scientific reports

Check for updates

OPEN

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

Seungseok Lee[1,6], Do Wan Kim[2,6], Na-eun Oh[1], Hayeon Lee[3], Seoyoung Park[4], Dong Keon Yon[4], Wu Seong Kang[5,7]✉ & Jinseok Lee[1,7]✉

Artificial intelligence (AI) is being increasingly applied in healthcare to improve patient care and clinical outcomes. We previously developed an AI model using ICD-10 (International Classification of Diseases, Tenth Revision) codes with other clinical variables to predict in-hospital mortality among trauma patients from a nationwide database. This study aimed to externally validate the performance of the AI model. Validation was conducted using a multicenter retrospective cohort study design, analyzing patient data from January 2020 to December 2021. The study included trauma patients based on specific ICD-10 codes, with other clinical variables. The performance of the AI model was evaluated against conventional metrics, including the ISS, and the ICISS (ICD-based ISS), using sensitivity, specificity, accuracy, balanced accuracy, precision, F1-score, and area under the receiver operating characteristic curve (AUROC) analyses. Data from 4,439 patients were analyzed. The AI model demonstrated high overall performance, achieving an AUROC of 0.9448 and a balanced accuracy of 85.08%, thereby outperforming traditional scoring systems such as ISS, or ICISS. Furthermore, the model accurately predicted mortality across datasets from each hospital (AUROCs of 0.9234 and 0.9653, respectively) despite significant differences in hospital characteristics. In the subset of patients with ISS < 9, the model showed a robust AUROC of 0.9043, indicating its effectiveness in predicting mortality, even in cases with lower-severity injuries. For patients with ISSs ≥ 9, the model maintained high sensitivity (93.60%) and balanced accuracy (77.08%), proving its reliability in more severe injury cases. External validation demonstrated the AI model's high predictive accuracy and reliability in assessing in-hospital mortality risk among trauma patients across different injury severities and heterogeneous cohorts. These findings support the model's potential integration into emergency departments and offer a significant tool for enhancing patient triage and treatment protocols.

Keywords Artificial Intelligence, In-Hospital mortality, Trauma patients, ICD-10, External validation, Injury Severity score

Abbreviations

AI	Artificial intelligence
ISS	injury severity score
ICISS	International Classification of Diseases, Tenth Revision–based Severity Score
TRISS	Trauma and Injury Severity Score
ICD-10	International Classification of Diseases, Tenth Revision
NEDIS	National Emergency Department Information System
AUROC	area under the ROC curve

AVPU	Alert/Verbal/Painful/Unresponsive
KTAS	Korean Triage and Acuity Scale
ED	emergency department
ICU	intensive care unit
KTDB	Korean Trauma Data Bank
TRIPOD	Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis
STROCSS	Strengthening the Reporting of Cohort, Cross-sectional, and Case–Control Studies in Surgery
SRRs	survival using survival risk ratios
Ps	probability of survival
RTS	Revised Trauma Score
CNUH	Chonnam National University Hospital
CHH	Cheju Halla General Hospital

[1]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. [2]Department of Thoracic and Cardiovascular Surgery, Chonnam National University Hospital, Chonnam National University Medical School, Gwangju, Republic of Korea. [3]Department of Electronics and Information Convergence Engineering, Kyung Hee University, Yongin-si, Republic of Korea. [4]Center for Digital Health, Medical Science Research Institute, Kyung Hee University College of Medicine, Seoul, Republic of Korea. [5]Department of Trauma Surgery, Jeju Regional Trauma Center, Cheju Halla General Hospital, 65, Doryeong-ro, Jeju-si, Jeju-do, Republic of Korea. [6]Seung Seok Lee and Do Wan Kim have contributed equally to this work and should be considered co-first authors. [7]Wu Seong Kang and Jinseok Lee contributed equally and should be considered as corresponding authors. ✉email: wuseongkang@naver.com; gonasago@khu.ac.kr

Trauma is one of the leading causes of in-hospital deaths and has been a burden on patients and healthcare providers[1]. Estimating severity and predicting prognosis is a prerequisite for reducing this burden, as it enables clinicians to allocate resources more effectively and prioritize interventions such as angioembolization or surgery[2,3]. Despite advances in trauma care, the ability to accurately predict outcomes in trauma patients remains a challenge. To date, several scoring systems have been used to estimate trauma mortality and severity, such as the Injury Severity Score (ISS), International Classification of Diseases, Tenth Revision–based Severity Score (ICISS), or Trauma and Injury Severity Score (TRISS)[4–6]. However, the performance and availability of the conventional metrics remain limited[7–9].

Recent AI models designed to predict in-hospital mortality have demonstrated potential to aid critical decision-making processes, although their integration into routine clinical practice remains limited[10,11]. We previously developed and trained an AI model that leveraged the International Classification of Diseases, Tenth Revision (ICD-10) coding system to predict in-hospital mortality in trauma patients, with promising results[11]. This model has demonstrated its efficacy within a comprehensive national database, exhibiting substantial accuracy, sensitivity, and specificity against established trauma-scoring systems[11]. This model was initially trained and tested on a comprehensive National Emergency Department Information System (NEDIS) dataset encompassing over 778,111 patients from more than 400 hospitals across South Korea. With an impressive area under the receiver operating characteristic curve (AUROC) of 0.9507, the model achieved not only high sensitivity and specificity but also demonstrated significantly balanced accuracy when compared to traditional trauma scoring systems.

However, the generalizability of any AI model must be tested through rigorous external validation before it can be used reliably in clinical practice. This study aimed to externally validate an ICD-10-based AI model for predicting in-hospital mortality in trauma patients using a multicenter retrospective cohort study. By leveraging datasets from heterogeneous centers, this study sought to assess the generalizability and accuracy of the AI model across different settings, patient populations, and practices. Trauma prediction models have clinical applicability across various points in the trauma care system, such as for triage in the emergency department and prognosis during hospitalization. The focus of our study is to predict in-hospital mortality, providing a critical prognostic tool at the point of admission or during early care stages, which can inform resource allocation and treatment prioritization.

Materials and methods

Validation of AI model for predicting in-hospital mortality in diverse hospital settings

In this study, we validated our previously trained AI model for predicting in-hospital mortality[11]. In the previous study[11], we constructed a 9-layer deep neural network model (DNN) and incorporated a comprehensive set of variables, including age, gender, intentionality of injury, mechanism of injury, presence of emergent symptoms, and the Alert/Verbal/Painful/Unresponsive (AVPU) scale[12], in addition to the initial and altered Korean Triage and Acuity Scale (KTAS)[13], specific ICD-10 codes (international version), and categorized procedure codes for surgical or interventional radiology procedures. The DNN comprised an input layer, followed by 7 fully connected (FC) hidden layers with 512, 256, 128, 64, 32, 16, and 8 nodes, respectively, and an output layer. Dropout with a rate of 0.3 and L2 regularization were applied to the FC hidden layers to prevent overfitting. Based on the one-hot encoding process for all variables, we integrated 866 ICD-10 codes (beginning with letters S or T) to derive 914 distinct input features for the AI model, as summarized in Supplementary Table 1. The KTAS serves as a standardized severity triage tool in emergency department (ED) to minimize complexity and uncertainty, categorizing patients into five levels: level 1, resuscitation; level 2, emergent; level 3, urgent; level 4, less urgent; and level 5, non-urgent[13]. According to the NEDIS policy, KTAS assessments are performed

by certified personnel within two minutes of ED admission, with alterations made as necessary based on the patient's condition prior to transition to the operating room, intensive care unit (ICU), or general ward. For ICD-10 codes, we used 866 codes starting with S or T. Procedure codes required for billing through the National Health Insurance Review & Assessment Service, covering surgical and angioembolization interventions, are more specifically categorized as follows and summarized in Supplementary Table 2: head procedures, vascular torso procedure, abdominal torso procedures, chest torso procedures, heart torso procedures, and extracorporeal membrane oxygenation.

Our AI model was constructed with seven fully connected layers serving as hidden layers, and it features an output layer that provides the probability of patient mortality. In a previous study, we trained the model using 778,111 patients from the NEDIS dataset, which was collected from 2016 to 2019 from 400 + hospitals in South Korea. The model provided an AUROC of 0.9507, with 87.68% sensitivity, 86.25% specificity, and 86.97% balanced accuracy. In this study, we validated the model using 4,439 patients from two regional trauma centers in South Korea. The training set for the development of the AI model comprised all types of EDs in South Korea, while the external validation set comprised two regional trauma centers.

Study design and dataset for external validation of AI model

This study was conducted at two regional trauma centers in South Korea, corresponding to level 1 trauma centers in the US. This study was approved by the relevant institutional review board approval was obtained, and the requirement for informed consent was waived because of the observational nature of this study. All patient data were coded anonymously to ensure privacy and confidentiality. Institutional review board (IRB) approval was obtained from Cheju Halla General Hospital and Chonnam National University Hospital (IRB numbers: CHH-2023-L16-01 and CNUH-2022-L02-01, respectively).

This multicenter retrospective cohort study was conducted in accordance with the TRIPOD (Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis) statement[14]and STROCSS (Strengthening the Reporting of Cohort, Cross-sectional, and Case–Control Studies in Surgery) criteria[15]. The dataset was collected from the Korean Trauma Data Bank (KTDB) data of two regional trauma centers in South Korea from January 2020 to December 2021. This study aimed to externally validate the AI model for predicting in-hospital mortality in trauma patients. The primary outcome of our study was the in-hospital mortality rate.

The inclusion criteria for the study were as follows: (1) trauma patients identified by a diagnostic code starting with S or T according to the Korean version of the ICD-10; (2) patients who were admitted to the ICU or a general ward directly from the ED; and (3) patients who were admitted to the ICU or a general ward following surgery or a procedure initiated in the ED. According to the KTDB policy, patients were excluded from the KTDB for the following reasons: (1) absence of a diagnostic code starting with S or T; (2) presence of diagnostic codes related to frostbite (T33-T35.6), intoxication (T36-T65), or unspecified injuries or complications (T66-T78, T80-T88). To focus the analysis on in-hospital mortality, we excluded patients who (1) died in the ED or were transferred to another facility before admission ($n = 686$) and (2) were transferred to another facility or left the hospital against medical advice after admission ($n = 2,061$). Consequently, the data of 4,439 patients were used for the external validation of the model. The selection process is illustrated in Fig. 1.

In South Korea, regional trauma centers are the highest-level trauma centers, corresponding to Level-1 trauma centers. CNUH, as the highest-level tertiary university hospital in the region, experienced a high patient transfer rate. However, transfers to other tertiary hospitals were rare (3%, 52/1,726), with approximately 97% of transferred patients sent to lower-level hospitals, likely for conservative treatment rather than critical care.

Statistical analysis

All statistical analyses were performed using R version 4.1.2 (R Foundation for Statistical Computing, Vienna, Austria). Proportions were compared using the chi-squared test or Fisher's exact test, as appropriate. Statistical significance was set at $p < 0.05$. AI models were implemented using Python (version 3.7.13) with TensorFlow (version 2.8.0), Keras (version 2.8.0), NumPy (version 1.21.6), Pandas (version 1.3.5), Matplotlib (version 3.5.1), and Scikit-learn (version 1.0.2). The performance of the prediction model was evaluated using seven metrics: sensitivity, specificity, accuracy, balanced accuracy, precision, F1-score, and AUROC. Sensitivity is a crucial measure of the ability to correctly identify positive cases, which is particularly important in serious conditions (e.g., death) in trauma, to avoid missing positive cases. Balanced accuracy accounts for both sensitivity and specificity, making it useful when dealing with imbalanced datasets, whereas AUROC provides a summary of model performance across all thresholds, which is helpful for comparing models in a general sense. In the evaluation of our AI model, calibration was assessed on the dataset using the Brier score[16], calibration slope, and calibration intercept. To further ascertain the clinical utility of the model, we performed decision curve analysis[17].

Conventional metrics for comparison

For comparison with traditional metrics, we employed ISS[4], and ICISS[5].]. ICISS calculates the probability of survival using survival risk ratios (SRRs)[5]. In our previous study, we determined the SRR for each diagnostic code and applied these SRRs in the current study. We performed a correlation analysis to evaluate variability and complementarity between models. Models with low correlations are more likely to capture different patterns in the data, making them useful for improving predictive accuracy across diverse data points[18–21].

Results

Patient characteristics

Notably, patients excluded due to death in the ED, being transferred, or leaving against medical advice were predominantly found at Chonnam National University Hospital (CNUH) compared to Cheju Halla General

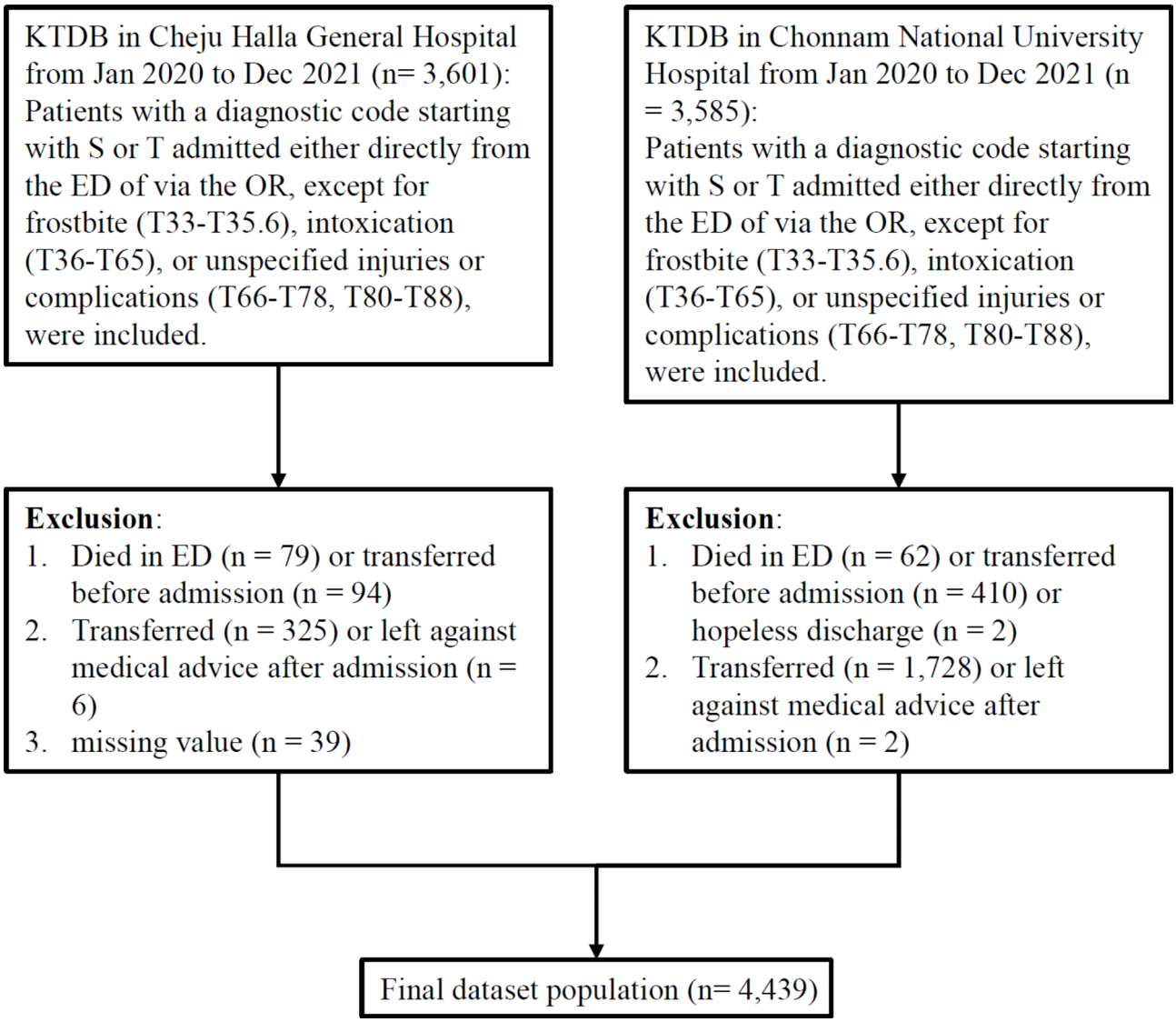

Fig. 1. Patient selection flow chart: This flowchart illustrates the process of selecting patients for the study from the Korean Trauma Data Bank (KTDB) in two hospitals.

Hospital (CHH) (61.4% (2204/3585) at CNUH vs. 15.1% (543/3058) at CHH, $p<0.001$). Other patient characteristics, according to hospital and mortality rates, are summarized in Table 1. The in-hospital mortality was significantly higher at CNUH (8.8% vs. 3.2%, $p<0.001$). Significant differences were observed in terms of vascular procedures (3.5% at CHH vs. 5.3% at CNUH, $p=0.008$), abdominal procedures (2.2% at CHH vs. 5.6% at CNUH, $p<0.001$), head procedures (0% at CHH, 0.4% at CNUH, $p=0.004$). Initial KTAS levels showed significant differences across levels 1 (0.8% at CHH vs. 3.2% at CNUH), 3 (59.6% at CHH, 37.1% at CNUH, $p<0.001$), 4 (20.6% at CHH vs. 37.9% at CNUH, $p<0.001$), and 5 (0.0% at CHH, 2.8% at CNUH, $p<0.001$). Further analysis showed the difference in intentionality and mechanisms of injury. The AVPU scale responses, alert (91.7% at CHH vs. 88.9% at CNUH, $p=0.004$), semi-coma (3.1% at CHH vs. 4.4% at CNUH, $p=0.036$), and coma (1.0% at CHH vs. 2.7% at CNUH, $p<0.001$), as well as sex distribution (62.3% male at CHH vs. 73.6% at CNUH, $p<0.001$) demonstrated significant differences. For all cases, including both hospitals, the comparison of patient characteristics between surviving and deceased patients is summarized in Supplementary Table 3. The comparison of patient characteristics between patients with ISS < 9 and ISS ≥ 9 is summarized in Supplementary Table 4. The distribution of ICD-10 codes between the two hospitals is summarized in Supplementary Table 5. A comparison of ICD-10 codes between deceased and surviving patients is summarized in Supplementary Table 6. The comparison of ICD-10 codes between patients with ISS < 9 and ISS ≥ 9 is summarized in Supplementary Table 7.

Validation results: AI model performance

Our AI model for predicting in-hospital mortality demonstrated high accuracy across the entire dataset (Table 2). The TRISS could not be calculated for 156 patients due to factors such as intubation and injuries to

	Cheju Halla General Hospital			Chonnam National University Hospital		
	Total	Survived	Deceased	Total	Survived	Deceased
Hospital	(*N* = 3058)	(*N* = 2960)	(*N* = 98)	(*N* = 1382)	(*N* = 1261)	(*N* = 121)
Mortality rate	98 (3.2%)			121 (8.8%)		
Age (year)	53.4 ± 21.5	52.8 ± 21.4	71.9 ± 17.1	49.0 ± 23.8	47.9 ± 23.9	60.2 ± 20.4
Procedure code						
Head procedure	74 (2.4%)	48 (1.6%)	26 (26.5%)	34 (2.5%)	8 (0.6%)	26 (21.5%)
Torso procedure-vascular	108 (3.5%)	91 (3.1%)	17 (17.3%)	73 (5.3%)	61 (4.8%)	12 (9.9%)
Torso procedure-abdomen	68 (2.2%)	51 (1.7%)	17 (17.3%)	78 (5.6%)	54 (4.3%)	24 (19.8%)
Torso procedure-chest	88 (2.9%)	78 (2.6%)	10 (10.2%)	32 (2.3%)	22 (1.7%)	10 (8.3%)
Torso procedure-heart	0 (0.0%)	0 (0.0%)	0 (0.0%)	5 (0.4%)	4 (0.3%)	1 (0.8%)
ECMO	2 (0.1%)	0 (0.0%)	2 (2.0%)	2 (0.1%)	1 (0.1%)	1 (0.8%)
Initial KTAS						
Level 1	24 (0.8%)	10 (0.3%)	14 (14.3%)	44 (3.2%)	10 (0.8%)	34 (28.1%)
Level 2	580 (19.0%)	538 (18.2%)	42 (42.9%)	262 (19.0%)	196 (15.5%)	66 (54.5%)
Level 3	1823 (59.6%)	1791 (60.5%)	32 (32.7%)	513 (37.1%)	497 (39.4%)	16 (13.2%)
Level 4	630 (20.6%)	620 (20.9%)	10 (10.2%)	524 (37.9%)	519 (41.2%)	5 (4.1%)
Level 5	1 (0.0%)	1 (0.0%)	0 (0.0%)	39 (2.8%)	39 (3.1%)	0 (0.0%)
Not classified	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
Missing data	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
Altered KTAS						
Level 1	34 (1.1%)	12 (0.4%)	22 (22.4%)	53 (3.8%)	14 (1.1%)	39 (32.2%)
Level 2	427 (14.0%)	386 (13.0%)	41 (41.8%)	331 (24.0%)	269 (21.3%)	62 (51.2%)
Level 3	2549 (83.4%)	2514 (84.9%)	35 (35.7%)	943 (68.2%)	926 (73.4%)	17 (14.0%)
Level 4	47 (1.5%)	47 (1.6%)	0 (0.0%)	54 (3.9%)	51 (4.0%)	3 (2.5%)
Level 5	1 (0.0%)	1 (0.0%)	0 (0.0%)	1 (0.1%)	1 (0.1%)	0 (0.0%)
Missing data	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
Intentionality						
Accidental, unintentional	2792 (91.3%)	2710 (91.6%)	82 (83.7%)	1091 (78.9%)	1003 (79.5%)	88 (72.7%)
Suicide, intentional self-harm	65 (2.1%)	64 (2.2%)	1 (1.0%)	4 (0.3%)	3 (0.2%)	1 (0.8%)
Assault, violence	77 (2.5%)	76 (2.6%)	1 (1.0%)	2 (0.1%)	2 (0.2%)	0 (0.0%)
Other specified	0 (0.0%)	0 (0.0%)	0 (0.0%)	116 (8.4%)	109 (8.6%)	7 (5.8%)
Unspecified	48 (1.6%)	45 (1.5%)	3 (3.1%)	160 (11.6%)	135 (10.7%)	25 (20.7%)
Missing data	76 (2.5%)	65 (2.2%)	11 (11.2%)	9 (0.7%)	9 (0.7%)	0 (0.0%)
Injury mechanism						
Car accident	331 (10.8%)	324 (10.9%)	7 (7.1%)	106 (7.7%)	95 (7.5%)	11 (9.1%)
Bicycle accident	67 (2.2%)	67 (2.3%)	0 (0.0%)	34 (2.5%)	32 (2.5%)	2 (1.7%)
Motorcycle accident	236 (7.7%)	227 (7.7%)	9 (9.2%)	93 (6.7%)	70 (5.6%)	23 (19.0%)
Traffic accident-pedestrian, train, airplane, ship, etc.	237 (7.8%)	216 (7.3%)	21 (21.4%)	100 (7.2%)	70 (5.6%)	30 (24.8%)
Traffic accident-unknown	0 (0.0%)	0 (0.0%)	0 (0.0%)	4 (0.3%)	3 (0.2%)	1 (0.8%)
Fall	626 (20.5%)	616 (20.8%)	10 (10.2%)	193 (14.0%)	171 (13.6%)	22 (18.2%)
Slip	716 (23.4%)	679 (22.9%)	37 (37.8%)	234 (16.9%)	217 (17.2%)	17 (14.0%)
Struck by person or object	212 (6.9%)	209 (7.1%)	3 (3.1%)	295 (21.3%)	286 (22.7%)	9 (7.4%)
Firearm/cut (sharp or object)/piece	253 (8.3%)	252 (8.5%)	1 (1.0%)	263 (19.0%)	258 (20.5%)	5 (4.1%)
Machine	107 (3.5%)	107 (3.6%)	0 (0.0%)	18 (1.3%)	18 (1.4%)	0 (0.0%)
Fire, flames, or heat	69 (2.3%)	65 (2.2%)	4 (4.1%)	9 (0.7%)	9 (0.7%)	0 (0.0%)
Drowning or nearly drowning	3 (0.1%)	3 (0.1%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
Poisoning	0 (0%)	0 (0.0%)	0 (0.0%)	0 (0%)	0 (0.0%)	0 (0.0%)
Choking, hanging	0 (0.0%)	0 (0.0%)	0 (0.0%)	1 (0.1%)	1 (0.1%)	0 (0.0%)
Other-rape, electric	129 (4.2%)	128 (4.3%)	1 (1.0%)	26 (1.9%)	25 (2.0%)	1 (0.8%)
Unknown	72 (2.4%)	67 (2.3%)	5 (5.1%)	6 (0.4%)	6 (0.5%)	0 (0.0%)
Missing data	0 (0%)	0 (0.0%)	0 (0.0%)	0 (0%)	0 (0.0%)	0 (0.0%)
Emergency presentation						
Yes	3020 (98.8%)	2922 (98.7%)	98 (100.0%)	1382 (100.0%)	1261 (100.0%)	121 (100.0%)
No	36 (1.2%)	36 (1.2%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
Unspecified	2 (0.1%)	2 (0.1%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
AVPU scale						
Continued						

Hospital	Cheju Halla General Hospital			Chonnam National University Hospital		
	Total (*N* = 3058)	Survived (*N* = 2960)	Deceased (*N* = 98)	Total (*N* = 1382)	Survived (*N* = 1261)	Deceased (*N* = 121)
Alert	2803 (91.7%)	2758 (93.2%)	45 (45.9%)	1229 (88.9%)	1190 (94.4%)	39 (32.2%)
Drowsy	128 (4.2%)	115 (3.9%)	13 (13.3%)	55 (4.0%)	43 (3.4%)	12 (9.9%)
Semi-coma	95 (3.1%)	76 (2.6%)	19 (19.4%)	61 (4.4%)	25 (2.0%)	36 (29.8%)
Coma	32 (1.0%)	11 (0.4%)	21 (21.4%)	37 (2.7%)	3 (0.2%)	34 (28.1%)
Unknown response	0 (0%)	0 (0.0%)	0 (0.0%)	0 (0%)	0 (0.0%)	0 (0.0%)
Sex (male)	1904 (62.3%)	1847 (62.4%)	57 (58.2%)	1017 (73.6%)	922 (73.1%)	95 (78.5%)
Injury severity score	8.2 ± 7.3	7.7 ± 6.6	20.9 ± 13.1	8.5 ± 8.9	7.0 ± 7.0	24.9 ± 9.9
Injury severity score < 9	1695 (55.4%)	1682 (56.8%)	13 (13.3%)	821 (59.4%)	819 (64.9%)	2 (1.7%)

Table 1. Patient characteristics and comparison between hospitals according to in-hospital mortality. KTAS, Korean Triage and Acuity Scale; ECMO, extracorporeal membrane oxygenation, AVPU, alert/verbal/pain/unresponsive.

the extremities. Therefore, we did not use TRISS to evaluate model performance. The model's performance is illustrated in Fig. 2(A), showing an AUROC of 0.9448, outperforming traditional scoring systems such as ISS (AUROC of 0.8807) and ICISS (AUROC of 0.7978). The model achieved a sensitivity of 89.95%, surpassing ISS-16 (78.08%), ICISS (76.71%), and ISS-25 (53.88%). The AI model's specificity was 80.21%, balancing sensitivity and specificity better than traditional models. The balanced accuracy of our model was the highest at 85.08%. Additionally, our model's precision was 95.39%, indicating exceptional reliability and precision in predictions. Overall, the AI model outperformed ISS and ICISS across various datasets, including hospital and injury severity variations. (Fig. 2. A-I) Supplementary Table 8 provides a comparative analysis with other machine learning models.

AI model performance in patients with ISS < 9

For patients with lower-severity injuries (ISS < 9), the AI model continued to demonstrate commendable accuracy (Fig. 2B,E,H). Figure 2B shows an AUROC of 0.8979, significantly better than ISS (AUROC = 0.5455) and ICISS (AUROC = 0.5). As summarized in Table 2, our model balanced sensitivity and specificity better, with a sensitivity of 40.0% compared to zero sensitivity for ISS and ICISS. The model's balanced accuracy was 68.74%, indicating precise risk stratification for less severe injuries. The model also showed exceptional precision (99.01%), minimizing false positives and optimizing resource allocation and patient management in emergency care. Characteristics and predictive probabilities for deceased patients with ISS < 9 are summarized in Table 3.

AI model performance in patients with ISS ≥ 9

For patients with more severe injuries (ISS ≥ 9), the AI model's capability was highlighted again (Fig. 2C,F,I). Figure 2C shows an AUROC of 0.9143, superior to ISS (AUROC = 0.8171) and ICISS (AUROC = 0.7164). Table 2 shows a high sensitivity of 93.63%, indicating strong performance in identifying in-hospital mortality. Although specificity was 60.56%, lower than traditional metrics, the balanced accuracy of 77.09% suggests a favorable equilibrium between sensitivity and specificity. This balance is crucial for high-risk patients, where accurate risk identification is essential. The lower specificity compared to ISS-16, ISS-25, and ICISS is offset by the higher sensitivity, preventing potentially fatal oversights in patient care.

Correlation of ISS and ICISS vs. our AI model

Figure 3 shows the correlation between the AI model's predictive probabilities and traditional scoring systems: ISS and ICISS. The regression for ISS versus the AI model yielded a slope of 15.62, an intercept of 4.97, and a moderate R-squared value of 0.34 (Fig. 3A). For ICISS versus the AI model, the regression showed a negative correlation with a slope of −0.37, an intercept of 0.96, and an R-squared value of 0.25 (Fig. 3B).

Calibration of AI model and decision curve analysis

Figure 4A shows the calibration of the AI model, with a Brier score of 0.10, indicating good accuracy. The calibration slope of 0.89 and intercept of −3.10 reflect the model's calibration accuracy. The R-squared value of 0.52 indicates the model explains a substantial portion of the variance in observed outcomes.

Figure 4B presents the decision curve analysis for the AI model, ISS and ICISS. The AI model shows the highest net benefit around threshold probabilities of 8.8% (mortality at CNUH) and 3.2% (mortality at CHH), indicating significant clinical advantages in decision-making processes, particularly at low threshold probabilities for intervention. Traditional scoring systems show lower net benefits across the same threshold range.

Discussion

In this study, we leveraged a large, diverse cohort to externally validate an ICD-10-based AI model pre-trained on a comprehensive national dataset of over 778,111 patients[11]. Despite notable heterogeneity between the two hospitals, our AI model exhibited outstanding performance. This study highlights several key points in evaluating medical AI applications. Firstly, the two hospitals had distinct characteristics. One hospital, situated on an island

Model	Sensitivity	Specificity	Accuracy	Balanced accuracy	Precision	F1-score	AUROC
Total dataset in both hospitals							
AI model	0.8995	0.8021	0.8069	0.8508	0.9539	0.8954	0.9448
ISS-16	0.7808	0.8665	0.8623	0.8237	0.9498	0.8951	0.8807
ISS-25	0.5388	0.9642	0.9432	0.7515	0.9493	0.9460	0.8807
ICISS	0.7671	0.7440	0.7452	0.7556	0.9421	0.8169	0.7978
Patients with ISS < 9 in both hospitals							
AI model	0.4000	0.9372	0.9340	0.6686	0.9905	0.9604	0.8979
ISS	0.0000	1.0000	0.9940	0.5000	0.9881	0.9911	0.5455
ICISS	0.0000	1.0000	0.9940	0.5000	0.9881	0.9911	0.5000
Patients with ISS ≥ 9 in both hospitals							
AI model	0.9363	0.6056	0.6407	0.7709	0.9062	0.7089	0.9143
ISS-16	0.8431	0.6725	0.6906	0.7578	0.8947	0.7498	0.8171
ISS-25	0.5784	0.9122	0.8768	0.7453	0.894	0.8840	0.8171
ICISS	0.6618	0.7376	0.7296	0.6997	0.8722	0.7781	0.7164
Total dataset at CHH							
AI model	0.8469	0.7919	0.7937	0.8598	0.9656	0.8598	0.9234
ISS-16	0.6327	0.8733	0.8656	0.7530	0.9592	0.9041	0.8168
ISS-25	0.4592	0.9645	0.9483	0.7119	0.9483	0.9535	0.8168
ICISS	0.5204	0.8416	0.8313	0.6810	0.9535	0.8824	0.6905
Patients with ISS < 9 at CHH							
AI model	0.4615	0.9334	0.9298	0.6975	0.9883	0.9568	0.8987
ISS-16	0.0000	1.0000	0.9923	0.5000	0.9847	0.9885	0.5110
ISS-25	0.0000	1.0000	0.9923	0.5000	0.9847	0.9885	0.5110
ICISS	0.0000	1.0000	0.9923	0.5000	0.9847	0.9885	0.5000
Patients with ISS ≥ 9 at CHH							
AI model	0.9059	0.6056	0.6244	0.7558	0.9363	0.7190	0.8969
ISS-16	0.7294	0.7066	0.7080	0.7180	0.9232	0.7831	0.7621
ISS-25	0.5294	0.9178	0.8936	0.7236	0.9254	0.9069	0.7621
ICISS	0.5647	0.7300	0.7197	0.6474	0.9095	0.7908	0.6446
Total dataset at CNUH							
AI model	0.9421	0.8262	0.8364	0.8842	0.9363	0.8670	0.9653
ISS-16	0.9091	0.8508	0.8559	0.8799	0.9355	0.8809	0.9351
ISS-25	0.6033	0.9635	0.9319	0.7834	0.9314	0.9317	0.9351
ICISS	0.8512	0.7762	0.7828	0.8137	0.9193	0.8267	0.8629
Patients with ISS < 9 at CNUH							
AI model	0.0000	0.9451	0.9428	0.4725	0.995	0.9682	0.8761
ISS-16	0.0000	1.0000	0.9976	0.5000	0.9951	0.9963	0.6868
ISS-25	0.0000	1.0000	0.9976	0.5000	0.9976	0.9963	0.6868
ICISS	0.0000	1.0000	0.9976	0.5000	0.9976	0.9963	0.5000
Patients with ISS ≥ 9 at CNUH							
AI model	0.9580	0.6054	0.6804	0.7817	0.8571	0.7088	0.9303
ISS-16	0.9244	0.5737	0.6482	0.7490	0.8389	0.6789	0.8381
ISS-25	0.6134	0.8957	0.8357	0.7546	0.8357	0.8357	0.8381
ICISS	0.7227	0.7959	0.7804	0.7593	0.8237	0.7940	0.7593

Table 2. Predictive performance of AI model, ISS, and ICISS. AI, artificial intelligence; ISS, Injury Severity Score; ICISS, ICD-10–based Injury Severity Score; AUROC, area under the receiver operating characteristic curve; CHH, Cheju Halla General Hospital; CNUH, Chonnam National University Hospital.

with few nearby alternatives, had a high transfer rate (48.2%), while the other, inland and surrounded by many hospitals, had a low transfer rate (9.2%). It underscores the importance of considering the generalizability of AI models across various data distributions. We anticipate that our model will be applicable in diverse clinical settings. Secondly, we validated the AI model for patients with low ISS (< 9). Elderly patients and those with multiple comorbidities are at increased risk of mortality even from minor injuries, which traditional ISS cannot estimate. Our AI model demonstrated potential as an alternative to conventional metrics such as ISS and ICISS in predicting mortality. Despite its higher predictive performance for mortality, ISS is based on various measures, including organ damage and potential for recovery. The complementary nature of ISS, ICISS, and our

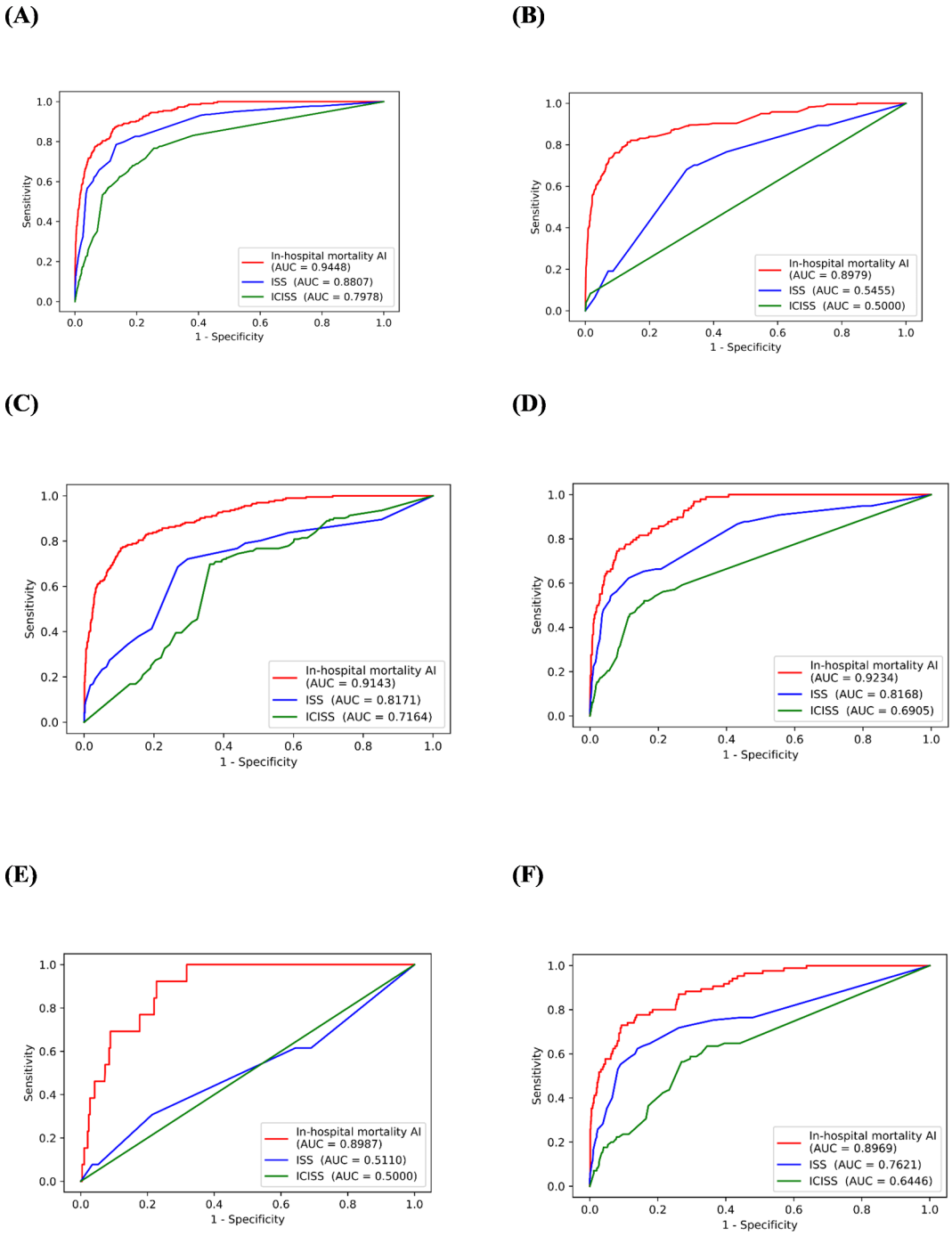

Fig. 2. Comparative ROC curves of our AI model and traditional models (ISS and ICISS) for predicting in-hospital mortality (**A**) from the entire dataset, (**B**) among patients with ISS < 9 from the entire dataset, (**C**) among patients with ISS ≥ 9 from the entire dataset, (**D**) from the CHH dataset, (**E**) among patients with ISS < 9 from the CHH dataset, (**F**) among patients with ISS ≥ 9 from the CHH dataset, (**G**) from the CNUH dataset, (**H**) among patients with ISS < 9 from the CNUH dataset, and (**I**) among patients with ISS ≥ 9 from the CNUH dataset.

AI model is highlighted by their wide distribution and variance. Thirdly, our AI model can predict mortality risk as soon as a diagnosis is made, even before surgery, making it useful at the admission stage in emergency rooms. The model can be used in various hospital stages and is especially beneficial in non-trauma centers, given its nationwide applicability. Nonetheless, prospective validation in real-world emergency department settings is necessary, as assigning ICD codes may be limited during the initial stages of care, and some injuries are only diagnosed intraoperatively. Additionally, further validation in non-trauma centers is required to assess the model's generalizability and feasibility in broader clinical environments.

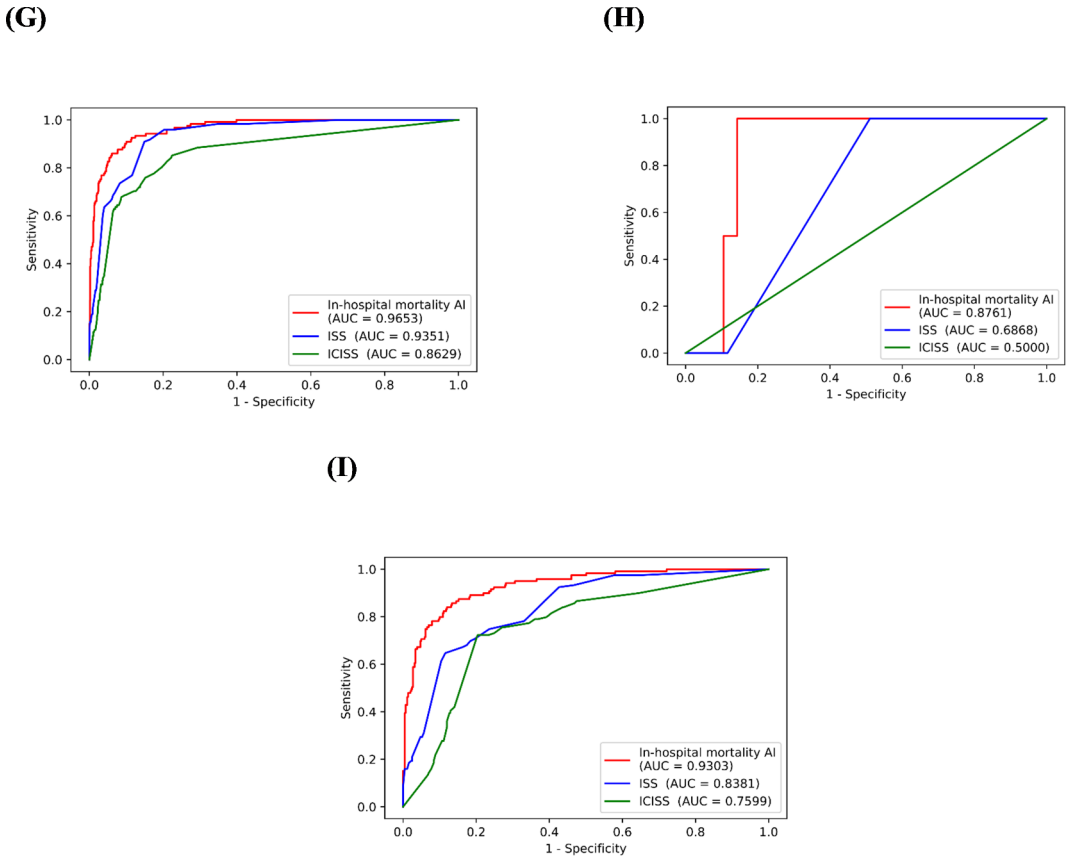

Figure 2. (continued)

Several systematic reviews have reported AI models for predicting in-hospital mortalities. Zhang reported that only 12% of studies performed external validation[22], which is crucial for model robustness, identifying overfitting, unveiling hidden biases, and assessing real-world applicability. The appropriate methodology for external validation remains unclear. For example, Gorczyca et al. used the Nationwide Readmission Database[23], while Kwon et al. validated an AI model using a single hospital dataset[24]. Our study used two diverse cohorts, potentially increasing the AI model's robustness. A reporting guideline for AI is under development[25], indicating the need for further discussion on external validation methodologies.

The primary outcome was in-hospital mortality, focusing on patients with low ISS (< 9). Deceased patients were predominantly older with high comorbidities. Our AI model accurately predicted deaths that traditional metrics missed, although it also had false negatives. The vulnerability of older patients with severe comorbidities to trauma needs further investigation. Recent studies have also used in-hospital mortality as a primary outcome, regardless of hospital stay duration[26–30]. In the real world, distinguishing trauma-related mortality is challenging. Our model discriminated mortality well, even in patients with low ISS, where age and comorbidities contribute to specific patient vulnerabilities.

Our AI model focused on in-hospital mortality. However, injury severity also includes tissue damage extent, hospitalization need, intensive care, treatment cost, treatment complexity, treatment length, temporary or permanent disability, and quality of life[31]. The comparison with ISS may not be entirely fair because the ISS does not include a physiological component, which limits its ability to comprehensively predict mortality. However, the ISS remains the most widely used conventional tool for evaluating injury severity and is the standard assessment tool in Korean regional trauma centers. Our study aimed to overcome the limitations of conventional tools, such as the ISS, by leveraging AI technology to provide a more robust and accurate prediction of in-hospital mortality. Our model addresses one aspect of injury severity—mortality—and aims to complement rather than replace conventional prediction models such as ISS. Notably, TRISS showed poor performance in previous studies in the US[8]and South Korea[9]. TRISS was not calculated for 156 patients who transferred from other hospitals and intubated without checked GCS, limiting its reliability. Vital signs at admission, used in TRISS, may not reflect the initial patient condition accurately. Our model development excluded vital signs due to their insignificance in predicting outcomes, highlighting TRISS's limitations in different clinical settings. Given that there were patients with uncheckable TRISS, an alternative AI model seems to be more useful. Notably, in our study, deceased patients with ISS < 9 showed a high TRISS (over 97%).

In our previous study, we did not consider dataset diversity during model development, leading to several critical issues. AI models trained on biased data may produce outcomes that favor certain groups while disadvantaging or discriminating against others. The NEDIS dataset encompasses various types of hospitals,

No.	Hospital	Age	Sex	AVPU scale	KTAS	ISS	ICISS	TRISS	DNN probability	Prediction	Result	Diagnosis	Underlying disease	Cause of death
1	CHH	61	F	Alert	3	1	1.000	98.5%	0.0335	FN	Death	Burn, Lt leg burn	Endometrial cancer stage IV, multiple metastases	MOF
2	CNUH	68	M	Alert	3	4	1.000	97.8%	0.0692	FN	Death	Right tibial fracture	HTN, DM	MOF
3	CHH	79	F	Alert	4	1	0.833	98.5%	0.0721	FN	Death	Pelvic contusion	COPD, pul. TB, bronchiectasis, destroyed lung	Respiratory failure, pneumoniae
4	CHH	79	F	Alert	4	8	1.000	97.5%	0.0785	FN	Death	Pelvic fracture	ESRD, Heart failure	MOF
5	CHH	68	F	Alert	3	4	1.000	98.2%	0.1149	FN	Death	Nasa bone fracture	Gall bladder cancer stage IV	Hepatic failure
6	CNUH	89	F	Alert	4	4	1.000	97.8%	0.1249	FN	Death	Right clavicle fracture	HTN, Old cerebral infarction, hypothyroidism	Respiratory failure, pneumoniae
7	CHH	56	M	Alert	3	5	1.000	98.0%	0.3336	FN	Death	Inhalation burns	Pul. TB with destroyed lung	Respiratory failure
8	CHH	90	F	Alert	3	5	1.000	98.0%	0.3452	FN	Death	Maxilla fracture	Old cerebral infarction, dementia	Respiratory failure, pneumoniae
9	CHH	88	F	Alert	3	4	1.000	98.2%	0.3766	FN	Death	T7, T12 compression fracture	Congestive heart failure, atrial fibrillation	MOF, sepsis, pneumonia
10	CHH	88	M	Alert	3	1	0.986	98.5%	0.5292	TP	Death	Cerebral contusion	Old cerebral infarction, Lt. hemiparesis, infective endocarditis, mitral valve regurgitation	MOF
11	CHH	77	F	Alert	2	5	0.938	96.3%	0.5954	TP	Death	Pneumothorax, mandible fracture	Pancreas cancer, chemotherapy	MOF
12	CHH	84	F	Alert	3	4	0.900	98.2%	0.6134	TP	Death	Pelvic fracture, acetabular fracture	ESRD, HTN, DM	MOF
13	CHH	92	M	Alert	3	4	0.986	98.2%	0.6447	TP	Death	T7 compression fracture	Lung cancer (untreated)	Respiratory failure
14	CHH	86	M	Verbal response	2	1	0.976	98.5%	0.7671	TP	Death	Facial laceration	S/p bladder cancer	MOF, pneumonia, sepsis
15	CHH	93	F	Painful response	4	1	1.000	97.3%	0.8586	TP	Death	Facial laceration	Cerebral infarction	Pneumoniae, respiratory failure

Table 3. Summary of deceased patients with ISS < 9. ISS, Injury Severity Score; AVPU, alert/verbal/pain/unresponsiveness; KTAS, Korean Triage and Acuity Scale; ICISS, ICD-10–based Injury Severity Score; TRISS, Trauma and Injury Severity Score; DNN, deep neural network; CHH, Cheju Halla General Hospital; CNUH, Chonnam National University Hospital; FN, false negative; TP, HTN, hypertension; DM, diabetes mellitus; COPD, chronic obstructive pulmonary disease; ESRD, end-stage renal disease; pul. TB, pulmonary tuberculosis; MOF, multiple organ failure.

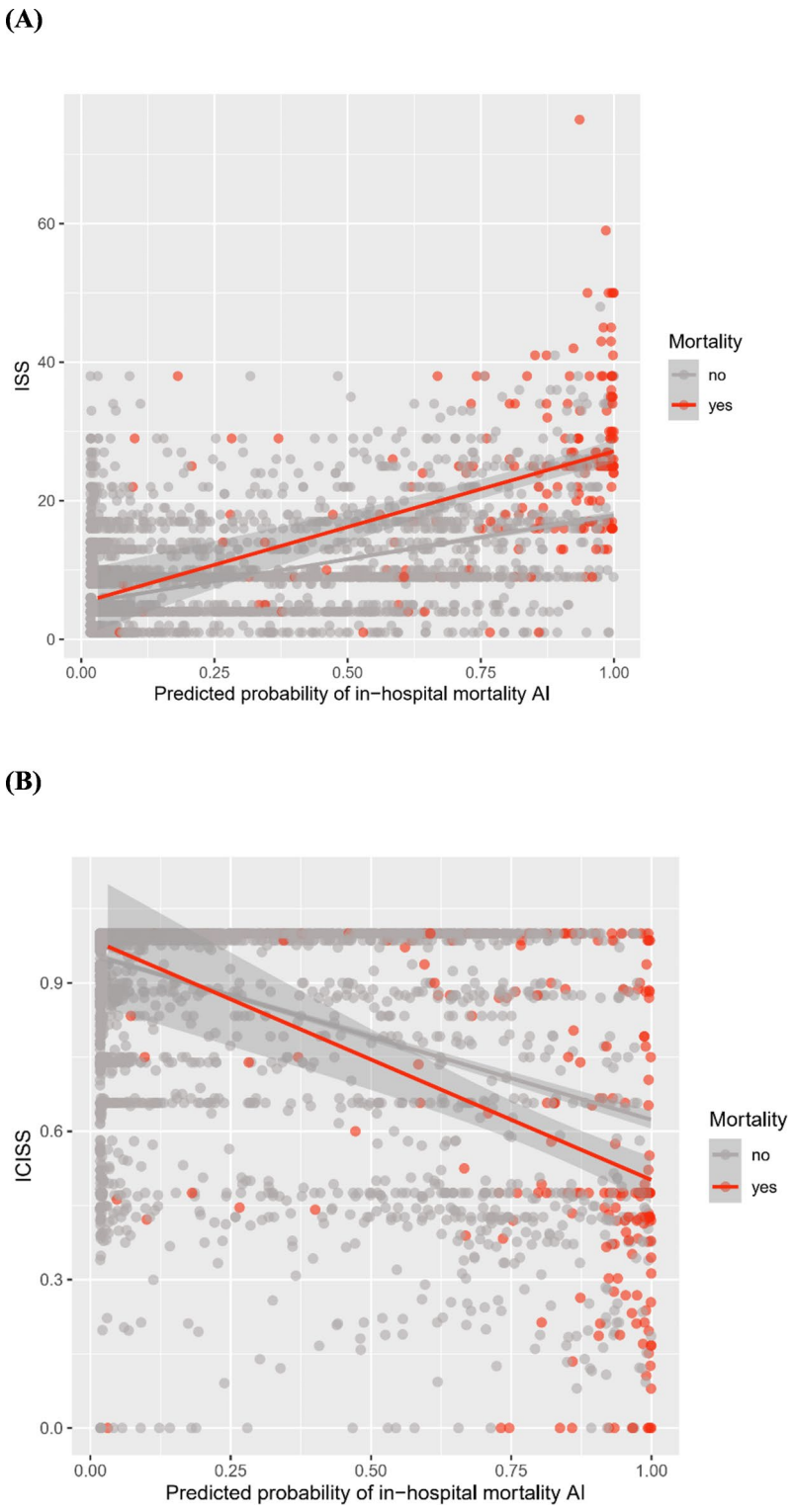

Fig. 3. Correlation between the predictive probabilities of in-hospital mortality as assessed by the AI model and three traditional scoring systems: (**A**) ISS, and (**B**) ICISS.

including trauma centers, non-trauma centers, and small hospitals. This is because the AI model was derived from a nationwide dataset. Consequently, the models may exhibit poor performance in specific situations or with certain inputs, highlighting the need for evaluation in real-world scenarios. Our AI model, derived from a comprehensive dataset, exhibited excellent performance in the context of trauma centers due to the data diversity in NEDIS. Christie et al. demonstrated that machine learning provided excellent discrimination

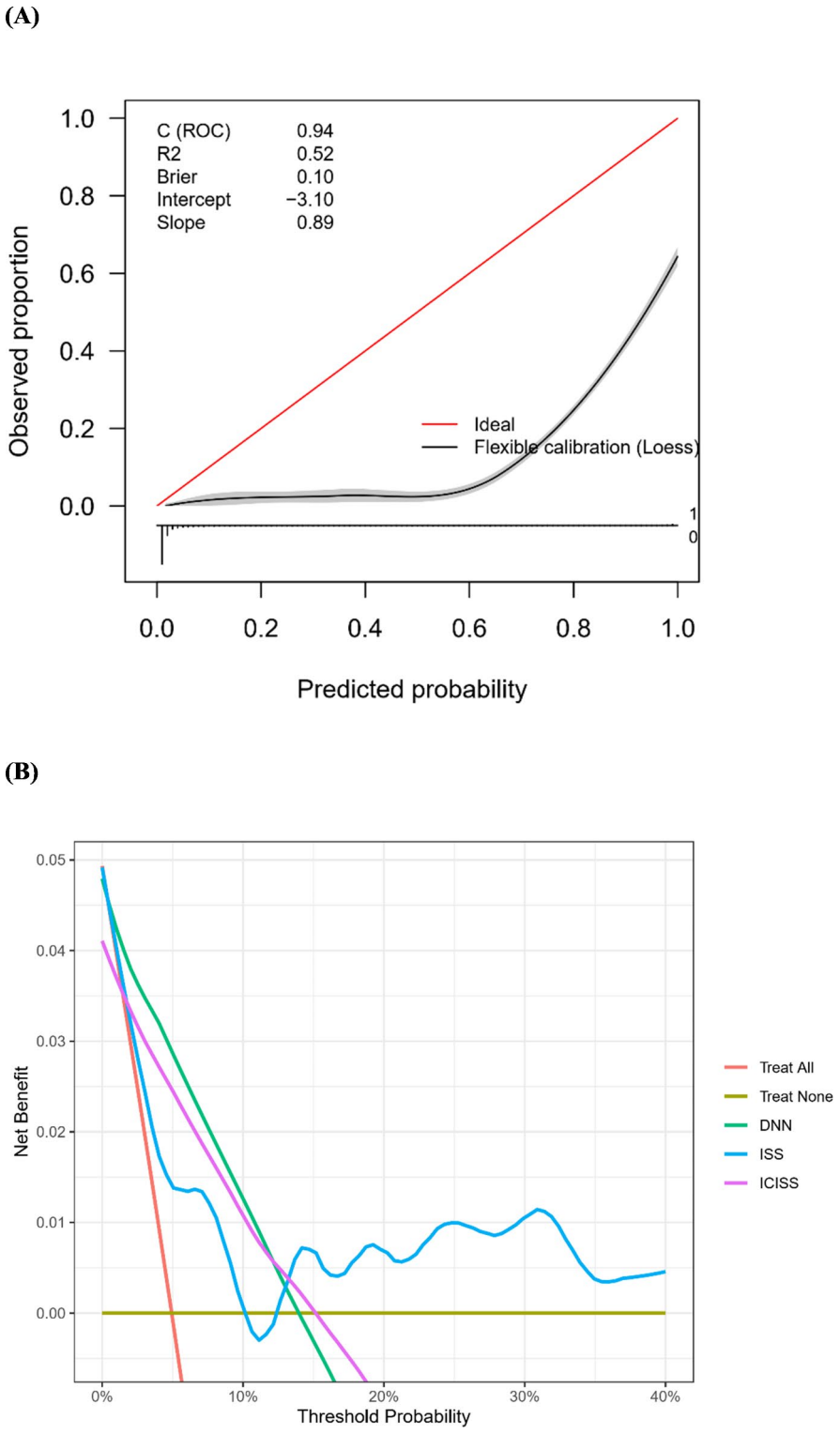

Fig. 4. Calibration and decision curve analysis: (**A**) Calibration of the DNN on the dataset. (**B**) Decision curve analysis of DNN, ISS, and ICISS.

of trauma mortality in diverse settings, including hospitals in the US and South Africa, with a high AUROC exceeding 0.9 in both cohorts[30]. The validation results of the diverse cohorts in our study were excellent as well.

In this study, we used a cutoff value of 9 on the ISS for minor trauma and analyzed the validation performances. Specifically, we intended to investigate the AI model's discriminative power in patients with minor injuries, focusing on those with 'very minor' injuries, and tested our AI model accordingly. However, we also recognize the importance of evaluating the model with a more widely accepted definition of major

trauma using a cutoff value of 16. Using the cutoff value of 16, the results show that our AI model consistently demonstrates high sensitivity, specificity, and balanced accuracy across all datasets, outperforming traditional models. This consistent performance highlights the AI model's potential as a reliable tool for improving patient triage and treatment protocols.

Our study had several limitations to be addressed in future work. First, our AI model did not incorporate significant confounders such as the precise extent of injuries, initial treatment, fluid resuscitation, infection control, or insurance. These factors can substantially influence the mortality outcomes of trauma patients. The exclusion of these confounders is a limitation of the NEDIS dataset used in this study. Future studies should aim to include these confounders to provide a more comprehensive analysis. Bias and risk of bias are important considerations when developing and validating AI models, particularly when applying them across diverse patient populations. Factors such as sex, age, or socioeconomic status can influence injury outcomes and model performance. Although we did not stratify results by sex in this study, we recognize this as a potential limitation. Future work should include subgroup analyses to evaluate performance across demographic groups, ensuring fairness and generalizability of the model. Addressing these biases is essential when transferring AI models to different clinical settings to avoid inequities in care. Using more appropriate datasets that capture these variables will be crucial for improving the accuracy and generalizability of the AI model. Second, the high rate of exclusion of patients may have caused selection bias. Nonetheless, the performance of the AI model was excellent even in hospitals where the number of excluded patients was 61.4%. Due to the excellent health insurance system in South Korea and the low medical costs for trauma patients, hospital stays can be long. Therefore, some trauma centers transfer patients to lower-level hospitals after acute care and stabilization. Although CNUH exhibited a high transfer rate, most transferred patients were sent to lower-level hospitals for conservative treatment, minimizing the possibility that the most severe trauma patients were transferred out and excluded from the study. However, CHH is located in an island area with only four other hospitals, which contributes to its lower transfer rate. Further prospective studies are required to mitigate this bias. Third, the ICD-10 and procedure codes in each hospital were used for billing and not for evaluating accurate diagnoses. Nonetheless, the ICD-10–based model can complement the weaknesses of ISS. AI can detect consistent patterns of human behavior. Fourth, we included children because our previous AI model incorporated children's data. Future tailored models for pediatric trauma patients are needed to improve accuracy and performance in this subgroup. Fifth, we did not consider serious injuries to the extremities, such as vascular injuries or extensive soft tissue damage. Future studies are needed to evaluate the impact of extremity injuries on mortality. Sixth, the calibration slope of 0.89, although close to 1.0, indicates slight miscalibration. This finding suggests that the model tends to slightly underestimate or overestimate probabilities at certain ranges. Additionally, the intercept of −3.10 and the R-squared value of 0.52 highlight areas for further improvement in calibration and overall accuracy. Despite achieving a moderate Brier score of 0.10, future efforts should focus on refining calibration to enhance the model's generalizability and robustness in real-world clinical settings. Finally, beyond validation research, a prospective study on the clinical decision support system is required for the practical use of AI. Future research should focus on prospective studies to evaluate how the AI model can be integrated into clinical workflows. Additionally, further investigations are needed to explore the model's performance across different healthcare settings and populations. These steps will help in refining the AI model and ensuring its robustness and generalizability in various clinical environments. Furthermore, determining the appropriate application of the model at various time points in the chain of care is crucial for identifying feasible predictors that are available at those specific stages. This will help align the model with real-world clinical workflows and ensure its utility in supporting relevant decisions, rather than relying solely on extensive data to drive performance improvements. Future work should focus on identifying practical predictors for early-stage decision-making in emergency departments while maintaining model accuracy and clinical relevance.

Conclusion

The external validation of the ICD-10–based AI model exhibited excellent performance. The AI model derived from a large nationwide dataset outperformed performance compared to conventional prediction models, despite the significant diversity of each cohort. It appears to serve as both a complement and alternative to the traditional model. Leveraging pre-existing big data is useful for development, validation, and implementation.

Data availability

Data availability statement: The 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.

Code availability

The code and input matrix files for the relevant analysis can be accessed at the following link: https://github.com/LeeSS96/NEDIS-Mortarlity.

Received: 16 August 2024; Accepted: 2 January 2025
Published online: 07 January 2025

References

1. Park, Y. et al. Major causes of preventable death in Trauma patients. *J. Trauma. Inj* **34**, 225–232 (2021).
2. Jung, P. Y. et al. Clinical practice guideline for the treatment of traumatic shock patients from the Korean Society of Traumatology. *J. Trauma. Inj.* **33**, 1–12 (2020).

3. Kim, O. H. et al. Part 2. Clinical Practice Guideline for Trauma Team Composition and Trauma Cardiopulmonary Resuscitation from the Korean Society of Traumatology. *J. Trauma. Inj.* **33**, 63–73 (2020).
4. Baker, S. P., O'Neill, B., Haddon, W. & Long, W. B. The injury severity score: A method for describing patients with multiple injuries and evaluating emergency care. *J. Trauma* **14**, 187–196 (1974).
5. Bergeron, E. et al. Canadian benchmarks in trauma. *J. Trauma* **62**, 491–497 (2007).
6. Jeong, T. S., Choi, D. H., Kim, W. K., Korea Neuro-Trauma Data Bank (KNTDB) Investigators2. The relationship between trauma scoring systems and outcomes in patients with severe traumatic brain injury. *Korean J. Neurotrauma* **18**, 169–177 (2022).
7. Gagné, M., Moore, L., Beaudoin, C., Kuimi, B., Sirois, M. J. & B. L. & Performance of international classification of diseases-based injury severity measures used to predict in-hospital mortality: A systematic review and meta-analysis. *J. Trauma. Acute Care Surg.* **80**, 419–426 (2016).
8. Demetriades, D. et al. TRISS methodology: An inappropriate tool for comparing outcomes between trauma centers. *J. Am. Coll. Surg.* **193**, 250–254 (2001).
9. Ha, M., Yu, S., Lee, J. H., Kim, B. C. & Choi, H. J. Does the probability of survival calculated by the trauma and injury severity score method accurately reflect the severity of neurotrauma patients admitted to regional trauma centers in Korea? *J. Korean Med. Sci.* **38**, e265 (2023).
10. Kang, W. S. et al. Artificial intelligence to predict in-hospital mortality using novel anatomical injury score. *Sci. Rep.* **11**, 23534 (2021).
11. Lee, 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.* **24**, e43757 (2022).
12. McNarry, A. F. & Goldhill, D. R. Simple bedside assessment of level of consciousness: Comparison of two simple assessment scales with the Glasgow Coma scale. *Anaesthesia* **59**, 34–37 (2004).
13. Ryu, J. H. et al. Changes in relative importance of the 5-level triage system, Korean triage and acuity scale, for the disposition of emergency patients induced by forced reduction in its level number: A multi-center registry-based retrospective cohort study. *J. Korean Med. Sci.* **34**, e114 (2019).
14. Moons, K. G. M. et al. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): Explanation and elaboration. *Ann. Intern. Med.* **162**, W1–73 (2015).
15. Agha, R. et al. STROCSS 2019 Guideline: Strengthening the reporting of cohort studies in surgery. *Int. J. Surg.* **72**, 156–165 (2019).
16. Rufibach, K. Use of Brier score to assess binary predictions. *J. Clin. Epidemiol.* **63**, 938–939 (2010).
17. Vickers, A. J., van Calster, B. & Steyerberg, E. W. A simple, step-by-step guide to interpreting decision curve analysis. *Diagn. Progn Res.* **3**, 18 (2019).
18. Wood, D. et al. A unified theory of diversity in ensemble learning. *J. Mach. Learn. Res.* **24**, 1–49 (2023).
19. Brown, G., Wyatt, J., Harris, R. & Yao, X. Diversity creation methods: A survey and categorisation. *Inf. Fusion* **6**, 5–20 (2005).
20. Zhou, Z. H. *Ensemble Methods: Foundations and Algorithms* (CRC Press, 2012).
21. Kuncheva, L. I. & Whitaker, C. J. Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. *Mach. Learn.* **51**, 181–207 (2003).
22. Zhang, T., Nikouline, A., Lightfoot, D. & Nolan, B. Machine learning in the prediction of Trauma outcomes: A systematic review. *Ann. Emerg. Med.* **80**, 440–455 (2022).
23. Gorczyca, M. T., Toscano, N. C. & Cheng, J. D. The trauma severity model: An ensemble machine learning approach to risk prediction. *Comput. Biol. Med.* **108**, 9–19 (2019).
24. Kwon, J. M. et al. Validation of deep-learning-based triage and acuity score using a large national dataset. *PloS ONE* **13**, e0205836 (2018).
25. Collins, G. S. et al. Protocol for development of a reporting guideline (TRIPOD-AI) and risk of bias tool (PROBAST-AI) for diagnostic and prognostic prediction model studies based on artificial intelligence. *BMJ Open* **11**, e048008 (2021).
26. Abujaber, A. et al. Prediction of in-hospital mortality in patients with post traumatic brain injury using National Trauma Registry and machine learning approach. *Scand. J. Trauma. Resusc. Emerg. Med.* **28**, 44 (2020).
27. Matsuo, K. et al. Machine learning to predict in-hospital morbidity and mortality after traumatic brain Injury. *J. Neurotrauma* **37**, 202–210 (2020).
28. Rau, C. S. et al. Mortality prediction in patients with isolated moderate and severe traumatic brain injury using machine learning models. *PLoS ONE* **13**, e0207192 (2018).
29. Ahmed, F. S. et al. A statistically rigorous deep neural network approach to predict mortality in trauma patients admitted to the intensive care unit. *J. Trauma Acute Care Surg.* **89**, 736–742 (2020).
30. Christie, S. A. et al. Machine learning without borders? An adaptable tool to optimize mortality prediction in diverse clinical settings. *J. Trauma. Acute Care Surg.* **85**, 921–927 (2018).
31. Loftis, K. L., Price, J. & Gillich, P. J. Evolution of the abbreviated Injury Scale: 1990–2015. *Traffic Inj Prev.* **19**, S109–S113 (2018).

Acknowledgements

This research was supported by the MSIT (Ministry of Science and ICT), Korea, under the ITRC (Information Technology Research Center) support program (IITP-2024-RS-2024-00438239), the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. RS-2024-00509257, Global AI Frontier Lab), and Chonnam National University Hospital Biomedical Research Institute (BCRI24068).

Author contributions

All the authors wrote the manuscript and created the figure. Concept and design: SSL, DWK, WSK, and JL. Statistical Analysis: SSL, NO, HL, SP, DKY, WSK, and JL. Interpretation of data: SSL, DWK, NO, HL, SP, DKY, WSK, and JL. All authors critically reviewed and agreed to the submission of the final manuscript. SSL and DWK 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.

Declarations

Competing interests

The authors declare no competing interests.

Institutional Review Board Statement

Institutional review board (IRB) approval was obtained from Cheju Halla General Hospital and Chonnam National University Hospital (IRB numbers: CHH-2023-L16-01 and CNUH-2022-L02-01, respectively).

Human ethics and consent to participate declarations

Not applicable. Informed consent was waived due to the study's observational nature and the de-identification of each patient.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-025-85420-5.

Correspondence and requests for materials should be addressed to W.S.K. or J.L.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.