npj | digital medicine

Article

Published in partnership with Seoul National University Bundang Hospital

https://doi.org/10.1038/s41746-025-01714-y

Predicting outcomes after moderate and severe traumatic brain injury using artificial intelligence: a systematic review

Check for updates

Armaan K. Malhotra[1,2,3], Husain Shakil[1,2,3], Christopher W. Smith[1,2], Yu Qing Huang[2,3,4], Jethro C. C. Kwong[5], Kevin E. Thorpe[6,7], Christopher D. Witiw[1,2,3], Abhaya V. Kulkarni[3,8], Jefferson R. Wilson[1,2,3,10] & Avery B. Nathens[3,9,10] ✉

Methodological standards of existing clinical AI research remain poorly characterized and may partially explain the implementation gap between model development and meaningful clinical translation. This systematic review aims to identify AI-based methods to predict outcomes after moderate to severe traumatic brain injury (TBI), where prognostic uncertainty is highest. The APPRAISE-AI quantitative appraisal tool was used to evaluate methodological quality. We identified 39 studies comprising 592,323 patients with moderate to severe TBI. The weakest domains were methodological conduct (median score 35%), robustness of results (20%), and reproducibility (35%). Higher journal impact factor, larger sample size, more recent publication year and use of data collected in high-income countries were associated with higher APPRAISE-AI scores. Most models were trained or validated using patient populations from high-income countries, underscoring the lack of diverse development datasets and possible generalizability concerns applying models outside these settings. Given its recent development, the APPRAISE-AI tool requires ongoing measurement property assessment.

Traumatic brain injury (TBI) is the leading cause of preventable trauma-related morbidity and mortality worldwide[1,2]. Outcome prediction for moderate to severe TBI patients remains a difficult task for clinicians[3–5]. With the advent of computational advancements, the number of automated clinical decision tools leveraging artificial intelligence (AI) has risen exponentially[6,7]. AI is an umbrella term and refers broadly to algorithms that learn from prior experiences and are capable of applying learned patterns to new data in the future. Together, these models have the potential to enhance patient care through improvements in predictive accuracy and identification of novel associations[8,9]. Using AI for TBI outcome prediction has the potential to integrate multimodal data sources to optimize prognostication. However, barriers to clinical translation stem from generalizability concerns and unknown risk of biases, which likely explain the low number of AI-based prediction models that have gained traction in real clinical practice[10,11]. For example, lack of diverse training data may degrade prediction accuracy in specific subpopulations and potentially lead to patient harm if applied to clinical practice without knowledge of biased performance[10,11].

There is a lack of systematically conducted critical appraisal for AI-based prognostic models, resulting in generally limited clinical translation. This has motivated the development of several reporting guidelines for clinical AI model development including the APPRAISE-AI tool, which was designed as a quantitative appraisal instrument that facilitates empirical evaluation of AI-based clinical decision support models with emphasis on model design, validation methodology, clinical utility and patient safety[12].

In this study, we sought to systematically evaluate the quality of AI-based tools developed to prognosticate patients with moderate to severe TBI. We aimed to (1) characterize the methods used to develop the AI tools (study design, validation techniques) and (2) determine the risk of bias,

[1]Division of Neurosurgery, Unity Health, Toronto, ON, Canada. [2]Li Ka Shing Knowledge Institute, Unity Health, Toronto, ON, Canada. [3]Institute for Health Policy, Management and Evaluation, University of Toronto, Toronto, ON, Canada. [4]Division of Geriatric Medicine, Department of Medicine, University of Toronto, Toronto, ON, Canada. [5]Division of Urology, Department of Surgery, University of Toronto, Toronto, ON, Canada. [6]Dalla Lana School of Public Health, University of Toronto, Toronto, ON, Canada. [7]Child Health Evaluative Sciences, Hospital for Sick Children, Toronto, Canada. [8]Division of Neurosurgery, Hospital for Sick Children, Toronto, ON, Canada. [9]Division of General Surgery, Sunnybrook Health Sciences Center, Toronto, ON, Canada. [10]These authors jointly supervised this work: Jefferson R. Wilson, Avery B. Nathens. ✉e-mail: avery.nathens@sunnybrook.ca

threats to validity and reporting standards of published models. Our goal was also to make recommendations that may enhance the quality of ongoing research and increase the likelihood of safe implementation with maximal clinical impact.

Results

Study characteristics

We identified 39 moderate to severe TBI prognostication studies meeting inclusion and exclusion criteria (Supplementary Fig. 1). There were 15 studies (39%) that predicted functional outcome using either the Glasgow Outcome Scale (GOS) or Glasgow Outcome Scale-Extended (GOS-E) measures (median follow-up duration 6 months)[13–27] 13 articles (33%) that predicted mortality[28–40] and 11 articles (28%) that predicted both functional outcome and mortality[41–51]. There was heterogeneity in functional outcome definitions using the GOS and GOS-E, with some studies binarizing these ordinal scales using different thresholds and a minority of studies conducting ordinal regression or treating these as continuous variables (Table 1)[17,22,51]. Among studies predicting both mortality and functional outcome, 3 studies (8%) did not meet functional outcome assessment inclusion criteria of ≥3 months after injury and were therefore considered only for mortality outcomes[46,48,49]. Data collection methods varied with 16 prospective (41%), 21 retrospective (54%) and 2 mixed retrospective and prospective studies (5%). Multicenter data was reported in 18 studies (46%), of which 8 studies (21%) included data from more than one country. The majority of articles were published between 2019–2024 ($n = 31$, 79%) with the remainder from 1997–2016 ($n = 8$, 21%).

Patient characteristics

There were 592,323 patients with moderate to severe TBI managed across over 20 countries in North America, Europe, Asia and the Middle East. Median sample size was 482 patients (interquartile range [IQR] 168–994). Cohorts from high-income countries predominated ($n = 30$, 77%), with a small number of upper middle-income country cohorts ($n = 9$, 23%); there were no low-income cohorts represented. Cohort composition was most frequently adult ($n = 32$, 82%), followed by mixed pediatric and adult ($n = 5$, 13%) and pediatric only ($n = 2$, 5%). There were 34 studies (87%) that reported sex composition; from these studies, the combined proportion of male patients was 62% ($n = 352{,}054/565{,}347$). Mean age was reported in 19 studies (pooled mean 42 years, 95% CI: 37–48, $I^2 > 90\%$) and median age in 21 studies (median age 47, IQR 32–52 years); age was unspecified in 2 studies. Mortality was reported descriptively (not specifically as a primary prediction outcome) in 30 studies at varying follow-up times. Moderate to severe TBI mortality reported within studies ranged from 6% to 64% with a median value of 24% (IQR 16–34%). Intracranial pressure (ICP) monitoring and surgical intervention (craniotomy or craniectomy) rates were infrequently reported with 11 studies (28%) reporting on either procedure type respectively.

Model characteristics

There was heterogeneity regarding AI model architectures, validation methods and comparator models utilized (Tables 1 and 2). Internal validation was the most common method reported with 16 studies utilizing (41%) cross-validation, 14 studies (36%) using random splits and 1 study (3%) utilizing a temporal data split. External validation was performed in only 8 studies (21%). Non-AI comparator multivariable logistic regression predictions were present in 24 studies (62%). Of these 24 studies, 10 studies (26%) used the International Mission for Prognosis and Analysis of Clinical Trials in TBI (IMPACT)[52] prognostic calculator (or IMPACT feature list) and 3 studies (8%) included the Corticoid Randomization after Significant Head Injury (CRASH) prognostic calculator (or CRASH feature list), both of which are previously validated TBI prediction models (multivariable logistic regression models) (Tables 1 and 2)[53]. Only 1 study (3%) obtained human-generated outcome predictions from clinical experts[47]. Absolute performance difference between AI and non-AI model predictions for performance metrics C-index, accuracy, sensitivity and specificity are summarized in Fig. 1. No studies included model equity assessments of pre-defined patient subgroups. Model explainability with variable importance rankings was provided in 28 studies (72%). Among the three most important features identified from each study, common variables were age ($n = 16$ studies, 41%), Glasgow Coma Scale score (GCS) ($n = 12$ studies, 31%), intracranial hematoma data ($n = 8$ studies, 21%) and pupil reactivity ($n = 7$ studies, 18%) (Supplementary Table 1).

Evidence appraisal

Intraclass correlation coefficient (ICC) for reviewers was 0.96 (95% CI: 0.93–0.98) for the overall APPRAISE-AI score, demonstrating excellent agreement. ICC ranged from 0.73–0.90 (moderate to good agreement) across domains (Supplementary Table 2). APPRAISE-AI scores were averaged between reviewers. Median overall score was 46 (out of 100 maximum points; IQR 39–52 points), reflecting an overall moderate study quality. The range of scores was broad from 27 to 66. There were 13 low-quality studies (33%), 21 moderate-quality studies (54%) and 5 high-quality studies (13%). As a fraction of domain-specific maximum points allocated, robustness of results, methodological conduct and reproducibility were the lowest scoring domains with median pooled percent scores of 20%, 35% and 35% respectively (Fig. 2). The two strongest domains were clinical relevance and reporting quality with pooled median scores of 88% and 87% respectively.

Review of individual item scores highlighted relative strength in overall title clarity, background information, problem and target population specification, ground truth definitions, specification of inclusion and exclusion criteria, AI model descriptions, critical analysis of results, acknowledgement of limitations and disclosure reporting (Fig. 3). Major weaknesses across studies were data source descriptions (often single center without explicit reporting of low/middle-income or community/rural patient populations), generally small sample sizes, sample size specification in only two studies, low-quality comparator models (absence of gold standard model, non-AI regression model or human expert predictions), absent bias assessments, lacking predictive error analyses, and poor reproducibility. Qualitative review of performance differences between AI versus non-AI comparator models across APPRAISE-AI score groups highlight a smaller magnitude difference in C-index difference among high-quality studies, with variability across accuracy, sensitivity and specificity metrics (Fig. 1).

Overall APPRAISE-AI scores were higher in studies utilizing data collected in high-income countries compared to upper middle-income countries (two-sample t-test; mean difference = 7.6 points, $p = 0.014$). Univariate linear regression demonstrated that high impact factor publications, sample size over 500 patients and more recent publication year were independently associated with higher mean overall APPRAISE-AI scores (Table 3 and Supplementary Figs. 2–4). After adjustment in a multivariable regression model, high impact factor journal publication, increasing sample size, more recent publication year and data collection in a high-income country were all independently associated with higher APPRAISE-AI overall scores (Table 3).

Discussion

We systematically reviewed studies predicting acute moderate to severe TBI mortality and functional outcomes using AI-based methods. Specific study strengths include strong study clinical rationales, robust ground truth specification, frequent use of functional outcomes and explicit definition of eligibility criteria. There were also notable weaknesses such as lack of sample size calculations, infrequent external validation, absent bias assessment in defined patient subgroups and a minority of studies including open-source data, source code and available models to generate single or bulk predictions. Collectively, these weaknesses threaten the validity, generalizability and potential safety of clinical decision support prediction models. These limitations with existing AI-based prognostic models also likely explain the implementation gap between model development and lack of meaningful clinical deployment. Further, we also demonstrate empirical associations between journal impact factor, study sample size, publication year and

Table 1 | Study details, model development and performance characteristics of studies predicting functional outcome following moderate to severe TBI at ≥3 months from injury

First author and year of publication	Study-specific TBI definition	Functional outcome assessment time	Outcome categorization	Sample size (adult, mixed, or pediatric)	AI model architecture	AI performance (top model performance reported)	Non-AI performance (top model; if existing nomogram or human experts included, all reported)
Jung 2023	Undergoing ICP monitoring in intensive care unit (ICU)	6 months	Functional outcome (GOS-E 1–3 vs. 4–8)	166 (adult)	Light Gradient Boosting Machine (GBM)	AUC: 0.858	Not included
Gravesteijn 2020	GCS ≤ 12	6 months	Functional outcome (GOS 1–3 vs. 4–5 or GOS-E 1–4 vs. 5–8)	12,397 (adult)	Support vector machine (SVM), artificial neural network (ANN), random forest and GBM	GBM: AUC: 0.78	Logistic regression (LR) and least absolute shrinkage and selection operator (LASSO) regression: AUC: 0.77
Hanko 2021	Undergoing primary decompressive craniectomy	6 months	Functional outcome (GOS 1–3 vs. 4–5)	40 (adult)	Random forest	AUC: 0.873	Not included
Lu 2015	Moderate to severe TBI surviving ≥14 days	6 months	Functional outcome (GOS 1–3 vs. 4–5)	115 (adult)	ANN, naïve Bayes, and decision tree	ANN: AUC: 0.961 Sens: 83.5% Spec: 89.7%	LR: AUC: 0.925 Sens: 81.1% Spec: 90.1%
Pease 2022	GCS ≤ 8 admitted to hospital	6 months	Functional outcome (GOS 1–3 vs. 4–5)	757 (adult)	Convolutional neural network (CNN)	Fusion model (clinical and imaging model) AUC: 0.68 Accuracy: 82% Sens: 77.7% Spec: Varied to match each human (mean 80.3%)	International Mission for Prognosis and Analysis of Clinical Trials in TBI (IMPACT) score: AUC: 0.83 Expert humans: Accuracy: 72.7% Sens: 70.3% Spec 80.3%
Arefan 2023	Blunt injury with post-resuscitation GCS ≤ 8	24 months	Functional outcome (GOS 1–3 vs. 4–5)	395 (with complete 24-month outcome data)	SVM, ANN, decision tree, naïve Bayes	SVM: AUC: 0.82 Accuracy: 74%	LR: AUC: 0.82 Accuracy: 71%
Guiza 2013	Undergoing ICP monitoring in ICU	6 months	Functional outcome (GOS 1–3 vs. 4–5)	160 (adult)	Gaussian processes (GP)	Core IMPACT variables with dynamic predictors: AUC: 0.87 Accuracy: 86% Sens: 88% Spec: 86%	LR: AUC: 0.68 Accuracy: 67% Sens: 67% Spec: 67% Corticoid Randomization after Significant Head Injury (CRASH): AUC: 0.67 Accuracy: 66% Sens: 67% Spec: 66% IMPACT: AUC: 0.67 Accuracy: 65% Sens: 72% Spec: 60%
Chen 2024	GCS ≤ 8 admitted to hospital with bloodwork drawn <12 h	6 months	Functional outcome (GOS 1–3 vs. 4–5)	800 (adult)	GBM, ANN, distributed random forest, generalized linear model	Distributed random forest with X1 feature (derived index with subdural hematoma thickness and	LR using CRASH variables (not exact prediction model) AUC: 0.844 LR using IMPACT variables

Table 1 (continued) | Study details, model development and performance characteristics of studies predicting functional outcome following moderate to severe TBI at ≥3 months from injury

First author and year of publication	Study-specific TBI definition	Functional outcome assessment time	Outcome categorization	Sample size (adult, mixed, or pediatric)	AI model architecture	AI performance (top model performance reported)	Non-AI performance (top model; if existing nomogram or human experts included, all reported)
						coagulopathy): AUC: 0.999	(not exact prediction model) AUC: 0.718
Baucher 2019	Acute SDH undergoing craniotomy/burr hole evacuation	6 months	Functional outcome (GOS 1–3 vs. 4–5)	82 (adult)	Classification and regression tree	Accuracy: 76%	Not included
Pourahmad 2019	Abbreviated injury severity score (AIS) score ≥3 in head region	6 months	Functional outcome (GOS-E 1–4 vs. 5–8)	741 (mixed; age >14 included)	SVM (linear kernel) with multiple variable selection methods	Optimal feature selection method: sequential forward selection AUC: 0.737 Accuracy: 79.1% Sens: 58.9% Spec: 89.2%	Not included
Pourahmad 2016	GCS ≤ 10 admitted to ICU	6 months	Functional outcome (GOS-E 1–4 vs. 5–8)	410 (mixed; age >11 included)	Decision tree and hybrid model using decision tree with ANN	Hybrid ANN and decision tree: AUC: 0.705 Sens: 55.1% Spec: 93.6%	Not included
Nourelahi 2022	Severe TBI without further specification	6 months	Functional outcome (GOS-E 1–4 vs. 5–8)	1682 (adult)	Random forest and SVM	SVM: AUC: 0.82 Accuracy: 78% Sens: 78% Spec: 78%	LR: AUC: 0.83 Accuracy: 78% Sens: 78% Spec: 78%
Haveman 2019	Moderate to severe TBI anticipated to stay in ICU > 24 h	12 months	Functional outcome (GOS-E 1–2 vs. 3–8 and 1–4 vs. 5–8; results for latter presented)	57 (adult)	Random forest	Electroencephalography (EEG), ICU and IMPACT variables: AUC: 0.76 Sens: 89% Spec: 80%	IMPACT score: AUC: 0.87 Sens: 100% Spec: 75%
Tewarie 2023	Moderate to severe TBI anticipated to stay in ICU > 24 h	12 months	Functional outcome (GOS-E 1–3 vs. 4–8)	104 (mixed; age >12 included)	Random forest	IMPACT and EEG features: AUC: 0.89 Sens: 83% Spec: 85%	IMPACT score: AUC: 0.81 Sens: 86% Spec: 70%
Pang 2007	Closed injury admitted to ICU with GCS ≤ 8	6 months	Functional outcome (GOS categorized as 3 levels and 5 levels; results for latter provided)	513 (adult)	Discriminant analysis, decision tree, Bayesian network, and ANN	Discriminant analysis: Accuracy: 69%	LR: Accuracy: 62%
Bark 2024	TBI all severities admitted to ICU (83% model development dataset moderate to severe TBI)	6–12 months	Functional outcome (GOS-E 1–4 vs. 5–8 and GOS-E as ordinal variable 1–8; results for latter presented)	1808 (adult)	Random forest & ANN	Top model ANN with IMPACT features: External Dataset 1: Accuracy: 52% Sens: 52% Spec: 75% External Dataset 2: Accuracy: 21% Sens: 21% Spec: 86%	Multivariable proportional odds LR with IMPACT features: External dataset 1: Accuracy: 44% Sens: 44% Spec: 83% External dataset 2: Accuracy: 18% Sens: 18% Spec: 84%

Table 1 (continued) | Study details, model development and performance characteristics of studies predicting functional outcome following moderate to severe TBI at ≥3 months from injury

First author and year of publication	Study-specific TBI definition	Functional outcome assessment time	Outcome categorization	Sample size (adult, mixed, or pediatric)	AI model architecture	AI performance (top model performance reported)	Non-AI performance (top model; if existing nomogram or human experts included, all reported)
Eiden 2019	Post-resuscitation GCS ≤ 8 undergoing ICP monitoring	6 months	Functional outcome (GOS as continuous outcome)	26 (adult)	Partial least squares regression and long-short-term memory network (LSTM)	LSTM: R^2 0.732 (since GOS maintained continuous)	Not included
Stein 2012	Severe TBI undergoing ICP monitoring, excluding severe polytrauma	3 months	Functional outcome (GOS-E 1–4 vs. 5–8)	52 (adult)	Compound covariate predictor, linear discriminant analysis, k-nearest neighbor (KNN) classifiers, and SVM (differing kernels)	SVM: Accuracy: 58%	Not included
Rubin 2019	Severe TBI surviving >24 h (excluding patients with fixed and dilated pupils)	6 months	Functional outcome (GOS 1–3 vs. 4–5)	630 (adult)	Random forest, linear discriminant analysis	Random forest: AUC: 0.83 Sens: 80% Spec: 80%	LASSO regression: AUC: 0.85 Sens/spec not included for comparator model
Farzaneh 2021	Blunt moderate to severe TBI (excluding patients with fixed and dilated pupils)	6 months	Functional outcome (GOS-E 1–4 vs. 5–8)	831 (adult)	eXtreme Gradient Boosting (XGBoost)	XGBoost: Accuracy: 74.9% AUC: 0.809 Sens: 71.3% Spec: 77.5%	Not included
Minoccheri 2022	GCS 4–12 meeting PROTECT trial[72] criteria	6 months	Functional outcome (GOS-E 1–4 vs. 5–8)	833 (adult)	Tropical geometry-based Fuzzy Neural Network[a], XGBoost, random forest, SVM	Random forest: AUC: 0.802 Accuracy: 74.4% Sens: 57.9% Spec not included	Not included
Folweiler 2020	Blunt TBI meeting COBRIT trial criteria	6 months	Functional outcome (GOS-E in unsupervised phenotype cluster analysis)	1598 (adult)	KNN	No outcome prediction performance metric. Reported correlation of phenotype clusters to GOS-E outcomes	
Rizoli 2016	Blunt severe (GCS ≤ 8) TBI without hypovolemic shock surviving > 24 h	6 months	Functional outcome (GOS-E 1–4 vs. 5–8)	1089 (adult)	Decision tree (derived using binary recursive partitioning)	Decision tree: AUC: 0.67 Accuracy: 68.3% Sens: 72.3% Spec: 62.5%	IMPACT (extended): AUC: 0.69 Accuracy: 73.2% Sens: 92.7% Spec: 44.3%

GCS Glasgow Coma Scale, *ICP* intracranial pressure, *SDH* subdural hematoma, *msTBI* moderate to severe TBI, *AIS* abbreviated injury severity, *GBM* gradient boosting machine, *SVM* support vector machine, *CNN/ANN* convolutional neural network/artificial neural network, *LASSO* least absolute shrinkage and selection operator, *KNN* K nearest neighbors, *LR* logistic regression, *sens* sensitivity, *spec* specificity, *EEG* electroencephalography, *LSTM* long-short-term memory network, *GP* Gaussian processes, *XGBoost* eXtreme Gradient Boosting, *IMPACT* International Mission for Prognosis and Analysis of Clinical Trials in TBI, *CRASH* Corticoid Randomization after Significant Head Injury.

[a] Fuzzy neural networks refer to a hybrid modeling approach whereby neural networks incorporate fuzzy logic to handle uncertainty, ambiguity, or imprecise input data—often using fuzzy membership functions, fuzzy rules, and rule inference.

Table 2 | Study details, model development and performance characteristics of studies predicting mortality following moderate to severe TBI

First author and year of publication	TBI definition	Time of mortality assessment	Sample size (adult, mixed, or pediatric)	Model architecture(s)	AI performance (top model performance reported)	Non-AI performance (top model; if existing nomogram or human experts included, all reported)
Yin 2024	GCS ≤ 12, admitted <24 h from injury; patients that died within 24 h excluded)	In-hospital (at discharge)	169 (adult)	Transductive support vector machine (SVM), light gradient boosting machine (GBM), feature tokenizer (FT) transformer, self-supervised learning	Self-supervised learning: AUC: 0.994 Accuracy: 99.4% Sens: 100% Spec: 99.4%	Logistic Regression (LR): AUC: 0.762 Accuracy: 88.2% Sens: 54.5% Spec: 90.5%
Zhang 2023	GCS ≤ 12, admitted to ICU; excluded patients that died within 24 h	In-hospital (at discharge)	482 (adult)	eXtreme Gradient Boosting (XGBoost), light GBM, and FT transformer	Light GBM: AUC: 0.953 Accuracy: 94.8% Sens: 86.8% Spec: 95.8%	LR: AUC: 0.813 Accuracy: 86.7% Sens: 58.5% Spec: 90.2%
Matsuo 2020	Blunt injury, abnormal CT head requiring admission (not strict msTBI, but median GCS 10)	In-hospital (at discharge)	232 (mixed; age >10 included)	SVM (multiple kernels), Gaussian/ multinomial naïve Bayes, ridge regression, extra trees, random forest, least absolute shrinkage and selection operator (LASSO) regression, GBM, decision tree	Random forest: AUC: 0.818 Accuracy: 95.5% Sens: 63.6% Spec: 100%	Not included
Jung 2023	Undergoing ICP monitoring in ICU	6 months	166 (adult)	LightGBM	AUC: 0.893	Not included
Gravesteijn 2020	GCS ≤ 12	6 months	12397 (adult)	SVM, artificial neural network (ANN), random forest and GBM	GBM: AUC: 0.83	LR and LASSO regression: AUC: 0.82
Hanko 2021	Undergoing primary decompressive craniectomy	6 months	40 (adult)	Random forest	AUC: 0.811	Not included
Lu 2015	Moderate to severe TBI surviving ≥14 days	6 months	115 (adult)	ANN, naïve Bayes, and decision tree	Naïve Bayes: AUC: 0.901 Sens: 81.2% Spec: 90.7%	LR: AUC: 0.873 Sens: 68.4% Spec: 91.0%
Pease 2022	GCS ≤ 8 admitted to hospital	6 months	757 (adult)	Convolution neural network (CNN)	Fusion model (clinical and imaging model) AUC: 0.80 Accuracy: 86% Sens: 78% Spec: Varied to match each human (mean 81.7%)	International Mission for Prognosis and Analysis of Clinical Trials in TBI (IMPACT) score: AUC: 0.83 Expert humans: Accuracy: 71.3% Sens: 55% Spec: 81.7%
Arefan 2023	Blunt injury with post-resuscitation GCS ≤ 8	24 months	395 (with complete 24-month outcome data)	SVM, ANN, decision tree, naïve Bayes	SVM: AUC: 0.85 Accuracy: 79%	LR: AUC: 0.85 Accuracy: 79%
Daley 2022	GCS ≤ 8 and Abbreviated injury severity score (AIS) ≥ 4 admitted to ICU	1 month (in-hospital)	196 (pediatric)	Random forest	AUC: 0.90	Not included
Cui 2021	Undergoing decompressive craniectomy, surviving hospitalization	12 months	230 (adult)	Random forest and LR with synthetic minority oversampling for training data	Random forest: AUC: 0.83 Accuracy: 81% Sens: 83.3% Spec: 80%	LR: AUC: 0.765 Accuracy: 67.2% Sens: 100% Spec: 52.5%
Wang 2022	GCS ≤ 12 arriving <6 h from injury	In-hospital	368 (adult)	XGBoost	AUC: 0.955 Accuracy: 95.5% Sens: 94.5% Spec: 96.4%	LR: AUC: 0.805 Accuracy: 70.3% Sens: 73.8% Spec: 75.4%

Table 2 (continued) | Study details, model development and performance characteristics of studies predicting mortality following moderate to severe TBI

First author and year of publication	TBI definition	Time of mortality assessment	Sample size (adult, mixed, or pediatric)	Model architecture(s)	AI performance (top model performance reported)	Non-AI performance (top model; if existing nomogram or human experts included, all reported)
Feng 2019	Severe TBI without further specification	Not well defined (timing poorly defined)	117 (adult)	22 models total including SVM, KNN, tree-based models, GBM	Linear SVM: AUC: 0.94 Accuracy: 92.5% Sens: 97% Spec: 58%	LR: AUC: 0.84 Accuracy: 87.7% Sens: 91% Spec: 65%
Wu 2023	Clinical indication of TBI undergoing CT with GCS ≤ 8	In-hospital	3917 (adult)	XGBoost, SVM	XGBoost (26-feature model): AUC: 0.87 Sens: 78% Spec: 82%	LASSO LR: AUC: 0.86 Sens: 72% Spec: 85% IMPACT and Corticoid Randomization after Significant Head Injury (CRASH) models no AUC/clinical utility included
Raj 2019	ICP monitoring for ≥24 h	30 days	472 (adult)	Dynamic LR with rolling time windows (coefficient C tuned as a hyperparameter)	Dynamic LR features selected were ICP, mean arterial pressure (MAP), cerebral perfusion pressure (CPP), GCS: AUC: 0.72 day 1; 0.84 day 5 Sens: 79% Spec: 94%	LR using IMPACT score variables: AUC: 0.78 Sens: 100% Spec: 50%
vanderPloeg 2016	Moderate to severe TBI not further specified	6 months	11026 (adult)	Classification and regression trees, random forests, SVM and ANN	Random forest with IMPACT core, laboratory and imaging features: AUC: 0.735	LR using IMPACT score variables: AUC: 0.757
Mekkodathil 2023	Blunt TBI direct to trauma center; not strict GCS definition but >90% msTBI	In-hospital	922 (mixed)	SVM, random forest, XGBoost	SVM: AUC 0.86 Accuracy: 81% Sens: 70% Spec: 81%	LR: AUC: 0.84 Accuracy: 80% Sens: 71% Spec: 80%
Yao 2020	GCS 4-12 according to the Progesterone for Traumatic Brain Injury, Experimental Clinical Treatment (PROTECT) trial[72] criteria	6 months	828 (adult)	CNN for automated hematoma segmentation, random forest for outcome prediction	Optimal model used IMPACT variables without computed tomography (CT) features & automated CT volume: AUC: 0.853 Sens: 70% Spec: 87.2%	LR using IMPACT score variables: AUC: 0.795 Sens: 71.2% Spec: 74.7%
Lang 1997	GCS ≤ 8 for 24 h post admission	6 months	1066 (adult)	ANN	Accuracy: 85.8% Sens: 90.6% Spec: 77.1%	LR: Accuracy: 83.1% Sens: 87.7% Spec: 75%
Fonseca 2022	GCS ≤ 12 admitted to ICU for ≥24 h or had neurosurgical intervention	In-hospital	300 (pediatric)	XGBoost, ANN, k-nearest neighbors (KNN), random forest	XGBoost: AUC: 0.91	Not included
Raj 2022	Undergoing ICP monitoring	1 month	1324 (adult)	Dynamic LR with rolling time windows (coefficient C tuned as a hyperparameter)	External Dataset 1: AUC 0.66 first 24 h; AUC 0.79 at 120 h Accuracy 120 h: 90.3% Sens 24/120 h: 93%/93% Spec 24/120 h: 79.7%/ 97.6% External dataset 2: 0.67 first 24 h; AUC 0.73 at 120 h Accuracy 120 h: 75.9% Sens 120 h: 77%/ 77% Spec 24/ 120 h: 89.7%/ 98.9%	Not included

Table 2 (continued) | Study details, model development and performance characteristics of studies predicting mortality following moderate to severe TBI

First author and year of publication	TBI definition	Time of mortality assessment	Sample size (adult, mixed, or pediatric)	Model architecture(s)	AI performance (top model performance reported)	Non-AI performance (top model; if existing nomogram or human experts included, all reported)
Cao 2023	Head region AIS ≥ 3 and AIS ≤ 1 in other regions (excluding AIS = 6)	In-hospital	545,388 (adult)	XGBoost Cox proportional hazard regression	AUC: 0.896 Mean time-dependent AUC: 0.713	Not included
Stein 2012	Severe TBI undergoing ICP monitoring, excluding severe polytrauma	In-hospital	52 (adult)	Compound covariate predictor, linear discriminant analysis, KNN classifiers, and SVM (differing kernels)	Top model KNN: Accuracy: 88%	Not included
Bark 2024	TBI all severities admitted to ICU (83% model development dataset moderate to severe TBI)	6–12 months	1808 (adult)	Random forest & ANN	Top model random forest with IMPACT features for External Dataset 1: AUC: 0.86 Top model ANN with IMPACT features for External Dataset 2: AUC: 0.69	Multivariable proportional odds LR with IMPACT features: External dataset 1: AUC: 0.88 External dataset 2: AUC: 0.63

GCS Glasgow Coma Scale, *ICP* intracranial pressure, *ICU* intensive care unit, *SDH* subdural hematoma, *msTBI* moderate to severe TBI, *AIS* abbreviated injury severity, *GBM* gradient boosting machine, *SVM* support vector machine, *CNN/ANN* convolutional neural network/artificial neural network, *LASSO* least absolute shrinkage and selection operator, *KNN* K nearest neighbors, *LR* logistic regression, *XGBoost* eXtreme gradient boosting, *MAP* mean arterial pressure, *CPP* cerebral perfusion pressure, *IMPACT* International Mission for Prognosis and Analysis of Clinical Trials in TBI, *CRASH* Corticoid Randomization after Significant Head Injury, *PROTECT* Progesterone for Traumatic Brain Injury, Experimental Clinical Treatment, *CT* computed tomography.

World Bank country classifications with overall APPRAISE-AI scores. There was a lack of studies representing low- and middle-income cohorts as well as rural or community-dwelling patient populations. This systematic lack of representation could culminate in worsening performance and potential for biased predictions if these models were applied in low- and middle-income country healthcare systems, where local injury epidemiology, care processes and treatment timing may differ. Our findings underscore a current lack of AI-based TBI prognostication models for low- and middle-income patient populations.

Three APPRAISE-AI domains had median scores within the low range across identified studies including methodological conduct, robustness of results, and reproducibility. Interestingly, similar low-quality domains have been previously identified in a methodological review of AI-based bladder cancer prognostication models, suggesting consistency and potential relevance of these inferences across more general medical prediction models[54]. The following focused recommendations consider domain-specific and item-specific scores to maximize the yield of targeted methodological improvements within AI-based TBI prediction models.

From a methodological conduct perspective, sample size calculations were reported in only 2 of the 39 included studies[28,51]. Lack of sample size specification may introduce biased estimation in the setting of insufficient event rates relative to candidate features used in modeling. In this setting, conclusions made about model performance may be attributable to study power, rather than data handling, model training or AI architectures (higher chance of type I or II errors)[55,56]. We demonstrated an association between sample size and APPRAISE-AI composite score, providing evidence to further support this claim. Prior simulation studies have demonstrated machine learning modeling approaches are data hungry and often need over 10 times the number of events per variable compared to logistic regression, meaning the necessary sample size for appropriate clinical prediction model development may be much larger than was reported in most of the included studies[57]. In addition to sample size, selection of a comparator model affords investigators an internal control to benchmark AI modeling choices and rigorously demonstrate they improved prediction performance compared to an existing gold standard. In this review, there were 15 studies that lacked the presence of a comparator model, 1 study that used a team of clinician experts and only 10 studies that included a previously validated prediction model for moderate to severe TBI (CRASH or IMPACT score)[52,53,58]. Specific steps to enhance methodological conduct in future work would be a priori sample size estimation, use of data collected from diverse patient populations and inclusion of comparator models as benchmarks when evaluating AI model performance.

The two items driving low scores for robustness of results were limited bias assessments and lack of error analyses. A major safeguard against unintended patient harm remains robust exploration of clinically relevant subgroups and task-specific applications. Out of 39 included studies, 27 did not investigate task-specific or subgroup-specific discrimination and clinical utility. In TBI patients, clinically relevant mismatch in clinical exam findings (ex: GCS) and severity of injury has been well-established for older adults[59]. Lack of age-stratified performance assessment across age strata is one potential threat to clinical workflow integration[60]. Similarly, 32 studies did not conduct a predictive or surprise error analysis through review of misclassified results. This can be a useful method to build model trust, understand potential impacts of model deployment and identify features influencing decision-making that may be nonsensical based on conventional clinical knowledge (such as detection of a ruler to define a malignant skin lesion)[61,62].

Transparency and reproducibility were low across included studies due to infrequent inclusion of data dictionaries, source code, publication of models that make single or bulk predictions and low rates of data availability (or specification of data access procedures). Investigators and journals should endeavor to provide scientific audiences with this information to facilitate maximum model usage, feedback and troubleshooting. This may take the form of accessible source code or an available trained model that other researchers can use to make individual predictions. Further, from a

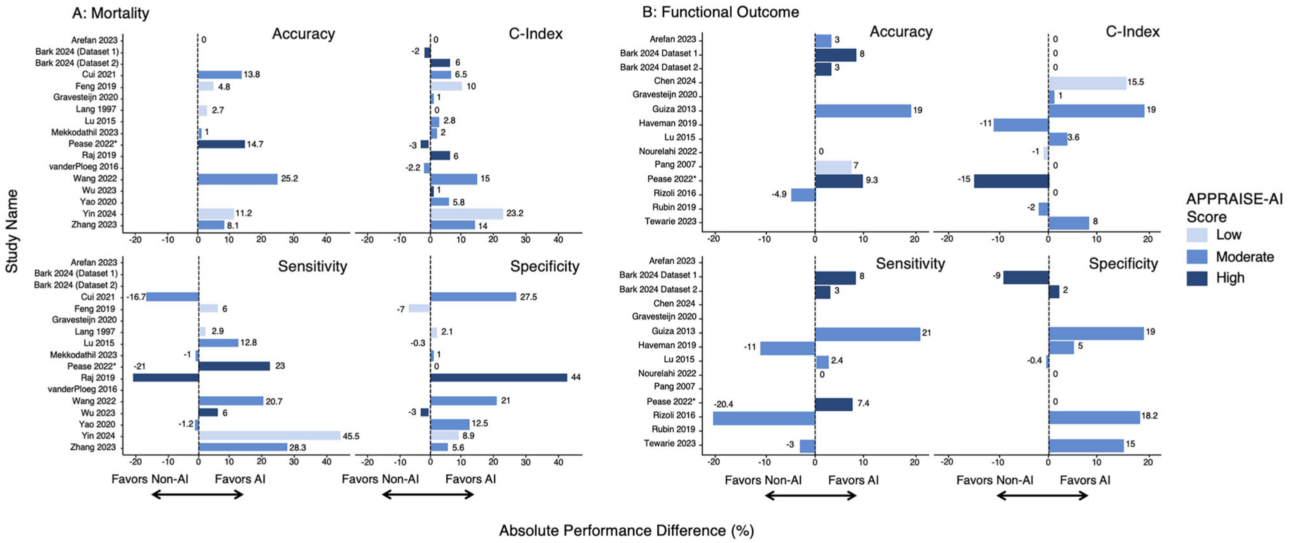

Fig. 1 | Absolute performance difference between AI and non-AI models across c-index (AUC), accuracy, sensitivity and specificity (reported as %). Positive absolute performance difference values mean AI model performance was higher than non-AI model for the given metric. Stratification corresponds to study-specific APPRAISE-AI score (low, moderate or high). **A**, **B** depict results for studies predicting mortality and functional outcome respectively. Note: absolute performance differences reflect comparisons of study-specific performance point estimates, not confidence intervals, which were inconsistently reported in included studies and models. Listed comparisons between AI and non-AI models may therefore overstate performance differences due to unreported confidence intervals quantifying uncertainty. Pease 2022 accuracy, sensitivity and specificity results from AI compared to average of three human experts (neurosurgeons).

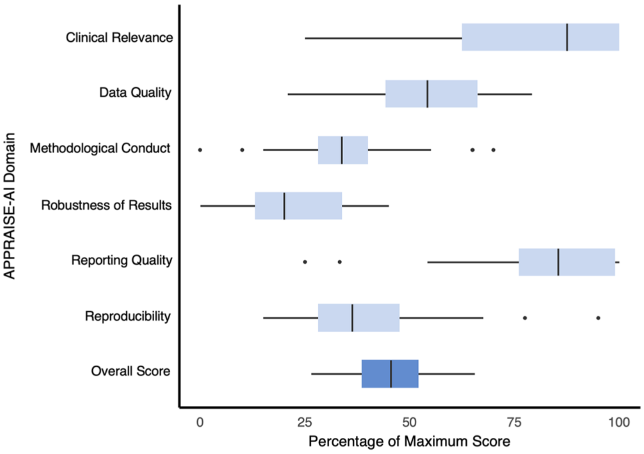

Fig. 2 | Box plot depicting consensus APPRAISE-AI domain-specific scores, and overall scores determined from review of included studies ($n = 39$). Scores were normalized as a proportion of the maximum domain-specific or overall score (percentages). Vertical bars show median values, boxes demonstrate interquartile range (25th to 75th percentile) and whiskers the bounds of 5th and 95th percentiles. Outliers are shown as individual points.

model specification perspective, investigators can enhance reproducibility by also explicitly outlining hyperparameter tuning steps (such as whether a grid search or random search was used, which final hyperparameters were selected and ranges of hyperparameter searching) and final model specifications. These steps, which ultimately determine final model parameters, are essential for other investigators to attempt to fully understand the model development process.

The current regulatory landscape of AI-based decision support systems in healthcare is complex and evolving. While we don't aim to provide a comprehensive overview of regulatory frameworks, most of these prognostic TBI models would fall into the category of Software as a Medical Device (SaMD). Given the complex nature of prognostication in TBI, these models would be decision support tools if implemented, where the ultimate decision rests with the treating clinical team, not any direct agency for an AI model. The Food and Drug Administration in the United States introduced the Proposed Regulatory Framework for Modifications to AI/ML-based SaMD, which outlines responsibilities for developers such as a requirement to monitor the performance of models in the real world[63]. Some of the highlighted weaknesses identified in existing AI-based TBI models in this review highlight crucial model development areas that may limit real-world performance; these results will be useful for regulators and model developers alike. In fact, our findings agree with several key considerations outlined by the recent World Health Organization *Regulatory considerations on artificial intelligence for health* including an emphasis on transparency, robust external validation and adherence to high data quality standards[64].

This study should be interpreted in light of a few limitations. We were unable to perform an across-study meta-analysis of AI and non-AI model performance owing to heterogeneity in outcome definitions, study designs, model architectures and included features. As a result, there is no pooled estimate for task-specific prediction performance across studies, only study-specific absolute performance differences where possible. Notably, these absolutely performance differences (shown in Fig. 1) compared study-specific performance point estimates, and did not incorporate confidence intervals since these were variably reported across studies. We additionally acknowledge the APPRAISE-AI tool has been recently introduced, and there remains limited validation of its measurement properties such as construct validity and responsiveness, temporal validity with evolving AI research standards and floor/ceiling effects with scoring. As ongoing psychometric evaluation of these measurement properties continue, there may be ongoing item modification or alterations to domain score weights. We also recommend future additional validation of APPRAISE-AI, especially with the rapidly evolving field of medical AI, which may quickly advance beyond existing appraisal instruments. The majority of included studies utilized tabular clinical data only, meaning that the multimodal benefits of AI may have been underutilized. We also did not encounter applications of natural language processing, despite explicitly including this as a search term. Finally, if a component of the APPRAISE-AI tool was completed, but not reported by the authors, we may have underestimated study quality (e.g. if a study included low-income patients, but this was not specified). To ensure fairness of methodological appraisals, we limited our review to only accessible information contained within studies, rather than contacting authors specifically, which may have yielded higher scores.

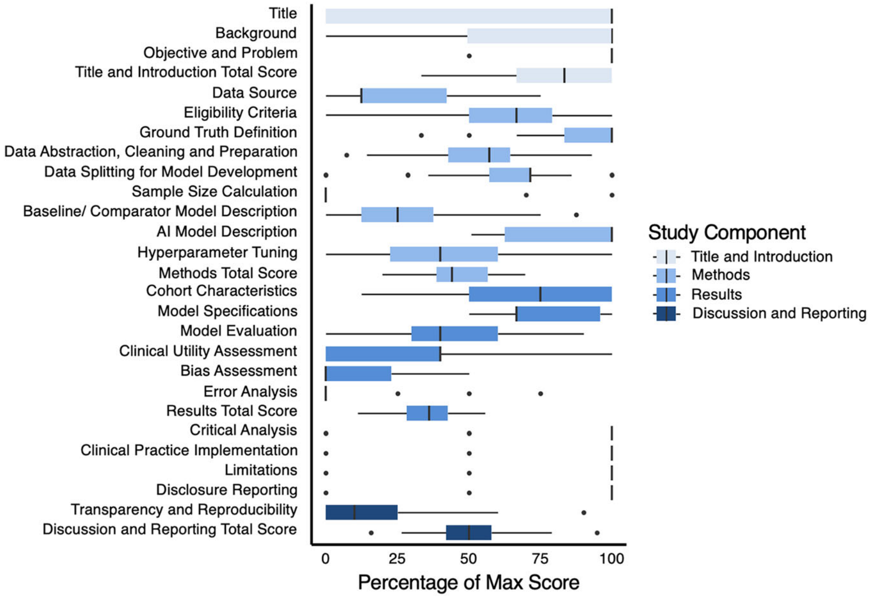

Fig. 3 | Box plot depicting individual APPRAISE-AI item-specific scores across study components determined from review of included studies (*n* = 39). Scores were normalized as a proportion of the maximum item-specific score (percentages). Vertical bars show median values, boxes demonstrate interquartile range (25th to 75th percentile; no range shown if score distribution for item is narrow) and whiskers the bounds of 5th and 95th percentiles. Outliers are shown as individual points.

The findings of this systematic review have the potential to increase methodological rigor of neurotrauma research, enhance model translation to clinical settings and reduce potential patient harm. Future prognostic neurotrauma research could specifically benefit from evaluation of model performance in pre-specified patient subgroups, explicit consideration of sample size requirements, evaluation of AI models against comparator benchmark prediction models and release of open-source models to maximize transparency.

Methods

The following systematic review was performed adhering to the CHARMS[65] (Checklist for Critical Appraisal and Data Extraction for Systematic Reviews of Prediction Modeling Studies) for data extraction and PRISMA[66] (Preferred Reporting Items for Systematic Reviews and Meta-analyses) guidelines for reporting (Supplementary Note 1). Our search strategy and protocol were registered in advance through PROSPERO (CRD42024553288).

Search strategy

We searched electronic databases including OVID Medline, Embase, Scopus, Web of Science and Evidence-Based Medicine Reviews-Cochrane library from inception to June 1, 2024. The search strategy utilized Boolean operators across three content domains: (1) traumatic brain injury, (2) artificial intelligence/machine learning (subfield of AI) and (3) prognostication (Supplementary Note 2).

Eligibility criteria

Study inclusion criteria included: (1) original studies that reported on AI-based models for patients with acute moderate to severe TBI, where prognostic uncertainty is greatest[4,5], defined by GCS < 13, inclusion ≤7 days from injury; (2) peer-reviewed journal articles; (3) sample size ≥10 patients; (4) prediction of a future outcome state (prognostic studies). Considering our interest in prediction models, a sample of ten patients would be minimally sufficient (though still very limited) to assess a single predictor, hence was decided as an inclusion criterion[67]. A sample size of ten patients was felt to be the minimum number of patients required to potentially evaluate a single predictor variable, as suggested by simulation studies of logistic regression prediction models[68]. There were two early post hoc protocol modifications. To ensure outcome homogeneity, we narrowed the outcome definition to include either death or functional status. Functional outcome assessment must occur a minimum ≥3 months from injury since earlier assessments are highly susceptible to under-estimating recovery trajectories (overly nihilistic)[3].

There was no restriction imposed regarding age, presence of extracranial injuries or type of data input (e.g. ICP monitoring, neuroimaging, or tabular clinical data as examples). We defined AI as utilizing computer systems to replicate human-like cognitive processing tasks for clinical decision support. A wide range of AI model architectures were permitted including tree-based models, transformers, neural networks, support vector machines and natural language processing. Studies reporting on patients with concussion or mild TBI only were excluded, unless ≥75% of the included cohort had moderate or severe TBI. We excluded reviews, commentaries, editorials, conference abstracts or proceedings, meta-analyses, and case reports or series of fewer than ten patients. We also excluded studies that did not report on development of a moderate to severe TBI subgroup-specific prognostic tool (i.e. if other acute brain injuries were included). Diagnostic studies alone (such as detection of intracranial hemorrhage) were excluded in addition to studies using AI methods to identify a covariate for a conventional regression analysis.

Screening and data collection

Abstract screening and full-text review were conducted by two independent authors. The following data were extracted from each study: study type, country of data collection, sample size, baseline characteristics such as age, sex, outcome characteristics including follow-up duration, outcome type and counts, as well as model details validation method, dataset size, architecture and presence of specified subgroup or fairness testing (to assess model equity). Continuous variable means with standard deviation or median with IQR and categorical counts were collected. Pooled mean age was determined directly using random effects meta-analysis with a maximum likelihood estimator from studies reporting means with standard deviation due to anticipated high heterogeneity[69].

Additional metrics of model performance were collected including area under receiver operating characteristic curve (AUC-ROC), accuracy, sensitivity and specificity. For studies comparing AI models to non-AI models (statistical models such as the previously validated IMPACT prognostic score[52] or clinician judgment), we determined absolute performance difference for reported metrics. In studies that reported performance metrics across different cohorts, we prioritized use of estimates according to the following hierarchy: external validation groups, testing (internal validation) and training[54]. We additionally collected the top three influential variables from variable importance rankings reported in included studies.

Table 3 | Linear regression results highlighting univariate associations between sample size (categorized as <100, 100–500, 500–1000, >1000), journal impact factor (categorized as <3, 3–5, 5–10 and >10) and publication year with APPRAISE-AI overall scores

Variable	Mean difference in APPRAISE-AI overall score	95% confidence interval	p value	R^2
Univariate linear regression				
Impact factor [Reference IF < 3]			**0.026**	0.21
IF 3–4.9	5.8	−1.5–13.1		
IF 5–9.9	8.1	−0.97–17.1		
IF > 10	**15.5**	**4.5–26.5**		
Sample size [Reference: <100 patients]			**0.002**	0.31
100–499 patients	8.4	−1.0–17.7		
500–1000 patients	**10.3**	**0.2–20.4**		
>1000 patients	**18.3**	**8.4–28.2**		
Publication year	**0.65**	**0.04–1.27**	**0.034**	0.11
Multivariable linear regression				
Impact factor [Reference IF < 3]			**0.041**	0.65
IF 3–4.9	0.49	−5.9–6.9		
IF 5–9.9	5.9	−2.6–12.3		
IF > 10	**10.4**	**0.8–20.0**		
Sample size [Reference: <100 patients]			**<0.001**	
100–499 patients	**11.4**	**3.6–19.1**		
500–1000 patients	**8.8**	**0.09–17.6**		
>1000 patients	**18.7**	**10.9–26.6**		
Publication year (per 1-year increase)	**0.70**	**0.23–1.16**	**0.002**	
Country of data collection [high-income country with reference to upper middle-income country]	**8.6**	**1.7–15.4**	**0.011**	

Multivariable linear regression model was additionally adjusted for country of data collection (high-income compared to upper-middle income). *p* values were determined for variables using likelihood ratio testing (to determine overall association, rather than association per variable level). Publication year ranged from 1997 to 2024 in the study. In the regression model, publication year was kept as a continuous variable and centered around 0. Bolded values represent statistically significant associations with $p < 0.05$.

Quality assessment

The APPRAISE-AI tool is comprised of 24 items summing to a maximum total score of 100 across six domains: clinical relevance, data quality, methodological conduct, robustness of results, reporting quality, and reproducibility[54]. Score interpretation follows pre-defined ranges from 0–19 (very low-quality), 20–39 (low-quality), 40–59 (moderate-quality), 60–79 (high-quality) and 80–100 (very high-quality). Use of this tool provides investigators the resolution to examine domain-specific study weaknesses as well as compare global between-study scores quantitatively.

Study risk of bias assessment was evaluated independently by two reviewers, each of whom was an expert in the field of brain injury and AI methodology. To measure agreement, we determined interrater reliability using the ICC. Since two reviewers independently scored each article at a single time, a two-way random effects ICC with absolute agreement was reported. Interrater reliability was poor, moderate, good or excellent according to the following thresholds: <0.50, 0.50–0.75, 0.75–0.90 and >0.90[70].

To assess the relationship between country of data collection and APPRAISE-AI scores (country categorizations as high-income, upper-middle-income and low or lower-middle-income per World Bank groupings)[71], we used a two-sample *t*-test assuming unequal variance (Welch's *t*-test) since there were only two country designations identified. To assess whether APPRAISE-AI scores changed with study sample size, journal impact factor or year of publication, we performed univariate linear regressions to quantify these relationships. We then constructed a multivariable regression model to obtain adjusted associations between above-mentioned variables with APPRAISE-AI scores. Multicollinearity was assessed, and variance inflation factor was <3 for all variables.

All other statistical analyses and plotting were performed using packages available through R Statistical Programming (V.4.2.1) with two-sided values for statistical significance less than 0.05.

Data availability

Data were extracted from peer-reviewed published articles and are accessible. Any additional data used and APPRAISE-AI scoring forms analyzed during the current study can be made available from the corresponding author on reasonable request.

Code availability

The underlying code for this study is not publicly available but may be made available to qualified researchers on reasonable request from the corresponding author.

Received: 28 November 2024; Accepted: 10 May 2025;
Published online: 18 June 2025

References

1. Dewan, M. C. et al. Estimating the global incidence of traumatic brain injury. *J. Neurosurg.* **130**, 1080–1097 (2018).
2. Maas, A. I. R. et al. Traumatic brain injury: progress and challenges in prevention, clinical care, and research. *Lancet Neurol.* **21**, 1004–1060 (2022).
3. McCrea, M. A. et al. Functional outcomes over the first year after moderate to severe traumatic brain injury in the prospective, longitudinal TRACK-TBI study. *JAMA Neurol.* **78**, 982–992 (2021).
4. Malhotra, A. K. et al. Withdrawal of life-sustaining treatment for pediatric patients with severe traumatic brain injury. *JAMA Surg.* **159**, 287–296 (2024).
5. Malhotra, A. K. et al. Admitting hospital influences on withdrawal of life-sustaining treatment decision for patients with severe traumatic brain injury. *Neurosurgery* https://doi.org/10.1227/neu.0000000000002840 (2024).
6. Hibi, A. et al. Automated identification and quantification of traumatic brain injury from CT scans: are we there yet?. *Medicine* **101**, e31848 (2022).
7. Smith, C. W. et al. Vision transformer–based decision support for neurosurgical intervention in acute traumatic brain injury: automated surgical intervention support tool. *Radiology Artif. Intell.* **6**, e230088 (2024).
8. Lin, E. & Yuh, E. L. Computational approaches for acute traumatic brain injury image recognition. *Front. Neurol.* **13**, 791816 (2022).
9. Khera, R. et al. AI in medicine—JAMA's focus on clinical outcomes, patient-centered care, quality, and equity. *JAMA* **330**, 818–820 (2023).
10. Yu, A. C. & Eng, J. One algorithm may not fit all: how selection bias affects machine learning performance. *RadioGraphics* **40**, 1932–1937 (2020).
11. Andaur Navarro, C. L. et al. Risk of bias in studies on prediction models developed using supervised machine learning techniques: systematic review. *BMJ* **375**, n2281 (2021).
12. Kwong, J. C. C. et al. APPRAISE-AI tool for quantitative evaluation of AI studies for clinical decision support. *JAMA Netw. Open* **6**, e2335377 (2023).
13. Baucher, G. et al. Predictive factors of poor prognosis after surgical management of traumatic acute subdural hematomas: a single-center series. *World Neurosurg.* **126**, e944–e952 (2019).

14. Chen, L. et al. The role of coagulopathy and subdural hematoma thickness at admission in predicting the prognoses of patients with severe traumatic brain injury: a multicenter retrospective cohort study from China. *Int. J. Surg*. https://doi.org/10.1097/JS9.0000000000001650 (2024).
15. Eiden, M. et al. Discovery and validation of temporal patterns involved in human brain ketometabolism in cerebral microdialysis fluids of traumatic brain injury patients. *EBioMedicine* **44**, 607–617 (2019).
16. Farzaneh, N., Williamson, C. A., Gryak, J. & Najarian, K. A hierarchical expert-guided machine learning framework for clinical decision support systems: an application to traumatic brain injury prognostication. *NPJ Digital Med.* **4**, 78 (2021).
17. Folweiler, K. A., Sandsmark, D. K., Diaz-Arrastia, R., Cohen, A. S. & Masino, A. J. Unsupervised machine learning reveals novel traumatic brain injury patient phenotypes with distinct acute injury profiles and long-term outcomes. *J. Neurotrauma* **37**, 1431–1444 (2020).
18. Guiza, F., Depreitere, B., Piper, I., Van den Berghe, G. & Meyfroidt, G. Novel methods to predict increased intracranial pressure during intensive care and long-term neurologic outcome after traumatic brain injury: development and validation in a multicenter dataset. *Crit. Care Med.* **41**, 554–564 (2013).
19. Haveman, M. E. et al. Predicting outcome in patients with moderate to severe traumatic brain injury using electroencephalography. *Crit. Care* **23**, 401 (2019).
20. Minoccheri, C. et al. An interpretable neural network for outcome prediction in traumatic brain injury. *BMC Med. Inform. Decis. Mak.* **22**, 203 (2022).
21. Nourelahi, M., Dadboud, F., Khalili, H., Niakan, A. & Parsaei, H. A machine learning model for predicting favorable outcome in severe traumatic brain injury patients after 6 months. *Acute Crit. Care* **37**, 45–52 (2022).
22. Pang, B. C. et al. Hybrid outcome prediction model for severe traumatic brain injury. *J. Neurotrauma* **24**, 136–146 (2007).
23. Pourahmad, S., Hafizi-Rastani, I., Khalili, H. & Paydar, S. Identifying important attributes for prognostic prediction in traumatic brain injury patients. A hybrid method of decision tree and neural network. *Methods Inf. Med.* **55**, 440–449 (2016).
24. Pourahmad, S., Rasouli-Emadi, S., Moayyedi, F. & Khalili, H. Comparison of four variable selection methods to determine the important variables in predicting the prognosis of traumatic brain injury patients by support vector machine. *J. Res. Med. Sci.* **24**, 97 (2019).
25. Rubin, M. L., Yamal, J. M., Chan, W. & Robertson, C. S. Prognosis of six-month glasgow outcome scale in severe traumatic brain injury using hospital admission characteristics, injury severity characteristics, and physiological monitoring during the first day post-injury. *J. Neurotrauma* **36**, 2417–2422 (2019).
26. Tewarie, P. K. B. et al. Early EEG monitoring predicts clinical outcome in patients with moderate to severe traumatic brain injury. *NeuroImage Clin.* **37**, 103350 (2023).
27. Rizoli, S. et al. Early prediction of outcome after severe traumatic brain injury: a simple and practical model. *BMC. Emerg. Med.* **16**, 32 (2016).
28. Cao, Y., Forssten, M. P., Sarani, B., Montgomery, S. & Mohseni, S. Development and validation of an XGBoost-algorithm-powered survival model for predicting in-hospital mortality based on 545,388 isolated severe traumatic brain injury patients from the TQIP database. *J. Pers. Med.* **13**, https://doi.org/10.3390/jpm13091401 (2023).
29. Cui, W. et al. Death after discharge: prognostic model of 1-year mortality in traumatic brain injury patients undergoing decompressive craniectomy. *Chin. Neurosurg. J.* **7**, 24 (2021).
30. Daley, M. et al. Pediatric severe traumatic brain injury mortality prediction determined with machine learning-based modeling. *Injury* **53**, 992–998 (2022).
31. Feng, J.-Z. et al. Comparison between logistic regression and machine learning algorithms on survival prediction of traumatic brain injuries. *J. Crit. Care* **54**, 110–116 (2019).
32. Fonseca, J., Liu, X., Oliveira, H. P. & Pereira, T. Learning models for traumatic brain injury mortality prediction on pediatric electronic health records. *Front. Neurol.* **13**, 859068 (2022).
33. Lang, E. W., Pitts, L. H., Damron, S. L. & Rutledge, R. Outcome after severe head injury: an analysis of prediction based upon comparison of neural network versus logistic regression analysis. *Neurol. Res.* **19**, 274–280 (1997).
34. Mekkodathil, A., El-Menyar, A., Naduvilekandy, M., Rizoli, S. & Al-Thani, H. Machine learning approach for the prediction of in-hospital mortality in traumatic brain injury using bio-clinical markers at presentation to the emergency department. *Diagnostics* **13** https://doi.org/10.3390/diagnostics13152605 (2023).
35. Raj, R. et al. Machine learning-based dynamic mortality prediction after traumatic brain injury. *Sci. Rep.* **9**, 17672 (2019).
36. Raj, R. et al. Dynamic prediction of mortality after traumatic brain injury using a machine learning algorithm. *npj Digital Med.* **5**, 96 (2022).
37. van der Ploeg, T., Nieboer, D. & Steyerberg, E. W. Modern modeling techniques had limited external validity in predicting mortality from traumatic brain injury. *J. Clin. Epidemiol.* **78**, 83–89 (2016).
38. Wang, R., Wang, L., Zhang, J., He, M. & Xu, J. XGBoost machine learning algorism performed better than regression models in predicting mortality of moderate-to-severe traumatic brain injury. *World Neurosurg.* **163**, e617–e622 (2022).
39. Wu, X. et al. Mortality prediction in severe traumatic brain injury using traditional and machine learning algorithms. *J. Neurotrauma* **40**, 1366–1375 (2023).
40. Yao, H., Williamson, C., Gryak, J. & Najarian, K. Automated hematoma segmentation and outcome prediction for patients with traumatic brain injury. *Artif. Intell. Med.* **107**, 101910 (2020).
41. Arefan, D., Pease, M., Eagle, S. R., Okonkwo, D. O. & Wu, S. Comparison of machine learning models to predict long-term outcomes after severe traumatic brain injury. *Neurosurg. Focus* **54**, E14 (2023).
42. Gravesteijn, B. Y. et al. Machine learning algorithms performed no better than regression models for prognostication in traumatic brain injury. *J. Clin. Epidemiol.* **122**, 95–107 (2020).
43. Hanko, M. et al. Random forest-based prediction of outcome and mortality in patients with traumatic brain injury undergoing primary decompressive craniectomy. *World Neurosurg.* **148**, e450–e458 (2021).
44. Jung, M.-K. et al. Prediction of serious intracranial hypertension from low-resolution neuromonitoring in traumatic brain injury: an explainable machine learning approach. *IEEE J. Biomed. Health Inform.*https://doi.org/10.1109/JBHI.2023.3240460 (2023).
45. Lu, H.-Y. et al. Predicting long-term outcome after traumatic brain injury using repeated measurements of Glasgow Coma Scale and data mining methods. *J. Med. Syst.* **39**, 14 (2015).
46. Matsuo, K. et al. Machine learning to predict in-hospital morbidity and mortality after traumatic brain injury. *J. Neurotrauma* **37**, 202–210 (2020).
47. Pease, M. et al. Outcome prediction in patients with severe traumatic brain injury using deep learning from head CT scans. *Radiology* **304**, 385–394 (2022).
48. Yin, A.-A. et al. Machine learning models for predicting in-hospital outcomes after non-surgical treatment among patients with moderate-to-severe traumatic brain injury. *J. Clin. Neurosci.***120**, 36–41 (2024).
49. Zhang, Z. et al. Machine learning algorithms for improved prediction of in-hospital outcomes after moderate-to-severe traumatic brain injury: a Chinese retrospective cohort study. *Acta Neurochir.***165**, 2237–2247 (2023).

50. Stein, D. M. et al. Computational gene mapping to analyze continuous automated physiologic monitoring data in neuro-trauma intensive care. *J. Trauma Acute Care Surg.* **73**, 419–415 (2012).
51. Bark, D. et al. Refining outcome prediction after traumatic brain injury with machine learning algorithms. *Sci. Rep.* **14**, 8036 (2024).
52. Steyerberg, E. W. et al. Predicting outcome after traumatic brain injury: development and international validation of prognostic scores based on admission characteristics. *PLoS Med.* **5**, e165 (2008).
53. Eagle, S. R. et al. Performance of CRASH and IMPACT prognostic models for traumatic brain injury at 12 and 24 months post-injury. *Neurotrauma Rep.* **4**, 118–123 (2023).
54. Kwong, J. C. C. et al. Predicting non-muscle invasive bladder cancer outcomes using artificial intelligence: a systematic review using APPRAISE-AI. *npj Digital Med.* **7**, 98 (2024).
55. Rajput, D., Wang, W.-J. & Chen, C.-C. Evaluation of a decided sample size in machine learning applications. *BMC Bioinforma.* **24**, 48 (2023).
56. Balki, I. et al. Sample-size determination methodologies for machine learning in medical imaging research: a systematic review. *Can. Assoc. Radiol. J.* **70**, 344–353 (2019).
57. van der Ploeg, T., Austin, P. C. & Steyerberg, E. W. Modern modelling techniques are data hungry: a simulation study for predicting dichotomous endpoints. *BMC Med. Res. Methodol.* **14**, 137 (2014).
58. Wongchareon, K., Thompson, H. J., Mitchell, P. H., Barber, J. & Temkin, N. IMPACT and CRASH prognostic models for traumatic brain injury: external validation in a South-American cohort. *Inj. Prev.* **26**, 546–554 (2020).
59. Kehoe, A. et al. Older patients with traumatic brain injury present with a higher GCS score than younger patients for a given severity of injury. *Emerg. Med. J.* **33**, 381–385 (2016).
60. Smith, C. W. et al. Vision transformer-based decision support for neurosurgical intervention in acute traumatic brain injury: automated surgical intervention support tool. *Radio. Artif. Intell.* **6**, e230088 (2024).
61. Evans, H. & Snead, D. Understanding the errors made by artificial intelligence algorithms in histopathology in terms of patient impact. *npj Digital Med.* **7**, 89 (2024).
62. Naqvi, M., Gilani, S. Q., Syed, T., Marques, O. & Kim, H. C. Skin cancer detection using deep learning—a review. *Diagnostics* **13**, https://doi.org/10.3390/diagnostics13111911 (2023).
63. Palaniappan, K., Lin, E. Y. T. & Vogel, S. Global regulatory frameworks for the use of artificial intelligence (AI) in the healthcare services sector. *Healthcare* **12**, https://doi.org/10.3390/healthcare12050562 (2024).
64. Organization, W. H. *Regulatory Considerations on Artificial Intelligence for Health* (World Health Organization, 2023).
65. Moons, K. G. et al. Critical appraisal and data extraction for systematic reviews of prediction modelling studies: the CHARMS checklist. *PLoS Med.* **11**, e1001744 (2014).
66. Page, M. J. et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ* **372**, n71 (2021).
67. Harrell, F. E. Jr, Lee, K. L. & Mark, D. B. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Stat. Med.* **15**, 361–387 (1996).
68. Peduzzi, P., Concato, J., Kemper, E., Holford, T. R. & Feinstein, A. R. A simulation study of the number of events per variable in logistic regression analysis. *J. Clin. Epidemiol.* **49**, 1373–1379 (1996).
69. Dettori, J. R., Norvell, D. C. & Chapman, J. R. Fixed-effect vs random-effects models for meta-analysis: 3 points to consider. *Glob. Spine J.* **12**, 1624–1626 (2022).
70. Liljequist, D., Elfving, B. & Skavberg Roaldsen, K. Intraclass correlation—a discussion and demonstration of basic features. *PLoS ONE* **14**, e0219854 (2019).
71. The World Bank (IBRD, I). *World Bank Country and Lending Groups*. https://datahelpdesk.worldbank.org/knowledgebase/articles/906519-world-bank-country-and-lending-groups (2024).
72. Wright, D. W. et al. Very early administration of progesterone for acute traumatic brain injury. *N. Engl. J. Med.* **371**, 2457–2466 (2014).

Acknowledgements

A.K.M. acknowledges salary support from the Canada Graduate Scholarship (Doctorate) made available by the Canadian Institutes for Health Research to conduct this work. This study otherwise received no specific funding.

Author contributions

A.K.M.: manuscript conception, manuscript drafting, figure and table creation; A.K.M., H.S.: systematic review, abstract screening, extraction; A.K.M., H.S., Y.Q.H., C.W.S.: evidence appraisal; A.K.M., H.S., Y.Q.H., C.W.S., J.C.C.K., K.E.T., C.D.W., A.V.K., J.R.W., A.B.N. provided critical review of drafted manuscript, including editing and material scientific contribution; A.K.M., H.S., Y.Q.H., C.W.S., J.C.C.K., K.E.T., C.D.W., A.V.K., J.R.W., A.B.N. reviewed and approved the final manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41746-025-01714-y.

Correspondence and requests for materials should be addressed to Avery B. Nathens.

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

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