records and national trauma registry data.

After applying these criteria, a total of 121,418 adult patients with trauma were included from the KTDB for model development and internal validation. An additional 7458 patients were included in the external validation dataset, comprising 672 from CNUH, 1773 from CHGH, 721 from CBUH, 1203 from GUGC in South Korea, and 3089 from WH in Australia. The overall patient selection process is illustrated in Fig. 1.

Key questions: data scaling, demographic integration, fairness, and generalisability

Complete lists of injury descriptors and anatomical categories used as input features for the Region-1999-, Region-6-, and Region-46-based AI models are provided in Table S2.

Specifically, the Region-1999-based AI model used binary (one-hot) encoding of all 1999 individual AIS injury descriptors, resulting in highly detailed but sparse anatomical data. In contrast, the Region-6-based AI model aggregated these detailed descriptors into six broad anatomical regions (head/neck, face, thorax, abdomen, extremities, and external), substantially reducing data dimensionality at the potential cost of clinically meaningful anatomical detail. The Region-46-based AI model balanced these extremes by consolidating AIS descriptors into 46 anatomically coherent and clinically relevant categories, thus aiming to optimise both predictive accuracy and computational efficiency. We introduced a scheme (Region-46) by regrouping AIS codes into 46 organ-specific categories

to better reflect organ-level severity. Codes associated with frequently injured or clinically significant organs were aggregated under named organ categories, with the highest AIS score retained per organ. Less frequently encountered or clinically less relevant injuries were retained within their broader anatomical region. The full Region-46 classification is presented in Table S3. For example, in the head region, AIS codes were subdivided into intracranial haemorrhage, subdural haemorrhage, and epidural haemorrhage due to their clinical relevance, while less critical injuries were grouped as “head” or “skull.” Facial injuries lacked severe or high-frequency codes and were thus grouped under “face.” In the neck, codes were categorised by structure—vascular, oesophageal, tracheal, and cervical spine—while other injuries remained under “neck.” Thoracic injuries were assigned to specific categories such as thoracic vessels, bronchus, diaphragm, heart, lung, and ribs, while remaining codes were labelled as “thorax.” Abdominal injuries were assigned to organs including the liver, spleen, pancreas, stomach, and kidneys, among others, with less frequent injuries classified under “abdomen.” Upper extremity injuries were grouped together since all had AIS scores below 4 and were generally less severe. For the lower extremities, the pelvis was treated as a distinct category due to its clinical significance, while other injuries remained under “lower extremity.” Finally, injuries coded under the external region were grouped as “external” due to the absence of high-severity codes.

The primary outcome was in-hospital mortality. This study addressed five clinically relevant and AI-related critical questions, focussing on AIS-based prediction of in-hospital mortality. First, to determine whether the Region-46-based AI model maintained its predictive advantage over Region-6-based, Region-1999-based,

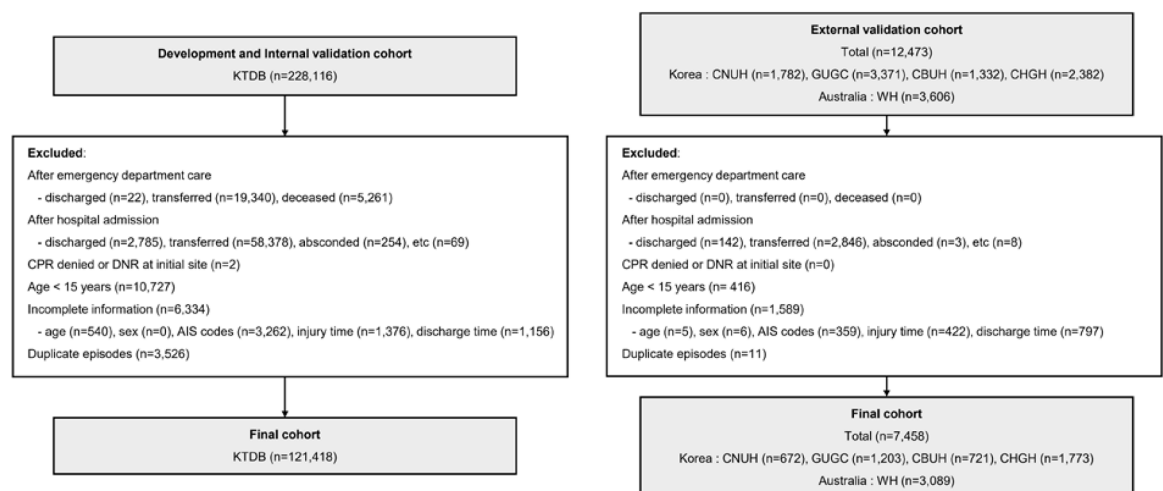


Fig. 1: Patient Selection Flowchart for KTDB and external validation datasets.

and ISS-based models when applied to a larger dataset, we trained and comparatively evaluated each model using the expanded KTDB cohort of 121,418 patients with trauma. For comparison, mortality prediction based on the ISS was performed using both a rule-based approach (labelling mortality by ISS thresholds) and a logistic regression model that combined the ISS score with age and sex. Second, to quantify predictive performance gains resulting from an increased dataset size, we compared the current Region-46-based AI model (trained on 121,418 patients; 2017–2022) directly against our previously published model (trained on 37,762 patients; 2014–2019). Third, to evaluate whether the addition of demographic variables (patient age and sex) provided incremental predictive value beyond anatomical injury features alone, we trained two variants of the Region-46-based AI model—one including and one excluding demographic variables—and compared their performance. Fourth, to assess the fairness of predictions across demographic subgroups defined by age (<65 vs ≥65 years) and sex (female vs male), we conducted subgroup analyses. Fairness was quantified by comparing differences in area under the receiver operating characteristic curve (AUROC) across these demographic categories between the Region-46-based AI model (with demographic inputs) and conventional ISS-based methods. For the ISS-based model, the optimal classification threshold was determined using Youden's J statistic,¹² and subgroup fairness was evaluated based on this threshold. Additionally, we evaluated the Equalised Odds (EO) difference as a fairness metric.¹³ This measure quantifies disparities in false positive rates (FPRs) and false negative rates (FNRs) between demographic groups, where smaller differences indicate higher fairness.¹³ Fifth, to assess model generalisability, we performed external validation on independent cohorts from four South Korean trauma centres and one Australian Level I trauma centre across all comparative analyses.

Machine learning model development and evaluation

We included 121,418 patients with trauma from the KTDB (113,433 survivors [93%], 7985 non-survivors [7%]), divided into training (80%; $n = 97,134$) and testing (20%; $n = 24,284$) cohorts. For model development, the KTDB dataset was stratified by trauma centre and survival status, then randomly divided into training and testing cohorts in an 8:2 ratio.¹⁴ To address class imbalance, the Synthetic Minority Over-sampling Technique (SMOTE) was applied to the training set to generate synthetic samples of the minority (deceased) class.¹⁵ The testing cohort was reserved exclusively for internal evaluation. External validation was conducted using independent datasets that were not involved in model training or internal testing. These data were used

solely to evaluate out-of-sample performance and assess the model's generalisability across institutions and geographic settings, in accordance with best practices in prediction modelling research.¹⁶ External validation comprised 7458 patients from five centres with full cohort details in [Table S4](#). To further assess model generalisability, centre-specific performance metrics were estimated with 2000 bootstrap iterations to obtain 95% confidence intervals.¹⁷

Seven machine learning algorithms were evaluated: adaptive boosting (AdaBoost), random forest (RF), gradient boosting machine (GBM), light gradient boosting machine (LightGBM), extreme gradient boosting (XGBoost), logistic regression (LR), and deep neural network (DNN).^{18–23} Hyperparameters were optimised for each model, with predictive performance assessed via five-fold cross-validation.²⁴ The optimal AI model was selected primarily based on balanced accuracy and AUROC, given the pronounced class imbalance. Model performance was further evaluated using the area under the receiver operating characteristic curve (AUROC), area under the precision–recall curve (AUPRC), sensitivity, specificity, accuracy, precision, and balanced accuracy. Balanced accuracy and AUROC served as the main selection criteria, while the remaining metrics provided complementary insights into classification performance. AUROC and AUPRC quantified overall discriminative ability, particularly under imbalanced outcome distributions. Sensitivity (true positive rate) and specificity (true negative rate) measured performance in identifying positive and negative outcomes, respectively. Accuracy represented the proportion of correctly classified cases, and precision indicated the proportion of true positives among predicted positives. To statistically compare model performance, bootstrap hypothesis testing (2000 iterations) and DeLong's test were performed to evaluate whether the area under the receiver operating characteristic curve (AUROC) of the best-performing model significantly differed from those of the other models.^{25,26} The bootstrap procedure generated empirical distributions of AUROC differences to estimate p -values, while DeLong's test provided analytic p -values. A two-sided p value < 0.05 was considered statistically significant.

To enhance model interpretability, SHapley Additive exPlanations (SHAP) analysis was conducted to quantify the contribution of each input variable to the DNN model's mortality predictions.²⁷ SHAP values were computed using the KernelExplainer algorithm with a k -means summarised background of the training data. The target dataset for explanation consisted of 2000 randomly selected samples from the test set. For each prediction, SHAP values were derived to estimate how individual features—46 anatomical indicators (Region-46 framework) and two demographic variables (age and sex)—influenced the predicted probability of in-hospital mortality.

Calibration and decision curve analysis

To assess model calibration and clinical utility, three post-hoc calibration methods—Platt scaling, Beta calibration, and isotonic regression—were applied using an out-of-fold strategy based on cross-validated training predictions.²⁸ Each method was fitted on the out-of-fold predicted probabilities from the training data and subsequently evaluated on both the internal test set and the external validation cohorts. Calibration curves were generated to assess the agreement between predicted and observed probabilities, and decision curve analysis (DCA) was performed to evaluate the net clinical benefit across a range of threshold probabilities.²⁹ All analyses were conducted to ensure consistent calibration and decision performance across datasets.

Association between predicted risk and in-hospital mortality

Cox proportional hazards regression analyses were conducted to evaluate the association between AI model-predicted mortality risk and observed in-hospital mortality, using both unadjusted and adjusted models (for age and sex).³⁰ Patients predicted to be deceased were stratified into tertiles according to predicted risk: T1 (lowest), T2 (intermediate), and T3 (highest). Hazard ratios (HRs) and 95% confidence intervals (CIs) were calculated for each tertile. Adjusted models provided insight into the independent predictive value of AI-derived risk categories, controlling for demographic factors.

Statistical Analysis

Survivor and non-survivor groups were compared using χ^2 tests, Fisher's exact test, Mann–Whitney U tests, or independent t-tests, as appropriate. Cox proportional hazards regression models assessed associations between predicted risk and observed mortality, adjusting for age and sex. Normality of continuous variables was assessed using the Kolmogorov–Smirnov test and visual inspection of histograms and Q–Q plots. Homogeneity of variance was evaluated using Levene's test. Based on these assessments, independent t-tests were used for normally distributed variables with equal variances, while the Mann–Whitney U test was applied for non-normally distributed data. Analyses used Python 3.10, with significance at two-sided $p < 0.05$.

Role of funders

The funder had no role in study design, data collection, analysis and interpretation, or the writing of the manuscript.

Results

Study cohort characteristics and dataset composition

Demographic characteristics and AIS scores for the KTDB cohort and aggregated external validation cohorts

are summarised in [Table S5](#). Overall, deceased patients were significantly older and exhibited higher AIS scores across all anatomical regions compared to survivors in both the train/test and external validation cohorts (all $p < 0.001$ [Mann–Whitney U test]). While male sex was significantly associated with mortality in the train/test cohort ($p < 0.001$ [χ^2 test]), this association was not observed in the external validation cohort ($p = 0.140$ [χ^2 test]). A total of 121,418 patients were included in the train-test dataset and 7458 patients in the external validation dataset. The overall in-hospital mortality rate was 7% in the train-test cohort and 6% in the external validation cohort. Median age was 57 in KTDB and 52 in external validation cohort. Median ISS was 8 in KTDB, and 4 in external validation cohort. Male was 65% in KTDB, 35% in external validation cohort. Institution-specific demographic characteristics and detailed AIS scores for external validation cohorts are provided in [Table S6](#).

Region-46-based AI model with age and sex vs. conventional ISS

The predictive performances of seven machine learning models trained using Region-46 anatomical features combined with demographic variables (age and sex), as well as two conventional ISS thresholds (≥ 16 and ≥ 25), is summarised in [Table S7](#). Across all evaluation datasets—training (five-fold cross-validation), internal testing, and external validation—the DNN consistently exhibited the highest predictive accuracy. The architecture of the proposed DNN is summarised in [Figure S1](#). The final DNN model included 48 input features (46 from Region-46 and two demographic variables), and consisted of two fully connected hidden layers (64 and 32 nodes) with LeakyReLU activation, dropout, and L2 regularisation. During five-fold cross-validation, the Region-46-based DNN model incorporating age and sex achieved superior AUROC (0.930 [SD, 0.005]) and balanced accuracy (85.4% [SD, 0.006]) compared with other algorithms. This performance advantage was maintained on the internal KTDB test set (AUROC, 0.934; balanced accuracy, 86.0%) ([Figure S2](#)). External validation cohorts further confirmed the robustness of the DNN model, with AUROCs of 0.920 (Korea) and 0.902 (Australia), consistently outperforming all ISS-based approaches.

Region-46 vs. Region-6 vs. Region-1999 DNN models with age and sex

We conducted additional cross-validation to identify the optimal algorithms and hyperparameters for Region-6- and Region-1999-based AI models combined with demographic variables, confirming that the DNN consistently demonstrated superior performance across these configurations. Predictive results from these models are summarised in [Table S8](#). During five-fold cross-validation, the Region-46-based DNN model incorporating

age and sex demonstrated superior predictive accuracy (AUROC, 0.930 [SD, 0.005]; balanced accuracy, 85.4% [SD, 0.006]) compared with the Region-6-based (AUROC, 0.922; balanced accuracy, 84.8%) and Region-1999-based models (AUROC, 0.896; balanced accuracy, 81.1%). This advantage persisted in the internal test dataset ([Figure S3](#)), with the Region-46-based model again showing superior results (AUROC, 0.934; balanced accuracy, 86.0%). Similar trends were observed in external validation cohorts, where the Region-46-based DNN consistently outperformed both Region-6 and Region-1999 models. Specifically, in the Korean external validation set, the Region-46-based model achieved an AUROC of 0.920 (balanced accuracy, 83.7%), surpassing Region-1999 (AUROC, 0.905; balanced accuracy, 82.3%) and Region-6 (AUROC, 0.901; balanced accuracy, 82.3%). Likewise, in the Australian cohort, the Region-46-based model showed higher performance (AUROC, 0.902; balanced accuracy, 81.5%) compared to Region-6 (AUROC, 0.893) and Region-1999 (AUROC, 0.872).

Impact of dataset expansion and demographic variable inclusion

Predictive performance of the current Region-46-based DNN model was compared with our previously published model,⁸ which was trained on a smaller dataset ($n = 37,762$) without age and sex ([Table S9](#) and [Figure S4](#)). We first evaluated the impact of expanding the training dataset from 37,762 patients to 121,418 patients while excluding demographic variables (age and sex). This dataset expansion alone improved overall predictive performance, increasing AUROC from 0.881 to 0.897 and balanced accuracy from 80.5% to 81.8%. Subsequently, we assessed the effect of incorporating age and sex into the larger dataset ($n = 121,418$), which further enhanced predictive accuracy, raising AUROC from 0.897 to 0.914 and balanced accuracy from 81.8% to 83.0%. Institution-specific results confirmed this pattern: at GUGC, AUROC improved from 0.904 (smaller dataset) to 0.919 (larger dataset), and further to 0.928 after adding demographic variables. Similarly, at WH, AUROC increased from 0.824 (smaller dataset) to 0.861 (larger dataset), and further rose to 0.902 with the addition of age and sex. These analyses demonstrate distinct and cumulative improvements from dataset expansion and the inclusion of demographic variables.

Model calibration and clinical utility

Calibration analyses demonstrated that all three post-hoc methods improved the reliability of predicted probabilities compared with the uncalibrated model. Among them, isotonic regression achieved the most consistent and well-aligned calibration across both internal and external datasets. In contrast, Platt and Beta calibration exhibited mild residual or unstable patterns. These results suggest that the non-parametric nature of

isotonic regression allowed more flexible correction of the DNN's overconfident probability outputs, particularly in the context of a highly imbalanced dataset (survival:death $\approx 93:7$).^{31,32}

Decision curve analysis (DCA) performed on both the internal test and external validation datasets further confirmed the clinical advantage of calibration. The isotonic-calibrated Region-46 DNN model consistently provided higher net clinical benefit across a wide range of threshold probabilities compared with the uncalibrated version. This finding demonstrates clinical applicability for mortality risk assessment. Collectively, these results indicate that post-hoc calibration substantially enhanced both the predictive reliability and the clinical utility of the model.

To visually illustrate these improvements, [Fig. 2](#) present the calibration curves of the uncalibrated and isotonic-calibrated DNN models in the internal test and external validation sets, respectively. [Fig. 3](#) shows the DCA results comparing the calibrated and uncalibrated DNN models using the internal test and overall external validation datasets. Furthermore, [Fig. 4](#) provides the DCA results for each of the five individual external validation centres, highlighting the consistent clinical benefit of calibration across diverse institutional cohorts. The detailed performance metrics of the calibrated models are summarised in [Tables 1–6](#). The calibrated DNN demonstrated strong generalisability across different hospitals. In the overall external validation, the model achieved an AUROC of 0.920 for the Korean cohort and 0.901 for the Australian cohort. When analysed by individual centres, the model maintained stable discrimination and classification balance, with balanced accuracy ranging from 74.6% to 89.1% and AUROC ranging from 0.819 to 0.961 ([Table 6](#)). These results indicate that the model's performance remained consistent across trauma centres of varying sizes and regional characteristics, supporting its reliability and applicability in diverse clinical settings.

Model interpretability analysis

We performed a comprehensive interpretability analysis using SHAP values to elucidate the contribution of individual features to the DNN's mortality prediction. [Fig. 5](#) depicts the overall feature-importance distribution, while [Table S11](#) lists detailed SHAP statistics for each variable. The five most influential predictors were AIS (cerebrum subdural), age, AIS (head), AIS (lower extremity), and AIS (pelvis). Notably, these factors are consistent with clinically recognised determinants of trauma-related mortality, thereby reinforcing the clinical plausibility of the DNN's learnt feature representations.

Model fairness evaluation by age and sex subgroups

Model fairness was assessed by comparing differences in AUROC between age (<65 vs ≥ 65 years) and sex (female vs male) subgroups for the Region-46-based

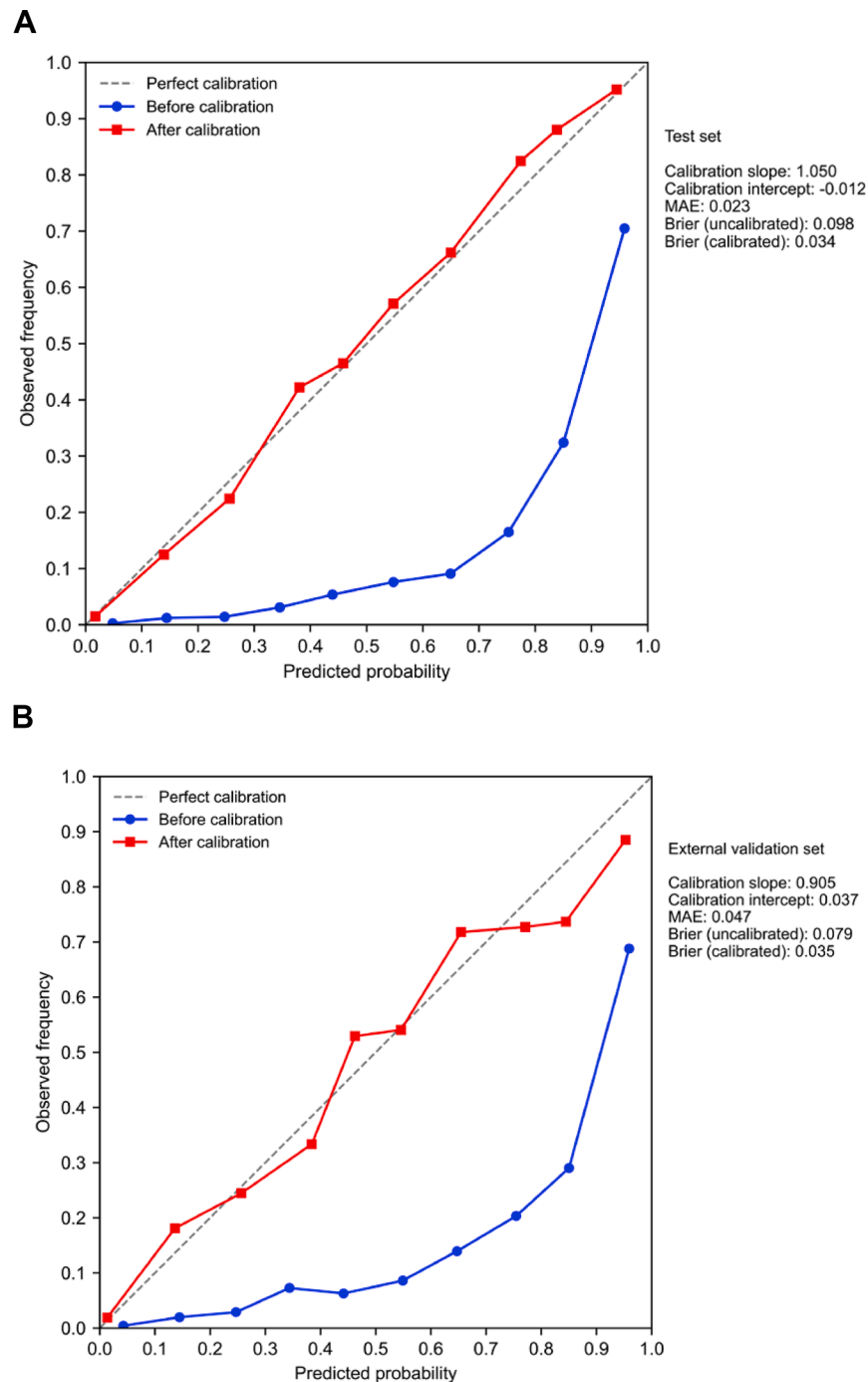


Fig. 2: Calibration curves of the DNN model using Region-46 features with age and sex before and after isotonic calibration in the internal test set (A) and external validation set (B).

DNN model incorporating age and sex, relative to the conventional ISS-based approach (Table S10). In both internal testing and external validation cohorts, the Region-46-based DNN model showed reduced performance disparities across subgroups compared with ISS. Specifically, for the external validation dataset, the difference in AUROC between age subgroups was notably

smaller in the DNN model (0.0679) compared with ISS (0.0909). Similarly, differences between female and male patients were substantially lower in the DNN model (0.0212) than with ISS (0.0462). The results indicate the Region-46-based DNN model, integrating demographic factors, provides more equitable predictions across age and sex subgroups compared with

conventional ISS-based risk stratification. We conducted additional analyses to evaluate the statistical significance of performance differences across subgroups. Specifically, we compared the performance of both the ISS and the AI model by age and sex. The results showed that both models demonstrated significantly different performance between patients aged <65 years and those aged ≥65 years. In contrast, no significant performance differences were observed between male and female groups for either model. Notably, in the external validation cohort, the ISS exhibited a significant difference in performance between sexes, whereas the AI model maintained consistent performance regardless of sex. Although significant differences were observed across age groups, the consistently decreasing trend in performance disparities across comparisons may indicate an improvement in model fairness. As a complementary analysis, model fairness was further evaluated using the EO difference, with the optimal ISS threshold (Youden's J = 16) applied for subgroup comparison. In the external validation cohort, the Region-46-based DNN model showed smaller EO differences across both age (0.164 vs 0.197) and sex (0.183 vs 0.202) subgroups compared with ISS-16, indicating more balanced error rates and improved fairness across demographic groups. Although minor variations were observed in the internal test set, the overall trend supported the superior fairness of the DNN model.

Stratified mortality risk by model prediction

Table S12 presents the HRs and adjusted HRs (aHRs) for in-hospital mortality according to model-predicted mortality probability tertiles. In the KTDB cohort (n = 93,317), the unadjusted HRs for tertiles T1, T2, and T3 were 1.30 (95% CI, 1.22–1.40), 2.53 (2.38–2.70), and 16.11 (15.38–16.87), respectively. After adjustment for age and sex, these remained strongly significant: aHRs were 1.26 (1.17–1.35), 2.43 (2.28–2.59), and 15.32 (14.62–16.05), indicating a strong association between higher predicted mortality probability and increased observed mortality. External validation cohorts from Korea (n = 4319) and Australia (n = 3089) demonstrated consistent trends. Mortality risk increased significantly across tertiles in both unadjusted and adjusted models. For example, in the Korean cohort, adjusted HRs for T1, T2, and T3 were 1.73 (1.25–2.40), 3.71 (2.79–4.95), and 17.24 (13.87–21.44), respectively. Similarly, the Australian cohort exhibited adjusted HRs of 1.03 (0.64–1.67), 2.21 (1.43–3.39), and 12.23 (8.86–16.87), respectively.

Web-based application for trauma mortality risk estimation

We have developed a user-friendly, web-based application that implements our trauma mortality risk prediction model (Fig. 6).

The calculator is publicly accessible at (http://ai-wm.khu.ac.kr/AIS_Trauma_Mortality).

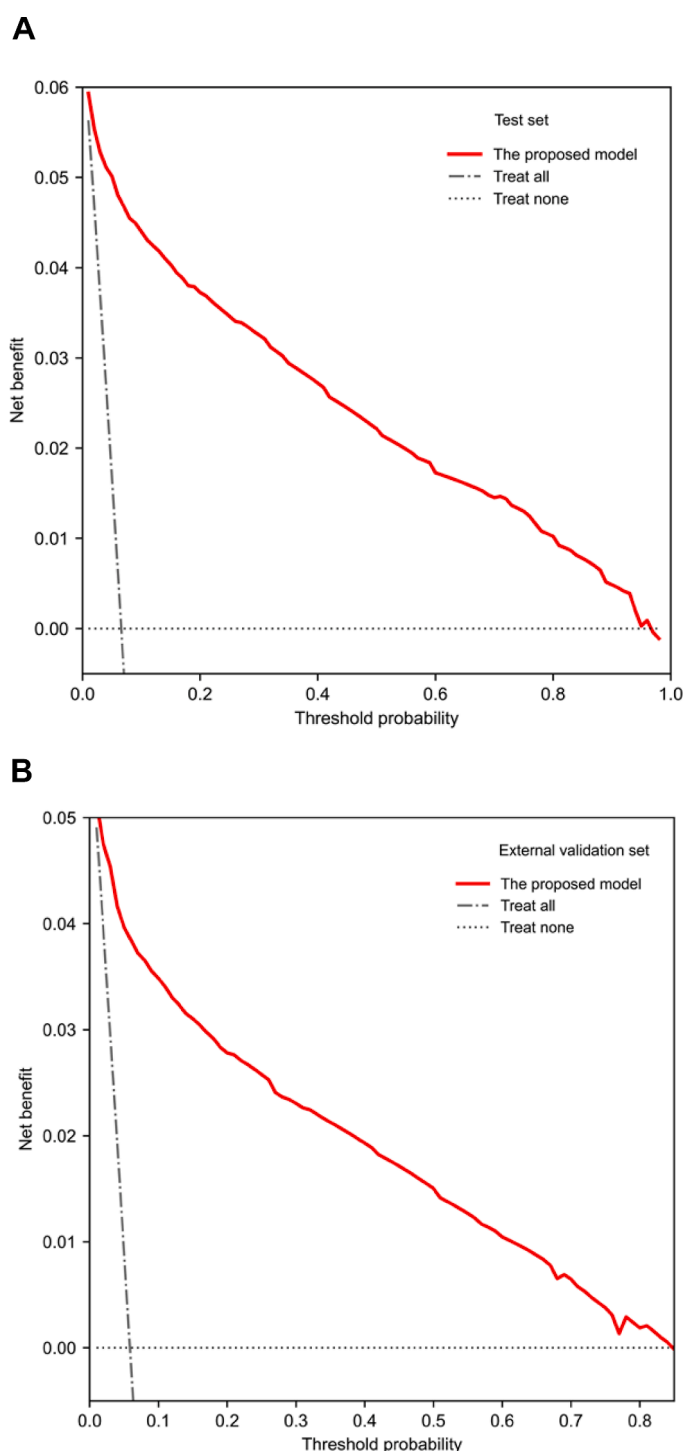


Fig. 3: Decision curve analysis of the proposed model in the internal test set (A) and external validation set (B).

Discussion

Key findings

This study provides evidence supporting five key findings regarding AI-based prediction of trauma mortality.

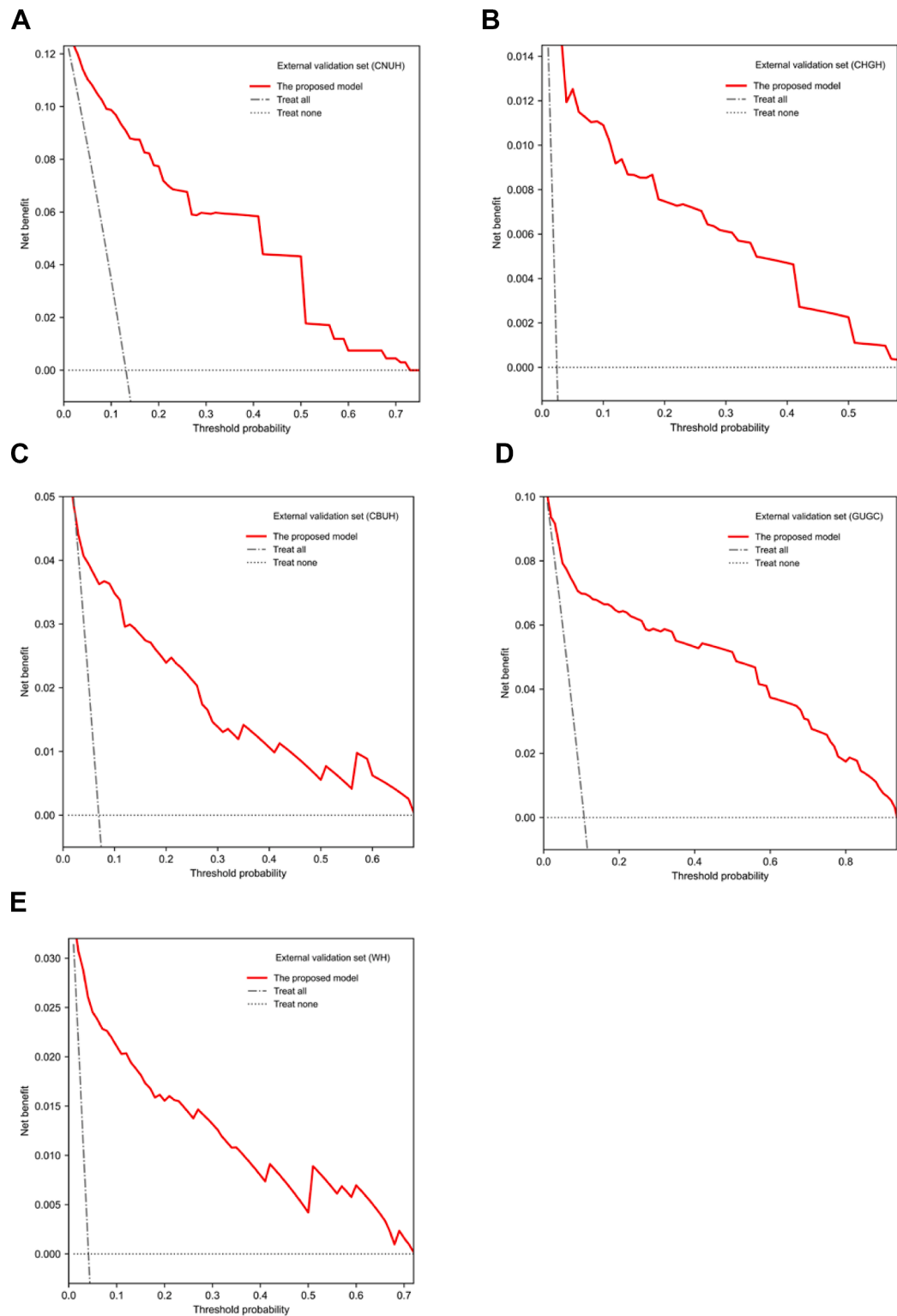


Fig. 4: Decision curve analysis of the proposed model across five external validation centres: (A) CNUH, (B) CHGH, (C) CBUH, (D) GUGC, and (E) WH.

	Sensitivity	Specificity	Precision	Accuracy	Balanced accuracy	AUROC	AUPRC	p-value (Bootstrap)	p-value (DeLong)
Test dataset (KTDB)									
ISS-16	81.2%	86.3%	29.4%	85.9%	83.7%	0.890	0.487	<0.001	<0.001
ISS-25	61.3%	96.0%	52.0%	93.7%	78.7%				
LR (ISS + age + sex)	84.1%	84.6%	27.7%	84.6%	84.3%	0.914	0.542	<0.001	<0.001
AdaBoost	82.8%	84.7%	27.7%	84.6%	83.8%	0.922	0.637	<0.001	<0.001
RF	85.8%	83.7%	27.0%	83.8%	84.7%	0.927	0.614	<0.001	<0.001
GBM	82.8%	84.7%	27.7%	84.6%	83.8%	0.917	0.619	<0.001	<0.001
LightGBM	82.7%	84.6%	27.4%	84.5%	83.7%	0.918	0.625	<0.001	<0.001
XGBoost	82.0%	85.8%	28.9%	85.5%	83.9%	0.919	0.629	<0.001	<0.001
LR	85.5%	85.1%	28.7%	85.1%	85.3%	0.930	0.647	<0.001	<0.001
DNN	86.8%	85.8%	30.1%	85.9%	86.3%	0.934	0.667	Ref.	Ref.

Abbreviations: AUPRC, Area Under the Precision–Recall Curve; AUROC, area under the receiver operating characteristic curve; ISS, Injury Severity Score; AdaBoost, adaptive boosting; RF, random forest; GBM, gradient boosting machine; XGBoost, extreme gradient boosting; LR, logistic regression; DNN, deep neural network; SD, standard deviation; KTDB, Korean Trauma Data Bank. Data represents (with SD for cross-validation). Performance metrics (sensitivity, specificity, accuracy, balanced accuracy, and AUROC) for predicting in-hospital mortality are reported. Models were trained and internally validated using the KTDB dataset (five-fold cross-validation and internal test set, 2017–2022). Results in bold indicate the best-performing model for each dataset. Ref. indicates the reference model used for pairwise statistical comparisons (p-values).

Table 1: Performance of calibrated machine learning models and ISS threshold for in-hospital mortality prediction using Region-46 features with age and sex in Test dataset (KTDB).

First, the Region-46-based AI model maintained superior predictive performance over Region-6–, Region-1999–, and ISS-based methods when trained on a large nationwide dataset exceeding 120,000 patients, demonstrating robustness and scalability. Second, among the machine learning algorithms evaluated, the Deep Neural Network (DNN) exhibited statistically superior discriminative performance compared with the other models. This finding suggests that the DNN was particularly well suited for modelling complex, non-linear feature interactions among numerous injury severity-related variables, effectively capturing latent patterns associated with in-hospital mortality even in tabular data settings. Third, expanding the training dataset to more than 120,000 patients further enhanced

predictive accuracy, confirming the scalability and stability of the Region-46-based DNN model when applied to a large nationwide cohort. Fourth, incorporating demographic variables (age and sex) alongside anatomical injury features resulted in meaningful performance improvements, highlighting the clinical value of comprehensive, patient-specific risk assessment. Fifth, the AI model exhibited reduced prediction bias and greater fairness across demographic subgroups, specifically age and sex, compared with conventional ISS-based methods. Finally, external validation using diverse, independent datasets from multiple trauma centres in South Korea and Australia confirmed the generalisability and reliability of the Region-46-based AI model across different clinical contexts.

	Sensitivity	Specificity	Precision	Accuracy	Balanced accuracy	AUROC	AUPRC	p-value (Bootstrap)	p-value (DeLong)
External validation dataset (Korea)									
ISS-16	81.3%	83.5%	27.4%	83.4%	82.4%	0.881	0.380	<0.001	<0.001
ISS-25	63.2%	94.6%	47.3%	92.4%	78.9%				
LR (ISS + age + sex)	80.6%	83.9%	27.7%	83.7%	82.3%	0.896	0.391	0.001	0.001
AdaBoost	71.6%	89.0%	33.3%	87.8%	80.3%	0.898	0.575	<0.001	<0.001
RF	80.3%	88.4%	34.5%	87.8%	84.3%	0.922	0.602	0.690	0.675
GBM	72.6%	89.4%	34.3%	88.2%	81.0%	0.898	0.576	<0.001	<0.001
LightGBM	72.6%	89.5%	34.5%	88.3%	81.0%	0.899	0.586	<0.001	<0.001
XGBoost	72.6%	89.8%	35.2%	88.6%	81.2%	0.902	0.589	<0.001	0.002
LR	76.5%	89.0%	34.6%	88.1%	82.7%	0.911	0.594	0.003	0.009
DNN	78.1%	89.6%	36.4%	88.8%	83.8%	0.920	0.628	Ref.	Ref.

Abbreviations: AUPRC, Area Under the Precision–Recall Curve; AUROC, area under the receiver operating characteristic curve; ISS, Injury Severity Score; AdaBoost, adaptive boosting; RF, random forest; GBM, gradient boosting machine; XGBoost, extreme gradient boosting; LR, logistic regression; DNN, deep neural network; SD, standard deviation; KTDB, Korean Trauma Data Bank. Performance metrics (sensitivity, specificity, accuracy, balanced accuracy, and AUROC) for predicting in-hospital mortality are reported. Models were trained and internally validated using the KTDB dataset (five-fold cross-validation and internal test set, 2017–2022) and externally validated on independent datasets from four South Korean trauma centres (CNUH, GUGC, CBUH, CHGH; 2023). Data represents (with SD for cross-validation). Results in bold indicate the best-performing model for each dataset. Ref. indicates the reference model used for pairwise statistical comparisons (p-values).

Table 2: Performance of calibrated machine learning models and ISS threshold for in-hospital mortality prediction using Region-46 features with age and sex in external validation dataset (South Korea).

	Sensitivity	Specificity	Precision	Accuracy	Balanced accuracy	AUROC	AUPRC	p-value (Bootstrap)	p-value (DeLong)
External validation dataset (Australia)									
ISS-16	65.4%	89.5%	21.1%	88.5%	77.4%	0.842	0.256	0.005	0.004
ISS-25	49.6%	97.0%	41.4%	95.1%	73.3%				
LR (ISS + age + sex)	91.6%	68.5%	25.8%	90.6%	80.0%	0.894	0.338	0.623	0.605
AdaBoost	72.4%	90.6%	24.9%	89.9%	81.5%	0.903	0.563	0.687	0.717
RF	72.4%	89.8%	23.3%	89.1%	81.1%	0.908	0.489	0.173	0.197
GBM	72.8%	89.5%	30.1%	88.5%	81.1%	0.899	0.533	0.950	0.920
LightGBM	74.8%	89.4%	23.2%	88.8%	82.1%	0.901	0.516	0.966	0.996
XGBoost	72.4%	90.1%	23.9%	89.4%	81.3%	0.900	0.509	0.974	0.916
LR	71.7%	90.9%	25.1%	90.1%	81.3%	0.900	0.554	0.755	0.746
DNN	74.0%	89.4%	23.1%	88.8%	81.7%	0.901	0.541	Ref.	Ref.

Abbreviations: AUPRC, Area Under the Precision–Recall Curve; AUROC, area under the receiver operating characteristic curve; ISS, Injury Severity Score; AdaBoost, adaptive boosting; RF, random forest; GBM, gradient boosting machine; XGBoost, extreme gradient boosting; LR, logistic regression; DNN, deep neural network; SD, standard deviation; KTDB, Korean Trauma Data Bank. Performance metrics (sensitivity, specificity, accuracy, balanced accuracy, and AUROC) for predicting in-hospital mortality are reported. Models were trained and internally validated using the KTDB dataset (five-fold cross-validation and internal test set, 2017–2022) and externally validated on independent datasets from four South Korean trauma centres (CNUH, GUGC, CBUH, CHGH; 2023) and one Australian trauma centre (Westmead Hospital; 2022–2024). Data represents (with SD for cross-validation). Results in bold indicate the best-performing model for each dataset. Ref. indicates the reference model used for pairwise statistical comparisons (p-values).

Table 3: Performance of calibrated machine learning models and ISS threshold for in-hospital mortality prediction using Region-46 features with age and sex in external validation dataset (Australia).

Clinical implication

The proposed AI-based model is particularly well-suited for routine clinical adoption due to its utilisation of the existing AIS framework, requiring minimal additional data input. Although physiological parameters are crucial indicators of a patient's condition, most trauma registries capture only a single snapshot of vital signs, which may not reflect the dynamic changes that occur during patient transport. Physiological data are inherently time-dependent and can vary significantly between prehospital and hospital settings. In contrast, our model does not rely on transient clinical variables such as physiological measurements or laboratory results.

While these parameters are undoubtedly important, the model's independence from them allows for stable prognostic estimation, robust performance over time, and consistent communication across institutions. Therefore, our model may serve as a reliable foundation for future, more comprehensive models that incorporate dynamic physiological data. Moreover, the integration of demographic variables—specifically patient age and sex—enhances predictive accuracy and simultaneously addresses fairness concerns associated with demographic biases in clinical decision-making.^{10,33} An AI model that incorporates all 46 AIS body regions and accounts for non-linear relationships may more

	Sensitivity	Specificity	Precision	Accuracy	Balanced accuracy	AUROC	AUPRC	p-value (Bootstrap)	p-value (DeLong)
Test dataset (KTDB)									
DNN (Region-1999)	88.1%	71.5%	17.9%	72.6%	79.8%	0.908	0.654	<0.001	<0.001
DNN (Region-6)	85.8%	84.4%	27.9%	84.5%	85.1%	0.923	0.584	<0.001	<0.001
DNN (Region-46)	86.8%	85.8%	30.1%	85.9%	86.3%	0.934	0.667	Ref.	Ref.
External validation dataset (Korea)									
DNN (Region-1999)	86.8%	72.2%	19.2%	73.2%	79.5%	0.905	0.631	0.113	0.118
DNN (Region-6)	74.8%	89.6%	35.5%	88.6%	82.2%	0.901	0.505	<0.001	<0.001
DNN (Region-46)	78.1%	89.6%	36.4%	88.8%	83.8%	0.920	0.628	Ref.	Ref.
External validation dataset (Australia)									
DNN (Region-1999)	83.7%	83.7%	16.8%	83.4%	80.4%	0.872	0.506	0.018	0.035
DNN (Region-6)	71.7%	88.7%	21.4%	88.0%	80.2%	0.892	0.534	0.132	0.114
DNN (Region-46)	74.0%	89.4%	23.1%	88.8%	81.7%	0.901	0.541	Ref.	Ref.

Abbreviations: AUPRC, Area Under the Precision–Recall Curve; AUROC, area under the receiver operating characteristic curve; SD, standard deviation; KTDB, Korean Trauma Data Bank. Performance metrics (sensitivity, specificity, accuracy, balanced accuracy, and AUROC) for DNN models trained using different anatomical feature encodings (Region-1999, Region-6, Region-46), all including patient age and sex as input features. Models were trained and internally validated on the KTDB dataset (2017–2022) using five-fold cross-validation and an internal test set and externally validated using independent datasets from four trauma centres in South Korea (2023) and one trauma centre in Australia (Westmead Hospital; 2022–2024). Data shown as means (with SD for cross-validation). Bold indicates the best-performing model for each dataset. Ref. indicates the reference model used for pairwise statistical comparisons (p-values).

Table 4: Performance of calibrated DNN Models Using Region-1999, Region-6, and Region-46 Features With Age and Sex for Predicting In-Hospital Mortality.

	Sensitivity	Specificity	Precision	Accuracy	Balanced accuracy	AUROC	AUPRC	p-value (Bootstrap)	p-value (DeLong)
All									
w/o age and sex with large dataset (n = 121,418)	75.5%	88.8%	29.6%	88.0%	82.2%	0.896	0.548	<0.001	0.001
w/age and sex with large dataset (n = 121,418)	76.9%	89.5%	31.4%	88.8%	83.2%	0.914	0.574	Ref.	Ref.
CNUH (Korea)									
w/o age and sex with large dataset (n = 121,418)	89.8%	89.6%	56.4%	89.6%	89.7%	0.965	0.825	0.454	0.444
w/age and sex with large dataset (n = 121,418)	87.5%	90.8%	58.8%	90.3%	89.1%	0.961	0.801	Ref.	Ref.
GUGC (Korea)									
w/o age and sex with large dataset (n = 121,418)	82.2%	85.7%	40.8%	85.3%	83.9%	0.919	0.729	0.289	0.307
w/age and sex with large dataset (n = 121,418)	79.8%	86.4%	41.4%	85.7%	83.1%	0.928	0.728	Ref.	Ref.
CBUH (Korea)									
w/o age and sex with large dataset (n = 121,418)	64.0%	81.4%	20.4%	80.2%	72.7%	0.801	0.455	0.320	0.332
w/age and sex with large dataset (n = 121,418)	74.0%	75.3%	18.2%	75.2%	74.6%	0.820	0.444	Ref.	Ref.
CHGH (Korea)									
w/o age and sex with large dataset (n = 121,418)	55.8%	95.3%	22.9%	94.4%	75.6%	0.852	0.359	0.344	0.333
w/age and sex with large dataset (n = 121,418)	58.1%	96.8%	30.9%	95.8%	77.5%	0.876	0.424	Ref.	Ref.
WH (Australia)									
w/o age and sex with large dataset (n = 121,418)	70.1%	87.7%	19.6%	87.0%	78.9%	0.862	0.477	0.001	0.001
w/age and sex with large dataset (n = 121,418)	74.0%	89.4%	23.1%	88.8%	81.7%	0.901	0.541	Ref.	Ref.

Abbreviations: AUPRC, Area Under the Precision-Recall Curve; AUROC, area under the receiver operating characteristic curve; DNN, deep neural network. Performance metrics (sensitivity, specificity, accuracy, balanced accuracy, and AUROC) comparing the previously published DNN model trained on a small KTDB dataset (n = 37,762; 2014–2019) without demographic variables (age and sex), against the current DNN models trained on a larger KTDB dataset (n = 121,418; 2017–2022), evaluated with and without age and sex. Model performance was assessed on external validation datasets from four South Korean trauma centres (2023) and one Australian trauma centre (Westmead Hospital; 2022–2024). Ref. indicates the reference model used for pairwise statistical comparisons (p-values).

Table 5: Performance Comparison of calibrated Current DNN Models Using Region-46 Features: With and Without Age and Sex, for Predicting In-Hospital Mortality.

accurately reflect clinically significant injury patterns. In contrast, the conventional six-region Injury Severity Score (ISS) assigns equal weight to injuries across diverse anatomical regions and considers only the three most severely injured regions, thereby excluding potentially relevant injuries such as rib fractures, spinal injuries, pancreatic trauma, or pelvic fractures. While this design helps prevent overfitting and aligns with the assumptions of traditional regression-based models, it inherently limits the model's ability to capture the full complexity of trauma severity. Given its accuracy, simplicity, and generalisability, our AI algorithm presents a viable alternative to the conventional ISS and could be readily incorporated into standard electronic health record (EHR) systems at trauma centres.

Comparison with prior research on injury severity models

Our previous AI model utilising the 46-region AIS structure showed modest but consistent improvements over the traditional six-region method.⁸ The current study expands upon these findings by demonstrating significant enhancement in predictive accuracy and reduced bias through larger datasets and incorporation of demographic variables. Previous study highlighted that elderly patients with low ISS scores could still face high mortality risks,³⁴ emphasising age as a critical prognostic factor inadequately captured by ISS alone. A systematic review indicated declining conventional ISS performance as dataset size increased (AUROC 0.77–0.87 for datasets >10,000).³⁵ Our findings align with and extend these results, demonstrating consistent

Centre	Sensitivity	Specificity	Precision	Accuracy	Balanced accuracy	AUROC	AUPRC
CNUH (Korea)	87.4 (80.5–93.9)	90.7 (88.3–93.0)	58.7 (50.4–67.2)	90.3 (88.1–92.4)	89.1 (85.4–92.5)	0.961 (0.943–0.976)	0.803 (0.719–0.873)
GUGC (Korea)	79.9 (73.2–86.4)	86.4 (84.3–88.5)	41.4 (35.3–47.8)	85.7 (83.6–87.7)	83.2 (79.5–86.6)	0.928 (0.901–0.951)	0.730 (0.652–0.798)
CBUH (Korea)	73.9 (61.0–85.7)	75.3 (71.9–78.5)	18.1 (13.0–23.6)	75.2 (71.8–78.1)	74.6 (67.9–80.8)	0.819 (0.743–0.888)	0.449 (0.304–0.586)
CHGH (Korea)	58.2 (42.9–72.7)	96.8 (95.9–97.6)	30.9 (21.1–41.3)	95.8 (94.9–96.8)	77.5 (69.8–84.8)	0.876 (0.806–0.938)	0.432 (0.275–0.580)
WH (Australia)	74.0 (66.4–81.2)	89.4 (88.2–90.5)	23.1 (19.1–27.3)	88.8 (87.7–89.9)	81.7 (77.8–85.3)	0.901 (0.869–0.930)	0.542 (0.455–0.627)

Abbreviations: AUPRC, Area Under the Precision-Recall Curve; AUROC, area under the receiver operating characteristic curve; DNN, deep neural network; CI, confidence interval. All performance metrics are shown as mean (95% CI) estimated by 2000 bootstrap iterations.

Table 6: Centre-specific external validation results of the calibrated DNN model using Region-46 features with age and sex.

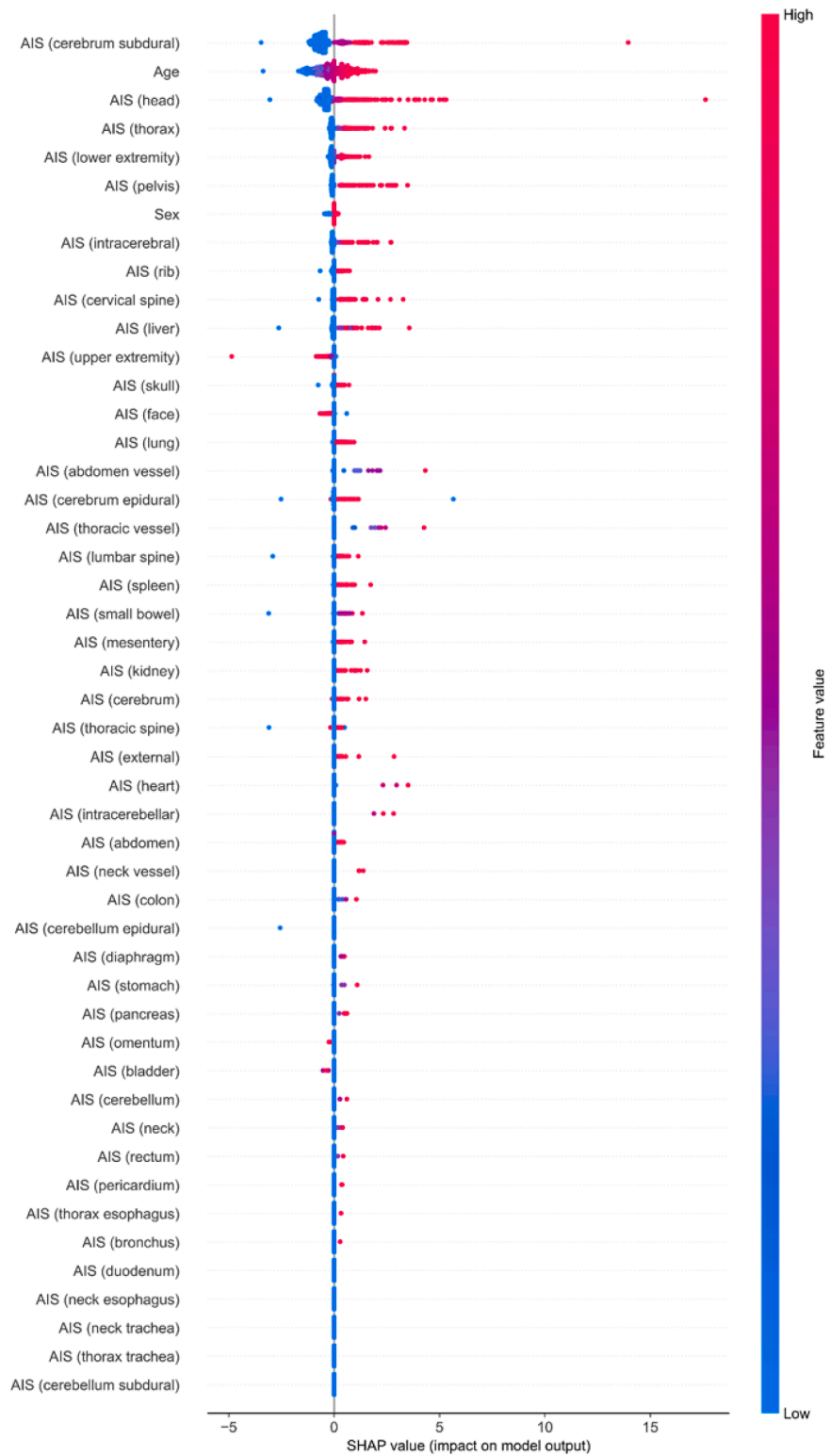
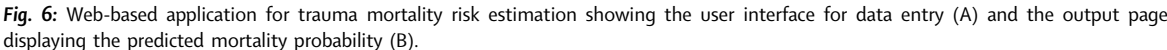


Fig. 5: Global feature importance of the proposed model derived from SHAP analysis.



Deep learning vs. tree-based ensembles in large-scale tabular trauma registries

While tree-based ensembles are often considered state-of-the-art for tabular datasets,⁴⁵ our results suggest that DNNs offer a robust and competitive alternative in the context of large-scale trauma modelling. This is likely attributable to our substantial sample size ($n > 120,000$) and the unique structural characteristics of AIS codes. Unlike the axis-aligned splits inherent in tree-based models, the DNN architecture is well-suited to learning distributed representations of high-dimensional, hierarchically organised injury patterns and their complex, non-linear interactions.⁴⁶ The consistent performance of our DNN across multiple

Limitations and future work

This study has several limitations. First, the retrospective, registry-based design may introduce selection bias and data-entry errors. Prospective validation and real-time integration into clinical workflows are therefore necessary to verify clinical utility and address data quality concerns. Despite being limited to 19 hospitals, the Korean Trauma Data Bank (KTDB) encompasses data from all 17 Regional Trauma Centres (RTCs) and two trauma education centres. The RTC system was established in 2012 through a national initiative by the Korean government to reduce preventable trauma-related mortality, supported by substantial public

funding. These RTCs are designated as the highest-level trauma centres in South Korea and are prioritised destinations for patients with severe trauma. Currently, 17 RTCs operate across the country, ensuring regional distribution and accessibility. Additionally, two trauma education centres, both located in Seoul—the largest metropolitan area in South Korea—contribute to the system. Collectively, these 19 centres offer near-comprehensive nationwide coverage and are considered representative of the trauma care system in South Korea. Second, the Region-46 anatomical encoding lacks international standardisation, potentially limiting interoperability across different healthcare systems. Future international collaborative efforts to standardise this schema could broaden its applicability and adoption. Third, our model relies exclusively on anatomical injury and demographic variables, omitting important physiological and temporal clinical information. Incorporating pre-hospital data, physiological parameters, and comorbidities would enhance predictive accuracy and real-time clinical relevance. Fourth, dependence on AIS coding restricts the model's generalisability to designated trauma centres. The AIS coding system necessitates the involvement of experienced coders working in close collaboration with medical staff, which may impose a financial burden on non-trauma centres. Indeed, AIS coding is susceptible to human error, which can affect the accuracy and reliability of trauma data. Extending validation to non-trauma centres or facilities with limited resources could further support broad clinical adoption. Fifth, Injury severity is a multidimensional construct that goes beyond mortality. It includes functional disability, hospital length of stay, healthcare costs, surgical needs, and rehabilitation requirements. Future artificial intelligence models may incorporate these diverse outcomes to enable more comprehensive and clinically relevant severity assessments. Sixth, Regarding the timing of the model's practical application in the clinical workflow, both limitations and potential exist. Obtaining an exact ISS or AIS score requires a certain amount of time, which may delay implementation in acute settings. Accurate calculation of full ISS and AIS scores requires experienced coders and access to comprehensive clinical information, such as radiologic reports and operative findings. These requirements may pose practical barriers to the rapid implementation of the model in real-world settings. However, in our country, most trauma surgeons receive formal training in AIS coding, and tentative ISS and AIS scores are often available early in the clinical course. The feasibility and utility of applying the model using preliminary AIS data warrant further investigation in future studies. We anticipate that this model will support clinicians even during the acute phase of patient care. In addition, we are currently developing an AI model capable of generating AIS codes from CT

images and text-based reports. The present study is part of a broader initiative to develop an AI-augmented trauma care system. Seventh, the decreased AUROC observed in age and sex subgroup analyses may be limited by insufficient statistical power. Although significant differences were observed across age groups, a consistent downward trend in performance disparities across all subgroup comparisons may indicate improving model fairness. In the external validation cohort, the increase in *p*-values for both age (<0.001 vs. 0.001) and sex (0.036 vs. 0.219) further supports this possibility. Notably, the AI model—designed with fairness in mind—demonstrated smaller EO differences across all subgroups in the external validation set. In contrast, it exhibited larger EO differences in the age subgroup analysis within the KTDB test set. These findings highlight the need for further investigation in larger and more diverse populations to more accurately assess and ensure fairness in model performance. Eighth, we prioritised sensitivity over precision, as avoiding 'missed' high-risk patients is clinically more critical than minimising false positives in mortality prediction. This aligns with our prior findings.⁴⁸ We suggest using calibrated probability scores to enable a staged response; this optimises resource allocation while ensuring high-level surveillance for all patients at risk. Finally, despite comprehensive external validation, evaluations in more diverse clinical contexts remain essential. Future studies should prioritise evaluating the model across varied healthcare settings to ensure robust generalisability. As a next step, we developed a user-friendly, web-based calculator that implements our predictive model. This tool enables clinicians and researchers to input AIS codes, along with patient age and sex, to estimate trauma mortality risk. By providing open access to this calculator, we aim to facilitate external validation, promote broader research applications, and support clinical decision-making. We believe that this translational approach will enhance the real-world impact of our findings and contribute to the improvement of trauma care systems. From a longer-term perspective, the model has the potential to contribute to the enhancement of the entire trauma care system, including interhospital transfers, rehabilitation, and follow-up care.

Conclusion

In this nationwide cohort study, the Region-46-based AI model demonstrated superior predictive performance compared with traditional ISS and other anatomical encoding methods, particularly when enhanced by demographic variables (age and sex). The proposed AI model, trained on a large-scale dataset and externally validated in diverse international trauma populations, offers accurate, fair, and generalisable prediction of trauma-related mortality. Given its simplicity, clinical interpretability, and ease of

integration into existing hospital information systems, this model represents a promising alternative to conventional ISS, supporting improved risk stratification and clinical decision-making in trauma care. We expect that our AI model will contribute to advancing the trauma care system and refining scoring methodologies, thereby enhancing prognostic accuracy and ultimately reducing preventable mortality.

Contributors

Concept and design: YC, JS, WSK, and JL. Statistical Analysis: YC, WSK, and JL. Interpretation of data: YC, JS, WSK, and JL. YC, JS, YO, JH, DWK, BY, JC, WJ, JK, NO, JA, RJF, PAT, DKY, DKS, WSK, and JL critically reviewed and approved to the submission of the final manuscript. YC and JS contributed equally to this work and should be considered co-first authors. WSK and JL contributed equally to this work and should be considered corresponding authors. YC, JS, WSK, and JL have directly accessed and verified the underlying data reported in the manuscript.

Data sharing statement

The data are not publicly available due to restrictions from KTDB policy and belong to the National Emergency Medical Centre (NEMC) of Korea. NEMC provides de-identified KTDB to researchers for nonprofit academic research. Details of accessing the raw data and guide are available at the NEMC website (<https://dw.nemc.or.kr>). The validation dataset used in this study contains potentially sensitive patient information and, therefore, will be made available upon reasonable request to ensure compliance with ethical guidelines and institutional privacy policies. Statistical code is available at https://github.com/yunjingchoi/In_hospital_mortality.

Declaration of interests

The authors declare no competing interests.

Acknowledgements

This research was supported by the MSIT (Ministry of Science and ICT), Korea, under the ITRC (Information Technology Research Center) support program (IITP-2025-RS-2024-00438239) and the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (RS-2024-00509257, Global AI Frontier Lab). In addition, this research was supported by the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health & Welfare, Republic of Korea (RS-2025-02220492).

Appendix A. Supplementary data

Supplementary data related to this article can be found at <https://doi.org/10.1016/j.ebiom.2026.106206>.

References

- Krug EG, Sharma GK, Lozano R. The global burden of injuries. *Am J Public Health*. 2000;90(4):523.
- Kwon J, Lee M, Kim Y, et al. Trauma system establishment and outcome improvement: a retrospective national cohort study in South Korea. *Int J Surg Lond Engl*. 2023;109(8):2293–2302.
- Baker SP, O'Neill B, Haddon W, Long WB. The injury severity score: a method for describing patients with multiple injuries and evaluating emergency care. *J Trauma*. 1974;14(3):187–196.
- Greenspan L, McLellan BA, Greig H. Abbreviated Injury Scale and Injury Severity Score: a scoring chart. *J Trauma*. 1985;25(1):60–64.
- Paffrath T, Lefering R, Flohé S, TraumaRegister DGU. How to define severely injured patients? – an Injury Severity Score (ISS) based approach alone is not sufficient. *Injury*. 2014;45(Suppl 3):S64–S69.
- Kang WS, Park YC, Jo YG, Kim JC. Pancreatic fistula and mortality after surgical management of pancreatic trauma: analysis of 81 consecutive patients during 11 years at a Korean trauma center. *Ann Surg Treat Res*. 2018;95(1):29–36.
- Kim H, Lee MS, Yoon SY, Han J, Lee JY, Seok J. Validation of chest trauma scoring systems in polytrauma: a retrospective study with 1,038 patients in Korea. *J Trauma Inj*. 2024;37(2):114–123.
- Kang WS, Chung H, Ko H, et al. Artificial intelligence to predict in-hospital mortality using novel anatomical injury score. *Sci Rep*. 2021;11(1):23534.
- Kheya TA, Bouadjenek MR, Aryal S. The Pursuit of Fairness in Artificial Intelligence Models: A Survey. arXiv [cited 2025 May 9]. Available from: <http://arxiv.org/abs/2403.17333>; 2024.
- Chu C, Donato-Woodger S, Khan SS, et al. Strategies to mitigate age-related bias in machine learning: scoping review. *JMIR Aging*. 2024 Mar 22;7:e53564.
- Collins GS, Moons KGM, Dhiman P, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods [cited 2025 Mar 14]; Available from: <https://www.bmj.com/content/385/bmj-2023-078378>; 2024.
- Youden WJ. Index for rating diagnostic tests. *Cancer*. 1950;3(1):32–35.
- Castelnovo A, Crupi R, Greco G, Regoli D, Penco IG, Cosentini AC. A clarification of the nuances in the fairness metrics landscape. *Sci Rep*. 2022;12:4209.
- Pawluszek-Filipiak K, Borkowski A. On the importance of train-test split ratio of datasets in automatic landslide detection by supervised classification. *Remote Sens*. 2020;12(18):3054.
- Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. *J Artif Intell Res*. 2002;16:321–357.
- Bleeker SE, Moll HA, Steyerberg EW, et al. External validation is necessary in prediction research: a clinical example. *J Clin Epidemiol*. 2003;56(9):826–832.
- Efron B. Bootstrap methods: another look at the jackknife. In: Kotz S, Johnson NL, eds. *Breakthroughs in Statistics*. New York, NY: Springer New York; 1992:569–593 [cited 2025 Nov 10]. (Springer Series in Statistics). Available from: http://link.springer.com/10.1007/978-1-4612-4380-9_41.
- Schapire RE. Explaining AdaBoost. In: Schölkopf B, Luo Z, Vovk V, eds. *Empirical Inference*. Berlin, Heidelberg: Springer Berlin Heidelberg; 2013:37–52 [cited 2025 Mar 13]. Available from: https://link.springer.com/10.1007/978-3-642-41136-6_5.
- Biau G, Scornet E. A random forest guided tour. *TEST*. 2016 Jun;25(2):197–227.
- Natekin A, Knoll A. *Frontiers in neurorobotics*. Gradient Boost Mach Tutor; 2013.
- Ke G, Meng Q, Finley T, et al. Lightgbm: a highly efficient gradient boosting decision tree. *Adv Neural Inf Process Syst*. 2017;30.
- Chen T, Guestrin C. Xgboost: a scalable tree boosting system. In: *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*. 2016:785–794.
- LaValley MP. Logistic regression. *Circulation*. 2008;117(18):2395–2399.
- Fushiki T. Estimation of prediction error by using K-fold cross-validation. *Stat Comput*. 2011;21(2):137–146.
- DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*. 1988;44:837–845.
- Bertail P, Cléménçon S, Vayatis N. On bootstrapping the ROC curve. *Adv Neural Inf Process Syst*; 21. [cited 2025 Nov 10]. Available from: <https://proceedings.neurips.cc/paper/2008/hash/30bb3825e8f631cc6075c0f87bb4978c-Abstract.html>; 2008.
- Lundberg SM, Lee SI. A unified approach to interpreting model predictions. *Adv Neural Inf Process Syst*; 30. [cited 2025 Nov 10]. Available from: <https://proceedings.neurips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html>; 2017.
- On behalf of Topic Group 'Evaluating diagnostic tests and prediction models' of the STRATOS initiative, Van Calster B, McLernon DJ, Van Smeden M, Wynants L, Steyerberg EW. Calibration: the Achilles heel of predictive analytics. *BMC Med*. 2019;17(1):230.
- Vickers AJ, van Calster B, Steyerberg EW. A simple, step-by-step guide to interpreting decision curve analysis. *Diagn Progn Res*. 2019;3:18.
- Harrell FE. Cox proportional hazards regression model. In: *Regression Modeling Strategies*. Cham: Springer International Publishing; 2015:475–519 (Springer Series in Statistics). [cited 2025 May 7]. Available from: https://link.springer.com/10.1007/978-3-319-19425-7_20.
- Niculescu-Mizil A, Caruana R. Predicting good probabilities with supervised learning. In: *Proceedings of the 22nd international*

- conference on Machine learning - ICML '05. Bonn, Germany: ACM Press; 2005:625–632 [cited 2025 Nov 10]. Available from: <http://portal.acm.org/citation.cfm?doid=1102351.1102430>.
- 32 Allikivi ML, Kull M. Non-parametric bayesian isotonic calibration: fighting over-confidence in binary classification. In: Brefeld U, Fromont E, Hotho A, Knobbe A, Maathuis M, Robardet C, eds. *Machine learning and knowledge discovery in databases*. Cham: Springer International Publishing; 2020:103–120 [cited 2025 Nov 10]. (Lecture Notes in Computer Science; vol. 11907). Available from: https://link.springer.com/10.1007/978-3-030-46147-8_7.
- 33 Hasanzadeh F, Josephson CB, Waters G, Adedinsowo D, Azizi Z, White JA. Bias recognition and mitigation strategies in artificial intelligence healthcare applications. *NPJ Digit Med*. 2025;8(1):1–13.
- 34 Lee S, Kim DW, Oh NE, 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(1):1100.
- 35 Tohira H, Jacobs I, Mountain D, Gibson N, Yeo A. Systematic review of predictive performance of injury severity scoring tools. *Scand J Trauma Resusc Emerg Med*. 2012;20:63.
- 36 Baur D, Gehlen T, Scherer J, et al. Decision support by machine learning systems for acute management of severely injured patients: a systematic review. *Front Surg*. 2022;9:924810.
- 37 Lee S, Kang WS, Seo S, et al. Model for predicting In-Hospital mortality of physical trauma patients using artificial intelligence techniques: nationwide population-based study in Korea. *J Med Internet Res*. 2022;24(12):e43757.
- 38 Lee S, Kang WS, Kim DW, et al. An artificial intelligence model for predicting trauma mortality among emergency department patients in South Korea: retrospective cohort study. *J Med Internet Res*. 2023;25:e49283.
- 39 Abe D, Inaji M, Hase T, et al. A prehospital triage system to detect traumatic intracranial hemorrhage using machine learning algorithms. *JAMA Netw Open*. 2022;5(6):e2216393.
- 40 Gebran A, Thakur SS, Maurer LR, et al. Development of a machine learning-based prescriptive tool to address racial disparities in access to care after penetrating trauma. *JAMA Surg*. 2023;158(10):1088–1095.
- 41 Tsiklides EJ, Sims C, Sinno T, Diamond SL. Using the National Trauma Data Bank (NTDB) and machine learning to predict trauma patient mortality at admission. *PloS One*. 2020;15(11):e0242166.
- 42 Maurer LR, Bertsimas D, Bouardi HT, et al. Trauma outcome predictor: an artificial intelligence interactive smartphone tool to predict outcomes in trauma patients. *J Trauma Acute Care Surg*. 2021;91(1):93–99.
- 43 Bonde A, Bonde M, Troelsen A, Sillesen M. Assessing the utility of a sliding-windows deep neural network approach for risk prediction of trauma patients. *Sci Rep*. 2023;13(1):5176.
- 44 Kwon JM, Lee Y, Lee Y, Lee S, Park H, Park J. Validation of deep-learning-based triage and acuity score using a large national dataset. *PloS One*. 2018;13(10):e0205836.
- 45 Grinsztajn L, Oyallon E, Varoquaux G. Why do tree-based models still outperform deep learning on tabular data?. arXiv [cited 2026 Jan 12]. Available from: <http://arxiv.org/abs/2207.08815>; 2022.
- 46 Miotto R, Li L, Kidd BA, Dudley JT. Deep patient: an unsupervised representation to predict the future of patients from the electronic health records. *Sci Rep*. 2016;6(1):26094.
- 47 Garcia N, Strzoda R, Lucca G, Borges E. A performance analysis of classifiers on imbalanced data. In: *Proceedings of the 24th International Conference on Enterprise Information Systems*. Online Streaming, — Select a Country —: SCITEPRESS - Science and Technology Publications; 2022:602–609 [cited 2026 Jan 25]. Available from: <https://www.scitepress.org/DigitalLibrary/Link.aspx?doi=10.5220/0011089100003179>.
- 48 Oh NE, Oh TYC, Hsu J, et al. Prehospital real-time AI for trauma mortality prediction: a multi-institutional and multi-national validation study. *Nat Commun*. 2026;17:1464.