The leap to ordinal: functional prognosis after traumatic brain injury using artificial intelligence

Shubhayu Bhattacharyay,1,2,3 Ioan Milosevic,1 Lindsay Wilson,4 David K. Menon,1 Robert D. Stevens,3,5 Ewout W. Steyerberg,6 David W. Nelson,7 Ari Ercole1,8 and the CENTER-TBI investigators and participants†

†A full list of the CENTER-TBI investigators and participants can be found in the Appendix.

Abstract

When a patient is admitted to the intensive care unit (ICU) after a traumatic brain injury (TBI), an early prognosis is essential for baseline risk adjustment and shared decision making. TBI outcomes are commonly categorised by the Glasgow Outcome Scale – Extended (GOSE) into eight, ordered levels of functional recovery at 6 months after injury. Existing ICU prognostic models predict binary outcomes at a certain threshold of GOSE (e.g., prediction of survival [GOSE > 1] or functional independence [GOSE > 4]). We aimed to develop ordinal prediction models that concurrently predict probabilities of each GOSE score.

From a prospective cohort (n = 1,550, 65 centres) in the ICU stratum of the Collaborative European NeuroTrauma Effectiveness Research in TBI (CENTER-TBI) patient dataset, we extracted all clinical information within 24 hours of ICU admission (1,151 predictors) and 6-month GOSE scores. We analysed the effect of two design elements on ordinal model performance: (1) the baseline predictor set, ranging from a concise set of ten validated predictors to a token-embedded representation of all possible predictors, and (2) the modelling strategy, from ordinal logistic regression to multinomial deep learning.

With repeated k-fold cross-validation, we found that expanding the baseline predictor set significantly improved ordinal prediction performance while increasing analytical complexity did not. Half of these gains could be achieved with the addition of eight high-impact predictors (2 demographic variables, 4 protein biomarkers, and 2 severity assessments) to the concise set. At best, ordinal models achieved 0.76 (95% CI: 0.74 – 0.77) ordinal discrimination ability (ordinal c-index) and 57% (95% CI: 54% – 60%) explanation of ordinal variation in 6-month GOSE (Somers’ $D_{xy}$). Model performance and the effect of expanding the predictor set
decreased at higher GOSE thresholds, indicating the difficulty of predicting better functional outcomes shortly after ICU admission.

Our results motivate the search for informative predictors that improve confidence in prognosis of higher GOSE and the development of ordinal dynamic prediction models.

**Author affiliations:**

1 Division of Anaesthesia, University of Cambridge, Cambridge CB2 0QQ, United Kingdom
2 Department of Clinical Neurosciences, University of Cambridge, Cambridge CB2 0QQ, United Kingdom
3 Laboratory of Computational Intensive Care Medicine, Johns Hopkins University, Baltimore, MD 21287, USA
4 Division of Psychology, University of Stirling, Stirling FK9 4LA, United Kingdom
5 Department of Anesthesiology and Critical Care Medicine, Johns Hopkins University, Baltimore, MD 21287, USA
6 Department of Biomedical Data Sciences, Leiden University Medical Center, Leiden 2333 ZA, The Netherlands
7 Department of Physiology and Pharmacology, Section for Perioperative Medicine and Intensive Care, Karolinska Institutet, Stockholm 171 77, Sweden
8 Cambridge Centre for Artificial Intelligence in Medicine, Cambridge CB3 0WA, United Kingdom

Correspondence to: Shubhayu Bhattacharyay
University Division of Anaesthesia, University of Cambridge
Box 93, Addenbrooke’s Hospital, Hills Road
Cambridge CB2 0QQ, United Kingdom

E-mail: sb2406@cam.ac.uk

**Running title:** Ordinal traumatic brain injury prognosis
**Keywords:** Traumatic Brain Injury; Precision Medicine; Machine Learning; Neural Network Models; Clinical Decision Rules

**Abbreviations:** AI = artificial intelligence; APM = all-predictor-based model; AUC = area under the receiver operating characteristic curve; CENTER-TBI = Collaborative European NeuroTrauma Effectiveness Research in Traumatic Brain Injury; CI = confidence interval; CPM = concise-predictor-based model; CSD3 = Cambridge Service for Data Driven Discovery; DeepMN = deep learning neural network with multinomial outcome encoding; DeepOR = deep learning neural network with ordinal outcome encoding; eCPM = extended concise-predictor-based model; EDH = extradural haematoma; FIV = full information value; GCS = Glasgow Coma Scale; GCSm = motor component score of the Glasgow Coma Scale; GOSE = Glasgow Outcome Scale – Extended; Hb = haemoglobin; ICI = integrated calibration index; ICU = intensive care unit; IMPACT = International Mission for Prognosis and Analysis of Clinical Trials in TBI; MNLR = multinomial logistic regression; NIV = no information value; ORC = ordinal c-index; POLR = proportional odds (i.e., ordinal) logistic regression; SHAP = Shapley additive explanation; TBI = traumatic brain injury; tSAH = traumatic subarachnoid haemorrhage

**Introduction**

Globally, traumatic brain injury (TBI) is a major cause of death, disability, and economic burden.\(^1\) The treatment of critically ill TBI patients is largely guided by an initial prognosis made within a day of admission to the intensive care unit (ICU).\(^2\) Early outcome prediction models set a baseline against which clinicians consider the effect of therapeutic strategies and compare patient trajectories. Therefore, well-calibrated and reliable prognostic models are an essential component of intensive care.

Outcome after TBI is most often evaluated on the ordered, eight-point Glasgow Outcome Scale – Extended (GOSE),\(^3\)\(^-\)\(^6\) which stratifies patients by their highest level of functional recovery according to participation in daily activities. Existing baseline prediction models used in the ICU dichotomise the GOSE into binary endpoints for TBI outcome. For example, the Acute Physiologic Assessment and Chronic Health Evaluation (APACHE) II\(^7\) model predicts in-hospital survival (GOSE > 1) while the International Mission for Prognosis and Analysis of Clinical Trials in TBI (IMPACT)\(^8\) models focus on predicting functional independence (GOSE > 4, or ‘favourable outcome’) and survival at 6 months post-injury.
Dichotomised GOSE prediction employs a fixed threshold of favourability among the eight levels of recovery for all patients. However, there is no empirical justification for an ideal treatment-effect threshold of GOSE. Ethically, each patient or caregiver should maintain the right to define his or her own threshold of favourability, which is inherently a personal decision. Moreover, by concealing the nuanced differences in outcome defined by the GOSE, dichotomisation limits the prognostic information made available during a shared treatment decision making process. For example, when clinicians, patients, or next of kin must together decide whether to withdraw life-sustaining therapy after severe TBI, knowing the probability of different levels of functional recovery in addition to the baseline probability of survival would enable better quality-of-life consideration and confidence in the decision (Fig. 1B). These problems of dichotomisation cannot be addressed simply by independently training a combination of binary prediction models at several GOSE thresholds. Without explicitly constraining model outputs across the thresholds (i.e., ensuring probabilities do not increase with higher thresholds) during training, one should not interpret multiple threshold outputs concurrently as combining outputs may result in nonsensical values. For example, the purported probability of survival (GOSE > 1) might be lower than that of recovering functional independence (GOSE > 4).

A practical solution would be to train ordinal outcome prediction models, which concurrently return probabilities at each GOSE threshold by learning the interdependent relationships between the predictor set and the possible levels of functional recovery (Fig. 1A). Ordinal GOSE prediction models would allow users to interpret the probability of different levels of functional recovery. Additionally, they can provide insight into the conditional probability of obtaining greater levels of recovery given lower levels (see Fig. 1B for a practical clinical application of this information). However, moving from binary to ordinal outcome prediction poses three key challenges. First, there is no guarantee that widely accepted TBI outcome predictor sets, validated either by binary or ordinal regression analysis, will be able to capture the nuanced differences between levels of functional recovery well enough for reliable prediction. Second, ordinal prediction models typically need to be more complicated than binary models to encode the possibility of and the constrained relationship between multiple outcomes. For GOSE prediction, ordinal models can either encode the outcomes as: (1) multinomial, in which nodes exist for each GOSE score and collectively undergo a softmax transformation (to constrain the sum of values to one) and probabilities are calculated by accumulating values up to each threshold, or (2) ordinal, in which nodes exist for each threshold
between consecutive GOSE scores, constrained such that output values must not increase with higher thresholds, and probabilities for each threshold are calculated with a sigmoid transformation (Fig. 1A). Third, assessment of prediction performance is not as intuitive with an ordinal outcome as with a binary outcome. Widely used dichotomous prediction performance metrics such as the $c$-index (i.e., the area under the receiver operating characteristic curve [AUC]) do not trivially extend to the ordinal case, so assessment of ordinal prediction models requires the consideration of multifactorial metrics and visualisations that may complicate interpretations of model performance.

As part of the Collaborative European NeuroTrauma Effectiveness Research in TBI (CENTER-TBI) project, we aim to address the challenges of ordinal outcome prediction. Our analyses cover a range of modelling strategies and predictors available within the first 24 hours of admission to the ICU.

**Materials and methods**

**Study population and dataset**

The study population was extracted from the ICU stratum of the core CENTER-TBI dataset (v3.0) using Opal database software. The project objectives and experimental design of CENTER-TBI have been described in detail by Maas et al. and Steyerberg et al. Study patients were prospectively recruited at one of 65 participating ICUs across Europe with the following eligibility criteria: admission to the hospital within 24 hours of injury, indication for CT scanning, and informed consent according to local and national requirements.

Per project protocol, each patient’s follow-up schedule included a GOSE assessment at 6 months post-injury, or, more precisely, within a window of 5-8 months post-injury. GOSE assessments were conducted using structured interviews and patient/carer questionnaires by the clinical research team of CENTER-TBI. The eight, ordinal scores of GOSE, representing the highest levels of functional recovery, are decoded in the heading of Table 1. Since patient/carer questionnaires do not distinguish vegetative patients (GOSE = 2) into a separate category, GOSE scores 2 and 3 (lower severe disability) were combined to one category (GOSE $\in \{2,3\}$) in our dataset. Of the 2,138 ICU patients in the CENTER-TBI dataset available for analysis, we excluded patients in the following order: (1) age less than 16 years at ICU admission ($n = 82$), (2) follow-up GOSE was unavailable ($n = 283$), and (3) ICU stay was less
than 24 hours \((n = 223)\). Our resulting sample size was \(n = 1,550\). For 1,351 patients (87.2%), either the patient died during ICU stay \((n = 205)\) or results from a GOSE evaluation at 5 – 8 months post-injury were available in the dataset \((n = 1,146)\). For the remaining 199 patients (12.8%), GOSE scores were imputed using a Markov multi-state model based on the observed GOSE scores recorded at different timepoints between 2 weeks to one-year post-injury.\(^{18}\) A flow diagram for study inclusion and follow-up is provided in Supplementary Fig. 1, and summary characteristics of the study population are detailed in Table 1.

**Repeated \(k\)-fold cross-validation**

We implemented the ‘scikit-learn’ module \((v0.23.2)\)^{19} in Python \((v3.7.6)\) to create 100 stratified partitions of our study population for repeated \(k\)-fold cross-validation (20 repeats, 5 folds). Within each of the partitions, approximately 80% of the population would constitute the training set \((n \approx 1,240\) patients) and 20% of the population would constitute the corresponding testing set \((n \approx 310\) patients). For parametric (i.e., deep learning) models, we implemented a stratified shuffle split on each of the 100 training sets to set 15% \((n \approx 46\) patients) aside for validation and hyperparameter optimisation.

**Selection and preparation of concise predictor set**

In selecting a concise predictor set, our primary aim was to find a small group of well-validated, widely measured clinical variables that are commonly used for TBI outcome prognosis in existing ICU practice. We selected the ten predictors from the extended IMPACT binary prediction model\(^{8}\) for moderate-to-severe TBI – defined by a baseline Glasgow Coma Scale (GCS)\(^{20,21}\) score between 3 and 12, inclusive – to represent our concise set. While 26.6% of our study population falls out of this GCS range (Table 1), we find that the IMPACT predictor set is the most rigorously validated\(^{22-26}\) baseline set available for the overall critically ill TBI population. The ten predictors, characterised in Table 2, are all measured within 24 hours of ICU admission and include demographic characteristics, clinical severity scores, CT characteristics, and laboratory measurements. The predictors as well as empirical justification for their inclusion in the IMPACT model have been described in detail.\(^{27}\) In this manuscript, each of the models trained on the IMPACT predictor set is denoted as a concise-predictor-based model (CPM).
Seven of the concise predictors had missing values for some of the patients in our study population (Supplementary Fig. 2). In each repeated cross-validation partition, we trained an independent, stochastic predictive mean matching imputation function on the training set and imputed all missing values across both sets using the ‘mice’ package (v3.9.0)\textsuperscript{28} in R (v4.0.0).\textsuperscript{29} The result was a multiply imputed \((m = 100)\) dataset with a unique imputation per partition, allowing us to simultaneously account for the variability due to resampling and the variability due to missing value imputation during repeated cross-validation.

Prior to the training of CPMs, each of the multi-categorical variables (i.e., GCSm, Marshall CT, and unreactive pupils in Table 2) were one-hot encoded and each of the continuous variables (i.e., age, glucose, and haemoglobin) were standardised based on the mean and standard deviation of each of the training sets with the ‘scikit-learn’ module in Python.

**Selection of concise-predictor-based models (CPMs)**

We tested four CPM types, each denoted by a subscript: (1) multinomial logistic regression (CPM\textsubscript{MNLR}), (2) proportional odds (i.e., ordinal) logistic regression (CPM\textsubscript{POLR}), (3) class-weighted feedforward neural network with a multinomial (i.e., softmax) output layer (CPM\textsubscript{DeepMN}), and (4) class-weighted feedforward neural network with an ordinal (i.e., constrained sigmoid at each threshold) output layer (CPM\textsubscript{DeepOR}). These models were selected because, in the setting of ordinal GOSE prediction, we wished to compare the performance of: (1) nonparametric logistic regression models (CPM\textsubscript{MNLR} and CPM\textsubscript{POLR}) to nonlinear, parametric deep learning networks (CPM\textsubscript{DeepMN} and CPM\textsubscript{DeepOR}), and (2) multinomial outcome encoding (CPM\textsubscript{MNLR} and CPM\textsubscript{DeepMN}) to ordinal outcome encoding (CPM\textsubscript{POLR} and CPM\textsubscript{DeepOR}). Each of these model types returns a predicted probability for each of the GOSE thresholds at 6 months post-injury from the concise set of predictors (Fig. 1A). A detailed explanation of CPM architectures, hyperparameters for the parametric CPMs, loss functions, and optimisation algorithms is provided in Supplementary Methods 1.

CPM\textsubscript{Best} denotes the optimal CPM for a given performance metric in the Results. CPM\textsubscript{MNLR} and CPM\textsubscript{POLR} were implemented with the ‘statsmodels’ module (dev. v0.14.0)\textsuperscript{30} in Python, and CPM\textsubscript{DeepMN} and CPM\textsubscript{DeepOR} were implemented with the ‘PyTorch’ (v1.10.0)\textsuperscript{31} module in Python.
Design of all-predictor-based models (APMs)

In contrast to the CPMs, we designed and trained prediction models on all baseline (i.e., available to ICU clinicians at 24 hours post-admission) clinical information (excluding high-resolution data such as full brain images or physiological waveforms) in the CENTER-TBI database. Each of these models is designated as an all-predictor-based model (APM).

For our study population, there are 1,151 predictors, each being in one of the 14 categories listed in Table 3, with variable levels of missingness and frequency per patient. This information also includes 81 predictors denoting treatments or interventions within the first 24 hours of ICU care (e.g., type and dose of medication administered) and 76 predictors denoting the explicit impressions or rationales of ICU physicians (e.g., reason for surgical intervention and expected prognosis with or without surgery).

To prepare this information into a suitable format for training APMs, we tokenised and embedded heterogenous patient data in a process visualised in Fig. 2. Predictor tokens were constructed in one of the following ways: (1) for categorical predictors, a token was constructed by concatenating the predictor name and value, e.g., ‘GCSTotalScore_04,’ (2) for continuous predictors, a token was constructed by learning the distribution of that predictor from the training set and discretising into 20 quantile bins, e.g., ‘SystolicBloodPressure_BIN17,’ (3) for text-based entries, we removed all special characters, spaces, and capitalisation from the text and treated them as categorical predictors, e.g., ‘InjuryDescription_skullfracture,’ and (4) for missing values, a separate token was created to designate missingness, e.g., ‘PriorMedications_NA’ (Fig. 2A). The unique tokens from a patient’s first 24 hours of ICU stay made up his or her individual predictor set, and the median number of unique tokens (excluding missing value tokens) per patient per predictor category are provided in Table 3. Notably, this process does not require any data cleaning, missing value imputation, outlier removal, or domain-specific knowledge for a large set of variables and imposes no constraints on the number or type of predictors per patients. Additionally, by including missing value tokens, models can discover meaningful patterns of missingness if they exist.

Taking inspiration from artificially intelligent (AI) natural language processing, all the predictor tokens from the training set (excluding the validation set) are used to construct a token dictionary. APMs learn a lower dimensional vector as well as a positive significance weight for each entry in the dictionary during training. The vectors for each of the tokens of a single patient are significance-weight-averaged into a single vector which is then fed into a
class-weighted feedforward neural network (Fig. 2B). If the neural network has no hidden layers, then the APM is analogous to logistic regression, while if it does have hidden layers, the APM corresponds to deep learning. In this work, we train APMs with one of two kinds of output layers: multinomial, i.e., softmax, (APM\textsubscript{MN}), or ordinal, i.e., constrained sigmoid at each GOSE threshold, (APM\textsubscript{OR}). Both model types output a predicted probability for each of the GOSE thresholds at 6 months post-injury. A detailed explanation of APM architectures, hyperparameters, loss functions, and optimisation algorithms is provided in Supplementary Methods 2.

APM\textsubscript{Best} denotes the optimal APM for a given performance metric in the Results. APM\textsubscript{MN} and APM\textsubscript{OR} were implemented with the ‘PyTorch’ module in Python.

**Predictor importance in all-predictor-based models (APMs)**

The relative importance of predictor tokens in the trained APMs was measured with absolute Shapley additive explanation (SHAP)\textsuperscript{37} values, which, in our case, can be interpreted as the magnitude of the relative contribution of a token towards a model output for a single patient. For APM\textsubscript{MN}, this corresponds to the predictor contributions towards each node (after softmax transformation, Fig. 1A) corresponding to the probability at a GOSE score. For APM\textsubscript{OR}, this corresponds to the predictor contributions towards each node (after sigmoid transformation, Fig. 1A) corresponding to the probability at a GOSE threshold. Absolute SHAP values were measured for each patient in the testing set of every repeated cross-validation partition, and we averaged these values over the partitions to derive our individualised importance scores per token. These scores were averaged, once again, over the entire patient set to calculate the mean absolute SHAP values of each token. Finally, to derive importance scores for each predictor, we calculated the maximum of the mean absolute SHAP values of the possible tokens from the predictor.

**Selection and preparation of extended concise predictor set**

We selected a small set of the most important APM predictors by mean absolute SHAP values to add to the concise predictor set and observe the change in model performance. Since the concise predictor set does not include any information on intervention decisions or physician impressions from the first day, we did not consider these predictor types. Moreover, for every multi-categorical predictor selected, we examined the mean absolute SHAP values
of each of the predictor’s possible tokens to determine which of the categories should be explicitly encoded (e.g., including 10 categories for employment status or just one indicator variable for retirement). The extended concise predictor set, including the 10 original concise predictors and the 8 added predictors, in our study population is listed and characterised in Supplementary Table 1. Each of the models trained on the concise set with these variables added is denoted as an extended concise-predictor-based model (eCPM).

The process of multiple imputation ($m = 100$), one-hot encoding, and standardisation of the extended concise predictor set was identical to that of the concise predictor set, as described earlier.

**Selection of extended concise-predictor-based models (eCPMs)**

The four eCPM model types we tested are identical to the four CPM model types, as described earlier and in Supplementary Methods 1 with, however, the extended concise predictor set: (1) multinomial logistic regression (eCPM$_\text{MNL}$), (2) proportional odds (i.e., ordinal) logistic regression (eCPM$_\text{POL}$), (3) class-weighted feedforward neural network with a multinomial (i.e., softmax) output layer (eCPM$_\text{DeepMN}$), and (4) class-weighted feedforward neural network with an ordinal (i.e., constrained sigmoid at each threshold) output layer (eCPM$_\text{DeepOR}$).

$eCPM_{\text{Best}}$ denotes the optimal eCPM for a given performance metric in the Results.

**Statistical analysis**

All model metrics, curves, and associated confidence intervals (CI) were calculated from testing set predictions using the repeated Bootstrap Bias Corrected Cross-Validation (BBC-CV) method with 1,000 resamples of unique patients for bootstrapping. The collection of metrics from the bootstrapped testing set resamples for each model then formed our unbiased estimation distribution for statistical inference (i.e., CI).

In this work, we assess model discrimination performance (i.e., how well do the models separate patients with different GOSE scores?) and probability calibration (i.e., how reliable are the predicted probabilities at each threshold?). The metrics and visualisations are explained in detail, with mathematical derivation and intuitive examples, in Supplementary Methods 3. In this section, we will only list the metrics, their interpretations, and their range of feasible values. Feasible values range from the value corresponding to no model information or random
guessing (i.e., the no information value [NIV]) to the value corresponding to ideal model performance (i.e., the full information value [FIV]).

Our primary metric of model discrimination performance is the ordinal c-index (ORC).\textsuperscript{13} ORC has two interpretations: (1) the probability that a model correctly separates two patients with two randomly chosen GOSE scores and (2) the average proportional closeness between a model’s functional outcome ranking of a set of patients (which includes one randomly chosen patient from each possible GOSE score) to their true functional outcome ranking. In addition, we calculate Somers’ $D_{xy}$,\textsuperscript{39,40} which is interpreted as the proportion of ordinal variation in GOSE that can be explained by the variation in model output. Our final metrics of model discrimination are dichotomous c-indices (i.e., AUC) at each threshold of GOSE. Each is interpreted as the probability of a model correctly discriminating a patient with GOSE above the threshold from one with GOSE below. The range of feasible values for each discrimination metric are: $\text{NIV}_{\text{ORC}} = 0.5$ to $\text{FIV}_{\text{ORC}} = 1$, $\text{NIV}_{\text{Somers’ } D_{xy}} = 0$ to $\text{FIV}_{\text{Somers’ } D_{xy}} = 1$, and $\text{NIV}_{\text{Dichotomous } c\text{-index}} = 0.5$ to $\text{FIV}_{\text{Dichotomous } c\text{-index}} = 1$. ORC is the only discrimination metric that is independent of the sample prevalence of each GOSE category.\textsuperscript{13}

To assess the calibration of predicted probabilities at each GOSE threshold, we use the logistic recalibration framework\textsuperscript{41} to measure calibration slope.\textsuperscript{42} A calibration slope less than one indicates overfitting (i.e., high predicted probabilities are overestimated while low predicted probabilities are underestimated) while a calibration slope greater than one indicates underfitting.\textsuperscript{43} We also examine smoothed probability calibration curves\textsuperscript{44} to detect miscalibrations that may be overlooked by the logistic recalibration framework.\textsuperscript{43} The ideal calibration curve is a diagonal line with slope one and $y$-intercept 0 while one indicative of random guessing would be a horizontal line with a $y$-intercept at the proportion of the study population above the given threshold. We accompany each calibration curve with the integrated calibration index (ICI),\textsuperscript{45} which is the mean absolute error between the smoothed and the ideal calibration curves, to aid comparison of curves across model types. $\text{FIV}_{\text{ICI}} = 0$, but $\text{NIV}_{\text{ICI}}$ varies based on the outcome distribution at each threshold (Supplementary Methods 3).

All metrics were calculated using the ‘scikit-learn’ and ‘SciPy’ (v1.6.2)\textsuperscript{46} modules in Python and figures were plotted using the ‘ggplot2’ package (v3.3.2)\textsuperscript{47} in R.
Computational resources

All computational and statistical components of this work were performed in parallel on the Cambridge Service for Data Driven Discovery (CSD3) high performance computer, operated by the University of Cambridge Research Computing Service (http://www.hpc.cam.ac.uk). The training of each APM was accelerated with graphical processing units and the ‘PyTorch Lightning’ (v1.5.0) module. The training of all parametric models (CPM_{DeepMN}, CPM_{DeepOR}, APM_{MN}, APM_{OR}, eCPM_{DeepMN}, and eCPM_{DeepOR}) was made more efficient by dropping out consistently underperforming parametric configurations, on the validation sets, with the Bootstrap Bias Corrected with Dropping Cross-Validation (BBCD-CV) method with 1,000 resamples of unique patients. The results of hyperparameter optimisation are detailed in Supplementary Results 1.

Ethical approval statement

The CENTER-TBI study has been conducted in accordance with all relevant laws of the European Union and all relevant laws of the country where the recruiting sites were located, including (but not limited to) the relevant privacy and data protection laws and regulations, the relevant laws and regulations on the use of human materials, and all relevant guidance relating to clinical studies from time in force including (but not limited to) the ICH Harmonised Tripartite Guideline for Good Clinical Practice (CPMP/ICH/135/95) and the World Medical Association Declaration of Helsinki entitled “Ethical Principles for Medical Research Involving Human Subjects.” Informed consent by the patients and/or the legal representative/next of kin was obtained (according to local legislation) for all patients recruited in the core dataset of CENTER-TBI and documented in the electronic case report form. Ethical approval was obtained for each recruiting site.

The list of sites, ethical committees, approval numbers and approval dates can be found on the website: https://www.center-tbi.eu/project/ethical-approval.

Data availability

All code used in this project can be found at the following online repository: https://github.com/sbhattacharyay/ordinal_GOSE_prediction (doi: 10.5281/zenodo.5933042).
Individual participant data, including data dictionary, the study protocol, and analysis scripts are available online, conditional to approved study proposal, with no end date. Interested investigators must provide a methodologically sound study proposal to the management committee. Proposals can be submitted online at https://www.center-tbi.eu/data. Signed confirmation of a data access agreement is required, and all access must comply with regulatory restrictions imposed on the original study.

Results

CPM and APM discrimination performance

The discrimination performance metrics for each CPM are listed in Supplementary Table 2. Deep learning models (CPM\textsubscript{DeepMN} and CPM\textsubscript{DeepOR}) made no significant improvement (based on 95% CI) over logistic regression models (CPM\textsubscript{MNLR} and CPM\textsubscript{POLR}). The only significant difference in discrimination among the model types was observed in CPM\textsubscript{DeepOR}, which had a significantly lower ORC and Somers’ $D_{xy}$ than the other models. The discrimination performance metrics for each APM are listed in Supplementary Table 3. APM\textsubscript{MN} had a significantly higher ORC, Somers’ $D_{xy}$, and dichotomous $c$-indices at lower GOSE thresholds (i.e., GOSE > 1 and GOSE > 3) than did APM\textsubscript{OR}. Moreover, in Supplementary Results 1, we see that the best-performing parametric configurations of APM\textsubscript{MN} did not contain additional hidden layers between the token embedding and output layers. Our results of performance within predictor sets consistently demonstrate that increasing analytical complexity, in terms of using deep learning (for CPMs) or adding hidden network layers (for APMs), did not improve discrimination of outcomes. In the case of deep learning models, multinomial outcome encoding significantly outperformed ordinal outcome encoding (Fig. 1A).

The discrimination performance metrics of the best-performing CPMs (CPM\textsubscript{Best}), compared with those of the best-performing APMs (APM\textsubscript{Best}), are listed in Table 4. In contrast to the case of analytical complexity, we observe that expanding the predictor set yielded a significant improvement in ORC, Somers’ $D_{xy}$, and each threshold-level dichotomous $c$-index except for those of the highest GOSE thresholds (i.e., GOSE > 6 and GOSE > 7). On average, models trained on the concise predictor set (CPMs) correctly separated two randomly selected patients from two randomly selected GOSE categories 70% (95% CI: 68% – 71%) of the time,
while models trained on all baseline predictors (APMs) in the CENTER-TBI dataset did so 76% (95% CI: 74% – 77%) of the time. These percentages also correspond to the average proportional closeness of predicted rankings to true GOSE rankings of patient sets. CPM_{Best} explained 44% (95% CI: 41% – 48%) of the ordinal variation in GOSE while APM_{Best} explained 57% (95% CI: 54% – 60%) in their respective model outputs. At increasing GOSE thresholds, the dichotomous c-indices of CPM_{Best} and APM_{Best}, as well as the gap between them, consistently decreased (Table 4). This signifies that predicting higher 6-month functional outcomes is more difficult than predicting lower 6-month functional outcomes. Moreover, the gains in discrimination earned from expanding the predictor set mostly come from improved performance at lower GOSE thresholds (i.e., predicting survival, return of consciousness, or recovery of functional independence).

**CPM and APM calibration performance**

The calibration slopes and calibration curves for each CPM are displayed in Supplementary Table 2 and Supplementary Fig. 3, respectively. Both logistic regression CPMs (CPM_{MNL} and CPM_{POL}) are significantly overfitted at the three highest GOSE thresholds (i.e., GOSE > 5, GOSE > 6, and GOSE > 7). The graphical calibration of CPM_{DeepOR} was significantly worse than that of the other CPMs (Supplementary Fig. 3). The calibration slopes and calibration curves for each APM are displayed in Supplementary Table 3 and Supplementary Fig. 4, respectively. APM_{OR} is poorly calibrated at each threshold of GOSE. APM_{MN} is significantly overfitted at the three highest GOSE thresholds (i.e., GOSE > 5, GOSE > 6, and GOSE > 7).

The calibration slopes and calibration curves for the best-calibrated CPMs (CPM_{Best}), compared against those for the best-calibrated APMs (APM_{Best}), are displayed in Table 4 and Fig. 3, respectively. Unlike CPM_{Best}, APM_{Best} could not avoid significant overfitting at the three highest GOSE thresholds (i.e., GOSE > 5, GOSE > 6, and GOSE > 7). At these thresholds, we observe that the calibration curve of APM_{Best} significantly veered off the diagonal line of ideal calibration for higher predicted probabilities. However, due to the relative infrequency of these predictions (comparative histograms in Fig. 3), the ICI of APM_{Best} is not significantly higher than that of CPM_{Best}. Our results suggest that APM_{Best} requires more patients with higher functional outcomes, in both the training and validation sets, to mitigate overfitting.43
Predictor importance

Given that $\text{APM}_{\text{MN}}$ significantly outperforms $\text{APM}_{\text{OR}}$ in discrimination and calibration, we focus the assessment of predictor importance to $\text{APM}_{\text{MN}}$. A bar plot of the mean absolute SHAP values associated with the 15 most important predictors in $\text{APM}_{\text{MN}}$ is provided in Fig. 4. We find that the subjective early prognoses of ICU physicians had the most important effect on $\text{APM}_{\text{MN}}$ predictions, particularly for the prediction of death ($\text{GOSE} = 1$) within 6 months. These impressions likely summarise information from a variable number of other predictors, along with the physician’s own experience-based judgement, and thus have high additive values. Our result suggests that integration of a physician’s interpretations of a patient’s baseline status may add important prognostic information, especially for the prediction of lower functional outcomes. However, this information would not be suitable for physician-agnostic baseline risk adjustment models (e.g., IMPACT). Thus, we do not consider physician impression predictors for our extended concise predictor set. We also observe that ‘age at admission’ was the only concise predictor among the 15 most important ones. The importance ranks (out of 1,151) of the concise predictors (Table 2) are: age = 5th, glucose = 23rd, Marshall CT = 25th, pupillary reactivity = 29th, GCSm = 42nd, haemoglobin = 50th, hypoxia = 284th, tSAH = 301st, EDH = 414th, and hypotension = 420th. The eight remaining predictors of the top 15 (Fig. 4) were added to the concise predictor set to form our extended concise predictor set. Within the tokens for “employment status before injury,” we found that the single token indicating retirement is much more important than the others. Thus, instead of encoding all 10 options for employment status, we included a single indicator variable for retirement in our extended concise predictor set. The eight added predictors included 2 demographic variables (retirement status and highest level of formal education), 4 protein biomarker concentrations (neurofilament light chain [NFL], glial fibrillary acidic protein [GFAP], total tau protein [T-tau], and S100 calcium-binding protein B [S100B]), and 2 clinical assessment variables (worst abbreviated injury score [AIS] among head, neck, brain, and cervical spine injuries and incidence of post-traumatic amnesia at ICU admission). The extended concise predictor set, including the 10 original concise predictors and the 8 added predictors, is statistically characterised in Supplementary Table 1.
eCPM discrimination and calibration

The discrimination and calibration metrics for the best-performing extended-predictor-based model (eCPM\textsubscript{Best}) are listed in Table 4. Inclusion of the eight selected predictors accounted for about half of the gains in discrimination performance achieved by APM\textsubscript{Best} over CPM\textsubscript{Best} according to ORC, Somers’ $D_{xy}$, and the dichotomous $c$-indices. Based on the difference in Somers’ $D_{xy}$, the eight added predictors allowed models to explain an additional 6% of the ordinal variation in GOSE at 6 months post-injury. Unlike APM\textsubscript{Best}, eCPM\textsubscript{Best} is not significantly overfitted at any threshold. The calibration curves of eCPMs (Supplementary Fig. 5) are largely similar to those of the corresponding CPMs (Supplementary Fig. 3), except at the highest threshold (i.e., GOSE > 7). Similar to those of APM\textsubscript{MN}, the calibration curves of eCPMs veer off the line of ideal calibration at higher predicted probabilities of GOSE > 7. The eCPM results support the finding that discrimination performance can be improved with the expansion of the predictor set. Furthermore, by limiting the number of added predictors and the analytical complexity of the model, eCPM avoided the significant miscalibration of APM at higher thresholds.

The discrimination and calibration metrics for each eCPM are listed in Supplementary Table 4.

Discussion

To our knowledge, this is the most comprehensive evaluation of early ordinal outcome prognosis for critically ill TBI patients. Our analysis cross-compares a range of ordinal prediction modelling strategies with a large range of available baseline predictors to determine the relative contribution of each towards model performance. Employing an AI tokenisation and embedding technique, we develop highly flexible ordinal prediction models that can learn from the entire, heterogeneous set of 1,151 predictors, available within the first 24 hours of ICU stay, in the CENTER-TBI dataset. This information includes not only all baseline clinical data currently deemed significant for ICU care of TBI but also advanced sub-study results (e.g., protein biomarkers, central haemostatic markers, genetic markers, and advanced MRI results) that represent the experimental frontier of clinical TBI assessment.\textsuperscript{1,15,16} Therefore, our work reveals the interpretable limits of baseline ordinal, 6-month GOSE prediction in the ICU at this time.
Our key finding is that augmenting the baseline predictor set was much more relevant for improving ordinal model prediction performance than was increasing analytical complexity with deep learning. Within this given predictor set, artificial neural networks did not perform better than logistic regression models (Supplementary Table 2, Supplementary Table 4), nor did models with additional hidden layers for the APMs (Supplementary Results 1). This result is consistent with findings in the binary prediction case. On the other hand, augmenting the predictor set, from CPM to APM, substantially improved ordinal discrimination (ORC: +8.6%, Table 4) and prediction at lower GOSE thresholds (e.g., GOSE > 1 c-index: +8.4%, Table 4). Just adding eight predictors to the concise predictor set accounted for about half of the gains in discrimination. However, the addition of predictors negatively affected model calibration, particularly at higher GOSE thresholds (Fig. 3, Table 4). This result underlines the need for careful consideration of probability calibration during model development (e.g., recalibrate with isotonic regression to mitigate overfitting).

At the same time, our results also indicate that ordinal early outcome prognosis for critically ill TBI patients is limited in capability. The best-performing model, which learns from all baseline information in the CENTER-TBI dataset, can only correctly discriminate two randomly chosen patients with two randomly chosen GOSE scores 76% (95% CI: 74% – 77%) of the time. This means, on average, a predicted GOSE ranking of seven randomly chosen patients, each of a different functional outcome at 6 months, will contain 5.10 (95% CI: 4.74–5.46) incorrect pairwise orderings out of 21. Currently, ordinal model outputs explain, at best, 57% (95% CI: 54% – 60%) of the ordinal variation in 6-month GOSE. Ordinal prediction models struggle to reliably predict full recovery (GOSE > 7 c-index: 75% [95% CI: 72% – 79%], Table 4), and gains from expanding the predictor set diminish with higher GOSE thresholds.

We recognise several limitations in our study. While the concise predictor set was originally designed for prognosis after moderate-to-severe TBI (i.e., baseline GCS 3 – 12), 26.6% of our study population had experienced mild (i.e., baseline GCS 13 – 15) TBI (Table 1). Predictor sets have been designed for mild TBI patients (e.g., UPFRONT study predictors). However, in line with the aims of the CENTER-TBI project, we focus the TBI population not by initial characterisation with GCS but by stratum of care (i.e., admission to the ICU). Therefore, we selected the single concise predictor set that was best validated for the majority of critically ill TBI patients. Our outcome categories (GOSE at 6 months post-injury) were statistically imputed for 13% of our dataset using available GOSE between 2 weeks and
one-year post-injury. Although this method was strongly validated on the same (CENTER-TBI) dataset,\textsuperscript{18} we do recognise that our outcome labels may not be precisely correct. The focus of this work is on the prediction of functional outcomes through GOSE; nonetheless, it is worth considering other outcomes, such as quality-of-life and psychological health, that are important for clinical decision making.\textsuperscript{51} Finally, before the AI models developed in this work and in subsequent iterations could be integrated into ICU practice, limitations of generalisability must be addressed.\textsuperscript{52} Our models were developed on a multicentre population, prospectively recruited between 2014 and 2017,\textsuperscript{24} across Europe, and may encode recruitment, collection, and clinical biases native to our patient set. AI models must continuously be updated, iteratively retrained on incoming information, to help fight the effect these biases may have on returned prognoses for a given patient.

In the setting of TBI prognosis, we encourage the use of AI not to add analytical complexity (i.e., make models “deeper”) but to expand the predictor set (i.e., make models “wider”). Studies have uncovered promising prognostic value in high-resolution TBI monitoring and imaging modalities (e.g., intracranial and cerebral perfusion pressure,\textsuperscript{53-55} accelerometry,\textsuperscript{56} MRI),\textsuperscript{57-59} and we recommend integrating these features into ordinal prognostic models, especially to improve prediction of higher functional outcomes. We also believe that there is a feasible performance limit to reliable ordinal outcome prognosis if only statically considering the clinical information from the first 24 hours of ICU stay. It would seem far-fetched to expect all relevant information pertaining to an outcome at 6 months to be encapsulated in the first 24 hours of ICU treatment. Heterogeneous pathophysiological processes unfold over time in patients after TBI,\textsuperscript{60,61} and dynamic prediction models, which return model outputs longitudinally with changing clinical information, are better equipped to consider these temporal effects on prognosis. Dynamic prognosis models have been developed for TBI patients\textsuperscript{62} and the greater ICU population,\textsuperscript{33,63,64} but none of them predict functional outcomes on an ordinal scale. We suggest that the next iteration of this work should be to develop ordinal dynamic prediction models on all clinical information available during the complete ICU stay.
Acknowledgements

We are grateful to the patients of our study for helping us in our efforts to improve TBI care and outcome. We gratefully acknowledge interactions and support from the International Initiative for TBI Research (InTBIR) investigators.

S.B. would like to thank: Abhishek Dixit (Univ. of Cambridge) for helping access the CENTER-TBI dataset, Jacob Deasy (Univ. of Cambridge) for aiding the development of modelling methodology, and Kathleen Mitchell-Fox (Princeton Univ.) for offering comments on the manuscript. All authors would like to thank Andrew I. R. Maas (Antwerp Univ. Hospital) for offering comments on the manuscript.

Funding

The research was supported by the National Institute for Health Research (NIHR) Brain Injury MedTech Co-operative based at Cambridge University Hospitals NHS Foundation Trust and University of Cambridge. The views expressed are those of the authors and not necessarily those of the NHS, NIHR or the Department of Health and Social Care.

CENTER-TBI was supported by the European Union 7th Framework programme (EC grant 602150). Additional funding was obtained from the Hannelore Kohl Stiftung (Germany), from OneMind (USA), and from Integra LifeSciences Corporation (USA).

CSD3 is supported by the United Kingdom Engineering and Physical Sciences Research Council (EPSRC Tier-2 capital grant EP/T022159/1).

S.B. is currently funded by a Gates Cambridge fellowship.

Competing interests

The authors report no competing interests.

Supplementary material

Supplementary material is attached below.
Appendix

The CENTER-TBI investigators and participants

Cecilia Åkerlund¹, Krisztina Amrein², Nada Andelic³, Lasse Andreassen⁴, Audny Anke⁵, Anna Antoni⁶, Gérard Audibert⁷, Philippe Azouvi⁸, Maria Luisa Azzolin⁹, Ronald Bartels¹⁰, Pál Barzó¹¹, Romuald Beauvais¹², Ronny Beer¹³, Bo-Michael Bellander¹⁴, Antonio Belli¹⁵, Habib Benali¹⁶, Maurizio Berardino¹⁷, Luigi Beretta⁹, Morten Blaabjerg¹⁶, Peter Bragge¹⁹, Alexandra Brazinova²⁰, Vibeke Brinck²¹, Joanne Brooker²², Camilla Brorsson²³, Andras Buki²⁴, Monika Bullinger²⁵, Manuel Cabeira²⁶, Alessio Caccioppola²⁷, Emiliana Calappi²⁷, Maria Rosa Calvi⁹, Peter Cameron²⁸, Guillermo Carbayo Lozano²⁹, Marco Carbonara²⁷, Simona Cavallo¹⁷, Giorgio Chevallard³⁰, Arturo Chieregato³⁰, Giuseppe Citerio³¹, Hans Clusmann³³, Mark Coburn³⁴, Jonathan Coles³⁵, Jamie D. Cooper³⁶, Peter Correia³⁷, Amra Ćović ³⁸, Nicola Curry³⁹, Endre Czeiter³, Claire Dahyot-Fizelier⁴⁰, Paul Dark⁴¹, Helen Dawes⁴², Véronique De Keyser⁴³, Vincent Degos¹⁶, Francesco Della Corte⁴⁴, Hugo den Boogert¹⁰, Bart Depreitere⁴⁵, Dula Dilvesi ⁴⁶, Abhishek Dixit⁴⁷, Emma Donoghue²², Jens Dreier⁴⁸, Guy-Loup Dulière⁴⁹, Ari Ercole⁵⁰, Patrick Esser⁴², Erzsébet Ezer⁵⁰, Martin Fabricius⁵¹, Valery L. Feigin⁵², Kelly Foks⁵³, Shirin Frisvold⁵⁴, Alex Furmanov⁵⁵, Pablo Gagliardo⁵⁶, Damien Galanaud⁶¹, Dashiell Gantner²⁸, Guo Yi Gao⁵⁷, Pradeep George⁵⁸, Alexandre Ghuysen⁵⁹, Lelde Giga⁶⁰, Ben Glocker⁶¹, Jagoš Golubovic⁶², Pedro A. Gomez⁶², Johannes Gratz⁶³, Benjamin Gravesteijn⁶⁴, Francesca Grossi⁴⁴, Russell L. Gruen⁶⁵, Deepak Gupta⁶⁶, Juanita A. Haagsma⁶⁴, Iain Haitsma⁶⁷, Raimund Helbok¹³, Eirik Helseth⁶⁸, Lindsay Horton ⁶⁹, Jilske Huijbens⁶⁴, Peter J. Hutchinson⁷⁰, Bram Jacobs⁷¹, Stefan Jankowski⁷², Mike Jarrett²¹, Ji-yao Jiang⁵⁸, Faye Johnson⁷³, Kelly Jones⁵², Mladen Karan⁶⁶, Angelos G. Koliats⁷⁰, Erwin Kompanje⁷⁴, Daniel Konidriella⁵¹, Evgenios Kornaropoulos⁴⁷, Lars-Owe Koskinen⁷⁵, Noémi Kovács⁷⁶, Ana Kowark⁷⁷, Alfonso Layares⁶², Linda Lanyon⁵⁸, Steven Laureys⁷⁸, Fiona Lecky⁷⁹, Didier Ledoux⁷⁸, Rolf Lefering⁸¹, Valerie Legrand⁸², Aurelie Lejeune⁸³, Leon Levi⁸⁴, Roger Lightfoot⁸⁵, Hester Lingsma⁶⁴, Andrew I.R. Maas⁴³, Ana M. Castaño-León⁶², Marc Maegle⁸⁶, Marek Majdan²⁰, Alex Manara⁸⁷, Geoffrey Manley⁸⁸, Costanza Martino⁶⁹, Hugues Maréchal⁴⁹, Julia Mattern⁹⁰, Catherine McMahon⁹¹, Béla Melegh⁹², David Menon⁴⁷, Tomas Menovsky⁴³, Ana Mikollic⁶⁴, Benoit Misset⁷⁸, Visakh Muraleedharan⁵⁸, Lynnette Murray²⁸, Ancuta Negrü³, David Nelson¹, Virginia Newcombe⁴⁷, Daan Nieboer⁴⁴, József Nyirádi², Otesile Olubukola⁷⁹, Matej Oresic⁹⁴, Fabrizio Orltolano²⁷, Aarno Palotie⁹⁵,⁹⁶,⁹⁷, Paul M. Parizel⁹⁸, Jean-François Payen⁹⁹, Natascha Perera¹², Vincent Perlberg¹⁶, Paolo
Persona\textsuperscript{100}, Wilco Peul\textsuperscript{101}, Anna Piippo-Karjalainen\textsuperscript{102}, Matti Pirinen\textsuperscript{95}, Dana Pisica\textsuperscript{64}, Horia Ples\textsuperscript{93}, Suzanne Polinder\textsuperscript{64}, Inigo Pomposo\textsuperscript{29}, Jussi P. Posti\textsuperscript{103}, Louis Puybasset\textsuperscript{104}, Andreea Radoi\textsuperscript{105}, Arminas Ragauskas\textsuperscript{106}, Rahul Raj\textsuperscript{102}, Malinka Rambadagalla\textsuperscript{107}, Isabel Retel Helmrich\textsuperscript{64}, Jonathan Rhodes\textsuperscript{108}, Sylvia Richardson\textsuperscript{109}, Sophie Richter\textsuperscript{47}, Samuli Ripatti\textsuperscript{95}, Saulius Rocka\textsuperscript{106}, Cecilie Roe\textsuperscript{110}, Olav Roise\textsuperscript{111,112}, Jonathan Rosand\textsuperscript{113}, Jeffrey V. Rosenfeld\textsuperscript{114}, Christina Rosenlund\textsuperscript{115}, Guy Rosenthal\textsuperscript{55}, Rolf Rossaint\textsuperscript{77}, Sandra Rossi\textsuperscript{100}, Daniel Rueckert\textsuperscript{61}, Martin Rusnák\textsuperscript{116}, Juan Sahuquillo\textsuperscript{105}, Oliver Sakowitz\textsuperscript{90,117}, Renan Sanchez-Porras\textsuperscript{117}, Janos Sandor\textsuperscript{118}, Nadine Schäfer\textsuperscript{81}, Silke Schmidt\textsuperscript{119}, Herbert Schoechl\textsuperscript{120}, Guus Schoonman\textsuperscript{121}, Rico Frederik Schou\textsuperscript{122}, Elisabeth Schwendenwein\textsuperscript{6}, Charlie Sewalt\textsuperscript{64}, Ranjit D. Singh\textsuperscript{101}, Toril Skandsen\textsuperscript{123,124}, Peter Smielewski\textsuperscript{26}, Abayomi Sorinola\textsuperscript{125}, Emmanuel Stamatakis\textsuperscript{47}, Simon Stanworth\textsuperscript{39}, Robert Stevens\textsuperscript{126}, William Stewart\textsuperscript{127}, Ewout W. Steyerberg\textsuperscript{64,128}, Nino Stocchetti\textsuperscript{129}, Nina Sundström\textsuperscript{130}, Riikka Takala\textsuperscript{131}, Viktória Tamás\textsuperscript{125}, Tomas Tamosuittis\textsuperscript{132}, Mark Steven Taylor\textsuperscript{20}, Braden Te Ao\textsuperscript{52}, Olli Tenovuo\textsuperscript{103}, Alice Theadom\textsuperscript{52}, Matt Thomas\textsuperscript{87}, Dick Tibboel\textsuperscript{133}, Marjolein Timmers\textsuperscript{74}, Christos Tolias\textsuperscript{134}, Tony Trapani\textsuperscript{28}, Cristina Maria Tudora\textsuperscript{93}, Andreas Unterberg\textsuperscript{90}, Peter Vajkoczy\textsuperscript{135}, Shirley Vallance\textsuperscript{28}, Egils Valeinis\textsuperscript{60}, Zoltán Vámos\textsuperscript{50}, Mathieu van der Jagt\textsuperscript{136}, Gregory Van der Steen\textsuperscript{43}, Joukje van der Naalt\textsuperscript{71}, Jeroen T.J.M. van Dijck\textsuperscript{101}, Inge A. M. van Erp\textsuperscript{101}, Thomas A. van Essen\textsuperscript{101}, Wim Van Hecke\textsuperscript{137}, Caroline van Heugten\textsuperscript{138}, Dominique Van Praag\textsuperscript{139}, Ernest van Veen\textsuperscript{64}, Thijs Vande Vyvere\textsuperscript{137}, Roel P. J. van Wijk\textsuperscript{101}, Alessia Vargiolu\textsuperscript{32}, Emmanuel Vega\textsuperscript{83}, Kimberley Velt\textsuperscript{64}, Jan Verheyden\textsuperscript{137}, Paul M. Vespa\textsuperscript{140}, Anne Vik\textsuperscript{23,141}, Rimantas Vileinis\textsuperscript{132}, Victor Volovic\textsuperscript{67}, Nicole von Steinbüchel\textsuperscript{38}, Daphne Voormolen\textsuperscript{64}, Petar Vulekovic\textsuperscript{46}, Kevin K.W. Wang\textsuperscript{142}, Daniel Whitehouse\textsuperscript{47}, Eveline Wiegers\textsuperscript{64}, Guy Williams\textsuperscript{47}, Lindsay Wilson\textsuperscript{69}, Stefan Winzeck\textsuperscript{47}, Stefan Wolf\textsuperscript{143}, Zhihui Yang\textsuperscript{113}, Peter Ylen\textsuperscript{44}, Alexander Youssi\textsuperscript{90}, Frederick A. Zeiler\textsuperscript{47,145}, Veronika Zelinkova\textsuperscript{20}, Agate Ziverte\textsuperscript{60}, Tommaso Zoerle\textsuperscript{27}

\textsuperscript{1} Department of Physiology and Pharmacology, Section of Perioperative Medicine and Intensive Care, Karolinska Institutet, Stockholm, Sweden

\textsuperscript{2} János Szentágothai Research Centre, University of Pécs, Pécs, Hungary

\textsuperscript{3} Division of Surgery and Clinical Neuroscience, Department of Physical Medicine and Rehabilitation, Oslo University Hospital and University of Oslo, Oslo, Norway

\textsuperscript{4} Department of Neurosurgery, University Hospital Northern Norway, Tromso, Norway
5 Department of Physical Medicine and Rehabilitation, University Hospital Northern Norway, Tromso, Norway

6 Trauma Surgery, Medical University Vienna, Vienna, Austria

7 Department of Anesthesiology & Intensive Care, University Hospital Nancy, Nancy, France

8 Raymond Poincare hospital, Assistance Publique – Hopitaux de Paris, Paris, France

9 Department of Anesthesiology & Intensive Care, S Raffaele University Hospital, Milan, Italy

10 Department of Neurosurgery, Radboud University Medical Center, Nijmegen, The Netherlands

11 Department of Neurosurgery, University of Szeged, Szeged, Hungary

12 International Projects Management, ARTTIC, Munchen, Germany

13 Department of Neurology, Neurological Intensive Care Unit, Medical University of Innsbruck, Innsbruck, Austria

14 Department of Neurosurgery & Anesthesia & intensive care medicine, Karolinska University Hospital, Stockholm, Sweden

15 NIHR Surgical Reconstruction and Microbiology Research Centre, Birmingham, UK

16 Anesthesie-Réanimation, Assistance Publique – Hopitaux de Paris, Paris, France

17 Department of Anesthesia & ICU, AOU Città della Salute e della Scienza di Torino - Orthopedic and Trauma Center, Torino, Italy

18 Department of Neurology, Odense University Hospital, Odense, Denmark

19 BehaviourWorks Australia, Monash Sustainability Institute, Monash University, Victoria, Australia

20 Department of Public Health, Faculty of Health Sciences and Social Work, Trnava University, Trnava, Slovakia

21 Quesgen Systems Inc., Burlingame, California, USA

22 Australian & New Zealand Intensive Care Research Centre, Department of Epidemiology and Preventive Medicine, School of Public Health and Preventive Medicine, Monash University, Melbourne, Australia
Department of Surgery and Perioperative Science, Umeå University, Umeå, Sweden

Department of Neurosurgery, Medical School, University of Pécs, Hungary and Neurotrauma Research Group, János Szentágothai Research Centre, University of Pécs, Hungary

Department of Medical Psychology, Universitätsklinikum Hamburg-Eppendorf, Hamburg, Germany

Brain Physics Lab, Division of Neurosurgery, Dept of Clinical Neurosciences, University of Cambridge, Addenbrooke’s Hospital, Cambridge, UK

Neuro ICU, Fondazione IRCCS Cà Granda Ospedale Maggiore Policlinico, Milan, Italy

ANZIC Research Centre, Monash University, Department of Epidemiology and Preventive Medicine, Melbourne, Victoria, Australia

Department of Neurosurgery, Hospital of Cruces, Bilbao, Spain

NeuroIntensive Care, Niguarda Hospital, Milan, Italy

School of Medicine and Surgery, Università Milano Bicocca, Milano, Italy

NeuroIntensive Care, ASST di Monza, Monza, Italy

Department of Neurosurgery, Medical Faculty RWTH Aachen University, Aachen, Germany

Department of Anesthesiology and Intensive Care Medicine, University Hospital Bonn, Bonn, Germany

Department of Anesthesia & Neurointensive Care, Cambridge University Hospital NHS Foundation Trust, Cambridge, UK

School of Public Health & PM, Monash University and The Alfred Hospital, Melbourne, Victoria, Australia

Radiology/MRI department, MRC Cognition and Brain Sciences Unit, Cambridge, UK

Institute of Medical Psychology and Medical Sociology, Universitätsmedizin Göttingen, Göttingen, Germany

Oxford University Hospitals NHS Trust, Oxford, UK

23

24

25

26

27

28

29

30

31

32

33

34

35

36

37

38

39
Intensive Care Unit, CHU Poitiers, Poitiers, France

University of Manchester NIHR Biomedical Research Centre, Critical Care Directorate, Salford Royal Hospital NHS Foundation Trust, Salford, UK

Movement Science Group, Faculty of Health and Life Sciences, Oxford Brookes University, Oxford, UK

Department of Neurosurgery, Antwerp University Hospital and University of Antwerp, Edegem, Belgium

Department of Anesthesia & Intensive Care, Maggiore Della Carità Hospital, Novara, Italy

Department of Neurosurgery, University Hospitals Leuven, Leuven, Belgium

Department of Neurosurgery, Clinical centre of Vojvodina, Faculty of Medicine, University of Novi Sad, Novi Sad, Serbia

Division of Anaesthesia, University of Cambridge, Addenbrooke’s Hospital, Cambridge, UK

Center for Stroke Research Berlin, Charité – Universitätsmedizin Berlin, corporate member of Freie Universität Berlin, Humboldt-Universität zu Berlin, and Berlin Institute of Health, Berlin, Germany

Intensive Care Unit, CHR Citadelle, Liège, Belgium

Department of Anaesthesiology and Intensive Therapy, University of Pécs, Pécs, Hungary

Departments of Neurology, Clinical Neurophysiology and Neuroanesthesiology, Region Hovedstaden Rigshospitalet, Copenhagen, Denmark

National Institute for Stroke and Applied Neurosciences, Faculty of Health and Environmental Studies, Auckland University of Technology, Auckland, New Zealand

Department of Neurology, Erasmus MC, Rotterdam, the Netherlands

Department of Anesthesiology and Intensive care, University Hospital Northern Norway, Tromso, Norway

Department of Neurosurgery, Hadassah-hebrew University Medical center, Jerusalem, Israel

Fundación Instituto Valenciano de Neurorrehabilitación (FIVAN), Valencia, Spain
57 Department of Neurosurgery, Shanghai Renji hospital, Shanghai Jiaotong University/school of medicine, Shanghai, China
58 Karolinska Institutet, INCF International Neuroinformatics Coordinating Facility, Stockholm, Sweden
59 Emergency Department, CHU, Liège, Belgium
60 Neurosurgery clinic, Pauls Stradins Clinical University Hospital, Riga, Latvia
61 Department of Computing, Imperial College London, London, UK
62 Department of Neurosurgery, Hospital Universitario 12 de Octubre, Madrid, Spain
63 Department of Anesthesia, Critical Care and Pain Medicine, Medical University of Vienna, Austria
64 Department of Public Health, Erasmus Medical Center-University Medical Center, Rotterdam, The Netherlands
65 College of Health and Medicine, Australian National University, Canberra, Australia
66 Department of Neurosurgery, Neurosciences Centre & JPN Apex trauma centre, All India Institute of Medical Sciences, New Delhi-110029, India
67 Department of Neurosurgery, Erasmus MC, Rotterdam, the Netherlands
68 Department of Neurosurgery, Oslo University Hospital, Oslo, Norway
69 Division of Psychology, University of Stirling, Stirling, UK
70 Division of Neurosurgery, Department of Clinical Neurosciences, Addenbrooke’s Hospital & University of Cambridge, Cambridge, UK
71 Department of Neurology, University of Groningen, University Medical Center Groningen, Groningen, Netherlands
72 Neurointensive Care, Sheffield Teaching Hospitals NHS Foundation Trust, Sheffield, UK
73 Salford Royal Hospital NHS Foundation Trust Acute Research Delivery Team, Salford, UK
74 Department of Intensive Care and Department of Ethics and Philosophy of Medicine, Erasmus Medical Center, Rotterdam, The Netherlands
75 Department of Clinical Neuroscience, Neurosurgery, Umeå University, Umeå, Sweden
Hungarian Brain Research Program - Grant No. KTIA_13_NAP-A-II/8, University of Pécs, Pécs, Hungary

Department of Anaesthesiology, University Hospital of Aachen, Aachen, Germany

Cyclotron Research Center, University of Liège, Liège, Belgium

Centre for Urgent and Emergency Care Research (CURE), Health Services Research Section, School of Health and Related Research (ScHARR), University of Sheffield, Sheffield, UK

Emergency Department, Salford Royal Hospital, Salford UK

Institute of Research in Operative Medicine (IFOM), Witten/Herdecke University, Cologne, Germany

VP Global Project Management CNS, ICON, Paris, France

Department of Anesthesiology-Intensive Care, Lille University Hospital, Lille, France

Department of Neurosurgery, Rambam Medical Center, Haifa, Israel

Department of Anesthesiology & Intensive Care, University Hospitals Southampton NHS Trust, Southampton, UK

Cologne-Merheim Medical Center (CMMC), Department of Traumatology, Orthopedic Surgery and Sportmedicine, Witten/Herdecke University, Cologne, Germany

Intensive Care Unit, Southmead Hospital, Bristol, Bristol, UK

Department of Neurological Surgery, University of California, San Francisco, California, USA

Department of Anesthesia & Intensive Care, M. Bufalini Hospital, Cesena, Italy

Department of Neurosurgery, University Hospital Heidelberg, Heidelberg, Germany

Department of Neurosurgery, The Walton centre NHS Foundation Trust, Liverpool, UK

Department of Medical Genetics, University of Pécs, Pécs, Hungary

Department of Neurosurgery, Emergency County Hospital Timisoara, Timisoara, Romania

School of Medical Sciences, Örebro University, Örebro, Sweden
Institute for Molecular Medicine Finland, University of Helsinki, Helsinki, Finland

Analytic and Translational Genetics Unit, Department of Medicine; Psychiatric & Neurodevelopmental Genetics Unit, Department of Psychiatry; Department of Neurology, Massachusetts General Hospital, Boston, MA, USA

Program in Medical and Population Genetics; The Stanley Center for Psychiatric Research, The Broad Institute of MIT and Harvard, Cambridge, MA, USA

Department of Radiology, University of Antwerp, Edegem, Belgium

Department of Anesthesiology & Intensive Care, University Hospital of Grenoble, Grenoble, France

Department of Anesthesia & Intensive Care, Azienda Ospedaliera Università di Padova, Padova, Italy

Dept. of Neurosurgery, Leiden University Medical Center, Leiden, The Netherlands and Dept. of Neurosurgery, Medical Center Haaglanden, The Hague, The Netherlands

Department of Neurosurgery, Helsinki University Central Hospital

Division of Clinical Neurosciences, Department of Neurosurgery and Turku Brain Injury Centre, Turku University Hospital and University of Turku, Turku, Finland

Department of Anesthesiology and Critical Care, Pitié -Salpêtrière Teaching Hospital, Assistance Publique, Hôpitaux de Paris and University Pierre et Marie Curie, Paris, France

Neurotraumatology and Neurosurgery Research Unit (UNINN), Vall d'Hebron Research Institute, Barcelona, Spain

Department of Neurosurgery, Kaunas University of technology and Vilnius University, Vilnius, Lithuania

Department of Neurosurgery, Rezekne Hospital, Latvia

Department of Anaesthesia, Critical Care & Pain Medicine NHS Lothian & University of Edinburg, Edinburgh, UK

Director, MRC Biostatistics Unit, Cambridge Institute of Public Health, Cambridge, UK

Department of Physical Medicine and Rehabilitation, Oslo University Hospital/University of Oslo, Oslo, Norway
111 Division of Orthopedics, Oslo University Hospital, Oslo, Norway

112 Institute of Clinical Medicine, Faculty of Medicine, University of Oslo, Oslo, Norway

113 Broad Institute, Cambridge MA Harvard Medical School, Boston MA, Massachusetts General Hospital, Boston MA, USA

114 National Trauma Research Institute, The Alfred Hospital, Monash University, Melbourne, Victoria, Australia

115 Department of Neurosurgery, Odense University Hospital, Odense, Denmark

116 International Neurotrauma Research Organisation, Vienna, Austria

117 Klinik für Neurochirurgie, Klinikum Ludwigsburg, Ludwigsburg, Germany

118 Division of Biostatistics and Epidemiology, Department of Preventive Medicine, University of Debrecen, Debrecen, Hungary

119 Department Health and Prevention, University Greifswald, Greifswald, Germany

120 Department of Anaesthesiology and Intensive Care, AUVA Trauma Hospital, Salzburg, Austria

121 Department of Neurology, Elisabeth-TweeSteden Ziekenhuis, Tilburg, the Netherlands

122 Department of Neuroanesthesia and Neurointensive Care, Odense University Hospital, Odense, Denmark

123 Department of Neuromedicine and Movement Science, Norwegian University of Science and Technology, NTNU, Trondheim, Norway

124 Department of Physical Medicine and Rehabilitation, St.Olavs Hospital, Trondheim University Hospital, Trondheim, Norway

125 Department of Neurosurgery, University of Pécs, Pécs, Hungary

126 Division of Neuroscience Critical Care, John Hopkins University School of Medicine, Baltimore, USA

127 Department of Neuropathology, Queen Elizabeth University Hospital and University of Glasgow, Glasgow, UK
Dept. of Department of Biomedical Data Sciences, Leiden University Medical Center, Leiden, The Netherlands

Department of Pathophysiology and Transplantation, Milan University, and Neuroscience ICU, Fondazione IRCCS Cà Granda Ospedale Maggiore Policlinico, Milano, Italy

Department of Radiation Sciences, Biomedical Engineering, Umeå University, Umeå, Sweden

Perioperative Services, Intensive Care Medicine and Pain Management, Turku University Hospital and University of Turku, Turku, Finland

Department of Neurosurgery, Kaunas University of Health Sciences, Kaunas, Lithuania

Intensive Care and Department of Pediatric Surgery, Erasmus Medical Center, Sophia Children’s Hospital, Rotterdam, The Netherlands

Department of Neurosurgery, Kings college London, London, UK

Neurologie, Neurochirurgie und Psychiatrie, Charité – Universitätsmedizin Berlin, Berlin, Germany

Department of Intensive Care Adults, Erasmus MC – University Medical Center Rotterdam, Rotterdam, the Netherlands

icometrix NV, Leuven, Belgium

Movement Science Group, Faculty of Health and Life Sciences, Oxford Brookes University, Oxford, UK

Psychology Department, Antwerp University Hospital, Edegem, Belgium

Director of Neurocritical Care, University of California, Los Angeles, USA

Department of Neurosurgery, St.Olavs Hospital, Trondheim University Hospital, Trondheim, Norway

Department of Emergency Medicine, University of Florida, Gainesville, Florida, USA

Department of Neurosurgery, Charité – Universitätsmedizin Berlin, corporate member of Freie Universität Berlin, Humboldt-Universität zu Berlin, and Berlin Institute of Health, Berlin, Germany
References

1. Maas AIR, Menon DK, Adelson PD, Andelic N, Bell MJ, Belli A, et al. Traumatic brain injury: integrated approaches to improve prevention, clinical care, and research. Lancet Neurol. 2017;16(12):987-1048.

2. Lingsma HF, Roozenbeek B, Steyerberg EW, Murray GD, Maas AI. Early prognosis in traumatic brain injury: from prophecies to predictions. Lancet Neurol. 2010;9(5):543-554.

3. Jennett B, Snoek J, Bond MR, Brooks N. Disability after severe head injury: observations on the use of the Glasgow Outcome Scale. J Neurol Neurosurg Psychiatry. 1981;44(4):285-293.

4. Horton L, Rhodes J, Wilson L. Randomized Controlled Trials in Adult Traumatic Brain Injury: A Systematic Review on the Use and Reporting of Clinical Outcome Assessments. J Neurotrauma. 2018;35(17):2005-2014.

5. McMillan T, Wilson L, Ponsford J, Levin H, Teasdale G, Bond M. The Glasgow Outcome Scale - 40 years of application and refinement. Nat Rev Neurol. 2016;12(8):477-485.

6. Wilson JT, Pettigrew LE, Teasdale GM. Structured interviews for the Glasgow Outcome Scale and the extended Glasgow Outcome Scale: guidelines for their use. J Neurotrauma. 1998;15(8):573-585.

7. Knaus WA, Draper EA, Wagner DP, Zimmerman JE. APACHE II: a severity of disease classification system. Crit Care Med. 1985;13(10):818-829.

8. Steyerberg EW, Mushkudiani N, Perel P, Butcher I, Lu J, McHugh GS, et al. Predicting Outcome after Traumatic Brain Injury: Development and International Validation of Prognostic Scores Based on Admission Characteristics. PLoS Med. 2008;5(8):e165.

9. Zuckerman DA, Giacino JT, Bodien YG. Traumatic Brain Injury: What Is a Favorable Outcome? J Neurotrauma. Published online 22 December 2021. doi:10.1089/neu.2021.0356

10. Turgeon AF, Lauzier F, Simard J, Scales DC, Burns KEA, Moore L, et al. Mortality associated with withdrawal of life-sustaining therapy for patients with severe traumatic brain injury: a Canadian multicentre cohort study. CMAJ. 2011;183(14):1581-1588.
11. Harrell Jr. FE, Margolis PA, Gove S, Mason KE, Mulholland EK, Lehmann D, et al. Development of a clinical prediction model for an ordinal outcome: the World Health Organization Multicentre Study of Clinical Signs and Etiological Agents of Pneumonia, Sepsis and Meningitis in Young Infants. Stat Med. 1998;17(8):909-944.

12. Hilden J. The Area under the ROC Curve and Its Competitors. Med Decis Making. 1991;11(2):95-101.

13. Van Calster B, Van Belle V, Vergouwe Y, Steyerberg EW. Discrimination ability of prediction models for ordinal outcomes: Relationships between existing measures and a new measure. Biom J. 2012;54(5):674-685.

14. Doiron D, Marcon Y, Fortier I, Burton P, Ferretti V. Software Application Profile: Opal and Mica: open-source software solutions for epidemiological data management, harmonization and dissemination. Int J Epidemiol. 2017;46(5):1372-1378.

15. Maas AIR, Menon DK, Steyerberg EW, Citerio G, Lecky F, Manley GT, et al. Collaborative European NeuroTrauma Effectiveness Research in Traumatic Brain Injury (CENTER-TBI): A Prospective Longitudinal Observational Study. Neurosurgery. 2014;76(1):67-80.

16. Steyerberg EW, Wiegens E, Sewalt C, Buki A, Citerio G, De Keyser V, et al. Case-mix, care pathways, and outcomes in patients with traumatic brain injury in CENTER-TBI: a European prospective, multicentre, longitudinal, cohort study. Lancet Neurol. 2019;18(10):923-934.

17. Wilson JTL, Edwards P, Fiddes H, Stewart E, Teasdale GM. Reliability of postal questionnaires for the Glasgow Outcome Scale. J Neurotrauma. 2002;19(9):999-1005.

18. Kunzmann K, Wernisch L, Richardson S, Steyerberg EW, Lingsma H, Ercole A, et al. Imputation of Ordinal Outcomes: A Comparison of Approaches in Traumatic Brain Injury. J Neurotrauma. 2021;38(4):455-463.

19. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. J Mach Learn Res. 2011;12:2825-2830.

20. Teasdale G, Jennett B. Assessment of coma and impaired consciousness. A practical scale. Lancet. 1974;304(7872):81-84.

21. Teasdale G, Maas A, Lecky F, Manley G, Stocchetti N, Murray G. The Glasgow Coma Scale at 40 years: standing the test of time. Lancet Neurol. 2014;13(8):844-854.

22. Dijkland SA, Foks KA, Polinder S, Dippel DWJ, Maas AIR, Lingsma HF, et al. Prognosis in Moderate and Severe Traumatic Brain Injury: A Systematic Review of Contemporary Models and Validation Studies. J Neurotrauma. 2020;37(1):1-13.
23. Han J, King NKK, Neilson SJ, Gandhi MP, Ng I. External Validation of the CRASH and IMPACT Prognostic Models in Severe Traumatic Brain Injury. *J Neurotrauma*. 2014;31(13):1146-1152.

24. Roozenbeek B, Lingsma HF, Lecky FE, Lu J, Weir J, Butcher I, *et al.* Prediction of outcome after moderate and severe traumatic brain injury: External validation of the International Mission on Prognosis and Analysis of Clinical Trials (IMPACT) and Corticoid Randomisation After Significant Head injury (CRASH) prognostic models. *Crit Care Med.* 2012;40(5):1609-1617.

25. Lingsma H, Andriessen, Teuntje M. J. C., Haitsema I, Horn J, van der Naalt J, Franschman G, *et al.* Prognosis in moderate and severe traumatic brain injury: External validation of the IMPACT models and the role of extracranial injuries. *J Trauma Acute Care Surg.* 2013;74(2):639-646.

26. Panczykowski DM, Puccio AM, Scruggs BJ, Bauer JS, Hricik AJ, Beers SR, *et al.* Prospective Independent Validation of IMPACT Modeling as a Prognostic Tool in Severe Traumatic Brain Injury. *J Neurotrauma.* 2012;29(1):47-52.

27. Murray GD, Butcher I, McHugh GS, Lu J, Mushkudiani NA, Maas AIR, *et al.* Multivariable Prognostic Analysis in Traumatic Brain Injury: Results from The IMPACT Study. *J Neurotrauma.* 2007;24(2):329-337.

28. van Buuren S, Groothuis-Oudshoorn CGM. mice: Multivariate Imputation by Chained Equations in R. *J Stat Softw.* 2011;45(3):1-67.

29. R Core Team. R: A Language and Environment for Statistical Computing. 2020; 4.0.0.

30. Seabold S, Perktold J. statsmodels: Econometric and statistical modeling with python. In: van der Walt S, Millman J, eds. *Proceedings of the 9th Python in Science Conference*. SciPy 2010:92-96.

31. Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, *et al.* PyTorch: An Imperative Style, High-Performance Deep Learning Library. In: Wallach H, Larochelle H, Beygelzimer A, d’Alché-Buc F, Fox E, Garnett R, eds. *Advances in Neural Information Processing Systems 32*. NeurIPS 2019:8024-8035.

32. CENTER-TBI Investigators and Participants. Data Dictionary. Accessed 26 January 2022. https://www.center-tbi.eu/data/dictionary.

33. Deasy J, Liò P, Ercole A. Dynamic survival prediction in intensive care units from heterogeneous time series without the need for variable selection or curation. *Sci Rep.* 2020;10(1):22129.
34. Ercole A, Dixit A, Nelson DW, Bhattacharyay S, Zeiler FA, Nieboer D, et al. Imputation strategies for missing baseline neurological assessment covariates after traumatic brain injury: A CENTER-TBI study. *PLoS ONE.* 2021;16(8):e0253425.

35. Bengio Y, Ducharme R, Vincent P, Jauvin C. A neural probabilistic language model. *J Mach Learn Res.* 2003;3:1137-1155.

36. Mikolov T, Sutskever I, Chen K, Corrado G, Dean J. Distributed Representations of Words and Phrases and their Compositionality. In: Burges CJC, Bottou L, Welling M, Ghahramani Z, Weinberger KQ, eds. *Advances in Neural Information Processing Systems 26.* NIPS 2013:3111–3119.

37. Lundberg SM, Lee S. A Unified Approach to Interpreting Model Predictions. In: Guyon I, Luxburg UV, Bengio S, Wallach H, Fergus R, Vishwanathan S, et al. eds. *Advances in Neural Information Processing Systems 30.* NIPS 2017:4765-4774.

38. Tsamardinos I, Greoudou E, Borboudakis G. Bootstrapping the out-of-sample predictions for efficient and accurate cross-validation. *Mach Learning.* 2018;107(12):1895-1922.

39. Somers RH. A New Asymmetric Measure of Association for Ordinal Variables. *Am Sociol Rev.* 1962;27(6):799-811.

40. Kim J. Predictive Measures of Ordinal Association. *Am J Sociol.* 1971;76(5):891-907.

41. Cox DR. Two further applications of a model for binary regression. *Biometrika.* 1958;45(3-4):562-565.

42. Miller ME, Langefeld CD, Tierney WM, Hui SL, McDonald CJ. Validation of Probabilistic Predictions. *Med Decis Making.* 1993;13(1):49-57.

43. Van Calster B, Nieboer D, Vergouwe Y, De Cock B, Pencina MJ, Steyerberg EW. A calibration hierarchy for risk models was defined: from utopia to empirical data. *J Clin Epidemiol.* 2016;74:167-176.

44. Austin PC, Steyerberg EW. Graphical assessment of internal and external calibration of logistic regression models by using loess smoothers. *Stat Med.* 2014;33(3):517-535.

45. Austin PC, Steyerberg EW. The Integrated Calibration Index (ICI) and related metrics for quantifying the calibration of logistic regression models. *Stat Med.* 2019;38(21):4051-4065.

46. Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat Methods.* 2020;17(3):261-272.

47. Wickham H. *ggplot2: Elegant Graphics for Data Analysis.* Springer; 2016.
48. Falcon W. PyTorch Lightning. Accessed 13 October 2021. https://github.com/PyTorchLightning/pytorch-lightning.

49. Gravesteijn BY, Nieboer D, Ercole A, Lingsma HF, Nelson D, van Calster B, et al. Machine learning algorithms performed no better than regression models for prognostication in traumatic brain injury. *J Clin Epidemiol.* 2020;122:95-107.

50. van der Naalt J, Timmerman ME, de Koning ME, van der Horn, Harm J., Scheenen ME, Jacobs B, et al. Early predictors of outcome after mild traumatic brain injury (UPFRONT): an observational cohort study. *Lancet Neurol.* 2017;16(7):532-540.

51. Kean J, Malec JF. Towards a Better Measure of Brain Injury Outcome: New Measures or a New Metric? *Arch Phys Med Rehabil.* 2014;95(7):1225-1228.

52. Futoma J, Simons M, Panch T, Doshi-Velez F, Celi LA. The myth of generalisability in clinical research and machine learning in health care. *Lancet Digit Health.* 2020;2(9):e489-e492.

53. Zeiler FA, Donnelly J, Smielewski P, Menon DK, Hutchinson PJ, Czosnyka M. Critical Thresholds of Intracranial Pressure-Derived Continuous Cerebrovascular Reactivity Indices for Outcome Prediction in Noncraniectomized Patients with Traumatic Brain Injury. *J Neurotrauma.* 2018;35(10):1107-1115.

54. Zeiler FA, Ercole A, Cabeleira M, Carbonara M, Stocchetti N, Menon DK, et al. Comparison of Performance of Different Optimal Cerebral Perfusion Pressure Parameters for Outcome Prediction in Adult Traumatic Brain Injury: A Collaborative European NeuroTrauma Effectiveness Research in Traumatic Brain Injury (CENTER-TBI) Study. *J Neurotrauma.* 2019;36(10):1505-1517.

55. Svedung Wettervik T, Howells T, Enblad P, Lewén A. Temporal Neurophysiological Dynamics in Traumatic Brain Injury: Role of Pressure Reactivity and Optimal Cerebral Perfusion Pressure for Predicting Outcome. *J Neurotrauma.* 2019;36(11):1818-1827.

56. Bhattacharyay S, Rattray J, Wang M, Dziedzic PH, Calvillo E, Kim HB, et al. Decoding accelerometry for classification and prediction of critically ill patients with severe brain injury. *Sci Rep.* 2021;11(1):23654.

57. Yuh EL, Mukherjee P, Lingsma HF, Yue JK, Ferguson AR, Gordon WA, et al. Magnetic resonance imaging improves 3-month outcome prediction in mild traumatic brain injury. *Ann Neurol.* 2013;73(2):224-235.

58. Griffin AD, Turtzo LC, Parikh GY, Tolpygo A, Lodato Z, Moses AD, et al. Traumatic microbleeds suggest vascular injury and predict disability in traumatic brain injury. *Brain.* 2019;142(11):3550-3564.
59. Wallace EJ, Mathias JL, Ward L. The relationship between diffusion tensor imaging findings and cognitive outcomes following adult traumatic brain injury: A meta-analysis. 
*Neurosci Biobehav Rev.* 2018;92:93-103.

60. Stocchetti N, Carbonara M, Citerio G, Ercole A, Skrifvars MB, Smielewski P, et al. Severe traumatic brain injury: targeted management in the intensive care unit. 
*Lancet Neurol.* 2017;16(6):452-464.

61. Wang KKW, Moghieb A, Yang Z, Zhang Z. Systems biomarkers as acute diagnostics and chronic monitoring tools for traumatic brain injury. In: Southern Š, ed. 
*Proceedings (Volume 8723) of SPIE Defense, Security, and Sensing: Sensing Technologies for Global Health, Military Medicine, and Environmental Monitoring III.* SPIE 2013.

62. Raj R, Luostarinen T, Pursiainen E, Posti JP, Takala RSK, Bendel S, et al. Machine learning-based dynamic mortality prediction after traumatic brain injury. 
*Sci Rep.* 2019;9(1):17672.

63. Meiring C, Dixit A, Harris S, MacCallum NS, Brealey DA, Watkinson PJ, et al. Optimal intensive care outcome prediction over time using machine learning. 
*PLoS ONE.* 2018;13(11):e0206862.

64. Thorsen-Meyer H, Nielsen AB, Nielsen AP, Kaas-Hansen B, Toft P, Schierbeck J, et al. Dynamic and explainable machine learning prediction of mortality in patients in the intensive care unit: a retrospective study of high-frequency data in electronic patient records. 
*Lancet Digit Health.* 2020;2(4):e179-e191.

65. Harrell Jr. FE. Ordinal Logistic Regression. In: 
*Regression Modeling Strategies: With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis.* Springer; 2015:311-325.

66. Licht C. New methods for generating significance levels from multiply-imputed data. Dr. rer. pol. thesis. The University of Bamberg, Faculty of Social Sciences, Economics, and Business Administration; 2010.
Comparison of ordinal outcome prediction to binary outcome prediction in terms of model architecture and clinical application. GOSE=Glasgow Outcome Scale – Extended at 6 months post-injury. ReLU=rectified linear unit. Pr(●)=Probability operator, i.e.,
“probability of ●.” Pr(●|○)=Conditional probability operator, i.e., “probability of ●, given ○.”

(A) Output layer architectures of binary and ordinal GOSE prediction models. Ordinal prediction models must not only have a more complicated output structure (in terms of learned weights and outcome encoding choices) but also constrain probabilities across the possible levels of functional outcome (indicated by ‘Constraint’ in the ordinal model representations). The constraint for multinomial outcome encoding is performed with a softmax activation function while the constraint for ordinal outcome encoding is performed with subtractions of output values (implemented with a negative ReLU transformation) from lower thresholds. In the provided legend formula for the softmax activation function, $z_i$ represents the outputted value of the $i^{th}$ node of the multinomial outcome encoding layer (i.e., the node representing the $i^{th}$ possible score of GOSE) preceding the softmax transformation.

(B) A sample patient case to demonstrate the difference in prognostic information between ordinal and binary GOSE prediction models. Binary models predict outcomes at one GOSE threshold while ordinal models predict outcomes at every GOSE threshold concurrently and provide conditional predictions of higher GOSE threshold outcomes given lower GOSE threshold outcomes. Bespoke conditional probability diagrams can be constructed between any number of GOSE thresholds, as desired by model users, so long as lower thresholds (e.g., GOSE > 1) precede higher thresholds (e.g., GOSE > 3) in directionality. Conditional probabilities are calculated by dividing the model probability at the higher threshold by the model probability at the lower threshold (e.g., $Pr(GOSE > 3|GOSE > 1) = Pr(GOSE > 3) / Pr(GOSE > 1)$).
Figure 2

**Figure 2** Tokenisation and embedding procedure for the development of ordinal all-predictor-based models (APMs). ICU=intensive care unit. ER=emergency room. Hx=history. SES=socioeconomic status. GOSE=Glasgow Outcome Scale – Extended at 6 months post-injury. (A) Process of converting all clinical information, from the first 24 hours of each patient, into an indexed dictionary of tokens during model training. The tokenisation process is illustrated with three example predictors and their associated values in step 2. The first entry in the trained token dictionary (‘0) <unrecognised>’) of step 3 is a placeholder token for any tokens encountered in the testing set that were not seen in the training set. (B) Visual representation of token embedding and significance-weighted averaging pipeline during APM prediction runs. After tokenising an individual patient’s clinical information, the vector of tokens is converted to a vector of the indices corresponding to each token in the trained token dictionary. The corresponding vectors and significance weights of the indices are extracted to

- ER care and ICU admission
- Brain imaging
- ICU monitoring & management
- Injury characteristics
- ICU assessments
- Labs and biomarkers
- Medical & behavioural Hx
- Medications
- Demographics and SES
- Surgery

1,151 possible predictors

ICU=intensive care unit. ER=emergency room.
weight-average the patient information into a single vector. The embedding layer and significance weights are learned through stochastic gradient descent during model training, and significance weights are constrained to be positive with an exponential function. While not explicitly shown, the weighted vectors are divided by the number of vectors during weight-averaging. The individual, weight-averaged vector then feeds into an ordinal prediction model to return probabilities at each GOSE threshold. The ordinal prediction model could either have multinomial output encoding (APM$_{MN}$) or ordinal outcome encoding (APM$_{OR}$), as represented in Fig. 1A.
Figure 3

Figure 3 Ordinal calibration curves of best-performing concise-predictor-based model (CPM_{Best}) and best-performing all-predictor-based model (APM_{Best}). GOSE=Glasgow Outcome Scale – Extended at 6 months post-injury. In each panel, a comparative histogram (200 uniform bins), centred at a horizontal line in the bottom quarter, displays the distribution of predicted probabilities for CPM_{Best} (above the line) and APM_{Best} (below the line) at the given GOSE threshold. CPM_{Best} and APM_{Best} correspond to the CPM (Supplementary Table 2) and APM (Supplementary Table 3), respectively, with the lowest unweighted average of integrated calibration indices (ICI) across the thresholds. Shaded areas are 95% confidence intervals derived using bias-corrected bootstrapping (1,000 resamples) to represent the variation across repeated k-fold cross-validation folds (20 repeats of 5 folds) and, for CPM_{Best}, 100 missing value imputations. The values in each panel correspond to the mean ICI (95% confidence interval) at the given threshold. The diagonal dashed line represents the line of perfect calibration (ICI = 0).
Figure 4

Figure 4 Mean absolute Shapley additive explanation (SHAP) values of most important predictors for multinomial-encoding all-predictor-based model (APM\textsubscript{MN}). ER=emergency room. GOS=Glasgow Outcome Scale (not extended). UO=unfavourable outcome, defined by functional dependence (i.e., GOSE \leq 4). AIS=Abbreviated Injury Scale. GOSE=Glasgow Outcome Scale – Extended at 6 months post-injury. CPM=predictors that are included in the original concise predictor set. eCPM=predictors that are added to the original concise predictor set to form the extended concise predictor set. The mean absolute SHAP value is interpreted as the average magnitude of the relative additive contribution of a predictor’s most important token towards the predicted probability at each GOSE score for a single patient. Predictor types are denoted by the coloured boundary around predictor names. Physician impression predictors denote predictors that encode the explicit impressions or rationales of ICU physicians and are not considered for the extended concise predictor set.
### Tables

#### Table 1 Summary characteristics of the study population at ICU admission stratified by ordinal 6-month outcomes

| Summary characteristics | Overall | Glasgow Outcome Scale – Extended (GOSE) at 6 months post-injury | p-value<sup>c</sup> |
|-------------------------|---------|-----------------------------------------------------------------|---------------------|
|                         |         | (1) Death (2 or 3) Vegetative or lower severe disability (4) Upper severe disability (5) Lower moderate disability (6) Upper moderate disability (7) Lower recovery (8) Upper good recovery |         |
| n<sup>a</sup>           | 1550    | 318 (20.5%) 262 (16.9%) 120 (7.7%) 227 (14.6%) 200 (12.9%) 206 (13.3%) 217 (14.0%) |         |
| Age [years]             | 51 (31–66) 66 (50–76) 55 (36–68) 48 (29–61) 44 (31–56) 41 (27–53) 48 (31–65) 41 (24–61) | <0.0001 | 0.59 |
| Sex                     |         | 409 (26.4%) 78 (24.5%) 71 (27.1%) 43 (35.8%) 64 (28.2%) 49 (24.5%) 59 (28.6%) 45 (20.7%) |         |
| Race (n<sup>b</sup> = 1427) |         | 1386 (97.1%) 281 (97.2%) 239 (96.8%) 106 (95.5%) 195 (96.5%) 183 (97.3%) 184 (98.4%) 198 (97.5%) | 0.13 |
| White                   | 21 (1.5%) 6 (2.1%) 4 (1.6%) 2 (1.8%) 2 (1.0%) 2 (1.1%) 1 (0.5%) 3 (1.5%) |         |
| Black                   | 21 (1.5%) 6 (2.1%) 4 (1.6%) 2 (1.8%) 2 (1.0%) 2 (1.1%) 1 (0.5%) 3 (1.5%) |         |
| Asian                   | 20 (1.4%) 6 (2.1%) 4 (1.6%) 2 (1.8%) 2 (1.0%) 2 (1.1%) 1 (0.5%) 3 (1.5%) |         |
| Baseline GCS (n<sup>b</sup> = 1465) |         | 8 (4–14) 5 (3–10) 6 (3–10) 8 (4–13) 8 (5–13) 9 (6–14) 13 (7–15) 13 (8–15) | <0.0001 |
| Mild [13–15]            | 390 (26.6%) 30 (10.3%) 38 (15.3%) 26 (23.4%) 42 (19.5%) 66 (34.9%) 91 (45.3%) 97 (46.4%) |         |
| Moderate [9–12]         | 331 (22.6%) 65 (22.3%) 41 (16.5%) 28 (25.2%) 65 (30.2%) 36 (19.0%) 40 (19.9%) 56 (26.8%) |         |
| Severe [3–8]            | 744 (50.8%) 196 (67.4%) 170 (68.3%) 57 (51.4%) 108 (50.2%) 87 (46.0%) 70 (34.8%) 56 (26.8%) |         |

Data are median (IQR) for continuous characteristics and n (% of column group) for categorical characteristics, unless otherwise indicated. Units or numerical definitions of characteristics are provided in square brackets. Baseline GCS=Glasgow Coma Scale at ICU admission, from 3 to 15. Conventionally, TBI severity is categorically defined by baseline GCS scores as indicated in square brackets.

<sup>a</sup>Percentages for sample size (n) represent proportion of study sample size in each GOSE group.

<sup>b</sup>Limited sample size of non-missing values for characteristic.

<sup>c</sup>p-values are determined from proportional odds logistic regression (POLR) coefficient analysis trained on all summary characteristics concurrently. For categorical variables with k > 2 categories (e.g., Race), p-values were calculated with a likelihood ratio test (with k-1 degrees of freedom) on POLR.
## Table 2: Concise baseline predictors of the study population stratified by ordinal 6-month outcomes

| Concise predictors | Overall (n = 1550) | Glasgow Outcome Scale – Extended (GOSE) at 6 months post-injury | p-value<sup>b</sup> |
|--------------------|-------------------|----------------------------------------------------------------|---------------------|
|                    | (n = 1550) | 1 (n = 318) | 2 or 3 (n = 262) | 4 (n = 120) | 5 (n = 227) | 6 (n = 200) | 7 (n = 206) | 8 (n = 217) |<p-value<sup>b</sup> |
| Age [years]         | 51 (31–66) | 66 (50–76) | 55 (36–68) | 48 (29–61) | 44 (31–56) | 41 (27–53) | 48 (31–63) | 41 (24–61) |<p-value<sup>b</sup> |
| GCSm (n= 1509)      | 5 (1–6) | 2 (1–5) | 3 (1–5) | 5 (1–6) | 5 (1–6) | 5 (2–6) | 5 (3–6) | 6 (3–6) |<p-value<sup>b</sup> |
| (1) No response     | 484 (32.1%) | 152 (50.0%) | 104 (40.6%) | 35 (29.9%) | 63 (28.5%) | 46 (23.6%) | 47 (23.0%) | 37 (17.5%) |<p-value<sup>b</sup> |
| (2) Abnormal extension | 54 (3.6%) | 17 (5.6%) | 20 (7.8%) | 4 (3.4%) | 6 (2.7%) | 3 (1.5%) | 2 (1.0%) | 2 (0.9%) |<p-value<sup>b</sup> |
| (3) Abnormal flexion | 63 (4.2%) | 14 (4.6%) | 12 (4.7%) | 8 (6.8%) | 11 (5.0%) | 8 (4.1%) | 4 (2.0%) | 6 (2.8%) |<p-value<sup>b</sup> |
| (4) Withdrawal from stimulus | 114 (7.6%) | 27 (8.9%) | 23 (9.0%) | 8 (6.8%) | 20 (9.0%) | 21 (10.8%) | 8 (3.9%) | 7 (3.3%) |<p-value<sup>b</sup> |
| (5) Movement localised to stimulus | 350 (20.2%) | 52 (17.1%) | 47 (18.4%) | 24 (20.5%) | 50 (22.6%) | 46 (23.6%) | 44 (21.6%) | 42 (19.8%) |<p-value<sup>b</sup> |
| (6) Obeyss commands | 489 (32.4%) | 42 (13.8%) | 50 (19.5%) | 38 (32.5%) | 71 (32.1%) | 71 (36.4%) | 99 (48.5%) | 118 (55.7%) |<p-value<sup>b</sup> |

Unreactive pupils (n= 1465) <p-value<sup>b</sup>

|                | (n = 1465) | 1 (n = 318) | 2 or 3 (n = 262) | 4 (n = 120) | 5 (n = 227) | 6 (n = 200) | 7 (n = 206) | 8 (n = 217) |<p-value<sup>b</sup> |
|----------------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|<p-value<sup>b</sup> |
| One            | 111 (7.6%) | 31 (10.5%) | 31 (12.3%) | 7 (6.3%) | 20 (9.3%) | 5 (2.6%) | 8 (4.1%) | 9 (4.4%) |<p-value<sup>b</sup> |
| Two            | 168 (11.5%) | 84 (28.5%) | 33 (13.0%) | 8 (7.2%) | 14 (6.5%) | 8 (4.2%) | 16 (8.2%) | 5 (2.4%) |<p-value<sup>b</sup> |
| Hypoxia        | 207 (13.4%) | 60 (18.9%) | 33 (12.6%) | 14 (11.7%) | 35 (15.4%) | 33 (16.5%) | 16 (7.8%) | 16 (7.4%) |<p-value<sup>b</sup> |
| Hypotension    | 210 (13.5%) | 56 (17.6%) | 51 (19.5%) | 21 (17.5%) | 32 (14.1%) | 22 (11.0%) | 15 (7.3%) | 13 (6.0%) |<p-value<sup>b</sup> |
| Marshall CT (n= 1255) | VI (II–VI) | III (II–VI) | II (II–VI) | II (II–VI) | II (II–III) | II (II–II) | VI (II–VI) |<p-value<sup>b</sup> |
| No visible pathology (I) | 118 (9.4%) | 8 (3.3%) | 11 (5.3%) | 5 (5.2%) | 17 (8.7%) | 25 (15.2%) | 24 (13.6%) | 28 (16.5%) |<p-value<sup>b</sup> |
| Diffuse injury II | 592 (47.2%) | 56 (22.8%) | 84 (40.6%) | 54 (56.2%) | 92 (47.2%) | 100 (60.6%) | 103 (58.5%) | 103 (60.6%) |<p-value<sup>b</sup> |
| Diffuse injury III | 108 (8.6%) | 42 (17.1%) | 17 (8.2%) | 10 (10.4%) | 14 (7.2%) | 9 (5.5%) | 6 (3.4%) | 10 (5.9%) |<p-value<sup>b</sup> |
| Diffuse injury IV | 16 (1.3%) | 7 (2.8%) | 1 (0.5%) | 1 (1.0%) | 4 (2.1%) | 1 (0.6%) | 1 (0.6%) | 1 (0.6%) |<p-value<sup>b</sup> |
| Mass lesion (V & VI) | 421 (33.5%) | 133 (54.0%) | 94 (45.4%) | 26 (27.1%) | 68 (34.9%) | 30 (18.2%) | 42 (23.9%) | 28 (16.5%) |<p-value<sup>b</sup> |
| sSAH (n= 1254) | 957 (76.3%) | 221 (90.2%) | 176 (84.2%) | 73 (76.0%) | 150 (76.9%) | 106 (63.9%) | 125 (71.4%) | 106 (63.1%) |<p-value<sup>b</sup> |
| EDH (n= 1257) | 244 (19.4%) | 31 (12.7%) | 32 (15.3%) | 21 (21.9%) | 46 (23.6%) | 32 (19.3%) | 42 (23.9%) | 40 (23.5%) |<p-value<sup>b</sup> |
| Glucose [mmol/L] (n= 1062) | 7.7 (6.6–9.4) | 8.8 (7.3–11) | 8.0 (6.5–9.8) | 7.6 (6.5–9.3) | 7.8 (6.6–9.6) | 7.7 (6.5–8.7) | 7.3 (6.3–8.5) | 7.1 (6.3–8.1) |<p-value<sup>b</sup> |
| Hb [g/dL] (n= 1140) | 13 (12–14) | 13 (11–14) | 14 (12–14) | 13 (12–14) | 14 (12–15) | 13 (12–15) | 14 (13–15) | 14 (13–15) |<p-value<sup>b</sup> |

Data are median (IQR) for continuous characteristics and n (% of column group) for categorical characteristics. Units of characteristics are provided in square brackets. GCSm= motor component score of the Glasgow Coma Scale. sSAH=traumatic subarachnoid haemorrhage. EDH=extradural haematoma. Hb=haemoglobin.<br>

<sup>a</sup>Limited sample size of non-missing values for characteristic.<br>

<sup>b</sup>p-values are determined from proportional odds logistic regression (POLR) analysis trained on all concise predictors concurrently<sup>44</sup> and are combined across 100 missing value imputations via z-transformation.<sup>46</sup> For categorical variables with k > 2 categories (e.g., GCSm), p-values were calculated with a likelihood ratio test (with k-1 degrees of freedom) on POLR.
Table 3  Predictor baseline tokens per patient in the CENTER-TBI dataset

| Predictor category                     | Types of tokens | Fixed at ICU admission | Continuous variable | Treatments and interventions | Physician impression or rationale |
|----------------------------------------|----------------|------------------------|---------------------|-------------------------------|----------------------------------|
| All                                    |                | 112 (103–121)          | 13 (10–16)          | 0 (0–0)                       | 7 (7–8)                          |
| Emergency care and ICU admission       |                | 112 (103–121)          | 13 (10–16)          | 0 (0–0)                       | 7 (7–8)                          |
| Brain imaging                          |                | 74 (68–83)             | 5 (2–8)             | 0 (0–0)                       | 9 (8–10)                         |
| ICU monitoring and management          |                | 63 (52–72)             | 3 (3–3)             | 10 (5–13)                     | 40 (34–46)                       | 13 (3–15)                       |
| Injury characteristics and severity    |                | 55 (49–62)             | 2 (2–2)             | 0 (0–0)                       | 0 (0–0)                          |
| End-of-day assessments                 |                | 0 (0–0)                | 19 (17–21)          | 0 (0–0)                       | 0 (0–0)                          |
| Laboratory measurements                |                | 14 (0–20)              | 42 (31–52)          | 0 (0–0)                       | 1 (1–1)                          |
| Medical and behavioural history        |                | 38 (32–51)             | 0 (0–1)             | 0 (0–0)                       | 0 (0–0)                          |
| Medications                            |                | 30 (21–40)             | 0 (0–0)             | 0 (0–0)                       | 22 (15–30)                      | 8 (5–11)                        |
| Bihourly assessments                   |                | 17 (0–32)              | 0 (0–0)             | 15 (0–27)                     | 1 (0–2)                         | 0 (0–0)                          |
| Demographics and socioeconomic status  |                | 15 (14–16)             | 2 (1–2)             | 0 (0–0)                       | 0 (0–0)                          |
| Protein biomarkers                     |                | 5 (5–5)                | 0 (0–0)             | 5 (5–5)                       | 0 (0–0)                          | 0 (0–0)                          |
| Surgery                                |                | 2 (1–6)                | 1 (1–2)             | 0 (0–0)                       | 0 (0–1)                         | 1 (0–3)                          |
| Haemostatic markers*                   |                | 0 (0–0)                | 0 (0–0)             | 0 (0–0)                       | 0 (0–0)                          |
| Transitions of care*                   |                | 0 (0–0)                | 0 (0–0)             | 0 (0–0)                       | 0 (0–0)                          |
| All predictors                         |                | 532 (486–580)          | 315 (288–341)       | 111 (90–132)                  | 64 (50–75)                      | 37 (29–44)                       |

Data represent median (IQR) number of non-missing, unique tokens per patient. Tokens were extracted from the clinical information available up to 24 hours after ICU admission for each study patient in the Collaborative European NeuroTrauma Effectiveness Research in TBI (CENTER-TBI) project dataset. Each token may be of only one predictor category (leftmost column) and of any number of token types (four rightmost columns). ICU=intensive care unit.

*Due to their relative infrequency in the CENTER-TBI dataset, these baseline predictor categories have a 3rd quartile of zero tokens per patient.
Table 4: Best ordinal model discrimination and calibration performance per predictor set

| Metric                        | Threshold | Model          | CPM<sub>Best</sub> | APM<sub>Best</sub> | eCPM<sub>Best</sub> |
|-------------------------------|-----------|----------------|---------------------|---------------------|---------------------|
| Ordinal c-index (ORC)         |           |                | 0.70 (0.68–0.71)    | 0.76 (0.74–0.77)    | 0.73 (0.71–0.74)    |
| Somers’ D<sub>xy</sub>         |           |                | 0.44 (0.41–0.48)    | 0.57 (0.54–0.60)    | 0.50 (0.46–0.54)    |
| Threshold-level dichotomous c-index* |           |                | 0.77 (0.75–0.78)    | 0.82 (0.80–0.83)    | 0.79 (0.78–0.80)    |
| GOSE > 1                      |           |                | 0.83 (0.81–0.85)    | 0.90 (0.88–0.92)    | 0.86 (0.84–0.87)    |
| GOSE > 3                      |           |                | 0.81 (0.79–0.83)    | 0.86 (0.84–0.88)    | 0.84 (0.83–0.86)    |
| GOSE > 4                      |           |                | 0.78 (0.76–0.80)    | 0.83 (0.80–0.85)    | 0.82 (0.80–0.83)    |
| GOSE > 5                      |           |                | 0.76 (0.74–0.77)    | 0.80 (0.78–0.83)    | 0.77 (0.75–0.79)    |
| GOSE > 6                      |           |                | 0.72 (0.70–0.74)    | 0.76 (0.73–0.79)    | 0.75 (0.73–0.77)    |
| GOSE > 7                      |           |                | 0.72 (0.69–0.74)    | 0.75 (0.72–0.79)    | 0.72 (0.70–0.75)    |
| Threshold-level calibration slope* |           |                | 0.98 (0.81–1.12)    | 0.84 (0.76–0.91)    | 1.00 (0.78–1.14)    |
| GOSE > 1                      |           |                | 0.95 (0.78–1.10)    | 0.98 (0.86–1.10)    | 0.98 (0.78–1.14)    |
| GOSE > 3                      |           |                | 0.97 (0.80–1.12)    | 0.90 (0.80–1.02)    | 1.05 (0.81–1.20)    |
| GOSE > 4                      |           |                | 1.06 (0.86–1.23)    | 0.89 (0.79–1.00)    | 1.10 (0.85–1.27)    |
| GOSE > 5                      |           |                | 1.01 (0.78–1.21)    | 0.82 (0.72–0.94)    | 1.01 (0.76–1.22)    |
| GOSE > 6                      |           |                | 0.98 (0.73–1.20)    | 0.74 (0.62–0.87)    | 0.97 (0.70–1.20)    |
| GOSE > 7                      |           |                | 0.92 (0.69–1.18)    | 0.68 (0.54–0.83)    | 0.89 (0.61–1.18)    |

Data represent mean (95% confidence interval) for the best-performing model, per predictor set, based on a given metric. For threshold-level metrics, a single best-performing model, per predictor set, was determined by the overall unweighted average across the thresholds. Interpretations for each metric are provided in the Materials and methods. Mean and confidence interval values were derived using bias-corrected bootstrapping (1,000 resamples) and represent the variation across repeated k-fold cross-validation folds (20 repeats of 5 folds) and, for the concise-predictor-based model (CPM) and the extended concise-predictor-based model (eCPM), 100 missing value imputations. CPM<sub>Best</sub>=CPM with best value for given metric (Supplementary Table 2). APM<sub>Best</sub>=all-predictor-based model (APM) with best value for given metric (Supplementary Table 3). eCPM<sub>Best</sub>=eCPM with best value for given metric (Supplementary Table 4). GOSE=Glasgow Outcome Scale – Extended at 6 months post-injury.

*Values in these rows correspond to the unweighted average across all GOSE thresholds.
Supplementary figures

Supplementary figure 1

Supplementary Figure 1 CONSORT-style flow diagram for patient enrolment and follow-up. CENTER-TBI=Collaborative European NeuroTrauma Effectiveness Research in TBI. ICU=intensive care unit. GOSE=Glasgow Outcome Scale – Extended. MSM=Markov multi-state model (see Materials and methods). The dashed, olive-green line in the lower-middle of the diagram divides the enrolment flow diagram (above) and the follow-up breakdown (below).
Supplementary figure 2

Characterisation of missingness among concise predictor set.

U.P.= unreactive pupils. GCSm= motor component score of the Glasgow Coma Scale. Hb= haemoglobin. Glu.= glucose. HoTN= hypotension. Marshall= Marshall computerised tomography classification. tSAH= traumatic subarachnoid haemorrhage. EDH= extradural haematoma. (A) Proportion of total sample size ($n = 1,550$) with missing values for each IMPACT extended model predictor. (B) Combinations of missing predictors (red) and non-missing predictors (blue) and the histogram (far right) of proportions of total sample size associated with each combination.
Supplementary Figure 3 Ordinal calibration curves of each concise-predictor-based model (CPM). GOSE=Glasgow Outcome Scale – Extended at 6 months post-injury. Shaded areas are 95% confidence intervals derived using bias-corrected bootstrapping (1,000 resamples) to represent the variation across repeated k-fold cross-validation folds (20 repeats of 5 folds) and 100 missing value imputations. The values in each panel correspond to the mean integrated calibration index (ICI) (95% confidence interval) at the given threshold. The diagonal dashed line represents the line of perfect calibration (ICI = 0). The CPM types (CPM_{MNL}, CPM_{POL}, CPM_{DeepMN}, and CPM_{DeepOR}) are decoded in the Materials and methods and described in Supplementary Methods 1.
**Supplementary Figure 4 Ordinal calibration curves of each all-predictor-based model (APM).** GOSE=Glascow Outcome Scale – Extended at 6 months post-injury. Shaded areas are 95% confidence intervals derived using bias-corrected bootstrapping (1,000 resamples) to represent the variation across repeated k-fold cross-validation folds (20 repeats of 5 folds). The values in each panel correspond to the mean integrated calibration index (ICI) (95% confidence interval) at the given threshold. The diagonal dashed line represents the line of perfect calibration (ICI = 0). The APM types (APM_{MN} and APM_{OR}) are decoded in the Materials and methods and described in Supplementary Methods 2.
Supplementary figure 5

Supplementary Figure 5 Ordinal calibration curves of each extended concise-predictor-based model (eCPM). GOSE=Glasgow Outcome Scale – Extended at 6 months post-injury. Shaded areas are 95% confidence intervals derived using bias-corrected bootstrapping (1,000 resamples) to represent the variation across repeated k-fold cross-validation folds (20 repeats of 5 folds) and 100 missing value imputations. The values in each panel correspond to the mean integrated calibration index (ICI) (95% confidence interval) at the given threshold. The diagonal dashed line represents the line of perfect calibration (ICI = 0). The eCPM types (eCPM_MNLR, eCPM_POLR, eCPM_DeepMN, and eCPM_DeepOR) are decoded in the Materials and methods and described in Supplementary Methods 1.
Supplementary tables

Supplementary Table 1. Extended concise baseline predictors of the study population stratified by ordinal 6-month outcomes

| Extended concise predictors                  | Overall (n = 1550) | Glasgow Outcome Scale–Extended (GOSE) at 6 months post-injury | p-value* |
|---------------------------------------------|--------------------|----------------------------------------------------------------|----------|
| Age [years]                                 | 51 (31–66)         | 66 (50–76) 55 (36–68) 48 (29–61) 44 (31–56) 41 (27–53) 48 (31–65) 41 (24–61) | <0.0001  |
| GCSm (n = 1509)                              | 5 (1–6)            | 2 (1–5) 3 (1–5) 5 (1–6) 5 (1–6) 5 (2–6) 5 (3–6) 6 (5–6)         | <0.0001  |
| (1) No response                             | 484 (32.1%)        | 152 (50.0%) 104 (40.6%) 35 (29.9%) 63 (28.5%) 46 (23.6%) 47 (23.0%) 37 (17.5%) |         |
| (2) Abnormal extension                      | 54 (3.6%)          | 17 (5.6%) 20 (7.8%) 4 (3.4%) 6 (2.7%) 3 (1.5%) 2 (1.0%) 2 (0.9%)         |         |
| (3) Abnormal flexion                        | 63 (4.2%)          | 14 (4.6%) 12 (4.7%) 8 (6.8%) 11 (5.0%) 8 (4.2%) 6 (2.4%)         |         |
| (4) Withdrawal from stimulus                | 114 (7.6%)         | 27 (8.9%) 23 (9.0%) 8 (6.8%) 20 (9.0%) 21 (10.8%) 8 (3.9%) 7 (3.3%) |         |
| (5) Movement localised to stimulus          | 305 (20.2%)        | 52 (17.1%) 47 (18.4%) 24 (20.5%) 50 (22.6%) 46 (23.6%) 44 (21.6%) 42 (19.8%) |         |
| (6) Obey's commands                         | 469 (32.4%)        | 42 (13.8%) 50 (19.5%) 38 (32.5%) 71 (32.1%) 71 (36.4%) 99 (48.5%) 118 (55.7%) |         |
| Unreactive pupils (n = 1465)                |                    | 111 (7.6%) 31 (10.5%) 31 (12.3%) 7 (6.3%) 20 (9.3%) 5 (2.6%) 8 (4.1%) 9 (4.4%) | <0.0001  |
| One                                         | 168 (11.5%)        | 84 (28.5%) 33 (13.0%) 8 (7.2%) 14 (6.5%) 8 (4.2%) 16 (8.2%) 5 (2.4%) |         |
| Hypoxia                                     | 207 (13.4%)        | 60 (18.9%) 33 (12.6%) 14 (11.7%) 35 (15.4%) 33 (16.5%) 16 (7.8%) 16 (7.4%) 0.6272 |
| Hypotension                                 | 210 (13.5%)        | 56 (17.6%) 51 (19.5%) 21 (17.5%) 32 (14.1%) 22 (11.0%) 15 (7.3%) 13 (6.0%) 0.0038 |
| Marshall CT (n = 1255)                      | VI (II–VI)         | III (II–VI) II (II–VI) II (II–II) II (II–III) II (II–II) VI (II–VI) | 0.0386  |
| No visible pathology (I)                    | 118 (9.4%)         | 8 (3.3%) 11 (5.3%) 5 (5.2%) 17 (8.7%) 25 (15.2%) 24 (13.6%) 28 (16.5%) |         |
| Diffuse injury II                           | 592 (47.2%)        | 56 (22.8%) 84 (40.6%) 54 (26.2%) 92 (47.2%) 100 (60.6%) 103 (58.5%) 103 (60.6%) |         |
| Diffuse injury III                          | 108 (8.6%)         | 42 (17.1%) 17 (8.2%) 10 (10.4%) 14 (7.2%) 9 (5.5%) 6 (3.4%) 10 (5.9%) |         |
| Diffuse injury IV                           | 16 (1.3%)          | 7 (2.8%) 1 (0.5%) 1 (1.0%) 4 (2.1%) 1 (0.6%) 1 (0.6%) |         |
| Mass lesion (V & VI)                        | 421 (33.5%)        | 133 (54.0%) 94 (45.4%) 26 (27.1%) 68 (34.9%) 30 (18.2%) 42 (23.9%) 28 (16.5%) |         |
| cSAH (n = 1254)                              | 957 (76.3%)        | 221 (90.2%) 176 (84.2%) 73 (76.0%) 150 (76.9%) 106 (63.9%) 125 (71.4%) 106 (63.1%) 0.4429 |
| EEDH (n = 1257)                              | 244 (19.4%)        | 31 (12.7%) 32 (15.3%) 21 (21.9%) 46 (23.6%) 32 (19.3%) 42 (23.9%) 40 (23.5%) 0.0035 |
| Glucose [mmol/L] (n = 1062)                  | 7.7 (6.6–9.4)      | 8.8 (7.3–11) 8.0 (6.5–9.8) 7.6 (6.5–9.3) 7.8 (6.6–9.6) 7.7 (6.5–8.7) 7.3 (6.3–8.5) 7.1 (6.3–8.1) 0.0123 |
| Outcome                              | n     | Median | IQR     |
|--------------------------------------|-------|--------|---------|
| Hb [g/dL] (n = 1140)                 | 13 (12–14) | 13 (11–14) | 13 (11–14) | 14 (12–14) | 13 (12–14) | 14 (12–15) | 13 (12–15) | 14 (13–15) | 0.3044 |
| Retired (n = 1372)                   | 353 (26.9%) | 136 (61.3%) | 74 (33.6%) | 23 (22.1%) | 12 (5.9%) | 13 (7.3%) | 52 (28.1%) | 43 (21.8%) | 0.0644 |
| Highest formal education (n = 1110)  |       |        |         |
| None                                 | 15 (1.4%) | 3 (2.4%) | 4 (2.1%) | 2 (2.0%) | 2 (1.1%) | 2 (1.1%) | 0 (0%) | 2 (1.1%) | 0.4897 |
| In degree program                    | 26 (2.3%) | 0 (0%) | 5 (2.6%) | 0 (0%) | 4 (2.1%) | 7 (4.1%) | 4 (2.5%) | 6 (3.4%)|
| Primary school                       | 155 (14.0%) | 31 (24.6%) | 44 (23.3%) | 14 (13.9%) | 17 (8.9%) | 16 (9.5%) | 14 (8.8%) | 19 (10.9%)|
| Secondary school                     | 458 (41.3%) | 50 (39.7%) | 63 (33.3%) | 46 (45.5%) | 80 (42.1%) | 59 (34.9%) | 75 (46.9%) | 85 (48.6%)|
| Technical certificate                | 235 (21.2%) | 16 (12.7%) | 38 (20.1%) | 21 (20.8%) | 57 (30.0%) | 43 (25.4%) | 32 (20.0%) | 28 (16.0%)|
| University degree                    | 221 (19.9%) | 26 (20.6%) | 35 (18.5%) | 18 (17.8%) | 30 (15.8%) | 42 (24.9%) | 35 (21.9%) | 35 (20.0%)|
| GFAP [ng/mL] (n = 1247)              | 17 (6–46) | 48 (13–96) | 32 (11–61) | 17 (6–43) | 13 (5–30) | 15 (3–30) | 10 (3–23) | 9 (3–22) | 0.0005 |
| T-tau [pg/mL] (n = 1248)             | 8 (4–19) | 17 (7–38) | 12 (6–23) | 9 (5–19) | 7 (3–14) | 7 (3–13) | 5 (3–12) | 6 (3–11) | 0.2568 |
| S100B [ng/mL] (n = 1267)             | 0.3 (2–6) | 0.6 (3–1.3) | 0.4 (2–6) | 0.3 (2–6) | 0.3 (2–4) | 0.2 (2–4) | 0.2 (1–5) | 0.2 (1–3) | 0.1929 |
| NFL [pg/mL] (n = 1247)               | 55 (28–127) | 121 (51–268) | 85 (46–150) | 61 (32–150) | 48 (28–87) | 41 (21–87) | 30 (17–60) | 35 (19–74) | 0.0354 |
| PTA (n = 1530)                       | 187 (12.2%) | 5 (1.6%) | 15 (5.8%) | 10 (8.5%) | 43 (19.3%) | 33 (16.8%) | 50 (24.4%) | 31 (14.4%) | 0.0010 |
| Worst head/neck, brain, or cervical spine AIS (n = 1523) |       |        |         |
| (1) Minor                            | 50 (3.2%) | 6 (1.9%) | 3 (1.1%) | 5 (4.2%) | 5 (2.2%) | 4 (2.0%) | 16 (7.8%) | 11 (5.1%) |
| (2) Moderate                         | 31 (2.0%) | 3 (0.9%) | 3 (1.1%) | 0 (0%) | 5 (2.2%) | 4 (2.0%) | 8 (3.9%) | 8 (3.7%) |
| (3) Severe                           | 112 (7.2%) | 6 (1.9%) | 6 (2.3%) | 7 (5.8%) | 21 (9.3%) | 19 (9.5%) | 25 (12.1%) | 28 (12.9%) |
| (4) Severe                           | 484 (31.2%) | 63 (19.8%) | 54 (20.6%) | 37 (30.8%) | 71 (31.3%) | 78 (39.0%) | 87 (42.2%) | 94 (43.3%) |
| (5) Critical                         | 846 (54.6%) | 216 (67.9%) | 195 (74.4%) | 70 (58.3%) | 125 (55.1%) | 94 (47.0%) | 70 (34.0%) | 76 (35.0%) |
| (6) Not survivable                   | 27 (1.7%) | 24 (7.5%) | 1 (0.4%) | 1 (0.8%) | 0 (0%) | 1 (0.5%) | 0 (0%) | 0 (0%) |

Data are median (IQR) for continuous characteristics and n (% of column group) for categorical characteristics. Units of characteristics are provided in square brackets. GCSm=motor component score of the Glasgow Coma Scale. Marshall CT=Marshall computerised tomography classification. tSAH=traumatic subarachnoid haemorrhage. EDH=extradural haematoma. Glu=glucose. Hb=haemoglobin. GFAP=glial fibrillary acidic protein. T-tau=total tau protein. S100B=S100 calcium-binding protein B.NFL=neurofilament light chain. PTA=incidence of post-traumatic amnesia. AIS=abbreviated injury scale.

*Limited sample size of non-missing values for characteristic.

*p-values are determined from proportional odds logistic regression analysis trained on all concise predictors concurrently and are combined across 100 missing value imputations via z-transformation. For categorical variables with k > 2 categories (e.g., GCSm), p-values were calculated with a likelihood ratio test (with k-1 degrees of freedom) on POLR.
## Supplementary Table 2. Ordinal concise-predictor-based model (CPM) discrimination and calibration performance

| Metric                          | Threshold | Model          | CPM\textsubscript{MNLR} | CPM\textsubscript{POLR} | CPM\textsubscript{DeepMN} | CPM\textsubscript{DeepOR} |
|---------------------------------|-----------|----------------|--------------------------|--------------------------|----------------------------|-----------------------------|
| Ordinal c-index (ORC)           |           |                |                          |                          |                            |                             |
|                                 |           |                | 0.69 (0.67–0.70)         | 0.69 (0.68–0.70)         | 0.70 (0.68–0.71)           | 0.59 (0.58–0.61)            |
| Somers' $D_{xy}$                |           |                | 0.43 (0.41–0.45)         | 0.43 (0.41–0.46)         | 0.44 (0.41–0.48)           | 0.23 (0.20–0.26)            |
| Threshold-level dichotomous c-index\(^a\) |           |                | 0.77 (0.75–0.78)         | 0.77 (0.75–0.78)         | 0.76 (0.74–0.78)           | 0.76 (0.73–0.78)            |
| GOSE > 1                        |           |                | 0.83 (0.81–0.85)         | 0.83 (0.81–0.84)         | 0.83 (0.80–0.86)           | 0.82 (0.79–0.85)            |
| GOSE > 3                        |           |                | 0.81 (0.79–0.83)         | 0.81 (0.79–0.82)         | 0.80 (0.78–0.83)           | 0.80 (0.77–0.82)            |
| GOSE > 4                        |           |                | 0.78 (0.76–0.80)         | 0.78 (0.76–0.79)         | 0.77 (0.74–0.80)           | 0.77 (0.74–0.79)            |
| GOSE > 5                        |           |                | 0.76 (0.74–0.77)         | 0.76 (0.74–0.77)         | 0.75 (0.72–0.78)           | 0.74 (0.71–0.77)            |
| GOSE > 6                        |           |                | 0.72 (0.70–0.74)         | 0.71 (0.69–0.73)         | 0.71 (0.68–0.74)           | 0.71 (0.67–0.74)            |
| GOSE > 7                        |           |                | 0.72 (0.69–0.74)         | 0.73 (0.70–0.75)         | 0.71 (0.67–0.75)           | 0.71 (0.67–0.75)            |
| Threshold-level calibration slope\(^*\) |           |                | 0.85 (0.78–0.91)         | 0.94 (0.88–1.01)         | 0.98 (0.81–1.12)           | 0.90 (0.79–1.02)            |
| GOSE > 1                        |           |                | 0.92 (0.84–1.00)         | 1.13 (1.04–1.23)         | 0.95 (0.78–1.10)           | 1.01 (0.85–1.18)            |
| GOSE > 3                        |           |                | 0.92 (0.85–1.00)         | 1.14 (1.05–1.23)         | 0.97 (0.80–1.12)           | 0.95 (0.83–1.09)            |
| GOSE > 4                        |           |                | 0.91 (0.84–1.00)         | 0.99 (0.91–1.08)         | 1.06 (0.86–1.23)           | 0.93 (0.80–1.06)            |
| GOSE > 5                        |           |                | 0.88 (0.80–0.97)         | 0.90 (0.82–0.99)         | 1.01 (0.78–1.21)           | 0.90 (0.76–1.06)            |
| GOSE > 6                        |           |                | 0.81 (0.71–0.91)         | 0.71 (0.63–0.80)         | 0.98 (0.73–1.20)           | 0.86 (0.67–1.06)            |
| GOSE > 7                        |           |                | 0.64 (0.50–0.80)         | 0.77 (0.67–0.88)         | 0.92 (0.69–1.18)           | 0.78 (0.57–1.02)            |

Data represent mean (95% confidence interval) for the CPM based on a given metric. Interpretations for each metric are provided in Materials and methods. Mean and confidence interval values were derived using bias-corrected bootstrapping (1,000 resamples) and represent the variation across repeated $k$-fold cross-validation folds (20 repeats of 5 folds) and 100 missing value imputations. GOSE=Glasgow Outcome Scale – Extended at 6 months post-injury. The CPM types (CPM\textsubscript{MNLR}, CPM\textsubscript{POLR}, CPM\textsubscript{DeepMN}, and CPM\textsubscript{DeepOR}) are decoded in the Materials and methods and described in Supplementary Methods 1.

\(^a\)Values in these rows correspond to the unweighted average across all GOSE thresholds.
Supplementary Table 3. Ordinal all-predictor-based model (APM) discrimination and calibration performance

| Metric                        | Threshold | Model   | APM<sub>MN</sub> | APM<sub>OR</sub> |
|-------------------------------|-----------|---------|------------------|------------------|
| Ordinal c-index (ORC)         |           |         | 0.76 (0.74–0.77) | 0.66 (0.65–0.68) |
| Somers' D<sub>xy</sub>         |           |         | 0.57 (0.54–0.60) | 0.37 (0.33–0.40) |
| Threshold-level dichotomous c-index<sup>a</sup> |           |         |                  |                  |
| GOSE > 1                      |           |         | 0.90 (0.88–0.92) | 0.83 (0.81–0.85) |
| GOSE > 3                      |           |         | 0.86 (0.84–0.88) | 0.82 (0.80–0.84) |
| GOSE > 4                      |           |         | 0.83 (0.80–0.85) | 0.80 (0.78–0.82) |
| GOSE > 5                      |           |         | 0.80 (0.78–0.83) | 0.78 (0.75–0.80) |
| GOSE > 6                      |           |         | 0.76 (0.73–0.79) | 0.74 (0.71–0.77) |
| GOSE > 7                      |           |         | 0.75 (0.72–0.79) | 0.71 (0.68–0.75) |
| Threshold-level calibration slope<sup>a</sup> |           |         |                  |                  |
| GOSE > 1                      |           |         | 0.98 (0.86–1.10) | 0.35 (0.31–0.38) |
| GOSE > 3                      |           |         | 0.90 (0.80–1.02) | 0.18 (0.16–0.21) |
| GOSE > 4                      |           |         | 0.89 (0.79–1.00) | 0.10 (0.09–0.12) |
| GOSE > 5                      |           |         | 0.82 (0.72–0.94) | 0.07 (0.06–0.09) |
| GOSE > 6                      |           |         | 0.74 (0.62–0.87) | 0.06 (0.05–0.07) |
| GOSE > 7                      |           |         | 0.68 (0.54–0.83) | 0.05 (0.04–0.06) |

Data represent mean (95% confidence interval) for the APM based on a given metric. Interpretations for each metric are provided in Materials and methods. Mean and confidence interval values were derived using bias-corrected bootstrapping (1,000 resamples) and represent the variation across repeated k-fold cross-validation folds (20 repeats of 5 folds). GOSE=Glasgow Outcome Scale – Extended at 6 months post-injury. The APM types (APM<sub>MN</sub> and APM<sub>OR</sub>) are decoded in the Materials and methods and described in Supplementary Methods 2.

<sup>a</sup>Values in these rows correspond to the unweighted average across all GOSE thresholds.
### Supplementary Table 4. Ordinal extended concise-predictor-based model (eCPM) discrimination and calibration performance

| Metric                          | Threshold | Model          | eCPM\textsubscript{MNLR} | eCPM\textsubscript{POLR} | eCPM\textsubscript{DeepMN} | eCPM\textsubscript{DeepOR} |
|---------------------------------|-----------|----------------|--------------------------|--------------------------|-----------------------------|-----------------------------|
| Ordinal c-index (ORC)           |           |                | 0.72 (0.71–0.73)         | 0.71 (0.70–0.72)         | 0.73 (0.71–0.74)            | 0.67 (0.65–0.68)            |
| Somers’ D\textsubscript{xy}     |           |                | 0.50 (0.48–0.52)         | 0.47 (0.45–0.49)         | 0.50 (0.46–0.54)            | 0.38 (0.35–0.41)            |
| Threshold-level dichotomous c-index\textsuperscript{a} | GOSE > 1  | 0.79 (0.78–0.80) | 0.79 (0.78–0.80) | 0.79 (0.77–0.81) | 0.77 (0.76–0.79) |
|                                 | GoSE > 3  | 0.86 (0.84–0.87) | 0.85 (0.84–0.87) | 0.86 (0.83–0.88) | 0.85 (0.82–0.87) |
|                                 | GoSE > 4  | 0.84 (0.83–0.86) | 0.84 (0.83–0.85) | 0.84 (0.82–0.86) | 0.83 (0.81–0.85) |
|                                 | GoSE > 5  | 0.82 (0.80–0.83) | 0.81 (0.80–0.83) | 0.81 (0.79–0.83) | 0.80 (0.77–0.82) |
|                                 | GoSE > 6  | 0.77 (0.75–0.79) | 0.77 (0.76–0.79) | 0.77 (0.74–0.80) | 0.76 (0.73–0.78) |
|                                 | GoSE > 7  | 0.75 (0.73–0.77) | 0.73 (0.71–0.75) | 0.74 (0.70–0.77) | 0.72 (0.69–0.75) |
| Threshold-level calibration slope\textsuperscript{a} | GOSE > 1  | 0.75 (0.70–0.81) | 0.89 (0.83–0.95) | 1.00 (0.78–1.14) | 0.59 (0.51–0.67) |
|                                 | GoSE > 3  | 0.81 (0.75–0.89) | 0.97 (0.87–1.10) | 0.98 (0.78–1.14) | 1.04 (0.90–1.20) |
|                                 | GoSE > 4  | 0.83 (0.77–0.90) | 1.12 (1.04–1.23) | 1.05 (0.81–1.20) | 0.79 (0.68–0.90) |
|                                 | GoSE > 5  | 0.81 (0.75–0.89) | 1.02 (0.94–1.11) | 1.10 (0.85–1.27) | 0.60 (0.52–0.69) |
|                                 | GoSE > 6  | 0.75 (0.67–0.82) | 0.86 (0.78–0.94) | 1.01 (0.76–1.22) | 0.47 (0.38–0.56) |
|                                 | GoSE > 7  | 0.72 (0.63–0.81) | 0.69 (0.62–0.77) | 0.97 (0.70–1.20) | 0.36 (0.27–0.46) |
|                                 |           |                | 0.72 (0.70–0.75) | 0.73 (0.70–0.75) | 0.72 (0.68–0.76) | 0.70 (0.66–0.74) |
|                                 |           |                | 0.89 (0.83–0.95) | 1.00 (0.78–1.14) | 0.59 (0.51–0.67) |
|                                 |           |                | 0.97 (0.87–1.10) | 0.98 (0.78–1.14) | 1.04 (0.90–1.20) |
|                                 |           |                | 1.12 (1.04–1.23) | 1.05 (0.81–1.20) | 0.79 (0.68–0.90) |
|                                 |           |                | 1.02 (0.94–1.11) | 1.10 (0.85–1.27) | 0.60 (0.52–0.69) |
|                                 |           |                | 0.86 (0.78–0.94) | 1.01 (0.76–1.22) | 0.47 (0.38–0.56) |
|                                 |           |                | 0.69 (0.62–0.77) | 0.97 (0.70–1.20) | 0.36 (0.27–0.46) |
|                                 |           |                | 0.68 (0.59–0.77) | 0.89 (0.61–1.18) | 0.28 (0.16–0.40) |

Data represent mean (95% confidence interval) for the eCPM based on a given metric. Interpretations for each metric are provided in Materials and methods. Mean and confidence interval values were derived using bias-corrected bootstrapping (1,000 resamples) and represent the variation across repeated k-fold cross-validation folds (20 repeats of 5 folds) and 100 missing value imputations. GOSE=Glasgow Outcome Scale – Extended at 6 months post-injury. The eCPM types (eCPM\textsubscript{MNLR}, eCPM\textsubscript{POLR}, eCPM\textsubscript{DeepMN}, and eCPM\textsubscript{DeepOR}) are decoded in the Materials and methods and described in Supplementary Methods 1. \textsuperscript{a}Values in these rows correspond to the unweighted average across all GOSE thresholds.
Supplementary methods

Supplementary Methods 1. Explanation of selected ordinal prediction models for CPM and eCPM

Multinomial logistic regression (MNLR)

CPM_{MNLR} and eCPM_{MNLR} were implemented using the ‘MNLogit’ class from the ‘statsmodels’ module (dev. v0.14.0) [1] in Python (v3.7.6). The GOSE score of 1 (death) was designated as the reference label, and, for each other GOSE score, a separate logistic model was trained to regress the logit of the ratio of the probability of that score to the reference score from a linear combination of the predictors. The logit outputs of each model feed into a softmax function, after which cumulative sums would determine the probability at each threshold. Model weights for MNLR were optimised using the Broyden–Fletcher–Goldfarb–Shanno (BFGS) algorithm [2] to maximize conditional likelihood.

Proportional odds (i.e., ordinal) logistic regression (POLR)

CPM_{POLR} and eCPM_{POLR} were implemented using the ‘OrderedModel’ class from the ‘statsmodels’ module in Python. The model maps GOSE scores to a latent, logit space where consecutive GOSE scores are separated by thresholds. Thus, the model trains only one set of linear predictor weights, but a separate intercept for each threshold. Model weights for POLR were optimised using the Broyden–Fletcher–Goldfarb–Shanno (BFGS) algorithm [2] to maximize conditional likelihood.

Class-weighted feedforward neural network with a multinomial output layer (DeepMN)

CPM_{DeepMN} and eCPM_{DeepMN} were implemented using the ‘PyTorch’ (v1.10.0) [3] module in Python. The network architecture of DeepMN included a hyperparametric number of dense hidden layers (either 1, 2, 3, 4, 5, or 6), each containing a hyperparametric number of nodes (either 128, 256, or 512) with a rectified linear unit (ReLU) activation function and a hyperparametric percentage (either 0% or 20%) dropout during training. The output layer of DeepMN was a softmax layer of 7 nodes, from which probabilities at each GOSE are calculated with cumulative sums (Fig. 1A). DeepMN was optimised using the Adam algorithm (γ [learning rate] = 0.001, β_1 = 0.9, β_2 = 0.999) [4] with categorical cross-entropy loss. In the loss function, classes were weighted inversely proportional to the frequency of each GOSE score in the training set to counter class imbalance.

Class-weighted feedforward neural network with an ordinal output layer (DeepOR)

CPM_{DeepOR} and eCPM_{DeepOR} were implemented using the ‘PyTorch’ (v1.10.0) [3] module in Python. The network architecture of DeepMN included a hyperparametric number of dense hidden layers (either 1, 2, 3, 4, 5, or 6), each containing a hyperparametric number of nodes (either 128, 256, or 512) with a rectified linear unit (ReLU) activation function and a hyperparametric percentage (either 0% or 20%) dropout during training. The output layer of DeepOR was a sigmoid layer of 6 nodes, where each node represented the binomial probability of the outcome being greater than a certain threshold, and each node is constrained to be less than or equal to lower-threshold nodes with a negative ReLU transformation (Fig. 1A). DeepOR was optimised using the Adam algorithm (γ [learning rate] = 0.001, β_1 = 0.9, β_2 = 0.999) [4].
0.999) with binary cross-entropy loss. In the loss function, classes were weighted inversely proportional to the frequency of each GOSE score in the training set to counter class imbalance.

| CPM or eCPM | Description | Hyperparameters | Total number of configurations |
|-------------|-------------|-----------------|-------------------------------|
| MNLR        | Multinomial logistic regression | 1                |                               |
| POLR        | Proportional odds (i.e., ordinal) logistic regression | 1                |                               |
| DeepMN      | Class-weighted feedforward neural network with a multinomial (i.e., softmax) output layer | 1, 2, 3, 4, 5, or 6 128, 256, or 512 0% or 20% | 2184 |
| DeepOR      | Class-weighted feedforward neural network with an ordinal (i.e., sigmoid at each threshold) output layer | 1, 2, 3, 4, 5, or 6 128, 256, or 512 0% or 20% | 2184 |

*Different hidden layers may have distinct numbers of neurons.

References

1. Seabold S, Perktold J. Statsmodels: Econometric and Statistical Modeling with Python. In: van der Walt S, Millman J, editors. Proceedings of the 9th Python in Science Conference (SciPy 2010). Austin: SciPy; 2010. pp. 92-96. doi: 10.25080/Majora-92bf1922-011
2. Fletcher R. Practical Methods of Optimization. 2nd ed. New York: John Wiley & Sons; 1987.
3. Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, et al. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In: Wallach H, Larochelle H, Beygelzimer A, d’Alché-Buc F, Fox E, Garnett R, editors. Advances in Neural Information Processing Systems 32 (NeurIPS 2019). Vancouver: NeurIPS; 2019.
4. Kingma DP, Ba J. Adam: A Method for Stochastic Optimization. arXiv:1412.6980v9 [Preprint]. 2017 [cited 2021 December 26]. Available from: [https://arxiv.org/abs/1412.6980](https://arxiv.org/abs/1412.6980)
Supplementary Methods 2. Explanation of APM for ordinal GOSE prediction

APM<sub>MN</sub> and APM<sub>OR</sub> were implemented using the ‘PyTorch’ (v1.10.0) [1] module in Python. Regarding hyperparameters, the embedding and weight-averaging layer (Fig. 2B) is considered to be the first hidden layer. Thus, the number of neurons for the first hidden layer can also be considered as the embedding dimension (i.e., the length of each of the embedding vectors trained on the token dictionary). The individual vector returned by the embedding and weight-averaging layer (Fig. 2B) then undergoes a hyperparametric number of dense hidden layers (either 0, 1, 2, 3, 4, or 5), each containing a hyperparametric number of nodes (either 128, 256, or 512) with a rectified linear unit (ReLU) activation function and a hyperparametric percentage (either 0% or 20%) dropout during training. The output layer of APM<sub>MN</sub> was a softmax layer of 7 nodes, from which probabilities at each GOSE are calculated with cumulative sums (Fig. 1A). APM<sub>MN</sub> was optimised using the Adam algorithm (γ [learning rate] = 0.001, β₁ = 0.9, β₂ = 0.999) [2] with categorical cross-entropy loss. In the loss function, classes were weighted inversely proportional to the frequency of each GOSE score in the training set to counter class imbalance. The output layer of APM<sub>OR</sub> was a sigmoid layer of 6 nodes, where each node represented the binomial probability of the outcome being greater than a certain threshold, and each node is constrained to be less than or equal to lower-threshold nodes with a negative ReLU transformation (Fig. 1A). APM<sub>OR</sub> was optimised using the Adam algorithm (γ [learning rate] = 0.001, β₁ = 0.9, β₂ = 0.999) with binary cross-entropy loss. In the loss function, classes were weighted inversely proportional to the frequency of each GOSE score in the training set to counter class imbalance.

| APM   | Description                                                                 | Hyperparameters                                    | Total number of configurations |
|-------|------------------------------------------------------------------------------|-----------------------------------------------------|-------------------------------|
| APM<sub>MN</sub> | Class-weighted embedding and weight-averaging layer followed by a feedforward neural network with a multinomial (i.e., softmax) output layer | Hidden layers: 1, 2, 3, 4, 5, or 6; Neurons per layer: 128, 256, or 512; Dropout: 0% or 20% | 2184                          |
| APM<sub>OR</sub> | Class-weighted embedding and weight-averaging layer followed by a feedforward neural network with an ordinal (i.e., sigmoid at each threshold) output layer | Hidden layers: 1, 2, 3, 4, 5, or 6; Neurons per layer: 128, 256, or 512; Dropout: 0% or 20% | 2184                          |

*The first hidden layer corresponds to the embedding and weight-averaging layer. Different hidden layers may have distinct numbers of neurons.

References

1. Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, et al. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In: Wallach H, Larochelle H, Beygelzimer A, d’Alché-Buc F, Fox E, Garnett R, editors. Advances in Neural Information Processing Systems 32 (NeurIPS 2019). Vancouver: NeurIPS; 2019.
2. Kingma DP, Ba J. Adam: A Method for Stochastic Optimization. arXiv:1412.6980v9 [Preprint]. 2017 [cited 2021 December 26]. Available from: [https://arxiv.org/abs/1412.6980](https://arxiv.org/abs/1412.6980)
Supplementary Methods 3. Detailed explanation of ordinal model performance and calibration metrics

In this appendix, we will describe each of our selected testing set discrimination, classification, and calibration metrics in mathematical and interpretive detail. Much of this information has already been published by Van Calster et al [1] and Austin et al [2], but we summarise and adapt it here for the ease of the reader. For each of the metrics, we derive the no information value (NIV), which corresponds to the metric value a model would theoretically achieve in the absence of predictive information, and the ideal, full information value (FIV).

Discrimination performance metrics

First, as a reference, let us define the dichotomous c-index, also known as the area under the receiver operating characteristic curve (AUC). Let us first assume a dichotomous prediction problem, in which there are $N_1$ patients with outcome 1 and $N_2$ patients with outcome 2. For a patient of outcome 1, let us denote the predicted probability of outcome 1 as $p_{1,n_1}$, where $n_1 \in [1, N_1]$. Likewise, for a patient of outcome 2, let us denote the predicted probability of outcome 1 as $p_{1,n_2}$, where $n_2 \in [1, N_2]$. The dichotomous c-index is then defined as:

$$c = \frac{1}{N_1N_2} \sum_{n_1=1}^{N_1} \sum_{n_2=1}^{N_2} I_{p_{1,n_1}>p_{1,n_2}}$$

where $I_{p_{1,n_1}>p_{1,n_2}}$ is an indicator variable defined by:

$$I_{p_{1,n_1}>p_{1,n_2}} = \begin{cases} 
1 & \text{if } p_{1,n_1} > p_{1,n_2}; \\
0.5 & \text{if } p_{1,n_1} = p_{1,n_2}; \\
0 & \text{otherwise}.
\end{cases}$$

Thus, the dichotomous c-index can be interpreted as the probability that a model correctly separates 2 patients of different outcome. The dichotomous c-index is the most widely used discrimination metric for binary outcome prediction; however, there is no trivial extension for ordinal outcome prediction [3]. In this appendix, we explore the extensions used for our study.

Ordinal c-index (ORC)

The ordinal c-index (ORC), developed by Van Calster et al [1], is the primary metric of model discrimination performance in our study. Consider a set of 7 randomly chosen patients, each of one of the GOSE scores in our study, such that each patient is represented by $n_o$ where $o \in \{1,2,3,4,5,6,7,8\}$. Now suppose an ordinal GOSE prediction model, such as one of those presented in Fig 1A, receives this set of patients and is tasked with ranking the patients in order of predicted functional outcome. Let $Pr^{(n_o)}(GOSE > t)$ represent the predicted probability, returned by our model, at threshold $t \in \{1,3,4,5,6,7\}$ for patient $n_o \in \{n_1, n_2, n_3, n_4, n_5, n_6, n_7, n_8\}$ in our set. One way the model could achieve this ranking is to start with the lowest threshold ($GOSE > 1$), select the patient with the lowest probability at this threshold (i.e., $\arg\min_{n_o} Pr^{(n_o)}(GOSE > 1)$), set that patient aside as the lowest ranked patient, move on to the subsequent threshold ($GOSE > 3$), repeat this process for the remaining
patients, and repeat at subsequent thresholds until a single patient remains for the highest rank. The ideal predicted ranking would be \( n_1 < n_2 \text{ or } 3 < n_4 < n_5 < n_6 < n_7 < n_8 \). The primary rationale behind ORC is to calculate the average proportional “closeness” between the model-predicted ranking and this ideal ranking. To achieve a mathematical definition for closeness, the developers of ORC considered a scenario: suppose the model-predicted ranking of the given set is: \( n_1 < n_4 < n_5 < n_2 \text{ or } 3 < n_6 < n_8 < n_7 \). From this predicted ranking, we would require at least 3 pairwise switching steps to achieve the target rank. For example:

- **Step 1**: switch \( n_4 \) and \( n_2 \text{ or } 3 \). Result: \( n_1 < n_2 \text{ or } 3 < n_5 < n_4 < n_6 < n_8 < n_7 \)
- **Step 2**: switch \( n_5 \) and \( n_4 \). Result: \( n_1 < n_2 \text{ or } 3 < n_4 < n_5 < n_6 < n_8 < n_7 \)
- **Step 3**: switch \( n_8 \) and \( n_7 \). Result: \( n_1 < n_2 \text{ or } 3 < n_4 < n_5 < n_6 < n_7 < n_8 \)

Let us define \( S \) as the number of necessary pairwise switching steps (i.e., the number of incorrect pairwise orderings) to reach the ideal ranking. Trivially, the ideal \( S(S_{\text{min}}) \) is 0. In the worst possible scenario, in which the predicted ranking is a complete reversal of the ideal ranking (i.e., \( n_8 < n_7 < n_6 < n_5 < n_4 < n_2 \text{ or } 3 < n_1 \)), one would require the maximum number of unique pairwise switching steps possible to achieve the ideal ranking. Since we have 7 possible outcome categories, this is equivalent to \( S_{\text{max}} = \binom{7}{2} = 21 \). In the case of a tie, we add 0.5 to \( S \). The definition of the proportion of closeness, denoted as \( C \), between the model-predicted ranking and the ideal ranking for a given set is thus:

\[
C = 1 - \frac{S}{S_{\text{max}}} = 1 - \frac{S}{21}
\]

In the example provided above, where \( S = 3 \), the proportional closeness between the predicted ranking and the ideal ranking is \( C = 1 - \frac{3}{21} \approx 0.86 \). Thus, to define ORC as the average proportional closeness in ranking over all possible sets,

\[
\text{ORC} = \frac{1}{N_1N_2\text{ or }3N_4N_5N_6N_7N_8} \sum_{n_1=1}^{N_1} \sum_{n_2\text{ or }3=1}^{N_2\text{ or }3} \sum_{n_4=1}^{N_4} \sum_{n_5=1}^{N_5} \sum_{n_6=1}^{N_6} \sum_{n_7=1}^{N_7} \sum_{n_8=1}^{N_8} C_{n_1,n_2\text{ or }3,n_4,n_5,n_6,n_7,n_8}
\]

where \( N_\alpha \forall \alpha \in \{1,2 \text{ or } 3,4,5,6,7,8\} \) denotes the number of patients of GOSE score \( \alpha \), and \( C_{n_1,n_2\text{ or }3,n_4,n_5,n_6,n_7,n_8} \) denotes the proportional closeness of the model ranking to the ideal ranking for patient set \( \{n_1,n_2\text{ or }3,n_4,n_5,n_6,n_7,n_8\} \). Furthermore, if we simplify this formula:

\[
\text{ORC} = \frac{1}{N_1N_2\text{ or }3N_4N_5N_6N_7N_8} \sum_{n_1=1}^{N_1} \sum_{n_2\text{ or }3=1}^{N_2\text{ or }3} \sum_{n_4=1}^{N_4} \sum_{n_5=1}^{N_5} \sum_{n_6=1}^{N_6} \sum_{n_7=1}^{N_7} \sum_{n_8=1}^{N_8} \left[ 1 - \frac{S_{n_1,n_2\text{ or }3,n_4,n_5,n_6,n_7,n_8}}{S_{\text{max}}} \right] 
\]

\[
= \frac{1}{N_1N_2\text{ or }3N_4N_5N_6N_7N_8} \sum_{n_1=1}^{N_1} \sum_{n_2\text{ or }3=1}^{N_2\text{ or }3} \sum_{n_4=1}^{N_4} \sum_{n_5=1}^{N_5} \sum_{n_6=1}^{N_6} \sum_{n_7=1}^{N_7} \sum_{n_8=1}^{N_8} \left[ S_{\text{max}} - \frac{S_{n_1,n_2\text{ or }3,n_4,n_5,n_6,n_7,n_8}}{S_{\text{max}}} \right] 
\]

\[
= \frac{1}{N_1N_2\text{ or }3N_4N_5N_6N_7N_8} \sum_{n_1=1}^{N_1} \sum_{n_2\text{ or }3=1}^{N_2\text{ or }3} \sum_{n_4=1}^{N_4} \sum_{n_5=1}^{N_5} \sum_{n_6=1}^{N_6} \sum_{n_7=1}^{N_7} \sum_{n_8=1}^{N_8} \left[ S_{\text{max}} - \frac{S_{n_1,n_2\text{ or }3,n_4,n_5,n_6,n_7,n_8}}{S_{\text{max}}} \right] 
\]

60
Upon simplification, Generalised index is defined as:

$$\text{ORC} = \frac{1}{N_1 N_2 \cdots N_4 N_5 N_6 N_7 N_8} \sum_{n_1=1}^{N_1} \sum_{n_2=1}^{N_2} \sum_{n_3=1}^{N_3} \sum_{n_4=1}^{N_4} \sum_{n_5=1}^{N_5} \sum_{n_6=1}^{N_6} \sum_{n_7=1}^{N_7} \sum_{n_8=1}^{N_8} \left[ \frac{1}{2} \sum_{i=1}^{N_1} \sum_{j=i+1}^{N_2} I_{n_i > n_j} \right]$$

which is equivalent to the unweighted average of all pairwise $c$-indices. Therefore, another interpretation of ORC is the probability of a model correctly separating 2 randomly selected patients of 2 randomly selected GOSE scores. Moreover, since the NIV of the $c$-index is 0.5 for random guessing and the FIV is 1, we know that ORC shares the same feasible range of values: $NIV_{\text{ORC}} = 0.5$ and $FIV_{\text{ORC}} = 1$. Finally, if there were only 2 possible ordinal outcome categories, we observe that ORC collapses into the dichotomous $c$-index.

The ORC is independent of the prevalence of each GOSE score in the dataset, as each possible set of patients is equally weighted regardless of frequency.

Somers’ $D_{xy}$

The generalised $c$-index, described by Harrell et al [4,5], is defined as the proportion of possible pairs of patients of different functional outcomes in the entire study population which the model correctly discriminates. A pair of patients of different outcomes is defined as a comparable pair and a pair of patients of different outcomes that is correctly discriminated is defined as a concordant pair. Let $N^{\text{comp}}$ denote the total number of comparable pairs in the study set and let $N^{\text{conc}}$ denote the total number of concordant pairs in the study set. Thus, the generalised $c$-index is defined as:

$$\text{Generalised } c\text{-index} = \frac{N^{\text{conc}}}{N^{\text{comp}}}$$

Upon simplification,
we find that the generalised c-index is equivalent to a prevalence-weighted average of pairwise c-indices. Therefore, the generalised c-index shares the same feasible range of values as the dichotomous c-index: \( NIV_{\text{Generalised c-index}} = 0.5 \) and \( FIV_{\text{Generalised c-index}} = 1 \). However, in contrast to ORC, generalised c-index is dependent on the prevalence of GOSE scores in the patient set.

Somers’ \( D_{xy} \) [6,7] is defined as the proportion of the difference between the number of concordant pairs and the number of discordant pairs to the total number of comparable pairs:

\[
\text{Somers’ } D_{xy} = \frac{N_{\text{conc}} - N_{\text{discord}}}{N_{\text{comp}}}
\]

Upon simplification,

\[
= \frac{N_{\text{conc}} - (N_{\text{comp}} - N_{\text{conc}})}{N_{\text{comp}}}
= \frac{2N_{\text{conc}} - N_{\text{comp}}}{N_{\text{comp}}}
= 2 \frac{N_{\text{conc}}}{N_{\text{comp}}} - 1
= 2(\text{Generalised c – index}) - 1
\]

we observe the relationship between Somers’ \( D_{xy} \) and the generalised c-index. Therefore, the feasible range of Somers’ \( D_{xy} \) is: \( NIV_{\text{Somers’ } D_{xy}} = 2(0.5) - 1 = 0 \) and \( FIV_{\text{Somers’ } D_{xy}} = 2(1) - 1 = 1 \). Moreover, Somers’ \( D_{xy} \) is also dependent on the prevalence of GOSE scores in the patient set. Somers’ \( D_{xy} \) can also be interpreted as the proportion of ordinal variation in the outcome that can be explained by the variation in model output.

**Threshold-level dichotomous c-index**

The threshold-level dichotomous c-indices represent the probability of the model correctly discriminating 2 randomly selected patients, one on each side of the threshold of functional recovery. The average of the threshold-level c-indices across the 6 possible GOSE thresholds represents the probability of the model correctly discriminating 2 patients, one on each side of a randomly selected GOSE threshold. The average threshold-level dichotomous c-index is also a prevalence-weighted form of the pairwise c-index, though weighting is not perfectly aligned with prevalence as with the generalised c-index [1]. The feasible range of dichotomous c-indices are: \( NIV_{\text{Dichotomous c-index}} = 0.5 \) to \( FIV_{\text{Dichotomous c-index}} = 1 \).
Probability calibration metrics

Threshold-level calibration slope

Let $Y \in \{0,1\}$ designate the true outcome at a threshold of GOSE and let $p_{pred} \in [0,1]$ designate the predicted probability value returned by a model at this threshold. The logistic recalibration framework \[8\] fits the following model from the testing set predictions: 

$$\text{logit}(Y) = \beta_0 + \beta_1 \text{logit}(p_{pred}).$$

$\beta_1$ represents the calibration slope \[9\]. When $\beta_0 = 0$ and $\beta_1 = 1$, the model is calibrated. When $\beta_1 < 1$, the model is overfitted and returns too extreme values: higher $p_{pred}$ are overestimated while lower $p_{pred}$ are underestimated. When $\beta_1 > 1$, the model is underfitted and the converse is true. We do not focus on $\beta_0$ in our study because, in the setting of cross-validation, $\beta_0$ is not relevant \[10\].

Threshold-level Integrated calibration index (ICI)

On the threshold-level probabilities and threshold-level outcomes of the testing set predictions, we fit a locally weighted scatterplot smoothing (LOWESS) function \[11\] to return the observed probability at each predicted probability value \[12\]. The range of corresponding observed probability for each predicted probability is visualised in a smoothed probability calibration plot (Fig 3B). Let $p_{pred} \in [0,1]$ denote a predicted probability value and $p_{obs}(p_{pred}) \in [0,1]$ denote the corresponding observed probability value. Then, the calibration error function, denoted as $E_{calibration}$, is defined as:

$$E_{calibration}(p_{pred}) = |p_{obs}(p_{pred}) - p_{pred}|.$$ 

The integrated calibration index (ICI) corresponds to the mean calibration error \[2\]. Since the ideal calibration error is 0, the FIVICI is trivially 0. However, the calculation of the NIV varies based on the outcome distribution at each threshold.

Consider the case of random guessing during prediction at a given threshold. This implies that the model returns predicted probabilities uniformly from 0 to 1, regardless of any patient information (S3A.1 Fig). Therefore, the corresponding observed probability at each predicted probability value equals $p_{above}$, the proportion of patients above the given threshold (S3A.1 Fig). In other words, there is no association between predicted and observed probabilities, and the model is completely uncalibrated.
S3A.1 Fig. Example of a probability calibration curve for a random-guessing prediction model at a given threshold of GOSE. The histogram (200 uniform bins), centred at the horizontal line in the bottom quarter, displays the uniform distribution of predicted probabilities for a random guessing model. This plot assumes that the proportion of patients above the threshold ($\pi_{above}$) is 0.8.

From the probability calibration curve (S3A.1 Fig), we derive a graphical representation of the probability density function of $E_{calibration}$ in S3A.2 Fig. This corresponds to an asymmetrical (if $\pi_{above} \neq 0.5$) distribution with density 2 up to $E_{calibration} = \min\{\pi_{above}, 1 - \pi_{above}\}$ and then density 1 from $E_{calibration} = \min\{\pi_{above}, 1 - \pi_{above}\}$ to $E_{calibration} = \max\{\pi_{above}, 1 - \pi_{above}\}$ (S3A.2 Fig).

S3A.2 Fig. Example of probability density of calibration error for a random-guessing prediction model at a given threshold of GOSE. This plot assumes that the proportion of patients above the threshold ($\pi_{above}$) is 0.8.
ICI is equivalent to the integral of the calibration error function over all returned probability prediction values:

\[
ICI = \frac{1}{\max\{p_{\text{pred}}\} - \min\{p_{\text{pred}}\}} \int_{\min\{p_{\text{pred}}\}}^{\max\{p_{\text{pred}}\}} f_{p_{\text{pred}}}(p_{\text{pred}}) E_{\text{calibration}}(p_{\text{pred}}) \, dp_{\text{pred}}
\]

where \( f_{p_{\text{pred}}}(p_{\text{pred}}) \) represents the probability density function over \( p_{\text{pred}} \) values. For the random-guessing model, we determined that \( p_{\text{obs}} \) is constant, i.e., \( p_{\text{obs}}(p_{\text{pred}}) = \pi_{\text{above}} \, \forall \, p_{\text{pred}} \in [0,1] \) at each threshold. Moreover, \( p_{\text{pred}} \) is distributed uniformly from 0 to 1. Therefore:

\[
NIV_{ICI} = \int_0^1 E_{\text{calibration}}(p_{\text{pred}}) \, dp_{\text{pred}}
\]

\[
= \int_0^1 |\pi_{\text{above}} - p_{\text{pred}}| \, dp_{\text{pred}}
\]

\[
= \int_0^{\pi_{\text{above}}} (\pi_{\text{above}} - p_{\text{pred}}) \, dp_{\text{pred}} + \int_{\pi_{\text{above}}}^1 (p_{\text{pred}} - \pi_{\text{above}}) \, dp_{\text{pred}}
\]

\[
= \frac{1}{2} \pi_{\text{above}}^2 + \frac{1}{2} (1 - \pi_{\text{above}})^2
\]

\[
= \pi_{\text{above}}^2 - \pi_{\text{above}} + \frac{1}{2}
\]

A graphical representation of cumulative distribution up to the NIV\(_{ICI}\) for our example is provided in S3A.3 Fig.
S3A.3 Fig. Example of cumulative probability density up to ICI for a random-guessing prediction model at a given threshold of GOSE. This plot assumes that the proportion of patients above the threshold ($\pi_{above}$) is 0.8. The ICI equals 0.34 in calibration error.

References

1. Van Calster B, Van Belle V, Vergouwe Y, Steyerberg EW. Discrimination ability of prediction models for ordinal outcomes: Relationships between existing measures and a new measure. Biom J. 2012;54: 674-685. doi: 10.1002/bimj.201200026.
2. Austin PC, Steyerberg EW. The Integrated Calibration Index (ICI) and related metrics for quantifying the calibration of logistic regression models. Statistics in Medicine. 2019;38: 4051-4065. doi: 10.1002/sim.8281.
3. Hilden J. The Area under the ROC Curve and Its Competitors. Med Decis Making. 1991;11: 95-101. doi: 10.1177/0272989X9101100204.
4. Harrell FE, Califf RM, Pryor DB, Lee KL, Rosati RA. Evaluating the Yield of Medical Tests. JAMA. 1982;247: 2543-2546. doi: 10.1001/jama.1982.03320430047030.
5. Harrell FE. Regression Modeling Strategies. 2nd ed. Cham: Springer; 2015. doi: 10.1007/978-3-319-19425-7.
6. Somers RH. A New Asymmetric Measure of Association for Ordinal Variables. Am Sociol Rev. 1962;27: 799-811. doi: 10.2307/2090408.
7. Kim J. Predictive Measures of Ordinal Association. Am J Sociol. 1971;76: 891-907. doi: 10.1086/225004.
8. Cox DR. Two further applications of a model for binary regression. Biometrika. 1958;45: 562-565. doi: 10.1093/biomet/45.3-4.562.
9. Miller ME, Langefeld CD, Tierney WM, Hui SL, McDonald CJ. Validation of Probabilistic Predictions. Med Decis Making. 1993;13: 49-57. doi: 10.1177/0272989X9301300107.
10. Van Calster B, Nieboer D, Vergouwe Y, De Cock B, Pencina MJ, Steyerberg EW. A calibration hierarchy for risk models was defined: from utopia to empirical data. J Clin Epidemiol. 2016;74: 167-176. doi: 10.1016/j.jclinepi.2015.12.005.
11. Cleveland WS. Robust Locally Weighted Regression and Smoothing Scatterplots. J Am Stat Assoc. 1979;74: 829-836. doi: 10.1080/01621459.1979.10481038.
12. Austin PC, Steyerberg EW. Graphical assessment of internal and external calibration of logistic regression models by using loess smoothers. Stat Med. 2014;33: 517-535. doi: 10.1002/sim.5941.
Supplementary results

Supplementary Results 1. Hyperparameter optimisation results

Training for each of the parametric models (CPM\textsubscript{DeepMN}, CPM\textsubscript{DeepOR}, APM\textsubscript{MN}, APM\textsubscript{OR}, eCPM\textsubscript{DeepMN}, and eCPM\textsubscript{DeepOR}) was made more efficient by dropping out consistently underperforming parametric configurations, on the validation sets, with the Bootstrap Bias Corrected with Dropping Cross-Validation (BBCD-CV) method [1]. During configuration dropout, the optimal configuration for each model was determined over all existing validation set predictions up to that point, and 1,000 resamples of unique patients were drawn to form bootstrapping resamples for the testing of suboptimal configurations versus the optimal configuration in terms of ordinal c-index (ORC) [2]. If a given suboptimal configuration was unable to match or outperform the optimal configuration in at least 5% of the resamples, it was dropped out from training in future repeated k-fold cross-validation partitions.

Each of the models began repeated k-fold cross-validation training with 2,184 parametric configurations (as detailed in Supplementary Methods 1 and Supplementary Methods 2). Under the repeated k-fold cross validation scheme of our study, models were trained in the order of repeats (from 1 to 20), and, within each repeat, in the order of folds (from 1 to 5). After training all viable configurations up to a certain partition, BBCD-CV was performed. The decision of which partitions was dependent on the number of remaining viable configurations and the availability of relevant cores (e.g., APM training required GPUs) on the high-performance computer (HPC), and thus varied by model. Since models of the same predictor set were trained together (i.e., CPM\textsubscript{DeepMN} and CPM\textsubscript{DeepOR}), BBCD-CV was performed for each of the models of a certain predictor set at after the same partition and a different optimal configuration was determined for each model.

In this appendix, we demonstrate the results of BBCD-CV hyperparameter optimisation by model type. First, we list the partitions after which BBCD-CV was performed, demonstrate the number of configurations dropped at these points, and characterise the variable hyperparameter distribution of the remaining viable configurations.

Concise-predictor-based models (CPMs)

BBCD-CV was performed thrice for CPM\textsubscript{DeepMN} and CPM\textsubscript{DeepOR}, after the end of: (1) repeat 1, (2) repeat 3, and (3) repeat 15. The number of remaining viable configurations after these dropouts is visualised, on a binary logarithmic scale, in S4A.1 Fig. The distribution of hyperparameters in the viable configurations, after each dropout, are listed in S4A.1 Table and S4A.2 Table for CPM\textsubscript{DeepMN} and CPM\textsubscript{DeepOR}, respectively.
S4A.1 Fig. Number of trained viable configurations for each CPM during repeated \(k\)-fold cross-validation.

S4A.1 Table. Variable hyperparameter distributions after each dropout for CPMDeepMN.

| Hyperparameter         | Value | Starting configurations \((n = 2184)\) | Remaining configurations after |
|------------------------|-------|--------------------------------------|-------------------------------|
|                        |       | Repeat 1 \((n = 521)\) | Repeat 3 \((n = 102)\) | Repeat 15 \((n = 52)\) |
| Training dropout per layer | 0   | 1092 (50.0%) | 221 (42.4%) | 19 (18.6%) | 8 (15.4%) |
|                        | 0.2  | 1092 (50.0%) | 300 (57.6%) | 83 (81.4%) | 44 (84.6%) |
| Number of layers       | 1    | 6 (0.3%) | 0 (0%) | 0 (0%) | 0 (0%) |
|                        | 2    | 18 (0.8%) | 3 (0.6%) | 2 (2.0%) | 1 (1.9%) |
|                        | 3    | 54 (2.5%) | 10 (1.9%) | 4 (3.9%) | 4 (7.7%) |
|                        | 4    | 162 (7.4%) | 32 (6.1%) | 12 (11.8%) | 8 (15.4%) |
|                        | 5    | 486 (22.3%) | 143 (27.4%) | 57 (55.9%) | 38 (73.1%) |
|                        | 6    | 1458 (66.8%) | 333 (63.9%) | 27 (26.5%) | 1 (1.9%) |

| Median number of neurons per layer | 128 | 284 (13.0%) | 90 (17.3%) | 32 (31.4%) | 18 (34.6%) |
|                                  | 192 | 320 (14.7%) | 67 (12.9%) | 8 (7.8%) | 3 (5.8%) |
|                                  | 256 | 920 (42.1%) | 230 (44.1%) | 44 (43.1%) | 25 (48.1%) |
|                                  | 320 | 56 (2.6%) | 9 (1.7%) | 2 (2.0%) | 0 (0%) |
|                                  | 384 | 320 (14.7%) | 58 (11.1%) | 5 (4.9%) | 2 (3.8%) |
|                                  | 512 | 284 (13.0%) | 67 (12.9%) | 11 (10.8%) | 4 (7.7%) |

S4A.2 Table. Variable hyperparameter distributions after each dropout for CPMDeepOR.

| Hyperparameter         | Value | Starting configurations \((n = 2184)\) | Remaining configurations after |
|------------------------|-------|--------------------------------------|-------------------------------|
|                        |       | Repeat 1 \((n = 146)\) | Repeat 3 \((n = 55)\) | Repeat 15 \((n = 22)\) |
| Training dropout per layer | 0   | 1092 (50.0%) | 42 (28.8%) | 13 (23.6%) | 5 (22.7%) |
|                        | 0.2  | 1092 (50.0%) | 104 (71.2%) | 42 (76.4%) | 17 (77.3%) |
| Number of layers       | 1    | 6 (0.3%) | 0 (0%) | 0 (0%) | 0 (0%) |
All-predictor-based models (APMs)

BBCD-CV was performed twice for APM\textsubscript{MN} and APM\textsubscript{OR}, after the end of: (1) the first fold of repeat 1, and (2) repeat 10. The number of remaining viable configurations after these dropouts is visualised, on a binary logarithmic scale, in S4A.2 Fig. The distribution of hyperparameters in the viable configurations, after each dropout, are listed in S4A.3 Table and S4A.4 Table for APM\textsubscript{MN} and APM\textsubscript{OR}, respectively.

S4A.2 Fig. Number of trained viable configurations for each APM during repeated $k$-fold cross-validation.

S4A.3 Table. Variable hyperparameter distributions after each dropout for APMs.

| Hyperparameter                      | Value | Starting configurations ($n = 2184$) | Remaining configurations after Repeat 1, Fold 1 ($n = 41$) | Remaining configurations after Repeat 10 ($n = 2$) |
|-------------------------------------|-------|-------------------------------------|----------------------------------------------------------|--------------------------------------------------|
| Training dropout per layer          | 0     | 1092 (50.0%)                        | 18 (43.9%)                                               | I (50.0%)                                       |
|                                     | 0.2   | 1092 (50.0%)                        | 23 (56.1%)                                               | I (50.0%)                                       |
S4A.4 Table. Variable hyperparameter distributions after each dropout for APM\textsubscript{OR}.

| Hyperparameter | Value | Starting configurations \((n = 2184)\) | Remaining configurations after Repeat 1, Fold 1 \((n = 161)\) | Remaining configurations after Repeat 10 \((n = 8)\) |
|----------------|-------|----------------------------------------|------------------------------------------------|---------------------------------------------|
| Training dropout per layer | 0.2 | 1092 (50.0%) | 139 (86.3%) | 8 (100.0%) |
| Number of layers | 1 | 6 (0.3%) | 1 (0.6%) | 0 (0%) |
| | 2 | 18 (0.8%) | 1 (0.6%) | 0 (0%) |
| | 3 | 54 (2.5%) | 5 (3.1%) | 0 (0%) |
| | 4 | 162 (7.4%) | 13 (8.1%) | 1 (12.5%) |
| | 5 | 486 (22.3%) | 36 (22.4%) | 2 (25.0%) |
| | 6 | 1458 (66.8%) | 105 (65.2%) | 5 (62.5%) |
| Median number of neurons per layer | 128 | 284 (13.0%) | 31 (19.3%) | 2 (25.0%) |
| | 192 | 320 (14.7%) | 29 (18.0%) | 4 (50.0%) |
| | 256 | 920 (42.1%) | 73 (45.3%) | 1 (12.5%) |
| | 320 | 56 (2.6%) | 6 (3.7%) | 0 (0%) |
| | 384 | 320 (14.7%) | 11 (6.8%) | 0 (0%) |
| | 512 | 284 (13.0%) | 11 (6.8%) | 1 (12.5%) |

Extended concise-predictor-based models (eCPMs)

BBCD-CV was performed twice for eCPM\textsubscript{DeepMN} and eCPM\textsubscript{DeepOR}, after the end of: (1) repeat 1, and (2) repeat 16. The number of remaining viable configurations after these dropouts is visualised, on a binary logarithmic scale, in S4A.3 Fig. The distribution of hyperparameters in the viable configurations, after each dropout, are listed in S4A.5 Table and S4A.6 Table for eCPM\textsubscript{DeepMN} and eCPM\textsubscript{DeepOR}, respectively.
S4A.3 Fig. Number of trained viable configurations for each eCPM during repeated $k$-fold cross-validation.

### S4A.5 Table. Variable hyperparameter distributions after each dropout for eCPM$_{DeepMN}$.

| Hyperparameter      | Value | Starting configurations ($n = 2184$) | Remaining configurations after Repeat 1 ($n = 121$) | Remaining configurations after Repeat 16 ($n = 10$) |
|---------------------|-------|-------------------------------------|-----------------------------------------------|-----------------------------------------------|
| Training dropout per layer | 0     | 1092 (50.0%)                        | 51 (42.1%)                                    | 4 (40.0%)                                    |
|                      | 0.2   | 1092 (50.0%)                        | 70 (57.9%)                                    | 6 (60.0%)                                    |
| Number of layers     | 1     | 6 (0.3%)                            | 3 (2.5%)                                      | 2 (20.0%)                                    |
|                      | 2     | 18 (0.8%)                           | 8 (6.6%)                                      | 3 (30.0%)                                    |
|                      | 3     | 54 (2.5%)                           | 15 (12.4%)                                    | 3 (30.0%)                                    |
|                      | 4     | 162 (7.4%)                          | 45 (37.2%)                                    | 2 (20.0%)                                    |
|                      | 5     | 486 (22.3%)                         | 48 (39.7%)                                    | 0 (0%)                                       |
|                      | 6     | 1458 (66.8%)                        | 2 (1.7%)                                      | 0 (0%)                                       |
| Median number of neurons per layer | 128   | 284 (13.0%)                         | 21 (17.4%)                                    | 3 (30.0%)                                    |
|                      | 192   | 320 (14.7%)                         | 14 (11.6%)                                    | 2 (20.0%)                                    |
|                      | 256   | 920 (42.1%)                         | 55 (45.5%)                                    | 4 (40.0%)                                    |
|                      | 320   | 56 (2.6%)                           | 5 (4.1%)                                      | 0 (0%)                                       |
|                      | 384   | 320 (14.7%)                         | 11 (9.1%)                                     | 0 (0%)                                       |
|                      | 512   | 284 (13.0%)                         | 15 (12.4%)                                    | 1 (10.0%)                                    |

### S4A.6 Table. Variable hyperparameter distributions after each dropout for eCPM$_{DeepOR}$.

| Hyperparameter      | Value | Starting configurations ($n = 2184$) | Remaining configurations after Repeat 1 ($n = 9$) | Remaining configurations after Repeat 16 ($n = 1$) |
|---------------------|-------|-------------------------------------|-----------------------------------------------|-----------------------------------------------|
| Training dropout per layer | 0     | 1092 (50.0%)                        | 1 (11.1%)                                     | 0 (0%)                                       |
|                      | 0.2   | 1092 (50.0%)                        | 8 (88.9%)                                     | 1 (100.0%)                                   |
| Number of layers     |       |                                     |                                               |                                              |
| Layer | Neurons | Median Percentage |
|-------|---------|-------------------|
| 1     | 6 (0.3%)| 1 (11.1%)         |
| 2     | 18 (0.8%)| 4 (44.4%)         |
| 3     | 54 (2.5%)| 2 (22.2%)         |
| 4     | 162 (7.4%)| 1 (11.1%)        |
| 5     | 486 (22.3%)| 0 (0%)          |
| 6     | 1458 (66.8%)| 1 (11.1%)      |

Median number of neurons per layer

| Neurons | Median Percentage |
|---------|-------------------|
| 128     | 284 (13.0%)       |
| 192     | 320 (14.7%)       |
| 256     | 920 (42.1%)       |
| 320     | 56 (2.6%)         |
| 384     | 320 (14.7%)       |
| 512     | 284 (13.0%)       |

References

1. Tsamardinos I, Greasidou E, Borboudakis G. Bootstrapping the out-of-sample predictions for efficient and accurate cross-validation. Mach Learning. 2018;107:1895-1922. doi: 10.1007/s10994-018-5714-4.

2. Van Calster B, Van Belle V, Vergouwe Y, Steyerberg EW. Discrimination ability of prediction models for ordinal outcomes: Relationships between existing measures and a new measure. Biom J. 2012;54:674-685. doi: 10.1002/bimj.201200026.