Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Medicine 2217

Refining outcome prediction in traumatic brain injury

DAVID BARK

ACTA UNIVERSITATIS
UPSALIENSIS
2025

ISSN 1651-6206
ISBN 978-91-513-2687-0
urn:nbn:se:uu:diva-571980

UPPSALA
UNIVERSITET

Dissertation presented at Uppsala University to be publicly examined in H:son-Holmdahlsalen, Dag Hammarskjölds Väg 8, Uppsala, Friday, 16 January 2026 at 12:15 for the degree of Doctor of Philosophy (Faculty of Medicine). The examination will be conducted in Swedish. Faculty examiner: Associate Professor Rahul Raj (Neurocenter, Division of Neurosurgery, Helsinki University).

Abstract
Bark, D. 2025. Refining outcome prediction in traumatic brain injury. *Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Medicine* 2217. 115 pp. Uppsala: Acta Universitatis Upsaliensis. ISBN 978-91-513-2687-0.

Traumatic brain injury (TBI) remains a leading cause of mortality and a major contributor to global disability-adjusted life years in young and middle-aged adults. While animal studies have shown potential, effective pharmacological treatments have failed to materialize in the clinical setting. A primary challenge in TBI research is the inherent heterogeneity of the condition.

Outcome prediction models can reduce this complexity by better risk stratification for clinical trials, optimizing resource allocation, and improving prognostic communication with patients and families. TBI outcomes are commonly measured by the Glasgow Outcome Scale Extended (GOSE), an 8-point scale ranging from death (1) to upper good recovery (8). However, current prediction models typically dichotomize this scale into favorable/unfavorable outcomes which limits nuanced prognostication.

The primary objective of this thesis was to develop prediction models capable of predicting the full 8-grade GOSE. A secondary objective was to evaluate the integration of machine learning in clinical care, specifically assessing an AI decision support system for detecting intracranial hemorrhage in computed tomography (CT) scans.

We found that advanced machine learning methods performed comparably to standard statistical models when limited to admission variables. However, the incorporation of dynamic physiological data captured prognostic signals that static admission models missed, thereby improving prediction accuracy.

Regarding the AI diagnostic tool, while it successfully identified most hemorrhages, we observed no significant clinical benefit for the patients. This underscores that in clinical settings, the implementation strategy is as critical as the technology itself.

Finally, this thesis emphasizes that predictive models are not static tools. They are sensitive to temporal changes in populations and healthcare protocols. Therefore, future implementation must prioritize the continuous updating and validation of these models.

Keywords: Traumatic brain injury, TBI, Acquired brain injury, Artificial Intelligence, Machine learning, prognosis, prognostic model, Neurosurgery, Neuroradiology.

David Bark, Neurosurgery, Akademiska sjukhuset, Uppsala University, SE-751 85 Uppsala, Sweden.

ISSN 1651-6206
ISBN 978-91-513-2687-0
URN urn:nbn:se:uu:diva-571980 (http://urn.kb.se/resolve?urn=urn:nbn:se:uu:diva-571980)

Dedicated to my family

List of Papers

This thesis is based on the following papers, which are referred to in the text by their Roman numerals.

I. Rostami, E., Gustafsson (Bark), D., Hånell, A., Howells, T., Lenell, S., Anders Lewén, Per Enblad. (2022). Prognosis in moderate-severe traumatic brain injury in a Swedish cohort and external validation of the IMPACT models. *Acta Neurochir (Wien)*: Mar;164(3):615-624.

II. Bark D., Boman M., Depreitere B., Wright D. W., Lewén A., Enblad P., Hånell A.*, & Rostami E.* (2024). Refining outcome prediction after traumatic brain injury with machine learning algorithms. *Scientific Reports* Apr 5;14(1):8036.

III. Bark, D., Basu, J., Toumpanakis, D., Burwick Nyberg, J., Bjerner, T., Rostami, E., and Fällmar, D. (2024). Clinical Impact of an AI Decision Support System for Detection of Intracranial Hemorrhage in CT Scans. *Neurotrauma Reports*: Oct 14;5(1):1009-1015.

IV. Bark, D., Hult, L., Hånell, A., Svedung-Wettervik, T., Lenell, S., Lewén, A., Enblad, P., Schön T., Rostami, E. Predicting ordinal GOSE using dynamic NICU physiological variables in moderate-to-severe TBI - A Temporal and Stratified Split Study. *Manuscript under revision*

*Contributed equally

Reprints were made with permission from the respective publishers.

The following papers, which are referred to in the text by their Roman numerals.

Contents

Abbreviations

ABA	Average balanced accuracy
ASDH	Acute subdural hematoma
AI	Artificial intelligence
AGI	Artificial general intelligence
AUROC	Area under the receiver operator curve
AW1	Accuracy within 1 category
AW2	Accuracy within 2 categories
BA	Balanced accuracy
BTF	Brain trauma foundation
CT	Computer tomography
CRASH	Corticosteroid Randomization after Significant Head injury
DAI	Diffuse axonal injury
DL	Deep learning
ECE	Expected Calibration Error
EDH	Epidural hematoma
FDA	Food and Drug Administration
GCS	Glasgow coma scale
GOS	Glasgow outcome scale
GOSE	Glasgow outcome scale extended
IMPACT	International mission for prognosis and analysis of clinical trials in TBI
ICH	Intracerebral hematoma
MAE	Mean average error
MCPD	Maximum category proportion discrepancy
MD	Mean discrepancy
MSE	Mean squared error
OOB	Out of bag
OLR	Ordinal logistic regression
POLR	Proportional odds logistic regression
PPV	Positive predictive value
ProTECT III	Progesterone for Traumatic Brain Injury: Experimental Clinical Treatment III
RF	Random Forest
ReLU	Rectified linear unit
TANH	Hyperbolic tangent

TCPD	Total category proportion discrepancy
TBI	Traumatic brain injury
tSAH	Traumatic subarachnoid hemorrhage
SAH	Subarachnoid hemorrhage
SDD	Standard deviation discrepancy
SHAP	SHapley Additive exPlanations
UUH	Uppsala university hospital
US	United states
QWK	Quadratic weighted Kappa
XGBoost	Extreme gradient boosting

Introduction

Traumatic brain injury

In 2018, it was estimated that approximately 69 million people sustain a traumatic brain injury (TBI) annually, and TBI is an especially burdensome disease in developing countries[1]. Globally, TBI is a leading cause of death in both young adults and the elderly. Patients with TBI may suffer from life-long disabilities such as neurological deficits, cognitive dysfunction, fatigue, headaches, and personality changes. This imposes a significant burden on patients, their families, and society at large[2]. In the US alone, the overall healthcare economic cost of non-fatal TBI in 2016 was $40.6 billion[3].

TBI is typically categorized by severity, pathoanatomical region, or by trauma mechanism. Trauma severity is usually divided into mild, moderate, and severe depending on the neurological state of the patient, as described by the Glasgow coma scale (GCS) at admission. Specifically, a GCS of 13-15 is classified as mild TBI, GCS 9-12 as moderate TBI, and GCS 3-8 as severe TBI (Table 1)[4].

Table 1. Interpretation of the Glasgow Coma Scale (GCS), which measures level of consciousness in 3 to 15 points.

GCS motor	**GCS verbal**	**GCS eye opening**
6. Obeys command	5. Oriented to person, place, and time	4. Spontaneous
5. Localizes pain	4. Confused	3. Opens to sound
4. Withdraws from pain	3. Inappropriate words	2. Opens to pain
3. Abnormal flexion (decorticate response)	2. Incomprehensible sounds	1. No response
2. Extension (decerebrate response)	1. No response	
1. No response		

Whereas mild TBI is usually associated with a better outcome, some reports suggest that up to 30 to 80 percent of patients with mild TBI suffer from persisting symptoms[5]. Moderate and severe TBI are associated with significantly worse outcomes.

A common way to categorize TBI by pathoanatomical classification is by the ordinal Marshall computer tomography (CT) score, which has been shown to have strong correlation with outcome (Table 2)[6].

Table 2. Summary of the Marshall computer tomography (CT) score which measures the radiological severity of traumatic brain injury.

Marshall CT score	**Interpretation**
1) Diffuse injury I	No visible pathology
2) Diffuse injury II	Present cisterns, midline shift < 5 mm, no lesion > 25 ml
3) Diffuse injury III	Compressed or absent cisterns, midline shift < 5 mm, no lesion >25 ml
4) Diffuse injury IV	Midline shift > 5 mm, no lesion > 25 ml
5) Evacuated mass lesion	Any surgically evacuated lesion
6) Non-evacuated mass lesion	Lesion > 25 ml that is not surgically evacuated

While the current gold standard of TBI classification is based on the GCS and trichotomized into mild, moderate, and severe, there is an increasing consensus that this classification is blunt and insufficient. In 2022, the US National Institutes of Health–National Institute of Neurological Disorders and Stroke launched an international initiative to address an improved characterization of acute TBI. They suggested a new framework based on four pillars: clinical, biomarker, imaging, and modifier, collectively called the CBI-M framework. The clinical pillar would be based on GCS, expanded by pupillary response as well as loss of consciousness, balance function, and cognitive function. The biomarker pillar includes GFAP, UCH-L1, and S100B. The basic imaging pillar includes types of hemorrhage and fracture, as well as the mass effect and lesion volume, and the binary presence of other types of injury such as penetrating injury and vascular injury. The modifier components include injury factors such as mechanism of injury and extracranial injury, presence of secondary insults, and patient-related factors such as age, sex, medical history, frailty, previous mental health/TBI/developmental history, living circumstances, social disadvantage, stressful circumstances, and social barriers. While the CBI-M framework is yet to be fully developed or validated, it highlights the importance of adopting a multifactorial approach to understanding the long-term outcomes after TBI[7].

Intracranial hemorrhage

Intracranial hemorrhage is frequently caused by traumatic brain injury, but it also arises from other etiologies such as tumors and hemorrhagic strokes.

According to the Swedish stroke register, 12% of strokes in Sweden were caused by intracerebral hemorrhages (ICH) in 2021[8]. Compared to ischemic strokes, hemorrhagic strokes are associated with higher morbidity and mortality. In fact, ICH is a serious medical event with a 1-month mortality rate of 40%[9]. Half of the resulting mortality occurs within the first 24 hours, and early treatment has been shown to improve patient outcomes[10]. In patients with intracranial hematoma, hematoma expansion and irreversible neurologic damage can occur as early as the first few hours after onset. Early surgical intervention can be life-saving[11], and making a prompt diagnosis is crucial for the appropriate management of these patients. The critical need for prompt diagnosis across these conditions has spurred the development of automated detection tools, a topic this thesis will also explore.

Outcome prediction

Improved outcome predictions lead to improved clinical trials by increasing statistical power. It also enables the identification of at-risk patients who require additional monitoring and serves as a benchmark for evaluating the quality of care.

Improving clinical trials

Despite major improvements in TBI recovery since the introduction of neurointensive care, progress in improving outcomes has plateaued in recent decades. While several animal studies have shown promising results in treating TBI, more than 30 phase-three clinical trials have failed to find any significant effect regarding their primary endpoint[12]. To mitigate patient heterogeneity, adjusting for baseline characteristics is known to increase statistical power[13]. However, standard analyses often rely on dichotomized outcomes, which discard valuable data regarding the degree of recovery. Transitioning to an 8- or 7-grade ordinal outcome retains this granular information, offering superior sensitivity to detect treatment effects and further maximizing statistical power. Development of such advanced models, with improved risk stratification across the full spectrum of recovery, could further reduce sample size requirements. They could also be used to identify patients with an outcome that is unexpectedly better or worse, which could be of interest for further analysis.

Identification of risk patients

Prediction models are developed on large cohorts, and care should be taken when applying predictions to individual patients. However, accurate prediction models could help identify patients at risk of deterioration who require intensified care. They could also provide statistical prognostic information that may be of interest to the patient and their relatives. A small survey study found that only 37% of doctors believed that they could predict the prognosis after TBI accurately[14]. While clinicians should not solely rely on the predictions of prognostic models in the individual case, accurate prognostic models could serve as a valuable complement to clinical judgement to improve patient outcome.

Evaluating care

Evaluation of intensive care is of utmost importance to ensure that optimal care continues to be provided. Using prediction models, the predicted outcome of patients might be compared to the actual outcome to reflect the quality of the care given and help identify suboptimal treatment[15].

Prediction of outcome after TBI is a relatively young field. The development of the Glasgow coma scale (GCS) in 1974 provided clinicians with a tool to systematically evaluate patient consciousness in order to predict outcomes and guide clinical management[4]. The introduction of the Glasgow outcome scale (GOS) in 1975 enabled a systematic way of primary outcome measurement and remains the second most cited neurosurgical article in history[16]. In order to assess the outcome more precisely, the extended 8-step version called the Glasgow Outcome Scale Extended (GOSE) tool was developed and published in 1998[17]. A summary of GOS and GOSE is found in table 3. GOSE has since become one of the most preferred instruments for assessing outcomes following TBI. This, together with the development of large TBI databanks, provided foundational tools for the modern era of TBI outcome prediction. Since then, outcome prediction after TBI has been of significant interest and is subject to extensive research.

Table 3. Interpretation of the Glasgow outcome scale (GOS) and the Glasgow outcome scale extended (GOSE). GOS is a five-level scale to evaluate outcome after traumatic brain injury and GOSE expands to eight levels reflecting death to practically full recovery.

GOS	**GOSE**	**Domain**	**Interpretation**
1) Dead	1) Dead		Dead
2) Vegetative	2) Vegetative	Consciousness	No interaction with the world
3) Severe disability	3) Lower severe disability	Function in the home	Unable to look after themselves for 8 hours
	4) Upper severe disability	Function in and outside the home	Unable to look after themselves for 24 hours OR shop OR travel
4) Moderate disability	5) Lower moderate disability	Function in work, social activities and relationships	Unable to return to any of the previous activities
	6) Upper moderate disability	Function in work, social activities and relationships	Some disabilities, but able to return to some of previous activities
5) Good recovery	7) Lower good recovery	Function in social activities and relationships	Minor disabilities that still affect daily activities
	8) Upper good recovery	Function in social activities and relationships	No or minor disabilities that do not affect daily activities

The limits of current prognostic models

In a systematic review by Perel et al. in 2005, a total of 53 reports were identified including 102 models[18]. They concluded that while there are many prognostic models, they typically have several limitations such as small sample sizes, no external validation, development mainly in high income countries, and limited clinical utility. In a major step forward in 2008, the Corticosteroid

Randomization after Significant Head injury (CRASH) Trial collaborators developed a logistic regression model on clinical data from the 10 008 patients recruited[19]. The CRASH model showed high discriminatory ability with an area under the receiver operator curve (AUROC) of 0.8. Following this, the International Mission for Prognosis and Analysis of Clinical Trials (IMPACT) investigators utilized the TBI database developed from their clinical studies and created the IMPACT prognostic calculator from 8,509 patients[20].

Both the CRASH and IMPACT models use hospital admission variables to predict death or unfavorable outcome (GOSE ≤ 4), as seen in table 4. The IMPACT prognostic model is made up of three modules: the core model consists of age, GCS motor score, and pupillary reactivity. The second module consists of the core model in addition to the Marshall CT score, whether there is traumatic subarachnoid hemorrhage (tSAH), and the presence of an epidural mass, hypoxia, or hypotension. The third module consists of the first two in addition to the laboratory variables of hemoglobin and glucose concentrations. The IMPACT calculator has been subject to extensive external validation of mortality prediction at 6 months, with AUROC levels reported between 0.6-0.9[21-23]. All report an improved prediction with increased complexity. Furthermore, the variables identified in the IMPACT study have been analyzed for their prognostic power. Age, Glasgow Coma Scale-motor and CT findings were found to have the highest R^2 value and high univariate odds ratios. [24] Combined, these factors have been found to have an R^2 value of around 35%, indicating that a majority of factors contributing to outcome are not included[13]

Table 4. Variables included in the IMPACT and CRASH outcome predictions modules.

IMPACT core	IMPACT - CT	IMPACT - lab	CRASH core	CRASH core + CT
Age	Hypoxia	Glucose	Country	Presence of petechial hemorrhages
GCS-m	Hypotension	Hemoglobin	Age	Obliteration of third ventricle or basal cisterns
Pupillary reaction	CT Marshall classification		GCS	Subarachnoid bleeding
			Pupillary reaction	Midline shift
			Major extracranial injury	Non-evacuated hematoma

A key limitation of these established models is their prediction of a dichotomized outcome, typically that the GOSE categories are split into either mortality (GOSE 1) or unfavorable outcome (GOSE 1-4) vs favorable outcome (GOSE 5-8). While statistically convenient, this dichotomization is a coarse simplification that loses considerable nuance. For a patient and their family, the difference between a vegetative state (GOSE 2) and severe disability with some independence (GOSE 4) is profound, yet both are classified as 'unfavorable'. Furthermore, improved outcome prediction should be attainable using more complex variables, such as intensive care data[25], genetics [26], biomarkers [27], and radiological data[28]. However, exploring this unexplained variance and predicting a granular, 8-level outcome requires a move towards more sophisticated and flexible statistical approaches.

Dynamic physiological data

After trauma, the primary injury triggers a cascade wherein the injured brain is susceptible to secondary insults. Intracranial hemorrhage and brain edema may increase the intracranial pressure, which potentially leads to decreased cerebral blood flow or brain herniation[29]. The cerebral perfusion pressure (CPP) reflects the cerebral blood flow and is calculated by subtracting the intracranial pressure (ICP) from the mean arterial blood pressure (MAP)[30]. In a healthy individual, the cerebral blood flow is maintained across a range of blood pressures through a process called autoregulation where the cerebral blood vessels constrict in response to higher blood pressure, which maintains stable cerebral blood flow. The capacity for autoregulation could be impaired after TBI which has been associated with poor outcome[31].

While the association between intracranial hypertension and poor outcome has been extensively verified, there is growing evidence that the tolerance for ICP insults depends on both cerebral perfusion pressure as well as the patient's autoregulation capacity. Autoregulation can be estimated continuously through the pressure reactivity index (PRx), which is a measure of the correlation coefficient between MAP and ICP over 5 minutes time windows[32]. A negative value indicates intact autoregulation, where increased blood pressure leads to vasoconstriction which in turns decreases ICP. On the other hand, positive values indicate dysfunction of autoregulation where MAP and ICP correlate positively[33].

In our NICU at Uppsala University Hospital, we adhere to a stepwise treatment escalation protocol for intracranial pressure with an ICP target of >20mmHg and CPP >60mmHg. The protocol consists of head elevation, sedation, mild hyperventilation, external ventricular drainage, barbiturates, and finally decompressive craniectomy (in accordance with Brain Trauma

Foundation (BTF) guidelines)[34]. While research into PRx and optimal CPP thresholds depending on autoregulatory status remains compelling, the evidence is considered insufficient for clinical use according to the current BTF guidelines[34]. It is however widely hypothesized that quantifying the burden of these secondary insults, such as intracranial hypertension, cerebral hypoperfusion and autoregulation status using dynamic physiological data, should refine prognostic accuracy[25,35-42].

The choice of which variables to include in a prediction model depends on the purpose of the model. If the main purpose is to stratify the patients before inclusion in a clinical trial, the model might solely rely on static admission variables such as the IMPACT model. However, if the purpose is to create the most accurate prediction model from the acute phase, it would be sensible to include dynamic physiological data from the NICU care as this information captures the evolving pathophysiological processes. In this thesis, we explore models that include only admission data as well as those incorporating dynamic physiological data.

The potential of machine learning

There is a current surge of general interest regarding machine learning and artificial intelligence (AI). Complex machine learning models such as neural networks have recently received increased attention in the medical field, for example in image recognition and segmentation which is particularly useful in radiology[43]. Recurrent neural networks that can interpret text or time-sensitive data have been used to develop prediction models from continuous intensive care data, such as intracranial pressure[44]. Finally, artificially intelligent models have been predicted to improve information collection and productivity in all parts of society. [45]

In essence though, artificial intelligence in its current form is no more than an advanced statistical model that fits a line to data points and draws conclusions from that line. In practice, this means that instead of telling the computer what to produce when given a set of variables, it “learns” patterns from previous data. As perhaps the simplest example, linear regression would be considered a form of machine learning. It is thus important to note that when referring to artificial intelligence, there is a spectrum ranging from a simple linear regression model (also known as narrow AI) to neural networks trained on billions of data points with broad applications that approach artificial general intelligence (AGI).

While linear and logistic regression have been the standard in outcome prediction after TBI, the purpose of this thesis is to explore other machine learning models to predict outcome after TBI in a more precise way. The main models explored were ordinal logistic regression, tree-based models such as Random Forest, and neural network models. Some important basic terminology is important to take into consideration. In machine learning terms, the independent variables are often called "features" and the dependent variables are called "labels". Less advanced machine learning models, such as linear regression, might fit a straight line to the data points, but this might not be enough if the relationship between the features and labels is not linear. The model would thus be "underfitted" or have a high "bias". On the contrary, advanced machine learning models such as neural networks might find a function that draws the line perfectly between training data points which makes it perfect in representing the relationship between the features and the labels in the training data, but the line might not be generalizable and thus the model would perform worse when evaluated in a novel dataset. That model would be "overfitted" or have a high "variance". The perfect model would have low bias and low variance, and an important part of developing machine learning models is to produce a model with a good balance between bias and variance (Figure 1)[46].

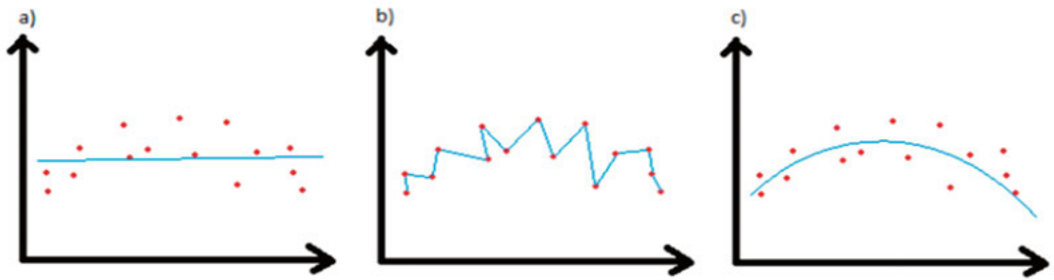


Figure 1. Graphical representation of under- and overfitting. a) A basic model that underfits the training data and has high bias, thus losing information regarding the relationship between features and label. b) An advanced model that fits perfectly to the training data is overfitted and has high variance, thus losing generalizability in new settings. C) A balanced model with low variance and low bias, providing a balance between data relationships and generalizability.

From established methods to advanced algorithms

Linear regression

Linear regression is a basic statistical technique used for predicting a numerical variable from one or more numerical or categorical variables. This is done by plotting a straight line with an intersection and slope that minimizes the squared sum of the vertical error from the data points, or the residuals. Linear regression is only appropriate when the relationship between the features and outcome is linear and there are no complex feature interactions[46].

Logistic regression

Logistic regression is a statistical technique used for predicting a binary outcome from one or more numerical or categorical variables. Logistic regression assumes a linear relationship between the features and the label and uses a logit function to transform the output into probabilities[46]. Logistic regression is a simple and easily interpretable model, which performs well with small datasets and a small number of features. However, logistic regression is less suitable with a large number of features, especially if there are complex interactions between them. Logistic regression has been the standard statistical model for outcome prediction after TBI, and both the CRASH and IMPACT models are multivariable logistic regression models. Logistic regression can also be extended to multiple categories using multinomial or ordinal logistic regression. When predicting the GOSE scale, which is an ordinal categorical scale, it is appropriate to use ordinal logistic regression. Ordinal logistic regression is divided into proportional and non-proportional odds regression depending on whether the odds of predicting lower or higher categories are constant across all categories, the so-called parallel regression assumption. In order to determine whether to apply proportional odds logistic regression (POLR) or non-proportional odds logistic regression, it is appropriate to conduct a test such as the Brant test to verify that the parallel regression assumption holds.

Decision trees and tree-based models

Decision trees are a tree-like statistical model for both classification and regression tasks. Decision trees divide the data into groups at each branch of the tree, representing a decision on a feature which sorts the data into groups with higher proportions of label group purity. For example, different GCS levels might be branches that divide the data into groups of purer GOSE outcomes. The decision tree attempts to maximize the label purity at each feature node until the groups either have 100% purity (containing only one label type), or when the model has exhausted the features[46]. Decision trees have several benefits such as being easily interpretable, and they can handle missing and non-linear relationships. However, they also have several disadvantages. The main disadvantage is that they are prone to overfitting the training set and are thus biased in testing data. Therefore, it is common to set up an ensemble of trees that work together to provide a prediction, such as a through Random Forest (RF)[47] or eXtreme Gradient Boosting (XGBoost) models[48].

Random forest is a decision tree-based ensemble method that combines multiple decision trees to produce more accurate predictions. Random Forest models build the decision trees by bootstrap aggregation (bagging) of the data, which means that each decision tree is set up using a random set of some of the features. Additionally, in the Random Forest model only a random subset of the features is considered for the next split, and the number of feature candidates at each split is referred to as the *mtry* number. In other words, *mtry* is the number of randomly selected variables used when forming each split in a tree. This reduces the correlation between the decision trees and improves model performance. Finally, in classification tasks, the prediction output is simply the class selected by the majority of trees[46]. Random Forests are powerful models that can handle a large number of features that do not have a linear relationship to the label, and that might have complex interactions between features. Specifically, reducing the mtry value will reduce the correlation between any pair of trees in the ensemble, and hence reduce the average variance, but this is only suitable when using a small number of relevant features. When the number of features is large, but the fraction of relevant features is small, Random Forests are likely to perform poorly with small *mtry* since at each split the chance is small that the relevant variables will be selected. Like other complex models, Random Forests might overfit the training data especially if there are a large number of features or if the forest is composed of a relatively high number of trees.

XGBoost models have gained popularity in recent years within medical machine learning research. Instead of training many individual trees on bootstrapped samples like in Random Forest, XGBoost builds an ensemble sequentially where each new tree is trained to correct the errors of previous ones[48]. It starts with a simple model, calculates the gradient of errors and their direction, and fits a new decision tree to those gradients. This tree is added to the ensemble with a learning rate weight, and this is then repeated many times. Instead of bagging, the model *boosts* by sequential training. This improves stability, and while more prone to overfitting than Random Forest, the model also supports regularization and optimization of learning rate to reduce overfitting. One of the reasons for the recent popularity of XGBoost is the model's excellent performance on tabular data, which is common in clinical research.

Neural networks

Neural networks are a type of machine learning algorithm inspired by the structure and function of the human brain. Neural networks have been around since the 1950s, yet only recently has computing power become powerful enough to allow sophisticated, high dimensional neural networks[46]. Neural networks are composed of layers of interconnected nodes (neurons) that learn to recognize patterns in the data. As mentioned before, there are many different types of neural networks, including convolutional networks used for image classification, and different kinds of recurrent neural networks used in text classification[49]. When using categorical and numerical data to predict outcome categories, we have applied a feedforward multilayered perceptron. "Feedforward" because the information travels in one direction, and "multilayered" referring to the multiple layers of neurons[50]. An essential part of the ability of this neural network to detect advanced patterns in data lies in the *activation function*, where each neuron takes the input data and passes it through a function which introduces non-linearity[46]. The most common activation function is the rectified linear unit (ReLU), which turns the input into a value between 0 and the positive input. Other common activation functions include the *sigmoid function* which turns the input into a value between 0 and 1, and is commonly used in binary classification problems, and *hyperbolic tangent* (TANH) which gives outputs between -1 and 1 and is commonly used when negative numbers are included. Each layer then passes it to the next by applying a coefficient, or weight, to that number, which changes it ever so slightly. The value of these weights is the way the network then "learns" from the data. The final layer, called the output layer, provides the results from the algorithm by passing the input from the previous layers through another type of activation function which depends on the purpose of the model. For our purpose, the commonly used SoftMax activation function takes the input from previous layers and converts it into a probability distribution between the eight GOSE categories. Interestingly, it turns out that a neural network can compute any function at all, meaning that they can capture complex relationships in data if the data is sufficient and of high quality.

Following model setup, the model is tested on the features using random weights. The model is then evaluated using a Loss function, which measures how far off the predictions are compared to the true values. An example of a loss-function used for GOSE is categorical cross entropy which measures the difference between the predicted probability distribution and the true probability distribution. At the beginning, it is expected that the model will not be better than random guessing. By a process called backpropagation, the weights are adjusted in a backwards fashion in order to minimize the loss-function using an optimizer algorithm. This process is repeated several times until the model reaches a satisfactory performance level in validation data[46]. One way to evaluate models in validation data in order to optimize generalizability is to use k-fold cross validation, where the training data is divided into “k” subsets of different training and validation parts. The model is then trained k times using different subsets of validation sets and the remaining subsets as training sets. The model parameters might then be optimized from the average performance in the k validation sets to make the model as generalizable as possible before evaluating it in a different test data set. Preferably, the test set should differ in time and space in a process called external validation.

Like Random Forests, neural networks can handle a large number of features and can detect complex, non-linear relationships between the features and labels. Complex neural networks are, however, computationally expensive, and when training on a large number of features, they require large amounts of data. Another problem is that they are typically not easily interpretable and are often called a “black box”, since you have limited insight into the decision-making process of the model.

To summarize, the purpose of this thesis is to explore machine learning models in order to improve outcome prediction following TBI. Although several models have been included, it is important to note that the intent was not to be exhaustive on all possible machine learning models. Examples of other models that could be explored in the future are k-nearest neighbor, support vector machines, and Naïve Bayes. The advantages of the neural network, Random Forest, and XGBoost models are, however, that they are very flexible in the input data and that they can handle non-linear and complex interacting relationships in large data sets.

Image recognition

One of the most promising fields of machine learning right now is image recognition. Image recognition is usually achieved by utilizing a type of neural network model called convolutional neural networks, or CNN. Image recognition is usually divided into object detection and image classification. In object recognition, the model might be trained to detect and locate hematomas in a CT image usually with the purpose of highlighting the scan in the radiologist's workflow so that hematoma-positive scans could be prioritized. In image classification, the hematomas might be further classified by size, location, and their mass effect, such as by midline shift.

Examples of such programs include AiDoc, Viz.Ai, and Heuron ICH, which are software platforms for automatic hematoma detection that have recently been approved by the FDA[51]. A similar software which was evaluated in this work was Zebra Medical Vision - HealthICH+. Interestingly, there are more experimental models that integrate radiological and clinical data for outcome prediction. In a recent study, Hibi et al. developed a multimodal prognostication model which utilized clinical data from CENTER-TBI as well as models developed from raw CT data to predict both binary and multiclassification outcomes based on GOSE and found that the models trained on CT data yielded improved predictions[52].

Model evaluation

Classically, binary outcome models have been evaluated using the receiver operator curve, which plots the true positive rate (sensitivity) against the false positive rate (1-specificity). The area under the receiver operator curve (AUROC) ranges from 0 to 1 and measures the performance of the model where 1 is perfect classification and 0.5 is equal to random guessing. While it is possible to calculate the AUROC for each predicted GOSE outcome category, or one-versus-rest, the results might not be easily interpreted. Instead, we have chosen to evaluate the models using other measures of discrimination and calibration.

Error rate and discrimination

The most intuitive way to measure the performance of a model is through error rate and discrimination, which measures ability to correctly predict positive cases as positive and negative cases as negative. In order to evaluate the error rate of a multicategory classification model, it is common to use accuracy, or the percentage of the time that the model predicts the correct category. When predicting outcomes in 8 categories, accuracy alone might not be the best way to evaluate the model as it doesn't confer information regarding how far off the model was from the correct category. In practice, it might be beneficial to have a model that is more likely to be "almost correct" on average even though it is less entirely correct. For that reason, we have also used accuracy within one and two categories to reflect how far off target the model is, as well as Quadratic Weighted Kappa (QWK)[53]. QWK is a chance-corrected metric which measures the agreement between the predictions and the observed outcomes, and ranges from -1 to 1, where 0 would represent total chance and 1 a perfect agreement. It is a weighted version of Cohen's Kappa, which is another common way to evaluate models, but the difference is that QWK weighs the results so that predictions that are farther from correct are penalized much more than near-misses. In other words, predicting GOSE 2 instead of 3 is less penalized than predicting GOSE 2 instead of 8. QWK is thus both an excellent metric for ordinal outcomes and rewards predictions close to the truth[53].

Other ways to evaluate a multiclass classification prediction model include, but are not limited to, average balanced accuracy, precision, recall, and specificity.

Average balanced accuracy measures the ability to correctly categorize examples from multiple categories, while considering any category imbalances in the data. Specifically, it computes the average of the true positive rates across all categories, regardless of their proportion in the dataset. Like accuracy, balanced accuracy produces a number between 0 and 1, and higher score indicates better classification performance overall.

Precision measures the proportion of true positive predictions out of all positive predictions made by the model. It is calculated by dividing the number of true positives by the sum of true positives and false positives and is thus a number between 0 and 1. Precision is a useful evaluation tool when the cost of a false positive prediction is high, such as in medical diagnoses where a false positive could lead to unnecessary treatments or surgeries. If the intention is to use a prediction model as a clinical tool, however, it is often better to have a higher false positive rate than false negative rate in order not to underestimate any risks. In that case, a high recall might be more important.

Recall, or sensitivity, measures the proportion of true positives out of all actual positive examples in the dataset. It is calculated by dividing the number of true positives by the sum of true positives and false negatives. Recall is a useful evaluation tool when the cost of a false negative prediction is high, such as when deciding to send a potentially severely ill patient home from the emergency department. There is often a trade-off between precision and recall, where the model either makes more false positive predictions at the cost of less false negatives or vice versa.

Finally, specificity is essentially the opposite of precision. Precision determines the proportion of true positives and is a determination of false positive accuracy, while specificity determines the proportion of true negatives and is a determination of the accuracy of negative predictions. Specificity is thus the proportion of true negative predictions out of all actual negative examples in the dataset and is calculated by dividing the number of true negatives by the sum of true negatives and false positives.

Confusion matrices

A confusion matrix can be utilized in order to evaluate models that predict multiple category outcomes. Confusion matrices plot the observed outcome category versus the predicted category outcome, and when predicting full

GOSE, it would thus produce a matrix with eight rows and eight columns. Within each box the number represents the times the model predicted a given category compared to which category was observed. The matrix then provides a heat map that represents how the model performs in each category, and the perfect model would produce a diagonal line with all predictions aligning with the corresponding observed outcome.

Calibration and balance

Model calibration represents how well the model's predicted probabilities of different outcomes reflect the probability of those categories in the observed outcomes. For example, if 100 patients have a predicted probability of approximately 10% for GOSE 3, then you expect 10 of those patients to have an observed outcome of GOSE 3 if the model is well calibrated. Calibration was displayed in calibration plots where the predicted probability is plotted against observed probability of that GOSE category.

A common way to numerically describe the overall calibration of a model is by expected calibration error (ECE), which summarizes the average discrepancy between predicted probabilities and observed outcomes. Patients are grouped into bins according to their predicted probabilities, and for each bin, the model's mean predicted probability is compared with the observed proportion of patients with that outcome. For example, if a group of patients all have a predicted probability of 70% for GOSE 4, good calibration means that roughly 70% of those patients have GOSE 4.

In a multi-class setting such as the 8-level GOSE, ECE can be computed in different ways. A one-vs-rest ECE treats each GOSE category separately (e.g. "GOSE 4" vs "not GOSE 4"), allowing us to see for each category whether the model tends to under- or over-predict that specific risk. A correct-class ECE instead aggregates over all patients and compares the probability assigned to each patient's true class with the model's empirical accuracy, providing a single global summary of how well the model's stated confidence matches how often it is correct.

Several measures might be used to determine model balance. Mean discrepancy (MD) is calculated as the mean of the predictions minus the mean of the observations. Negative values indicate that the model is too pessimistic, positive values that it is too optimistic, while values close to zero indicate that the model is balanced in this regard. Standard deviation discrepancy (SDD) was calculated as the standard deviation of the predictions minus the standard deviation of the observations. Negative values indicate that the model tends to make predictions closer to the center of the scale compared to the observations. Values close to zero indicate that the model is balanced in this regard. Finally, Predicted Category Proportions (PCP) describes the fraction of patients that are predicted to have a given GOSE level while the Observed Category Proportions (OCP) describes the fraction of patients that are observed to have a given GOSE level. Both will therefore consist of 8 values for each model. These values can be summarized by taking the sum of the absolute difference between the predicted category proportion and the observed category proportion for all GOSE levels, here referred to as the Total Category Proportion Discrepancy (TCPD). A value close to 0 indicates that the model, overall, doesn't favor or disfavor certain categories while a value close to 2 means that the model tends to either overpredict or underpredict certain categories. The Maximum Category Proportion Discrepancy (MCPD) is the highest category proportion discrepancy for all GOSE levels, where a value close to 0 indicates that the model neither favors nor disfavors any of the categories, while a value close to 1 indicates that the model is either favoring or disfavoring one of the categories.

Graphically, the balance of the models might be displayed with category proportion graphs and error distribution graphs. Category proportion graphs display the proportion of observed outcomes compared to predicted outcomes, complementing the confusion matrices. Error distribution graphs display each predicted GOSE category value subtracted from the corresponding observed GOSE value. A value of 0 means balanced prediction, more negative values mean that the model overestimated prognosis and positive values that the model underestimated prognosis.

Feature importance and SHAP analysis

The feature importance of classical linear models such as logistic regression might be evaluated by odds ratios and coefficients. These measures are not supported in tree-based and neural network models which utilize non-linear and interactive features, which is one of the reasons they are sometimes referred to as "black boxes". A common way to evaluate feature importance of these models is by SHapley Additive exPlanations (SHAP) analysis. SHAP analysis is a way to decompose a model's prediction for each patient into additive contributions from all predictors: starting from a baseline risk, each variable gets a SHAP value that shows how much it pushed the prediction up or down. SHAP gives both patient-level explanations ("for this patient, low GCS and high ICP drove the high predicted risk") and global importance rankings (mean absolute SHAP values across all patients). Mean absolute SHAP might be visualized as a bar plot which is a clean plot that disregards the direction of how the feature pushed the prediction, or as beeswarm plots where each dot represents a patient and it is possible to detect both direction and spread of how the features impacted predictions.

Training, testing and validation cohorts

When using data to train a model, cohort selection is as important as, if not more important than, the evaluation metrics. The training data must not overlap with the testing data, in order to prove generalizability. In order for the model to be robust, it must also be tested on validation cohorts. If the purpose of the model is to be generalizable to other geographic areas, the models must first be validated on those cohorts, and in order to test robustness over time the validation cohorts should also have temporal difference. When training a model, it is common to use a cross-validation strategy when optimizing the model variables, and the test set should preferably not be touched until the models are finalized. The data is thus usually split into two parts – a training set and a test set by a fraction of around 4:1. The training data is typically split into smaller cohorts where the model is trained on parts of the cohort and tested on a small slice of the test data, so called cross validation, and where changes to the models are done to optimize cross validation results. There are several ways to set up cross validation, but the most common is a 5- or 10-fold cross validation where the model trains on 4 or 9 parts and validates on the last. The folds are usually created so that the groups have a fairly equal distribution of the target variable, so called stratified splits. In this work, however, we also incorporated a temporal split where the folds are separated by time of trauma in order to test temporal robustness of the models over time.

Model availability

When developing a model, it is not only important that it has a low error rate and good balance, but also that it is practical. This means not only that the model is available in terms of the code and perhaps as a web application, but also that the features used are readily available in the clinical environment. It is also important to note that as more machine learning models are trained in developed countries, they are biased towards that population and might not be suitable in developing countries. It is essential that machine learning models are developed and validated in all environments that they are intended to be used.

Clinical evaluation and implementation of AI models

Even though there is a recent boom in machine learning algorithms, clinical implementation remains scarce. Many models show great promise in initial evaluation, but the rigorous external validation that is needed before clinical implementation has yet to be established.

Clinical implementation of prediction models could provide information on injury trajectories and guide clinical decisions and possibly aid in withdrawal of care.

Furthermore, urgent detection of intracranial hemorrhage and prioritization of these patients could improve the time to surgery or time to escalation of care, for instance transfer to an intensive care unit. Therefore, a decision support system that automatically detects ICH on CT scans immediately after image acquisition and highlights those cases in the radiologist's workflow could potentially save critical time for the patients.

As mentioned before, image analyzing software providing decision support for radiologists based on AI has made immense advances during recent years[54]. Especially the detection of hematomas on brain CT has been the subject of intense research [55,56] and very promising results have been reported [57,58] [59]. There is rapid progress on the development and refinement of such algorithms which is fueled by the high potential usefulness of automation that can potentially lead to the reduction of diagnosis-to-intervention time[60]. However, radiological applications of machine learning algorithms are susceptible to selection bias since their performance is largely dependent on the original training data[61]. If an algorithm is overfitted to the training data, then the algorithm may not perform well in a different clinical environment, and no clinical implementation of radiological AI model is yet in place in Sweden. Subsequently, there is a crucial need for the evaluation of these models in the clinical setting in order to find the best way of implementation. Recently, an AI radiological decision tool called Zebra Medical Vision - HealthICH+ was evaluated at Uppsala University Hospital. The software is an ensemble of three convolutional neural networks with the purpose of identifying hemorrhages in the CT-brain of patients which are then triaged for evaluation by a radiologist. In the third study, we analyzed whether this AI evaluation impacted the clinical care and outcome of the patients.

Aims

The overall aim of the thesis was to develop and evaluate machine learning models that predict multi-class functional outcomes after traumatic brain injury, as well as evaluate a clinically implemented AI model for the detection of intracranial hematomas.

Specific aims

I. External validation of the IMPACT prediction model and feasibility of 8-grade GOSE prediction
II. Development and external validation of novel outcome prediction models that predict the full 8-grade GOSE using admission variables
III. Evaluation of a clinical AI support system that detects intracranial hematomas and the impact on patient care
IV. Development and external validation of novel outcome prediction models that predict 7-grade GOSE using admission and dynamic physiological variables from the NICU

Methods

Ethics statement

The models in study I, II, and IV were developed from Uppsala data which was collected in accordance with the ethics permit provided by the Uppsala university ethics committee (EPNU 2010/138) as well as the Swedish central ethics committee (CEPN 2010/138/1 and 2020-05462). Study III had approval from the Swedish Ethical Review Authority. Due to the acute nature of the examinations, the need for written patient consent was waivered by the Swedish Ethical Review Authority.

Study I

In study I, we utilized a retrospective analysis of a prospectively maintained single centre institutional TBI registry (Uppsala registry). Patients with moderate to severe TBI admitted to the neurointensive care unit (NICU) at the Department of Neurosurgery at the University Hospital in Uppsala, Sweden, between 2008 to 2020, were screened for this study. The available IMPACT variables were retracted from the register[62]. Complimentary radiology data, as well as laboratory data for hemoglobin and glucose in plasma, and capillary and blood gas tests within 24 h of trauma, respectively, were extracted from the medical records system used at Uppsala University hospital. Patients with an age ≥ 18 and GCS ≤ 12 were included in the study. The flowchart in Figure 2 presents the included patient. We could include 635 patients in the core model of IMPACT. In the CT model, 605 patients could be included (30 patients were excluded due to missing data). Unfortunately, due to a change in the medical records system, a significant amount of laboratory glucose data was lost from record, and thus 294 patients were excluded from the laboratory model due to missing data. Due to this, a total number of 311 patients were eligible for inclusion into the laboratory model. Clinical outcome was assessed at 6–12 months post-injury, by specially trained personnel with structured telephone interviews, using the Extended Glasgow Outcome Scale (GOSE), containing 8 categories of global outcome, from death to upper good recovery.

All data management, statistical calculations, and graphs were made using R (version 3.6.1). Outcome predictions were made using the IMPACT logistic regression model, and the calculations were verified to correspond to the results generated by the online IMPACT calculator (www.tbi-impact.org), created by the IMPACT investigators. Receiver operating characteristic (ROC) curves and the corresponding area under curve (AUC) were calculated for prediction of mortality (GOSE 1) and unfavorable outcome (GOSE 1–4) using the R package pROC[63]. In order to investigate how the IMPACT prediction (mortality and unfavorable outcome) was distributed between actual GOSE scores, the predicted outcomes were plotted within each GOSE score. When plotting outcome data, the GOSE values were randomly perturbed ± 0.2 units to avoid overplotting.

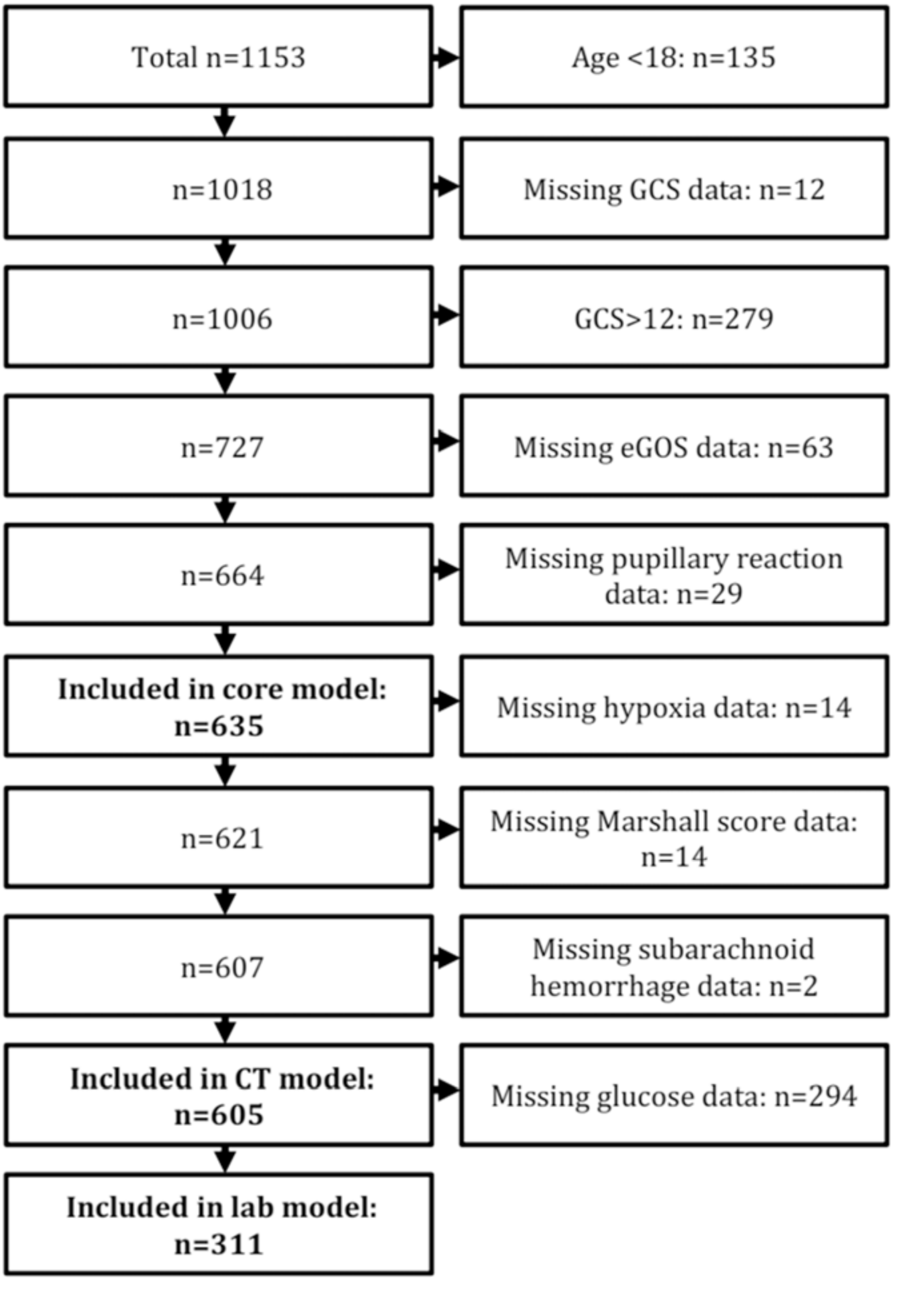


Figure 2. Flowchart of patient inclusion in study I.

Study II

In study II, we utilized a retrospective analysis of a prospectively maintained single centre institutional TBI registry (Uppsala registry) and could include 1157 patients with TBI admitted to the neurointensive care unit (NICU) at the Department of Neurosurgery at the University Hospital in Uppsala, Sweden, from 2008 to 2020, were screened for inclusion (Figure 3). Clinical variables were extracted from the Uppsala Traumatic Brain Injury register[62]. The IMPACT core and radiological variables upon admission were used for modelling: age, pupil reaction, GCS-motor score and computer tomography (CT) classification. Glasgow Outcome Scale (GOSE) was assessed at 6–12 months post-injury, by specially trained personnel with structured telephone interviews. The Uppsala cohort was divided into a training set (80%) and a test set (20%). In order to have balanced dependable variables within each group, the group assignment was randomized but constrained to have approximately the same fraction of patients in each GOSE category between the training and testing groups. However, some GOSE categories were generally overrepresented in the data. In the training data 122 patients (26%) had an outcome of GOSE 8, 78 patients (17%) GOSE 7 and 77 patients (16%) GOSE 3. The Leuven cohort had a large majority of less severe TBI, with 76% of the patients presenting with GCS-m 6 and 46% with the outcome of GOSE 8 while the PROTECTIII cohort included only 3% presenting with GCS-m of 6 and 10% with the outcome of GOSE 8.

Leuven cohort

Since 2013, core data related to TBI (including admission GCS, pupil reaction, age, accident mechanism and admission CT Marshall score) for all patients of all TBI severities admitted to the University Hospitals Leuven, with or without hospitalization, are prospectively registered and kept in a secure database. Inclusion criteria were thus all consecutive TBI patients from 2013 until August 2022. Similar to the Uppsala cohort, no exclusion criteria were set for this cohort except incomplete data. At 6 months, patients received the GOSE postal questionnaire at home and the questionnaire was scored as previously published by Wilson et al[64]. Response rate is regularly checked and was found to be around 38%. This project, intended for monitoring and benchmarking of quality of care and implemented in care standards, was approved by the UZ Leuven Ethics Committee. Data provided for the present study was fully anonymized. The Leuven cohort had a large majority of less severe TBI.

ProTECT III study cohort

ProTECT III was a phase 3 double blind placebo controlled multicenter clinical trial designed to determine the efficacy of administering intravenous (IV) progesterone (initiated within 4 h of injury and administered for 72 h, followed by an additional 24-h taper) versus placebo for treating patients with moderate to severe acute TBI (Glasgow coma scale score 12-4). The study was conducted in 49 high volume trauma centers participating in the National Institute of Neurological Disorders and Stroke (NINDS) and was conducted through the NINDS funded Neurological Emergencies Treatment Trials (NETT) network in the United States, with tight adherence to guidelines on the management of moderate to severe TBI. Emory University Institutional Review Board was the IRB of record for the study with the protocol # IRB00014409. The term used in the US is—EFIC (Exception from Informed Consent) with is strictly guided by the FDA under the waiver of informed consent (21 CFR 50.24). It required community notification and consultation in all the areas where the study was being conducted, including local IRB review and approval at all sites. ProTECT III EFIC Plan submitted with FDA IND #104188. The outcome of the trial was powered for a 10% increase in the proportion of patients with a favorable outcome by a 10% (absolute) difference, determined by the Glasgow Outcome Scale-Extended (GOSE) score at 6 months post injury when compared to placebo. The primary outcome analysis of the GOSE was a stratified dichotomy methodology for assessing improvement with GOSE scores as "favorable" versus "not favorable", based on the brain injury severity score measured at randomization (best pre- randomization GCS or iGCS). 840 out of the planned 1,140 patients were randomized, as the study was stopped for futility at a preplanned analysis. The study was conducted under an exception from consent waiver, approved by the local IRBs and the FDA (Investigational New Drug application 104188). Patient inclusion and exclusion in this cohort were much stricter than the Uppsala and Leuven cohorts. Specifically, inclusion criteria were moderate to severe injury (GCS 3–12 or motor response 2–5 if intubated), age ≥ 18, blunt and closed injury, as well as the ability to start drug infusion within 4 h from time of injury. Exclusion criteria were non-survivable trauma as determined by treating team, bilateral dilated unresponsive pupils, spinal cord injury with neurological deficits, pre-injury paralysis, inability to perform activities of daily living without assistance, cardiopulmonary arrest, status epilepticus on arrival or concern for post ictal state, systolic blood pressure < 90 at least 5 min apart, O2 saturation < 90 for at least 5 consecutive minutes, prisoner or ward of state, known active breast or reproductive organ cancers, known allergy to progesterone or Intralipid components, known history of blood clotting disorder or history of pulmonary embolism, blood or serum ethanol ≥ 250 mg %, positive qualitative pregnancy test, concern for inability to follow up at 6 months (residence in foreign country, homeless, undocumented immigrant, high likelihood of becoming incarcerated during study period etc.), and patient opt out. Excluded in this study were also patients with incomplete data.

The models were set up using R (version 4.1.2). The proportional odds logistic regression (POLR) model was set up using the MASS package in R. The random forest model was set up using the Random Forest package in R. The optimal value chosen for the number of predictors sampled for splitting at each node (mtry) was found by testing different values and choosing the one with the least error rate in cross-validation, so called "out of bag" (OOB) error rate. The mtry value was subsequently set to 1. The optimal number of trees for classification was found by plotting the error rate and the number of trees and was subsequently set to 500. The neural network was set up using the Keras package in R. The categorical input data was set up using one hot encoding, and the continuous variable (age) was scaled using the scale function. The model was set up with the Keras sequential model, including four hidden layers of 1600, 800, 200, and 100 nodes respectively with ReLU activation function and an output layer using the SoftMax activation function. The loss function was set to categorical cross-entropy, and the optimizer function was RMSprop. All models were evaluated by 10-fold cross validation and optimized to minimize the average error in the validation sets before testing the models in the external datasets.

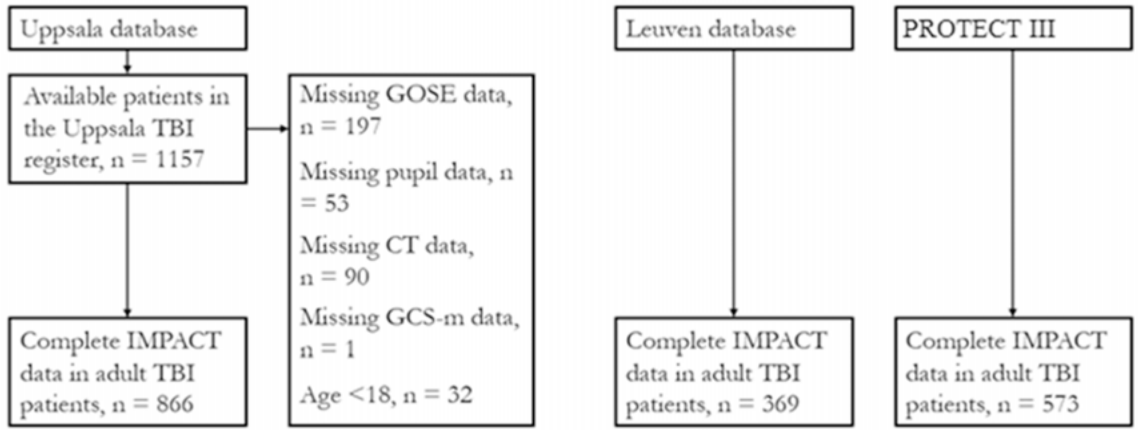


Figure 3. Flowchart of patient inclusion in study II.

Study III

The general aim of study III was to evaluate the clinical impact of a decision support system concerning detection of intracranial hemorrhage, both from a radiological and neurosurgical perspective. The evaluated software was Zebra Medical Vision - HealthICH+, an ensemble of three convolutional neural networks. The software was connected with a picture archiving and communication system (PACS) and actively analyses each incoming/performed head CT scan for signs of hemorrhage while marking the positive patients in the radiological work list and alerting the radiologist. In case of a positive finding, it provides an indicative CT slice highlighting the suspected area.

All CT head scans, except for stroke alarms, during two time periods during a total of five months between December 2019 and April 2020 were retrospectively and consecutively included and evaluated. The gap in inclusion was caused by maintenance on the AI server, without affecting the daily patient flow. Each patient was only included once (first instance). Of 2306 included cases, 288 individual patients were positively annotated by the program (AI-positive).

The referrals to CT were categorized as “routine”, “subacute” or “acute”, and the time point for scanning was defined as either “office hours” or “on call” – for a total of six categories.

Data was extracted regarding the time in minutes from when the scan was completed, and when the first report was signed by a radiologist. Median times were calculated for each patient category. Time from AI-positive scans were compared to negative, individually and on group level.

The presence of AI annotation (i.e., positive finding for hemorrhage) was compared to the content of the final report by a specialist in neuroradiology. AI annotations were defined as true positive and false positive, depending on the description of hemorrhage in the final report. Ambiguous cases were re-evaluated by a specialist in neuroradiology. Scans that were AI-negative (n = 2018) were not re-evaluated and negative predictive value was not calculated.

True positive findings were categorized into Parenchymal, Subdural, Epidural and Subarachnoid hemorrhage, with several simultaneous categories as a valid option. False positive findings were categorized into Tumors, Infarcts, Artifacts, Calcification, and Unknown reason. Positive predictive value was calculated for intracranial bleedings per se, and for “relevant diagnostic findings”, by combining true positives with cases with intracranial tumors without bleeding.

Differences between preliminary report and final report were tested for false positive preliminary reports. The specific hypothesis was that the less experienced primary radiologist may be influenced by the program to make a false positive preliminary report.

All AI-positive patients that had a time interval from scan to report that was shorter than the median time for negative patients in the same referral category ("acute", "subacute" or "routine") was defined as having "time benefit from AI-annotation" as well as having potential clinical benefit as a subsequent effect. The medical charts of these patients were systematically reviewed by a resident in neurosurgery, using an a priori designed data extraction form. Specifically, the charts were analyzed as to whether the diagnosis was previously known, (i.e., if the examination was a control of a previous known diagnosis), whether the on-call neurosurgeon was summoned after the examination, and whether the examination led to elevation of the level of care or to emergency surgery (within 24 hours). If surgery was performed, the time between the examination and time of surgery was noted.

Statistical parameters were produced using R version 4.2.2 and graphs were produced using the ggplot2 library. [65] Data skewness was determined optically by histogram as well as with Shapiro-Wilk test analysis. Because of data skewness, significant difference in time benefit between different patient groups was analyzed with two-sided Wilcoxon ranked sum test analysis. Correlation between continuous variables was determined with linear regression analysis. The boxplot graph (Figure 15) uses a square root y-scale to include extreme outliers.

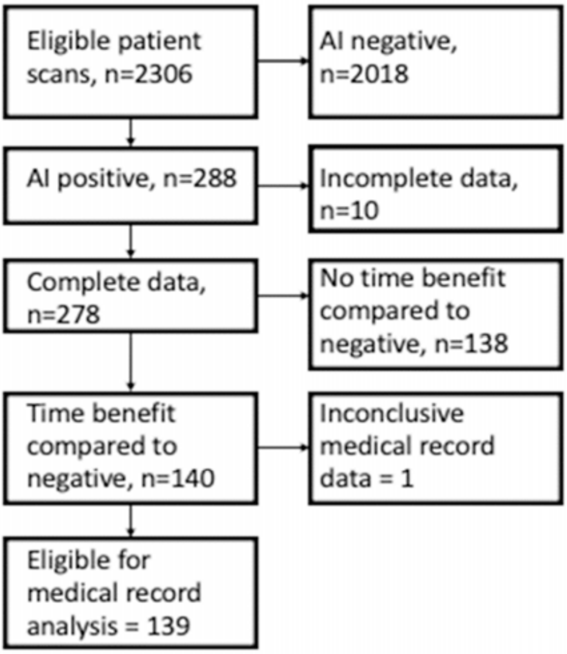


Figure 4. Inclusion flowchart of study III.

Study IV

In study IV, we utilized a retrospective analysis of a prospectively maintained single centre institutional TBI registry (Uppsala registry)[62] from patients who sustained a TBI during the years 2008-2024. Patients were included if they met the following criteria: moderate-severe TBI as classified by Glasgow Coma Scale (GCS) 3-12, a valid GOSE assessment of 6 to 12 months and a known timepoint of trauma (with ≤ 6 hours uncertainty). Patients with mild TBI, missing GOSE recordings, follow up time of GOSE >12 months, and uncertain trauma timepoint were excluded. In the NICU static/dynamic comparison group, we only included patients with correctly recorded timestamps and with at least one dynamic variable (ICP, MAP, CPP, PRx). Time of trauma was determined from the patient chart[62].

All patients with moderate or severe TBI treated at the neurointensive care unit (NICU) were subject to the same stepwise treatment escalation protocol of intracranial hypertension with the target of ICP between < 20 mmHg and

CPP > 60. The protocol consisted of elevated head, sedation, mild hyperventilation, external ventricular drainage, barbiturates, and finally decompressive craniectomy (in accordance with BTF guidelines)[34]. Arterial hypotension is treated with crystalloids and colloid fluid treatment, as well as vasoconstriction by norepinephrine.

The primary analyses involved developing and comparing prognostic models using data up to 7 days post-injury using daily timepoint (24 hours, 48 hours, 72 hours etc up to 168 hours). From the registry, we defined two distinct patient cohorts: the static data cohort and the static/dynamic cohort. The static cohort included all patients who had static admission data and the static/dynamic cohort included only patients with at least one dynamic variable (ICP, MAP, CPP, PRx) up to each daily timepoint.

Outcome definitions

The primary outcome was the ordinal GOSE outcome, recoded as eight ordered classes. Outcome 1 (death) and 2 (vegetative state/VS) was combined into one category due to their clinical severity and low prevalence of the VS outcome. The resulting 7-grade target variables were: 1/2=death/VS (GOSE 1–2), 3=lower severe disability, 4=upper severe disability, 5=lower moderate disability, 6=upper moderate disability, 7=lower good recovery and 8=upper good recovery. GOSE data was recorded by telephone interview of the patient or next-of-kin by a trained nurse[62].

For secondary analysis and comparison with traditional binary prediction models, we also calculated two binary outcomes: mortality/VS (GOSE 1–2) and favourable outcome (GOSE ≥ 5)/unfavourable outcome (GOSE <5).

Predictors

We collected data from the Uppsala TBI registry[62] including numeric, categorical and binary features such as demographics, injury characteristics, clinical variables, initial radiology findings, key timestamps (time of injury), and outcomes. Specifically, we included the following admission variables: age, gender, total Glasgow coma scale (GCS) score, GCS motor score, GCS eye score, GCS verbal score, pupillary size and reactivity, Marshall CT score and main radiological finding on the initial CT scan, injury mechanism, reaction level scale (RLS), referring hospital, presence and type of global cerebral ischemia, presence and type of paresis, presence and type of previous brain injury, whether the patients was intubated or not, presence and type of extracranial injury, extensive bleeding, and medical history (hypertension/cardiovascular disease/diabetes type 2).

We also collected high-frequency 100 Hz physiological data including intracranial pressure, mean arterial pressure, cerebral perfusion pressure which were available as averaged minute-by-minute recordings. Intracranial pressure was measured using intraparenchymal monitoring devices (Codman ICP MicroSensor; Codman & Shurtleff Inc., Raynham, MA, and Sophysa Pressio 2 ICP monitor system; Sophysa Inc, Orsay, France) and extra-ventricular drainage (HanniSet, Xtrans, Smith Medical GmbH, Glasbrunn, Germany or VentrEX, Neuromedex, Hamburg, Germany). Arterial blood pressure was recorded using an arterial needle in the radial artery, with the dome at the level of heart. The pressure reactivity index (PRx) was precomputed as a moving 5-min correlation of 10-sec averaged ICP and mean arterial pressure (MAP), updated every minute[32]. We also included data on whether the patient was treated with barbiturates or decompressive craniectomy (DC). Time to DC was recorded and the feature was only included if the surgery was performed within the time window to prevent the model from seeing features that would be in the future, what is known as data leakage.

Feature engineering

To translate the complex dynamic physiological data into meaningful predictors, we engineered a comprehensive set of features designed to capture different aspects of a patient's clinical course, such as overall stability, exposure to harmful events, and trends over time (see below).

Mean intracranial pressure (ICPm), mean arterial pressure (MAP), cerebral perfusion pressure (CPP) and pressure reactivity index (PRx) were precomputed minute-wise. For each patient in the static/dynamic cohort, NICU measurements were aligned to time-from-trauma by subtracting the patient’s trauma timestamp and then restricted to the interval for the given timeframe. Clinically implausible values were set to missing using variable-specific ranges for acceptance: ICPm -5–100 mmHg, CPP 0–150 mmHg, MAP 20–160 mmHg.

Short gaps were handled only within feature computation by aligning to the 1-minute grid with limited forward-fill of ≤ 15 minutes. We required at least 5 raw points per variable before deriving features to avoid imputation and feature engineering from extremely scarce datapoints. Per-patient NICU processing outcomes were logged and summarized. All NICU features were computed strictly from data within the timeframes.

A comprehensive set of dynamic features was engineered from the cleaned, time-aligned NICU data, including:

Data coverage: Indicators for the presence of each physiological variable and the fraction of time each was monitored and included as features.

Summary statistics of time windows: Features included summary statistics of the raw data points (mean, standard deviation, 10th/90th percentiles), as well as a time-weighted mean. To handle irregular sampling of time-series data, we resampled the data onto a uniform 1-minute grid. Missing episodes were imputed using a constrained forward-filling approach, where the last recorded measurement was carried forward for a maximum of 15 minutes to bridge short gaps in monitoring without imputing stale data over longer periods. The time-weighted mean was subsequently calculated as the mean of this imputed, gridded data. We also calculated the slope of physiological variables during the time windows.

Event burdens: The fraction of monitored time and absolute minutes spent beyond clinically relevant thresholds. Specifically, ICP above 15, 20, 22 and 35 and below 10 mm Hg. The thresholds of 20- and 22-mm Hg were chosen since they have previously been associated to outcome and are widespread clinical praxis thresholds (20 mm Hg according to local management and 22 mm Hg according to BTF guidelines)[34]. The motivation to also evaluate ICP > 15 mmHg, i.e., moderately elevated but slightly below strict management target thresholds, was that higher burden of ICP within 15-20 mmHg has been associated with worse outcome, when cerebral autoregulation is impaired[41]. The cutoff of 35 to specifically try to find those with extremely high ICP insults. The cutoff of below 10 was chosen to try to specifically find those patients with stable ICP which has been shown to be a predictor of improved outcome[25].

Furthermore, we chose the commonly used CPP thresholds of below 50 and 60. Previous studies at our centre have found worse outcome in TBI patients with high CPP, specifically above 80[39] so we also included CPP above 80 and 100 as thresholds. We also included thresholds MAP above 120 mmHg. These were computed only when a minimum of 3 hours of observed data existed for that variable for each timeframe to ensure stability.

Insult dose: Area-under-the-curve metrics quantifying the cumulative magnitude and duration of insults of ICP >20[40].

Event signatures: To quantify the burden of secondary insults, we identified adverse episodes from the time-series data. An episode of intracranial hypertension, for example, was defined as a period where ICP exceeded 20 mmHg for at least 5 minutes; brief interruptions of less than 2 minutes were merged to constitute a single event. From these episodes, we calculated a comprehensive set of features, including episode count, total duration, mean/median duration, time-to-first episode, peak insult severity, statistics on the inter-episode interval (mean and coefficient of variation), and the duration of physiological recovery following an episode. To ensure the integrity of these calculations, missing values were imputed via forward-fill limited to gaps of three minutes or less, preventing the artificial extension of insult episodes.

Joint burdens: The burden of simultaneous multivariable insults were quantified as both time-based burdens and as discrete episode signatures. This included “double-hit” events (e.g., concurrent high ICP and high PRx or high and low CPP) and “triple-hit” events (e.g., concurrent high ICP, high PRx, and low CPP). Different cut-offs for the definition of burden were explored, specifically ICP > 15, 20, 22 and 35, CPP <60 and 50, and PRx > 0.2, 0.3 and 0.4.

To summarize, data leakage was mitigated by restricting all NICU features to timeframe, censoring DC timing/type to information available by the timeframe and limiting forward-fill to ≤ 15 min for feature computation and ≤3 minutes for event burden. Excluding patients who died or were discharged within the landmark was considered but ultimately not performed in order to maximise data availability. Missingness was handled by several approaches. For linear models, numeric features were imputed using the training set median, and a binary missingness indicator was added. For tree-based models, missingness was handled internally. Categorical features were imputed with a constant "missing" value. NICU features were only derived if a patient had at least 5 raw data points for that variable.

Model development

The dataset was divided into a training set (80%) and a held-out test set (20%). We explored two splitting strategies: a stratified split based on ordinal outcome and a temporal split, which separated patients by time of trauma and allocated older patient data for training and more recent patient data for testing. All splits used a fixed random seed.

A sanity-check procedure was embedded to confirm that key dynamic features showed expected distributions between good and poor outcome groups in the training data (e.g., mean ICP higher among poor outcomes) to detect preprocessing anomalies.

We developed and evaluated machine-learning models to predict the 7-grade Glasgow Outcome Scale Extended (GOSE 1-2 combined) after TBI, using information available at seven prespecified timepoints: 24, 48, 72, 96, 120, 144 and 168 hours after trauma. For each timepoint, only data recorded up to and including that time were available to the model to avoid future-data leakage. For the models incorporating both admission and NICU data, a feature selection step was applied to the dynamic features to reduce dimensionality and improve generalization. We used recursive feature elimination (RFE) with a computationally efficient XGBoost model as the estimator. Only numeric

NICU features with ≥ 35% non-missingness were considered; targets were aligned and missing rows removed. RFE iteratively pruned the weakest features until only the top 50 NICU features remained. The selected dynamic features were then combined with the full set of admission predictors. To avoid instability from highly sparse predictors, features with less than 25% non-missingness across the entire training cohort was removed before model fitting.

We compared two modelling approaches:

1. Extreme Gradient Boosting (XGBoost): A multi-class gradient boosting classifier (objective: multi:softprob) that treats the 7-grade outcome as independent, non-ordinal categories. To address class imbalance, different balance weights were explored to address class frequencies.
2. Ordinal logistic regression: A proportional odds logistic regression model (mord.LogisticAT) that explicitly models the ordinal nature of the GOSE outcome.

In other words, we compared a powerful, non-ordinal gradient boosting model (XGBoost) with a model explicitly designed for ordered outcomes (Ordinal Logistic Regression) to assess whether respecting the ordinal structure was beneficial.

Categorical predictors were one-hot encoded, and numeric features for the ordinal regression model were median-imputed and standardized. For XGBoost, numeric features were used directly without scaling.

Hyperparameter tuning was performed for both models using 50 iterations of randomized search with the Quadratic Weighted Kappa (QWK) as the optimization metric. Category thresholds were tuned to maximize QWK and then fixed prior to test evaluation. Cross-validation was set up using 5 splits: for the temporal data split TimeSeriesSplit was used to preserve chronological order, while StratifiedKFold was used for the stratified split. After identifying the optimal hyperparameters, a post-hoc calibration of decision thresholds was applied to the entire training set. This calibration aimed to further maximize QWK of the final ordinal predictions using a hybrid approach that balanced QWK optimization with matching the outcome distribution priors. The hyperparameter search space for XGBoost included: n_estimators (400–1000), max_depth (1–8), learning_rate (0.03–0.1), subsample (0.5–1.0), colsample_bytree (0.5–1.0), gamma (0.5–5), reg_lambda (0.5–5), reg_alpha (0–3), and min_child_weight (1–30). The ordinal model tuned alpha over 10^{-3}–10^{2}.

Model evaluation

Model performance was primarily assessed on the hold-out test set using the QWK metric with 95% confidence intervals generated via 1000 bootstrap resamples. We also evaluated discrimination for secondary binary endpoints (mortality/VS and favourable outcome) using the Area Under the Receiver Operating Characteristic Curve (AUROC). Performance was further visualized using confusion matrices and plots comparing predicted versus observed outcome distributions. We also reported overall accuracy and accuracy within 1 and 2 GOSE categories.

We assessed model calibration by performing a one-vs-rest analysis for each of the seven outcome categories. For each category, we generated a reliability diagram (using 20 quantile-based bins), calculated the calibration slope with its 95% bootstrap confidence interval, and computed the Expected Calibration Error (ECE).

Finally, to enhance interpretability, feature importance was investigated using Shapley Additive exPlanations (SHAP), a method that quantifies the contribution of each input variable to the model's output.

Ethical approval

The data was collected in Sweden in accordance with the ethics permit approved by the National Ethical Review Authority (Dnr 2010/138, Ö 19-2010, 2015/224, Dnr 2024-05969-01). Informed consent by the patients and/or the legal representative/next of kin was obtained.

Software specifications

All analyses were performed in Python (v3.10+) using the libraries scikit-learn, xgboost, mord, SHAP, pandas, numpy, and matplotlib. To ensure reproducibility, each run generated a timestamped output directory, containing the full configuration file (JSON), detailed pipeline logs, data processing audits, exact features lists used for each model, generation of graphical figures (such as performance plots), and a final summary of results. A fixed random seed was used for all stochastic procedures such as data splitting and model training, to ensure reproducibility. The de-identified data is available upon reasonable request, and the full code is available in a GitHub repository. Exact feature list can be found in the supplement.

Results

Study I

The demographics and characteristics of patients are presented in Table 5. When analyzing the 635 patients with complete parameters for the core model, the core model performed an AUC of 0.85 for mortality. Analysis for unfavorable outcome (GOSE 1-4) showed an AUC of 0.79. Adding CT module to core did not increase AUC for mortality and slightly decreased AUC for unfavorable outcome to 0.78. When the 311 patients having lab parameters (blood hemoglobin and glucose) were included in the model, the AUC for mortality increased to 0.89 while for unfavorable outcome it decreased to 0.76 (Figure 5). In order to analyze whether the laboratory patient cohort differed from the whole cohort, we also made core and CT module ROC curves specifically for this patient group which yielded similar AUC results.

In order to investigate how well IMPACT could predict the full-scale of GOSE we compared predicted outcome from the IMPACT models to actual GOSE outcome. As seen in figure 6, the core, CT and laboratory models all generally predicted a low, but still overestimated, risk of mortality in the GOSE 3-8 groups. In the GOSE 1 group, all models predicted a wide range of mortality risk and thus generally overestimated the survival chance of the group. In the prediction of unfavorable outcome, the models produce a spread of outcome prediction in each GOSE group in particular in GOSE 3.

Nevertheless, there are clear tendencies that all models correctly predicted low risk of unfavorable outcome in GOSE 5-8 but also incorrectly predicted a low risk in the GOSE 3-4 groups. In the GOSE 1-2 groups, the models correctly predict a much higher risk of unfavorable outcome. It is also notable that a large number of patients ended up in the 40-60% range, which doesn't contribute much to outcome prediction.

Finally, calibration plots (Figure 7) also display fair calibration overall. The core model slightly overestimated the risk of mortality and underestimated unfavorable outcome. The CT model generally performed very well in predicting mortality although there was a slight tendency of underpredicting the risk. Finally, the lab model performed worst, overpredicting the risk of low mortality, underpredicting the risk of high mortality and generally underpredicting the risk of unfavorable outcome.

Table 5. Demographics and site of primary care in the patient cohort of study I. Abbreviations: ASDH = Acute subdural hematoma, EDH = Epidural hematoma, DAI = Diffuse axonal injury. UUH = Uppsala University Hospital. NA = Not available (data).

Parameter	Value	Radiology findings	n=	GCS on admission	n=	Pupillary reaction	n=	Trauma mechanism	n=
Number of patients (n)	635	ASDH	211	3	41	Both reacting	510	Cyclist hit by vehicle	22
Mean age (yrs)	53	EDH	37	4	30	One reacting	65	Fall accident	327
Median age (yrs)	57	Contusion	160	5	28	None reacting	60	Vehicle accident	165
Min age (yrs)	18	Subarachnoid hemorrhage	44	6	85			Pedestrian hit by vehicle	24
Max age (yrs)	93	DAI	40	7	55			Assault	19
Male (n)	490	Impression fracture	8	7.5	60			Sports accident	16
Male (%)	77	Mixed	116	8	33			Other	62
Female (n)	145	Other	13	9	15				
Female (%)	23	Normal	3	9.5	152				
Admitted to UUH (n)	127	NA	3	10	9				
Admitted to UUH (%)	20			11	12				
Transferred from local hospital (n)	508			11.5	100				
Local hospital (%)	80			12	15				

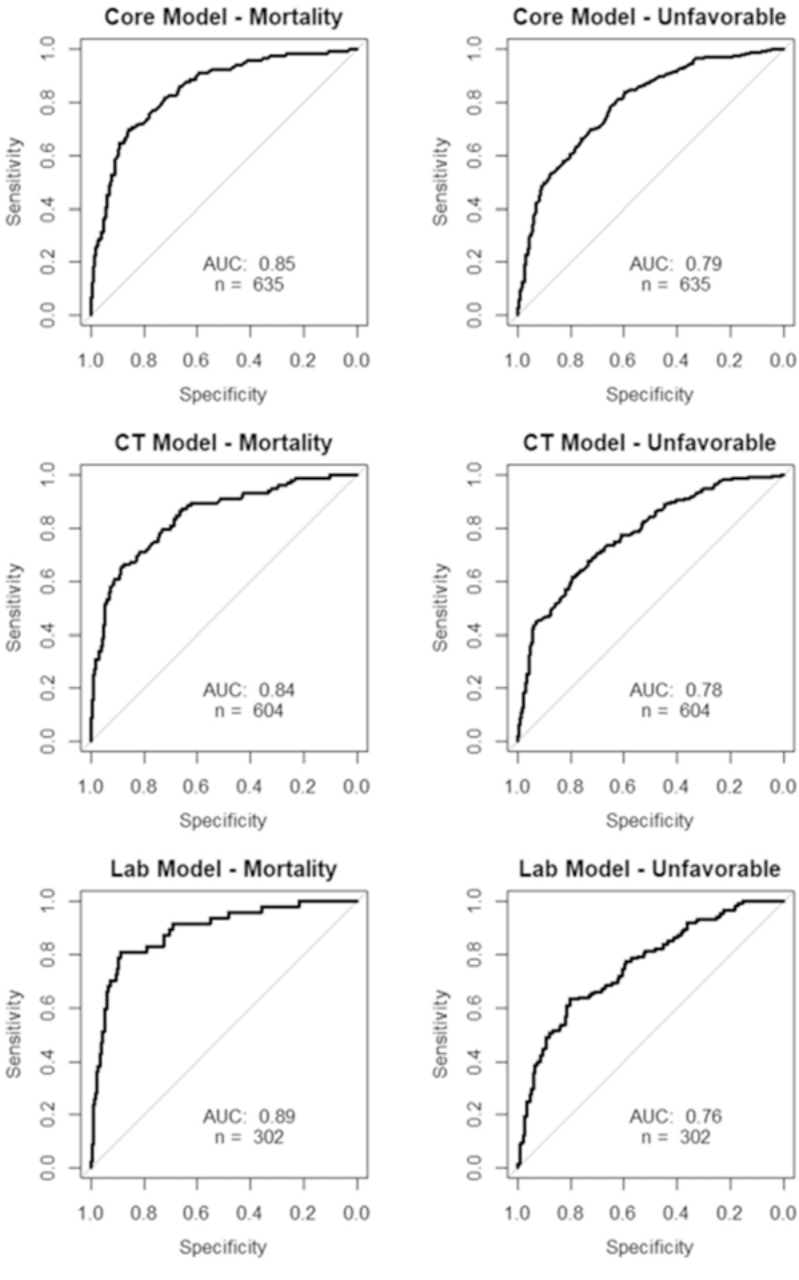


Figure 5. ROC curves for the core, CT, and laboratory IMPACT models in study I.

The IMPACT models were tested in full-scale GOSE outcome prediction, but the models performed inadequate discrimination since they were not developed for this purpose.

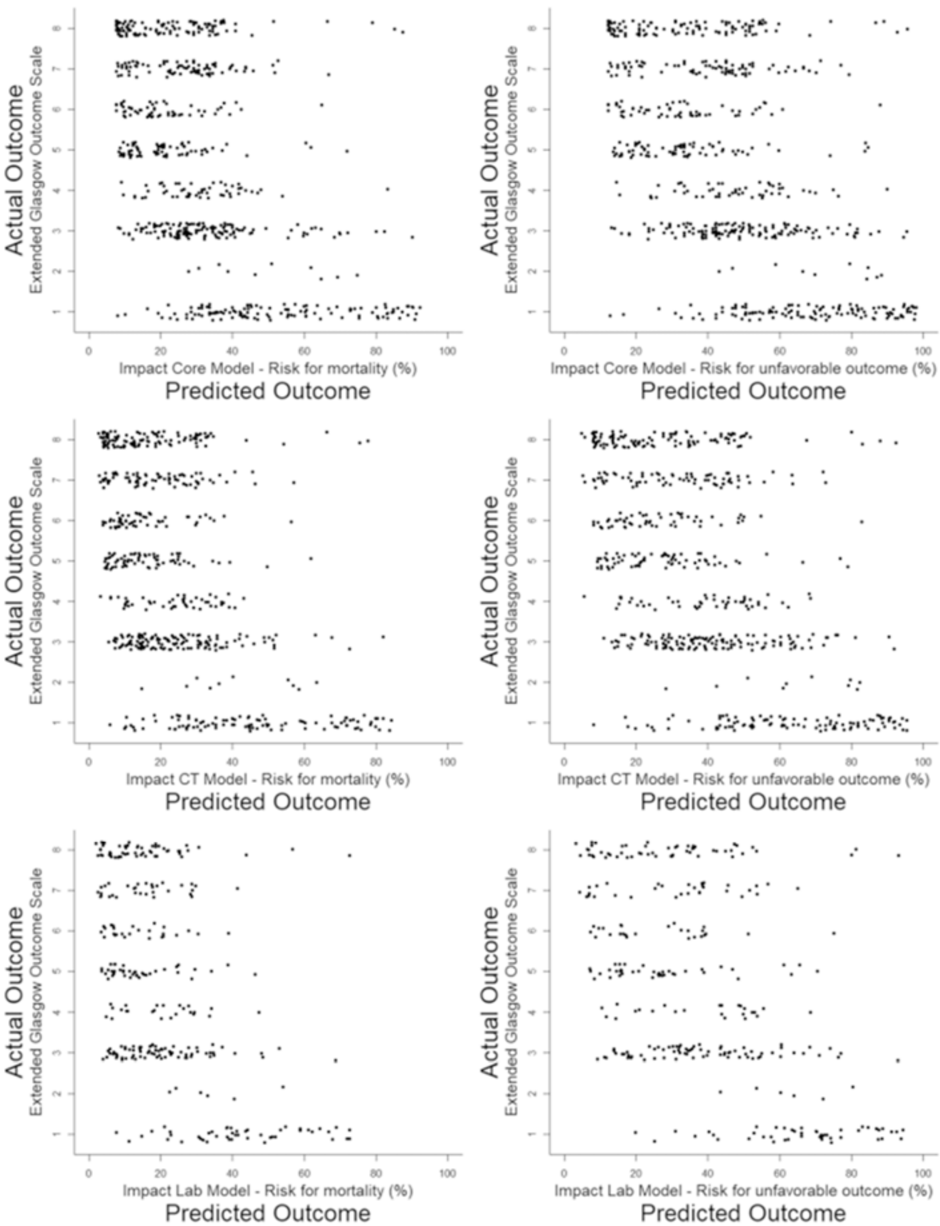


Figure 6. Actual GOSE outcome compared to predicted outcome by the IMPACT models in study I. The dots represent individual patients in their actual GOSE groups compared to percentual risk of mortality or unfavorable outcome respectively. GOSE scale extends from 1 = death up to 8 = upper good recovery. Unfavorable outcome = GOSE 1–4.

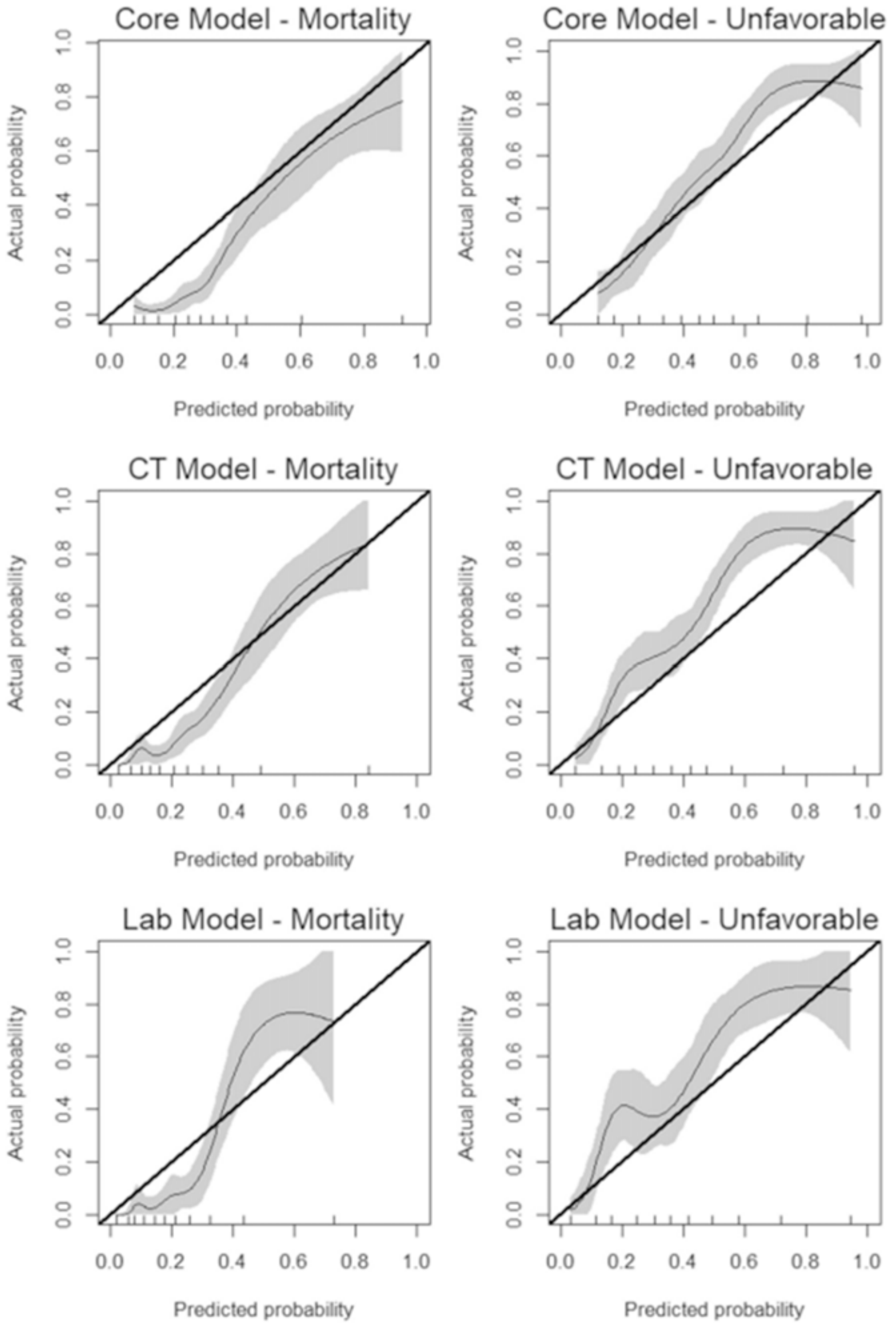


Figure 7. Calibration plots of the core, CT, and lab models in study I. The thin line describes the performance of the models, with the gray area representing two standard errors. The thick line represents the optimal model, where the area over the thick line represents underestimation, and the area under the thick line represents overestimation of risk.

Study II

A total of 1157 patients were considered for inclusion from the Uppsala TBI registry[62]. No data was imputed, and the data was restricted to cases with all input variables available. That resulted in 866 patients, 693 in the training set and 173 in the test set. External validation was performed in 369 TBI patients with complete dataset from Leuven, Belgium, and 573 TBI-patients extracted from the Progesterone for Traumatic Brain Injury: Experimental Clinical Treatment (ProTECT III) trial (US) (Figure 8)[66]. Patient demographics are presented in table 6.

Model performance

A summary of the performance of each model is presented in table 7. Accuracy was relatively similar in all three models, but POLR performs slightly worse than RF and NN. POLR performed an accuracy of 0.3, 0.44 and 0.18 in the Uppsala, Leuven and ProTECT III cohorts respectively. While RF and NN performed an accuracy of 0.35, 0.51-0.52, and 0.20-0.21 in the Uppsala, Leuven and ProTECT III cohorts. AW1 and AW2 follow a similar trend with RF and NN outperforming POLR and all models showing better results in the Leuven cohort, and worse in the ProTECT III cohort. Interestingly, ABA is quite similar between all models and cohorts around the suboptimal values 0.5-0.6. Contrarily, POLR has slightly better precision in the Uppsala and Leuven cohorts, while the precision of the RF model was best in the ProTECT III cohort, meaning that they had less risk of false positives in those cohorts. Recall was quite similar in all models with 0.3-0.35 in the Uppsala cohort, 0.44-0-52 in the Leuven cohort and 0.18-0.21 in the ProTECT III cohort meaning that all models had similar results in making false negative predictions. Specificity was similarly equal in the Uppsala cohort, 0.82-0.85, and ProTECT III cohort, 0.84-0.86. In the Leuven cohort POLR performed higher specificity (0.83) compared to RF (0.73) and NN (0.75). Furthermore, NN performed the best mean deviation (MD) at 0.23 followed closely by POLR at 0.35 and then RF at 0.89 in the Uppsala cohort, meaning that all models were fairly balanced but that RF was slightly more optimistic. In the external validation, NN and RF were performed similarly in the Leuven cohort at 0.53 and 0.64 respectively but POLR was negative 0.54 meaning that POLR was slightly more pessimistic. All models performed worse in the ProTECT III cohort with MDs ranging from -2.15 to -2.53 meaning that all models were too pessimistic in their predictions. The SDD of the models show that all models made predictions more to the edges of the scale in the Uppsala cohort and Leuven cohorts while in the ProTECT III cohort the NN model performed more predictions in the center of the scale (SDD -1.02) compared to RF (-0.8) and POLR (-0.58). TCPD shows that all models have similar risk of overprediction of some categories but that the neural network is slightly more balanced compared to the POLR and RF models in all cohorts. MCPD shows however that all models are quite equally likely to over- or underpredicting one specific category. Finally, AUROC for unfavorable outcome and mortality were quite similar in all models. AUROC for unfavorable outcome was in the range of 0.75-0.77, 0.86-0.9 and 0.51-0.59 in the Uppsala, Leuven and ProTECT III cohorts respectively. For mortality the values ranged from 0.8-0.83, 0.78-0.88 and 0.60-0.69 in the Uppsala, Leuven and ProTECT III cohorts respectively.

Graphically, the results are displayed as confusion matrices in figure 9, as category proportion graphs in figure 10 and as error distribution graphs in figure 11. The confusion matrices and category proportion graphs clearly show that POLR only predicts GOSE 1,3 and 8 with an overprediction of GOSE 8 in the Uppsala and Leuven cohorts and overprediction of GOSE 1 in the ProTECT III cohort. RF and NN similarly overpredicts GOSE 1,3 and 8 but at least makes some predictions in the other categories. They also tend to overpredict category 8 in the Uppsala and Leuven cohorts, and GOSE 3 in the ProTECT III cohort. The error distribution graph shows that the models, however, are fairly balanced in the Uppsala and Leuven cohorts but again being too pessimistic in the ProTECT III cohort.

The POLR, RF and NN models were also evaluated in the Uppsala cohort using calibration plots (figure 12) which show that the POLR model shows fair calibration in the GOSE categories 1 and 8 but underpredicting the probability of categories 3-7 and the models practically predicted no probability of GOSE 2. The RF models predicted probability is sufficiently calibrated in the GOSE categories 1 and 3 while overpredicting GOSE 8 and under predicting GOSE 4-7, and similar shows a very low predicted probability for GOSE 2. Likewise, the NN model is reasonably calibrated in the GOSE categories 1, 3, 8 while adding better calibration in predicted probability for GOSE 4 and 5 compared to RF and POLR. However, the NN still underpredicts the GOSE categories 2, 6 and 7.

Simulation of different sample sizes showed that increasing the sample size increased the accuracy of the POLR models up to a plateau at 400 before increasing again and plateauing at 600 (figure 13).

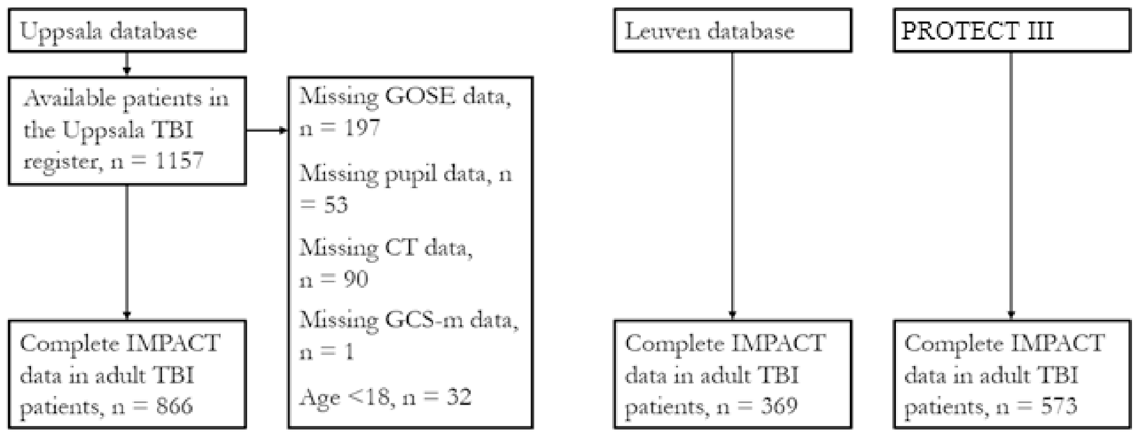


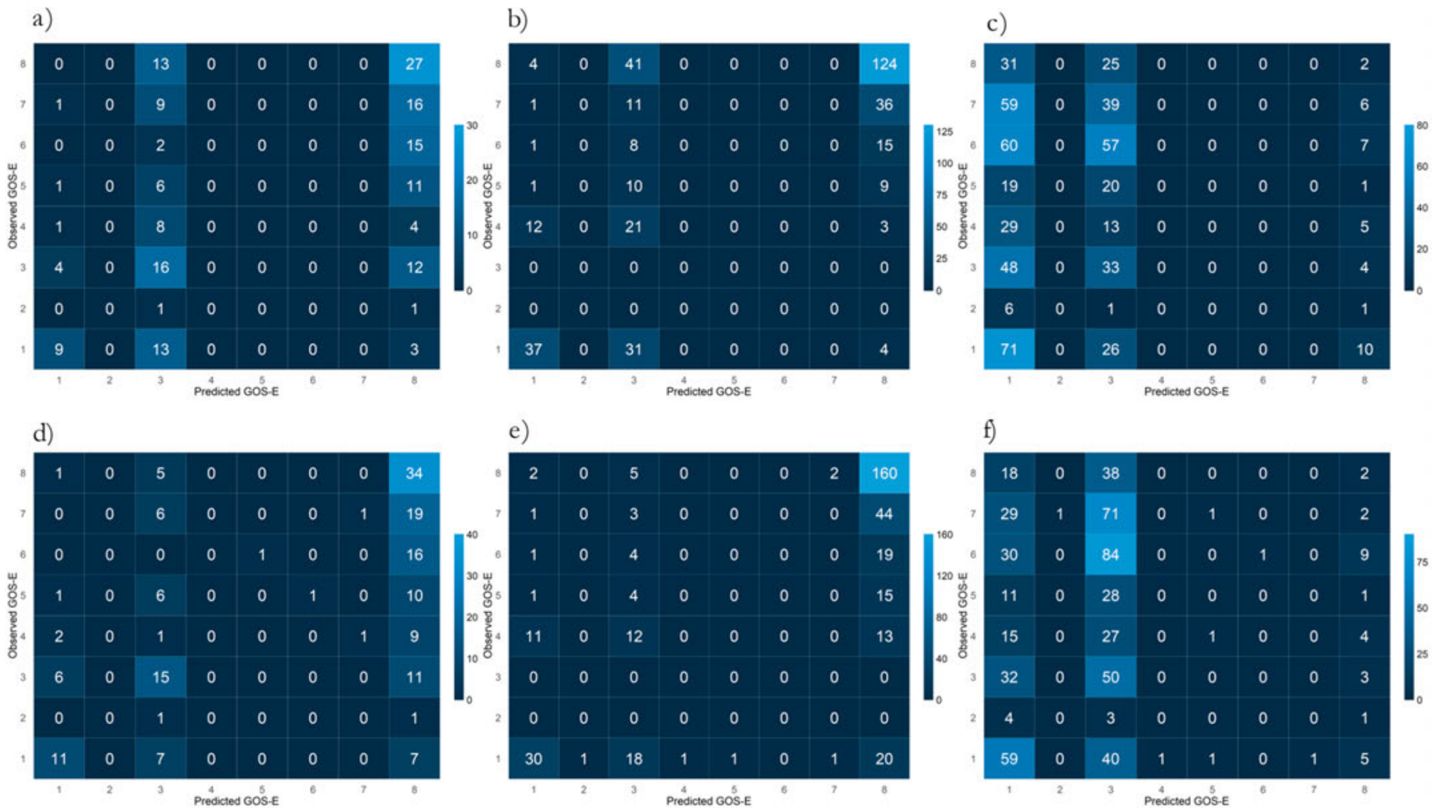
Figure 8. Flowchart of patient inclusion, including exclusion criteria of the Uppsala cohort in study II.

Table 6. Patient demographics of the included patients from the Uppsala traumatic brain injury register in study II. Abbreviations: GCS-m = Glasgow Coma Scale–motor. ASDH = Acute subdural hematoma. SAH = Subarachnoid hemorrhage. EDH =Epidural hematoma. DAI = Diffuse axonal injury.

Trauma mechanism	**n=**	**Admission data**	**n =**	**Radiology findings**	**n =**	**Marshall score**	**n =**	**GOSE**	**n =**
Fall accident	446 (52%)	Male	668 (77%)	ASDH	273 (32%)	Diffuse I	12 (1%)	1	126 (15%)
Vehicle accident	214 (25%)	Female	198 (23%)	Contusion	222 (26%)	Diffuse II	392 (45%)	2	10 (1%)
Pedestrian hit by vehicle	30 (3%)	Mean age	52.67	Traumatic SAH	72 (8%)	Diffuse III	113 (13%)	3	158 (18%)
Cyclist hit by vehicle	28 (3%)	Median age	56	EDH	67 (8%)	Diffuse IV	51 (6%)	4	65 (8%)
Assault	38 (4%)	GCS-m 1	32 (4%)	DAI	38 (4%)	Evacuated mass lesion	177 (20%)	5	88 (10%)
Sports accident	24 (3%)	GCS-m 2	27 (3%)	Impression fracture	21 (2%)	Non-evacuated mass lesion	121 (14%)	6	87 (10%)
Other	86 (10%)	GCS-m 3	268 (31%)	Mixed	141 (16%)			7	131 (15%)
		GCS-m 4	38 (4%)	Other	23 (3%)			8	201 (23%)
		GCS-m 5	103 (12%)	Normal	5 (1%)				
		GCS-m 6	398 (46%)						
		Both pupils reacting	748 (86%)						
		One pupil reacting	61 (7%)						
		No pupil reacting	57 (7%)						

Table 7. Summary of the models' performances in study II. POLR = Proportional odds logistic regression. RF = Random Forest regression. NN = Neural network. AW1 = Accuracy within one category. AW2 = Accuracy within two categories. ABA = Average balanced accuracy. MD = Mean discrepancy. SSD = Standard deviation discrepancy. TCPD = Total category proportion discrepancy. MCPD = Maximum category proportion discrepancy. AUROC-M = Area under the receiver operator curve for mortality (GOSE 1). AUROC-U = Area under the receiver operator curve for unfavorable outcome (GOSE 1-4).

Uppsala	Accuracy	AW1	AW2	ABA	Precision	Recall	Specificity	MD	SDD	TCPD	MCPD	AUROC-U	AUROC-M
POLR	0.30	0.45	0.66	0.54	0.35	0.30	0.82	0.35	0.29	0.98	0.28	0.77	0.80
RF	0.35	0.49	0.69	0.56	0.32	0.35	0.83	0.89	0.32	0.88	0.39	0.77	0.83
NN	0.35	0.50	0.64	0.57	0.31	0.35	0.85	0.23	0.38	0.64	0.23	0.75	0.83
Leuven	**Accuracy**	**AW1**	**AW2**	**ABA**	**Precision**	**Recall**	**Specificity**	**MD**	**SDD**	**TCPD**	**MCPD**	**AUROC-U**	**AUROC-M**
POLR	0.44	0.59	0.74	0.57	0.65	0.44	0.83	-0.54	0.19	0.78	0.33	0.90	0.88
RF	0.51	0.67	0.79	0.56	0.43	0.51	0.73	0.64	-0.04	0.81	0.28	0.89	0.86
NN	0.52	0.69	0.78	0.56	0.50	0.52	0.75	0.53	-0.01	0.74	0.25	0.86	0.78
ProTECT III	**Accuracy**	**AW1**	**AW2**	**ABA**	**Precision**	**Recall**	**Specificity**	**MD**	**SDD**	**TCPD**	**MCPD**	**AUROC-U**	**AUROC-M**
POLR	0.18	0.23	0.41	0.51	0.16	0.18	0.84	-2.53	-0.58	1.20	0.38	0.59	0.63
RF	0.20	0.26	0.45	0.51	0.30	0.20	0.85	-2.15	-0.80	1.21	0.45	0.51	0.60
NN	0.21	0.29	0.44	0.52	0.12	0.21	0.86	-2.35	-1.02	1.14	0.34	0.59	0.69


a)
b)
c)
d)
e)
f)
Observed GOS-E
Predicted GOS-E

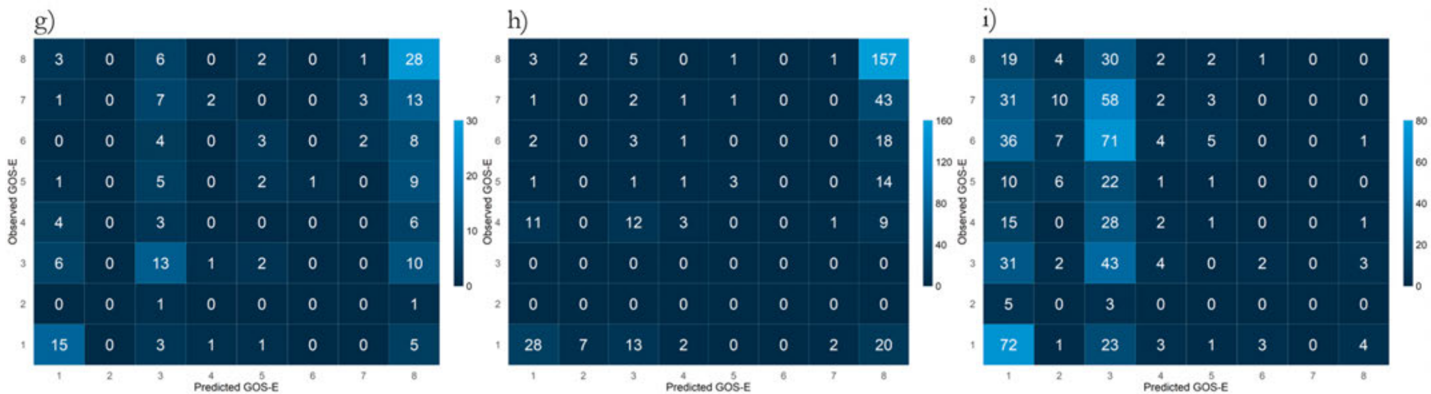
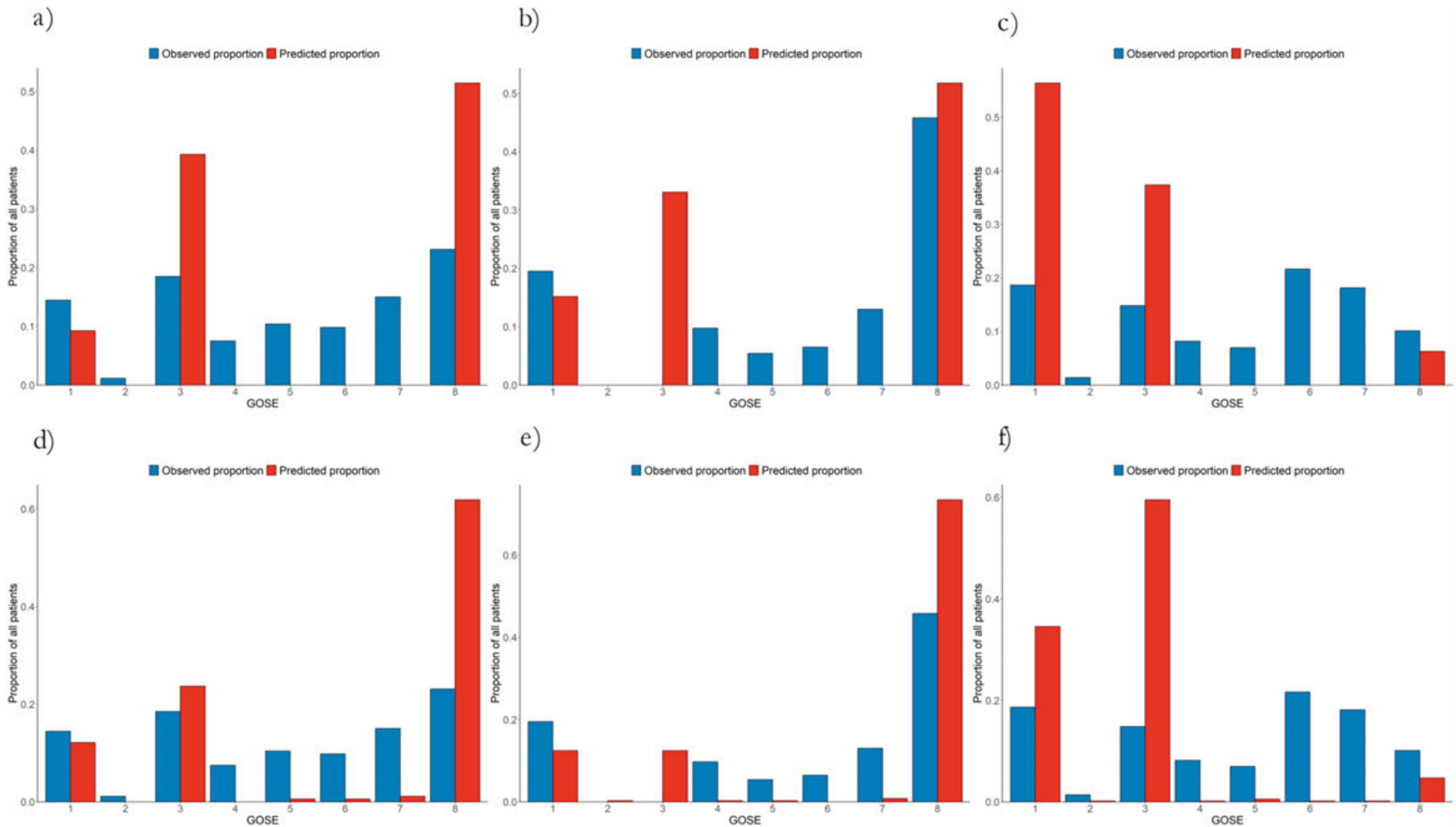


Figure 9. Confusion matrix evaluation of the models in study II. showing observed Glasgow outcome scale extended (GOSE) over predicted GOSE. All models overpredict GOSE categories 1, 3 and 8. While the proportional odds logistic regression (POLR) model only predicts those two categories, the random forest (RF) and neural network (NN) models predict some middle categories. a) POLR in the Uppsala cohort showing, albeit somewhat balanced, prediction of only GOSE 1,3 8. b) POLR in the Leuven cohort showing overprediction of GOSE 8. c) POLR in the ProTECT III cohort showing overprediction of GOSE 1 and 3. d) RF in the Uppsala cohort showing overprediction of GOSE 8, some in GOSE 1 and 3 and a few middle category predictions. e) RF in the Leuven cohort, similarly overprediction GOSE 8. f) RF in the ProTECT III cohort, overpredicting GOSE 1 and 3. Finally, g) shows the NN model with slightly more spread-out predictions but still overprediction of GOSE 1, 3 and 8. h) The NN model overpredicts GOSE 8 in the Leuven cohort. i) The NN model also overpredicts GOSE 1 and 3 in the ProTECT III cohort.


a)
b)
c)
d)
e)
f)
Observed proportion
Predicted proportion
Proportion of all patients
GOSE

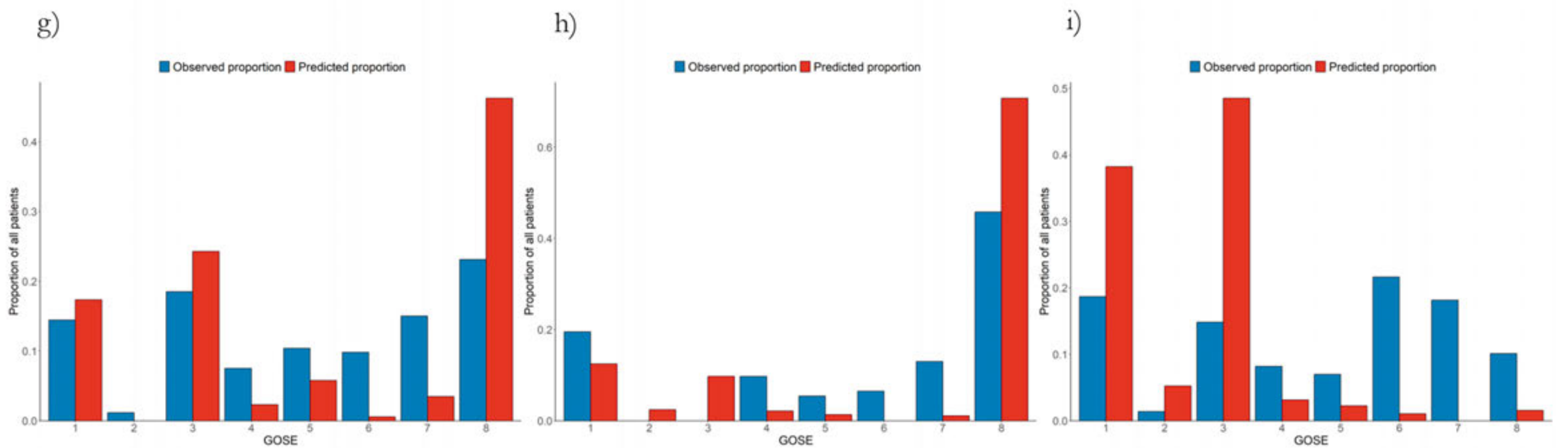
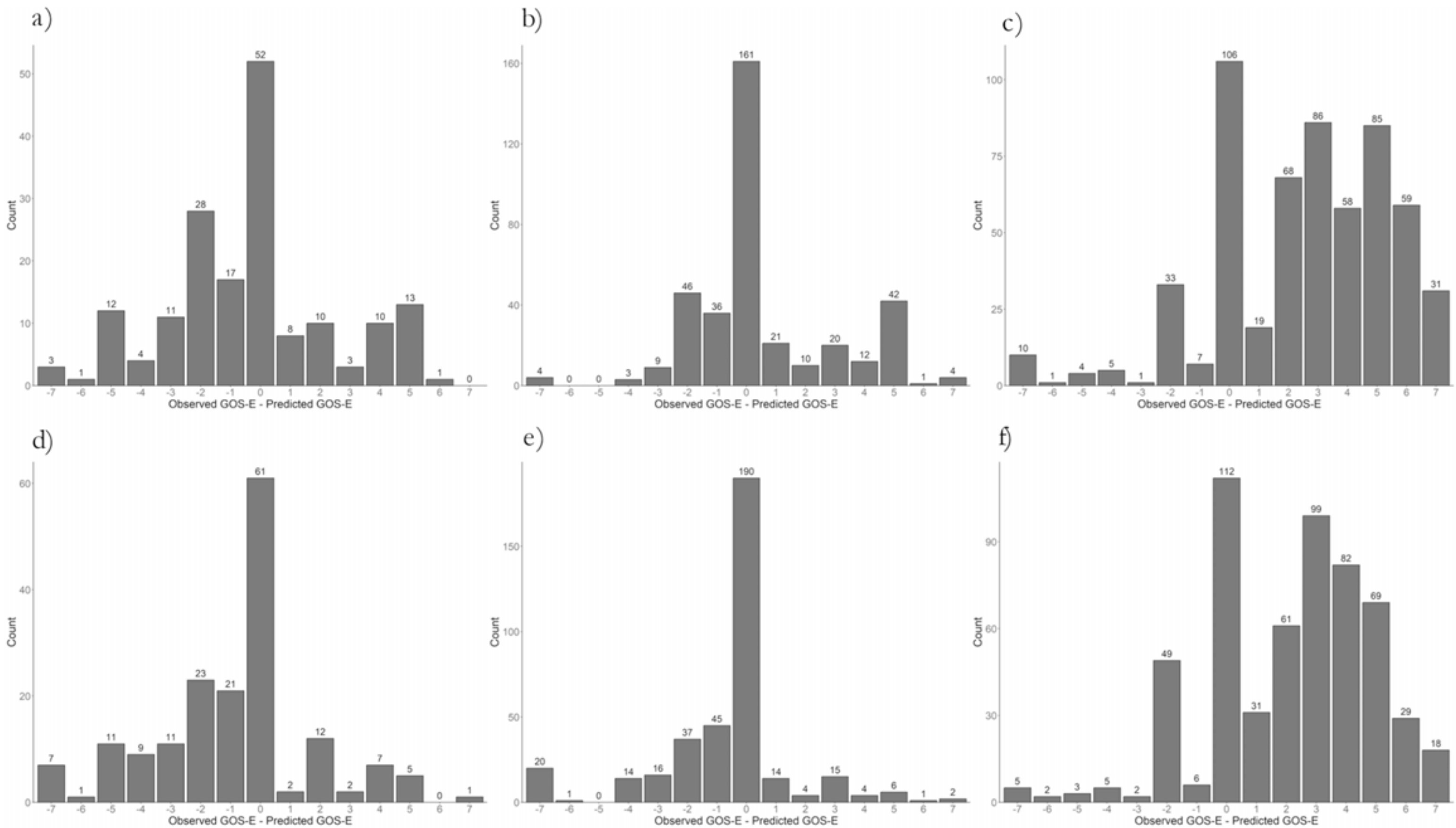


Figure 10. Category proportion graphs of the proportional odds logistic regression (POLR), random forest (RF) and neural network (NN) models complementing the confusion matrices in study II. The figure clearly shows the overprediction of Glasgow outcome scale extended (GOSE) categories 1 3, and 8 in all models with the neural network (g-i) showing slightly more spread-out predictions. The rows correspond to the POLR (a-c), RF (d-f) and NN (g-i) models and the columns to Uppsala, Leuven and ProTECT III cohorts respectively. a) POLR in the Uppsala cohort. b) POLR in the Leuven cohort. c) POLR in the ProTECT III cohort. d) RF in the Uppsala cohort. e) RF in the Leuven cohort. f) RF in the ProTECT III cohort. g) NN in the Uppsala cohort. h) NN in the Leuven cohort. i) NN in the ProTECT III cohort.


a)
b)
c)
d)
e)
f)
Count
Observed GOS-E - Predicted GOS-E

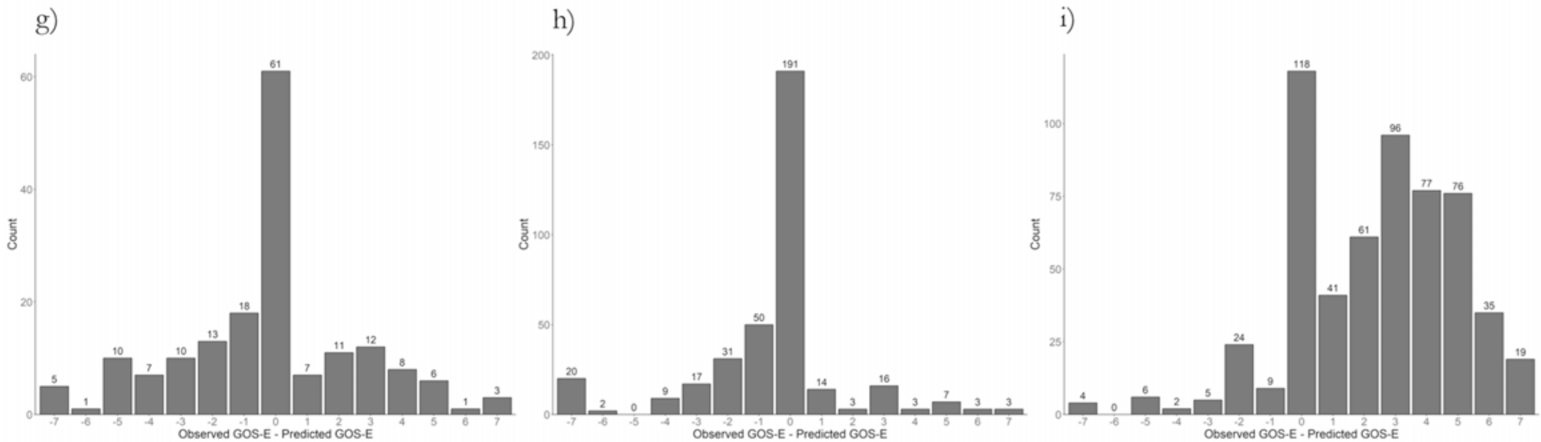
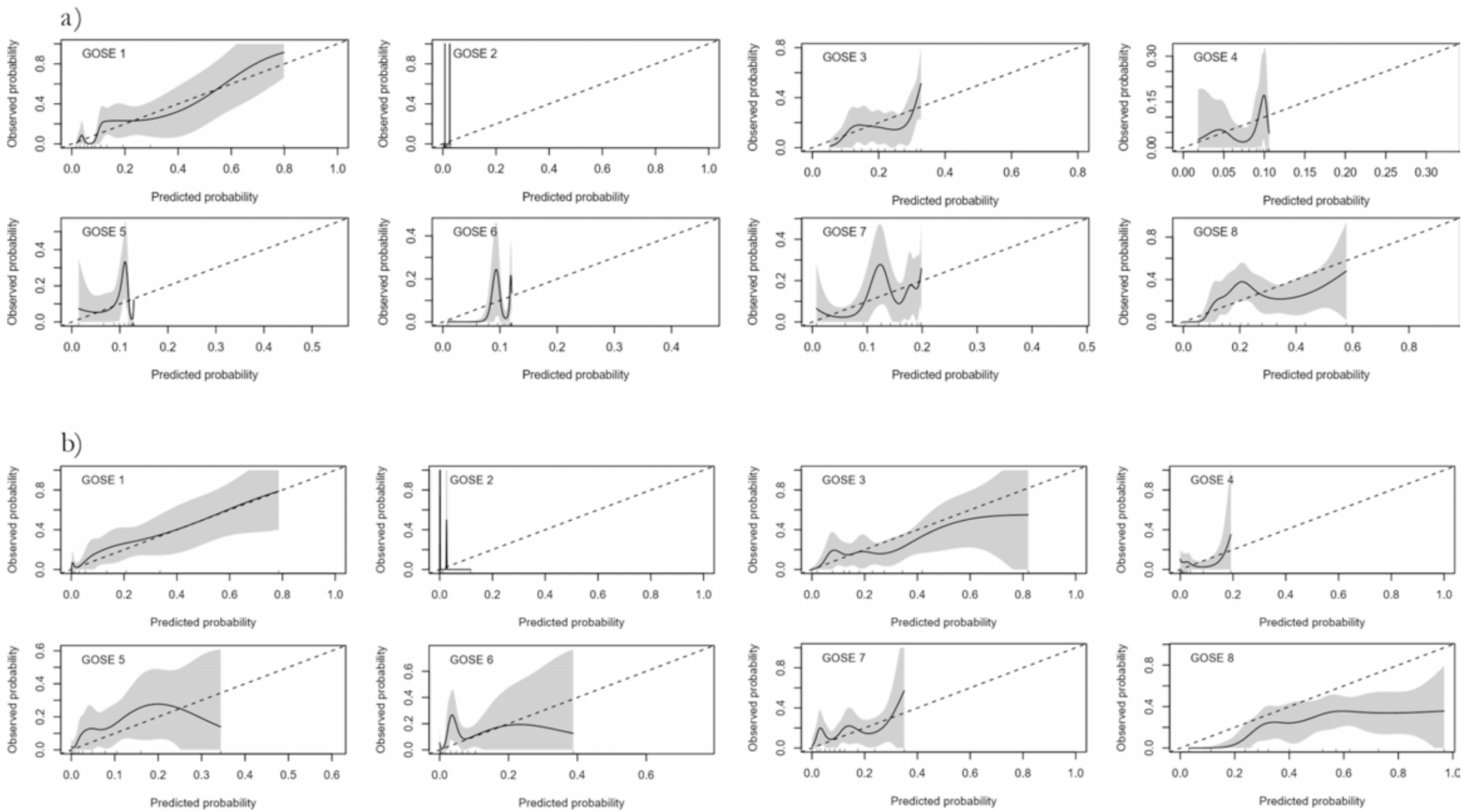


Figure 11. Error distribution graphs of the proportional odds logistic regression (POLR), random forest (RF) and neural network (NN) models in study II. The figure shows observed Glasgow outcome scale extended (GOSE) minus predicted GOSE and thus the balance of the models. All models slightly overestimated outcome in the Uppsala cohort (a, d and g), were fairly balanced in the Leuven cohort (b, e and h) and underestimated outcome in the ProTECT III cohort (c, f and i). a) POLR in the Uppsala cohort. b) POLR in the Leuven cohort. c) POLR in the ProTECT III cohort. d) RF in the Uppsala cohort. e) RF in the Leuven cohort. f) RF in the ProTECT III cohort. g) NN in the Uppsala cohort. h) NN in the Leuven cohort. i) NN in the ProTECT III cohort.


a)
b)
GOSE 1
GOSE 2
GOSE 3
GOSE 4
GOSE 5
GOSE 6
GOSE 7
GOSE 8
Observed probability
Predicted probability

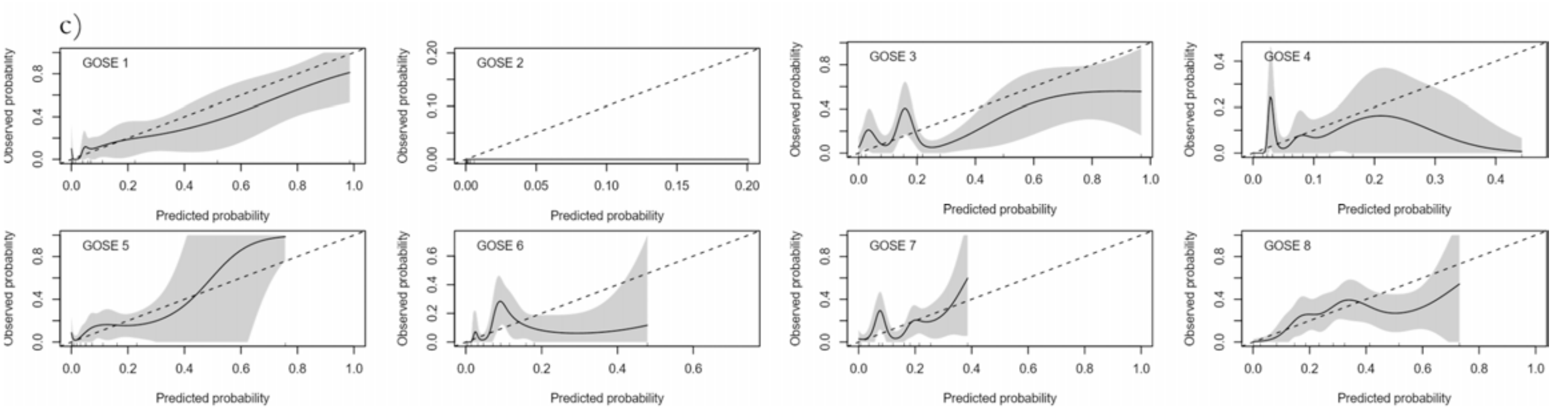


Figure 12. Calibration graphs of the proportional odds logistic regression (POLR), random forest (RF) and neural network (NN) models in the Uppsala cohort in study II. The solid line describes the performance of the models in relation to the observed probability for that category, with the grey area representing two standard errors. The dashed line represents an optimally calibrated model, where the area over the thick line represents underestimation, and the area under the thick line represents overestimation of the probability for a patient to end up in the respective category. Note that the y-axis differs in between some of the Glasgow outcome scale extended (GOSE) categories. a) The POLR models predicted probability is fairly calibrated in GOSE 1 and 8 categories but cuts short at around 0.1-0.2 in category 3-7 and practically never predicts GOSE 2. b) The RF models predicted probability is fairly calibrated in GOSE categories 1 and 3 while overpredicting GOSE 8 and under predicting GOSE 4-7 and giving a very low predicted probability for GOSE 2. c) Similarly, the NN model is fairly calibrated in GOSE 1, 3, 8 while adding better predicted probability for GOSE 4 and 5. The NN still underpredicts GOSE 2, 6 and 7.

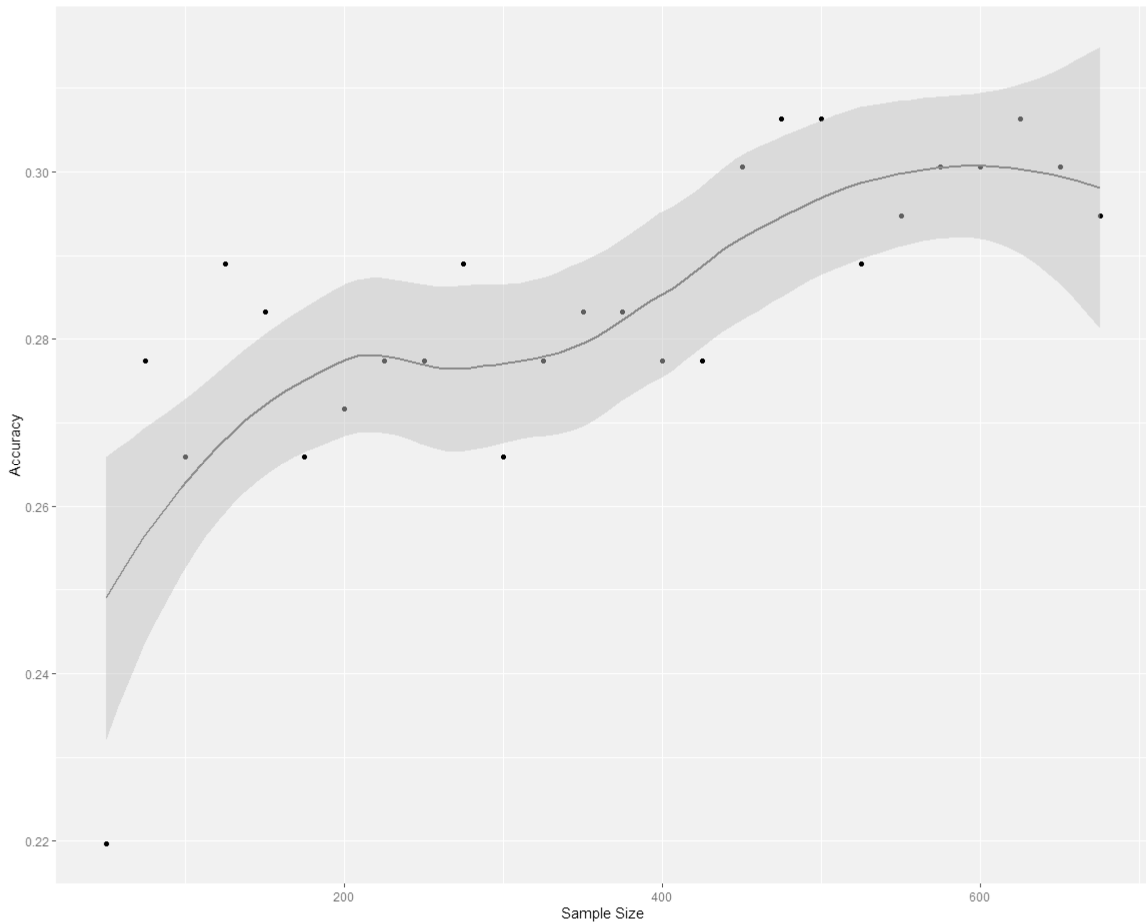


Figure 13. Simulation of increasing the training sample size in proportional odds logistic regression (POLR) models in study II. Sample size increased accuracy of the POLR models up to a plateau at 400 before increasing again and plateauing at 600

Study III

A summary of the exclusion criteria is displayed as a flowchart in figure 14. After exclusion of 10 cases with incomplete data, 278 AI-positive cases were evaluated by comparing the AI classification to the final report, interpreted by a resident in radiology. All cases with unclear or ambiguous findings (n = 45) were re-evaluated by a specialist in neuroradiology (D.F.) for a decisive assessment. Among the AI-positive cases, 10 (3.6 %) had epidural hematomas and 37 (13.3 %) had subarachnoid hemorrhages. 117 (42.1 %) had subdural and 65 (23.4%) intraparenchymal blood. This included postoperative bleedings and bleedings associated to tumors. When several types of bleeding were present, the most dominating finding decided the classification.

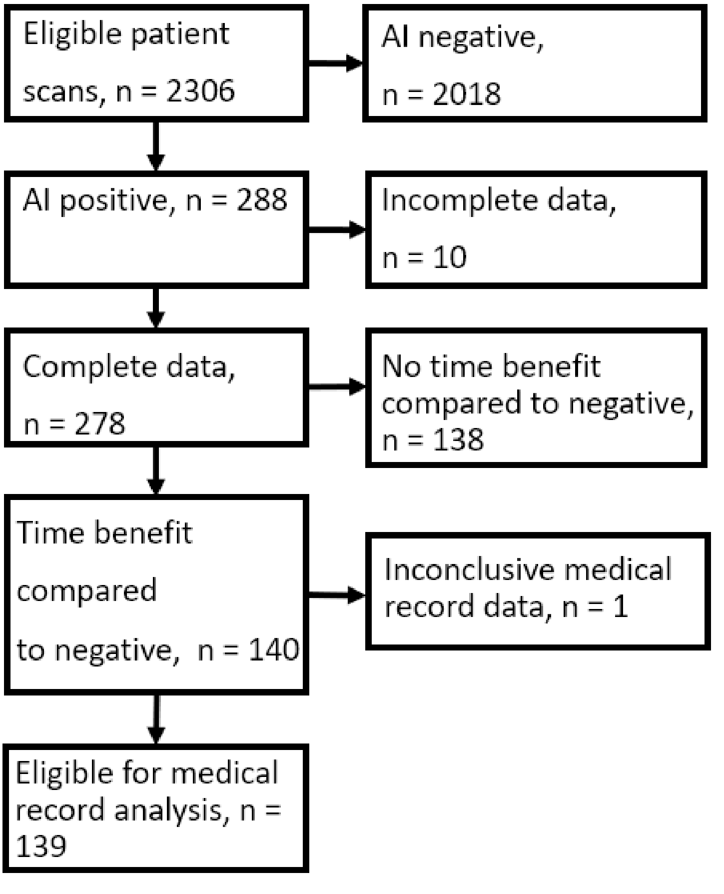


Figure 14. Flowchart of patient inclusion leading to 278 being eligible for positive predictive analysis and 139 being eligible for medical record analysis in study III.

In 49 (17.6 %) of the 278 AI-positive cases, no intracranial hemorrhage was found, leading to a positive predictive value (PPV) of 0.823 for detection of ICH. Sensitivity, specificity and negative predictive value cannot be assessed with the current study design.

Among the 49 false positive cases, many had other pathological findings that presumably triggered the artificially intelligent classification. Of these, 16 had intracranial tumors (without overt bleeding), four had infarcts and five were visually assessed as completely normal. The remaining 20 cases had some component (post operative remnants, calcifications, artefacts or abscess) that were assumed to be the source of the false positive classification. If tumor instead is considered a clinically relevant and thus a true positive finding for the AI classification, the PPV rises to 0.881.

In 3 of the AI-positive cases (1.1 %), the preliminary report described an intracranial bleeding (or suspicion thereof) that was later negated in the final report. These were thoroughly re-examined by a specialist in neuroradiology to evaluate if the primary radiologist (i.e., resident in radiology) was misled by the positive AI classification. In one case, stagnated contrast media from a previous intervention had been misinterpreted as blood by the AI and in the preliminary report (later disproved with virtual non contrast reconstructions). In two other cases, a large, demarcated infarct, and hyperdense metastasis, had both been respectively described as having "possible hemorrhagic components", which was considered plausible upon re-evaluation. Therefore, none of the false positive cases were suspected to have misled the primary radiologist.

Median time in minutes from scan to report are listed in Table 8. On the group level, none of the categories had a statistically significant difference (time benefit) between AI-positive and AI-negative examinations.

Table 8. Time benefits from AI-annotation in the different computer tomography priority categories. Median time in minutes for AI-positive/negative cases, on the group level for each of the six categories. Total n = 2306.

Time of day:	**On call**			**Office hours**		
Priority category:	Acute	Subacute	Routine	Acute	Subacute	Routine
Median time:	27.5/28.35	37.3/34.1	1421.9/1500.1	29.4/26.2	36.5/30.9	95.4/213.3
n =	112/997	58/215	9/199	20/204	58/139	17/278
p =	0.834	0.793	0.861	0.692	0.232	0.160

Time from scan to report was also examined on the individual level for each AI-positive case compared to the median time of negative scans in the corresponding category. The distribution of "time benefit from AI-annotation" in different categories is listed in Table 9.

Table 9. Display of time benefit windows in the different computer tomography priority categories. The number of patients with time benefit from AI-annotation in the three different priority categories on the individual level. Time benefit in minutes is stratified in three arbitrary cate-gories for presentational purposes.

Priority category:	**Total**	**Acute**	**Subacute**	**Routine**
Time benefit ≥30 min:	17	0	0	17
Time benefit 5-29 min:	99	49	49	1
Time benefit <5 min:	24	21	3	0
Sum of category:	140	70	52	18

After excluding one patient with inconclusive medical records, 139 patients with individual time benefit were individually screened (figure 1). The time benefit in these cases had a range from 0.05 to 624 minutes (median = 13.37, IQR = 12.91). Out of these 139 cases, 67 (48 %) had a previously known diagnosis. In 27 (19 %) patients the examination led to the on-call neurosurgeon being summoned, 17 (12 %) to elevation of the level of care and 8 (6 %) to emergency surgery (within 24 hours). Furthermore, only 17 of these 139 patients had a time benefit from AI-annotation of 30 minutes or more. All of these 17 were routine scans and were presumably rapidly assessed due to the conspicuous annotation in the work list.

Of the 17 patients with > 30 minutes of time benefit, none resulted in the on-call neurosurgeon being summoned, none led to elevation of level of care and none led to emergency surgery. Furthermore, out of these 17 patients, 10 were control scans of known diagnoses and 7 were newly diagnosed pathologies. Out of those 7, one was a newly diagnosed tumor, two were of small subdural hematomas, one was a post-surgical control, and three were false positives.

There was no significant time benefit from AI-annotation between the group that did not have emergency surgery compared to those who went through emergency surgery, $p = 0.4611$, or between the groups that had or did not have elevation of the level of care $p = 0.1886$ (figure 15). In one of the 8 patients that went through emergency surgery, it was not possible to extract the time to surgery from the medical records. Out of the remaining 7 patients, linear regression analysis showed no significant correlation between time benefit from AI-annotation compared to time to surgery ($p = 0.498$, $r^2 = 0.09619$), but there was a tendency of longer time to surgery in patients that had larger time benefit from AI-annotation.

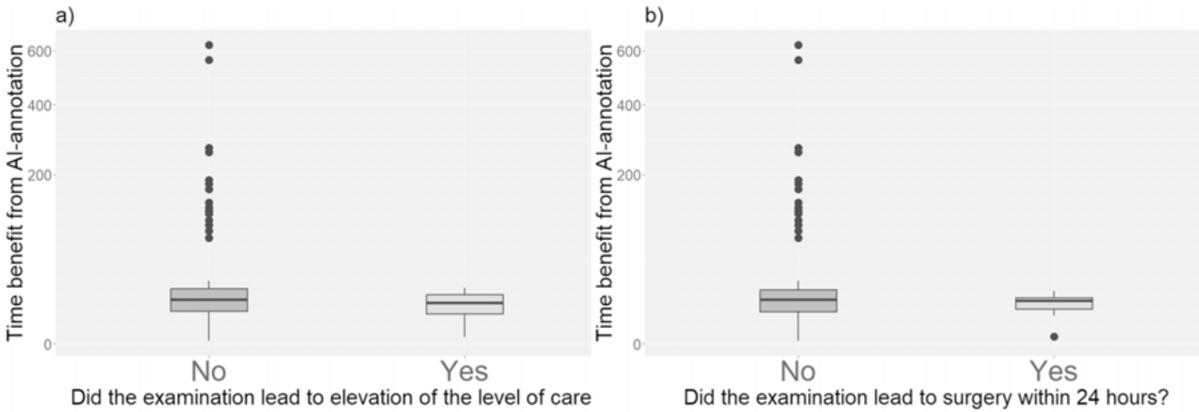


Figure 15. Time benefits from AI-annotation compared to clinical management of the patients in study III. Note that the y-scale is square root transformed to include the extreme outliers. a) Comparison between groups with and without elevation of the level of care. There was no significant time benefit between the group that had elevation of the level of care compared to the group who did not, $p = 0.1886$. b) Comparison between groups that did or did not go through emergency surgery. There was no significant time benefit between the group that had emergency surgery compared to the group who did not, $p = 0.4611$.

Study IV

A total of 1261 patients were considered for inclusion from the Uppsala TBI registry. After exclusion of patients with mild TBI (GCS 13-15, n=565), missing GOSE data (n=87), follow-up time >12 months (n=15) and unknown time of trauma (n=102), a total of 492 patients remained in the static cohort (Figure 1). The NICU static/dynamic cohort was derived by including only patients with available NICU data files containing valid timestamps. Patients were required to have recordings for at least one dynamic variable (ICP, MAP, CPP, or PRx) within the prespecified post-injury timepoint. After applying these criteria, the remaining seven groups were; 334 patients in the 24-hour cohort, 385 in the 48-hour cohort, 412 in the 72-hour cohort, 424 in the 96-hour cohort, 429 in the 120-hour cohort, 432 in the 144-hour cohort and 434 in the 168-hour cohort (Figure 16).

As described below, the best performing static/dynamic model was the XGBoost at 96-hour with both admission and NICU data. Therefore, demographics data will be presented for that cohort as well as the static cohort. Specifically, as seen in Table 1, in both the static patient cohort as well as in the 96-hour cohort the most common trauma mechanism was falling accidents (46% and 45% in the static and 96-hour cohorts, respectively), followed by vehicle accidents (29% and 30%, respectively). Mean age was 50 in the all admitted cohort and 49 in the 96-hour cohort, median was 52 in both cohorts. Trauma severity as measured by the full GCS in the static and 96-hour cohorts, was moderate (GCS 9-12) in 286 (58%) and 262 (62%) patients, respectively, and severe (GCS 3-8) in 206 (42%) and 162 (38%) patients, respectively. GCS-motor had similar distributions in both cohorts (Table 1) with the most common being GCS-motor 5 (45% and 47% in the static and 96-hour cohorts) followed by GCS-motor 4 (20% and 19% in the static and 96-hour cohorts respectively). The most prevalent main radiological finding was acute subdural hematoma (28% in the static and 27% in the 96-hour cohorts) followed by contusions and mixed. A notable proportion of patients had a missing Marshall score at admission (45% and 41% in the static and 96-hour cohorts, respectively), which was handled using imputation during modelling. Among patients with a known score, the most common Marshall score was 2 (28% and 30% in in the static and 96-hour cohorts) and 5 (18% and 20% at in the static and 96-hour cohorts). The most common GOSE outcomes were 3 (23% and 24% in the static and 96-hour cohorts) although followed closely by GOSE 1, 5, and 8. GOSE 2 outcome was much rarer at around 2% for both cohorts (Table 10).

The burden of secondary insults as measured by ICP, CPP and PRx generally increased over the monitoring time. Specifically, the duration of ICP > 20 increased from 211 minutes in the entire 24-hour cohort to 804 minutes at 48 hours, 1407 minutes at 72 hours, 1961 minutes at 96 hours, 2471 minutes at 120 hours, 2967 minutes at 144 hours and 3430 minutes at 168 hours. The burden of CPP <50 increased from 84 minutes in the entire 24-hour cohort to 396 minutes in the 96-hour cohort and 598 minutes in the 168-hour cohort. The burden of PRx, double and triple hit insults followed similar trends (full insult burden in supplement).

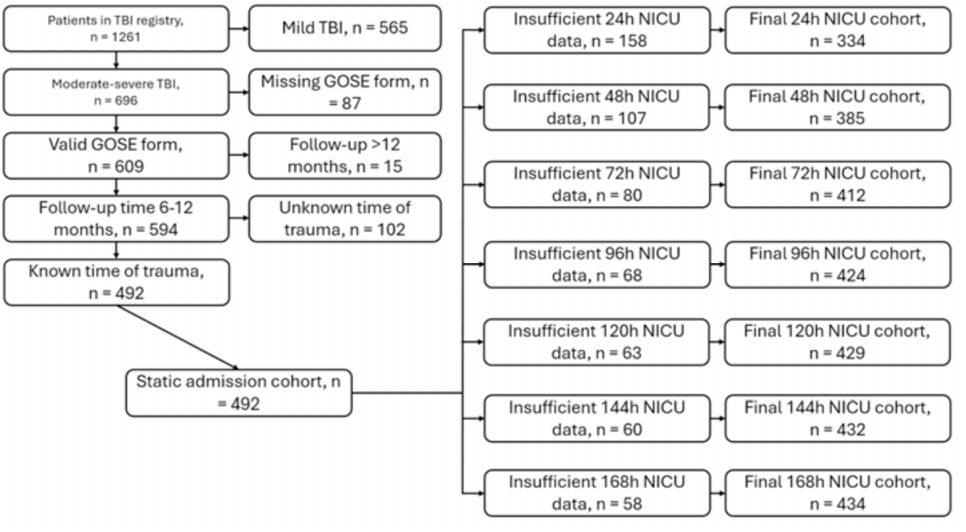


Figure 16. Flow diagram of patient inclusion/exclusion, resulting in n = 492 for the static cohort, n = 334 for the 24-hour cohort, n = 385 for the 48-hour cohort, n = 412 for the 72 hour cohort, n = 424 for the 96-hour cohort, n = 492 for the 120-hour cohort, n = 432 for the 144-hour cohort and n = 434 for the 168-hour cohort. TBI = traumatic brain injury, GOSE = Glasgow outcome scale extended, NICU = neurointensive care unit.

Table 10. Patient demographics of the static and 96-hour timeframe cohorts respectively. SD = standard deviation, IQR = interquartile range, GCS = Glasgow coma scale, ASDH = acute subdural hematoma, SAH = subarachnoid hemorrhage, EDH = epidural hematoma, DAI = diffuse axonal injury, GOSE = Glasgow outcome scale extended, CVD = cardiovascular disease, DM2 = diabetes mellitus type 2.

96-hour cohort	59 (13,9%)	8 (1,9%)	100 (23,6%)	37 (8,7%)	62 (14,6%)	44 (10,4%)	43 (10,1%)	71 (16,7%)							
Static	88 (17,9%)	10 (2%)	115 (23,4%)	39 (7,9%)	65 (13,2%)	49 (10%)	46 (9,3%)	80 (16,3%)							
GOSE	GOSE 1	GOSE 2	GOSE 3	GOSE 4	GOSE 5	GOSE 6	GOSE 7	GOSE 8							
96-hour cohort	175 (41.3%)	1 (0.2%)	126 (29.7%)	14 (3.3%)	9 (2.1%)	83 (19.6%)	16 (3.8%)								
Static	223 (45.3%)	1 (0.2%)	137 (27.8%)	17 (3.5%)	9 (1.8%)	87 (17.7%)	18 (3.7%)								
Marshall score	Marshall score: Missing/Unknown	Marshall score 1	Marshall score 2	Marshall score 3	Marshall score 4	Marshall score 5	Marshall score 6								
96-hour cohort	1 (0.2%)	115 (27.1%)	103 (24.3%)	39 (9.2%)	26 (6.1%)	31 (7.3%)	4 (0.9%)	94 (22.2%)	11 (2.6%)						
Static	3 (0.6%)	138 (28.0%)	117 (23.8%)	42 (8.5%)	29 (5.9%)	39 (7.9%)	4 (0.8%)	105 (21.3%)	6 (1.3%)						

Main admission radiological finding	Normal	ASDH	Contusion	Traumatic SAH	EDH	DAI	Impression fracture	Mixed	Other						
96-hour cohort	317 (74.8%)	107 (25.2%)	49.0 (19.4)	52 (32-65)	29 (6.8%)	24 (5.7%)	32 (7.5%)	81 (19.1%)	201 (47.4%)	57 (13.4%)	315 (74.3%)	41 (9.7%)	56 (13.2%)	109 (25.7%)	41 (9.7%)
Static	363 (73.8%)	129 (26.2%)	49.7 (19.7)	52 (32-66)	42 (8.5%)	32 (6.5%)	36 (7.3%)	97 (19.7%)	223 (45.3%)	62 (12.6%)	354 (72.0%)	53 (10.8%)	69 (14.0%)	128 (26.0%)	52 (10.6%)
Admission data	Male	Female	Mean age (SD)	Median age (IQR)	GCS-motor 1	GCS-motor 2	GCS-motor 3	GCS-motor 4	GCS-motor 5	GCS-motor 6	Both pupils reacting	One pupil reacting	No pupil reacting	CVD	DM2
96-hour cohort	190 (44.8%)	126 (29.7%)	15 (3.5%)	18 (4.2%)	18 (4.2%)	13 (3.1%)	44 (10.4%)								
Static	228 (46.3%)	142 (28.9%)	20 (4.1%)	20 (4.1%)	21 (4.3%)	13 (2.6%)	48 (9.8%)								
Trauma mechanism	Fall accident	Vehicle accident	Pedestrian hit by vehicle	Cyclist hit by vehicle	Assault	Sports accident	Other								

Model performance

In the selected static/dynamic cohort with physiological data, the inclusion of dynamic NICU features improved model performance for predicting 7-grade GOSE outcomes, particularly at later time points. This improvement was most pronounced for the XGBoost models (Table 2). For instance, in the 48-hour NICU cohort, the XGB model's quadratic weighted kappa (QWK) increased from 0.24 with admission features only to 0.39 with the addition of NICU features. Similarly, the OLR model's QWK improved from 0.38 to 0.43 at the same time point. In the static cohort, the quadratic weighted kappa (QWK) was higher than the static/dynamic cohorts, where the Ordinal Logistic Regression (OLR) performed a QWK of 0.51 and the XGBoost model of 0.53 (Table 11).

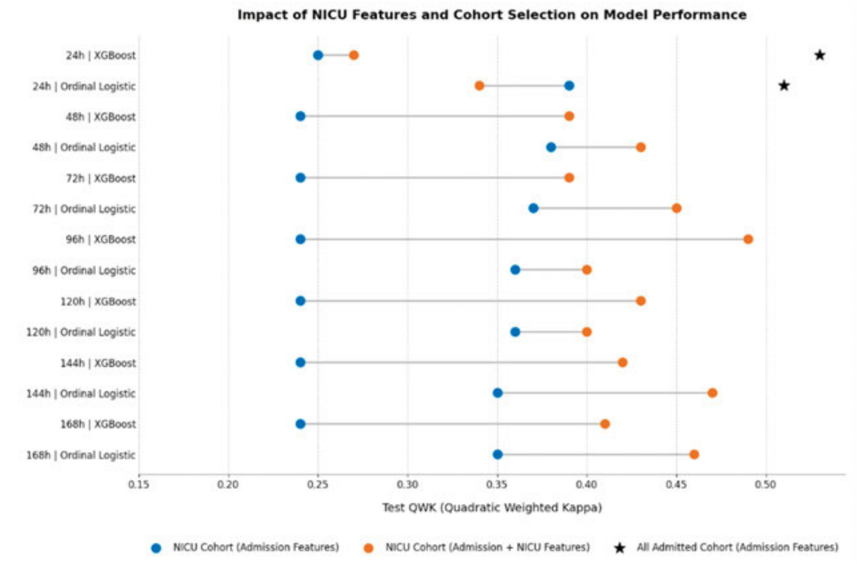
This trend of improved performance with the inclusion of dynamic data culminated in the best-performing model in the NICU cohorts: the 96-hour XGBoost model, which achieved a QWK of 0.49. This was an improvement over its admission-only counterpart, which had a QWK of 0.24. While this pattern was observed across most time points, the benefit of adding NICU features was generally more pronounced from 48 hours onwards (Figure 17). However, when the models were evaluated on a more challenging temporal split, which better reflects real-world deployment scenarios, overall performance was reduced in the XGBoost models while the OLR models were not as affected, indicating robustness to dataset shift. Similarly, in the temporal split, the inclusion of dynamic variables generally but not always improved performance. (Figure 17). The peak performance observed in the stratified analysis (QWK of 0.49 for the 96-hour XGBoost model) was not replicated. In the temporal split, and the best-performing model was instead the 120-hour Ordinal model with NICU features, which achieved a slightly lower QWK of 0.42 (Table 12).

In the stratified split, analysis of secondary binary endpoints similarly showed the best discrimination in the static cohort with an AUROC for mortality in the range of 0.91-0.92 for both models and an AUROC for favourable outcome of 0.79-0.80. In the selected static/dynamic cohort, we showed reduced discrimination for mortality (AUROC range: 0.75-0.88) and favourable outcome (AUROC range: 0.67-0.83). Unlike the primary ordinal outcome, the inclusion of NICU features did not consistently improve discrimination for these binary endpoints, with performance varying across different models and time points. The best-performing 96-hour XGBoost model achieved an AUROC of 0.80 for mortality and 0.78 for favourable outcome (Table 11).

In the stratified split, model calibration for dichotomized outcomes, as measured by Expected Calibration Error (ECE), was variable. For the NICU cohorts, ECE ranged from 0.04 to 0.10 for mortality and 0.07 to 0.18 for favourable outcome. The top-performing 96-hour XGBoost model demonstrated reasonable calibration, with an ECE of 0.08 for mortality and 0.10 for favourable outcome. Overall accuracy for the 7-grade outcome remained low across all models (Table 11). While all categories were represented in the predictions of the 96-hour XGBoost model, the model still showed some confusion between outcome categories (Figure 18). Furthermore, calibration slopes were superior in the edge categories (Dead/VS and Good-upper), compared to the middle categories showed poorer calibration (Figure 19). Furthermore, the category proportion plot shows that the model is overpredicting the middle categories of the GOSE scale and underpredicting the edge categories

Finally, SHAP analysis of the 96-hour XGBoost model showed that the top five features were time fraction of ICP < 10 in the time window hour 56-64, MAP standard deviation in the time window hour 64-72, age, and CPP mean in the time window hour 56-64, and time fraction of double hit insults with ICP>22 and PRx>0.3.

a)


Impact of NICU Features and Cohort Selection on Model Performance
24h | XGBoost
24h | Ordinal Logistic
48h | XGBoost
48h | Ordinal Logistic
72h | XGBoost
72h | Ordinal Logistic
96h | XGBoost
96h | Ordinal Logistic
120h | XGBoost
120h | Ordinal Logistic
144h | XGBoost
144h | Ordinal Logistic
168h | XGBoost
168h | Ordinal Logistic
0.15
0.20
0.25
0.30
0.35
0.40
0.45
0.50
Test QWK (Quadratic Weighted Kappa)
NICU Cohort (Admission Features)
NICU Cohort (Admission + NICU Features)
All Admitted Cohort (Admission Features)

b)

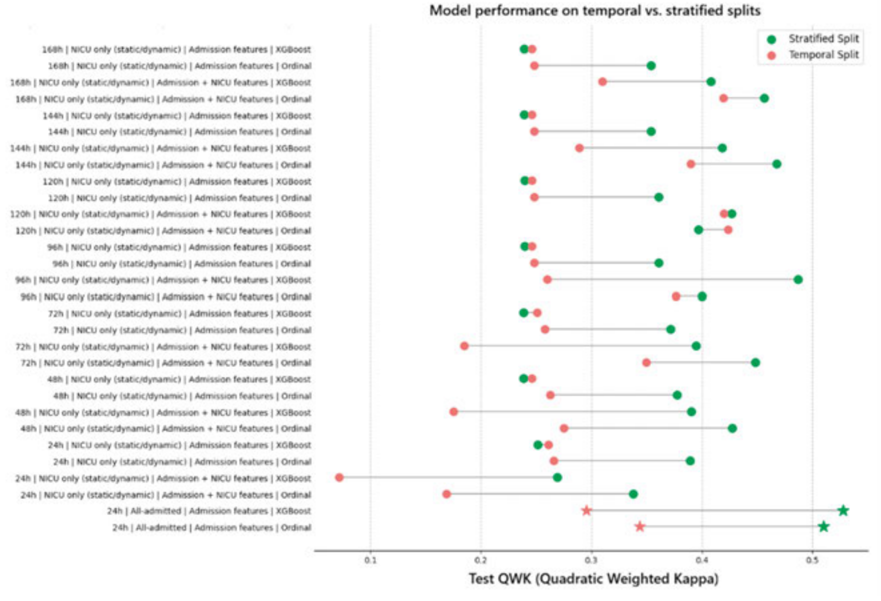

Model performance on temporal vs. stratified splits
Stratified Split
Temporal Split
168h | NICU only (static/dynamic) | Admission features | XGBoost
168h | NICU only (static/dynamic) | Admission features | Ordinal
168h | NICU only (static/dynamic) | Admission + NICU features | XGBoost
168h | NICU only (static/dynamic) | Admission + NICU features | Ordinal
144h | NICU only (static/dynamic) | Admission features | XGBoost
144h | NICU only (static/dynamic) | Admission features | Ordinal
144h | NICU only (static/dynamic) | Admission + NICU features | XGBoost
144h | NICU only (static/dynamic) | Admission + NICU features | Ordinal
120h | NICU only (static/dynamic) | Admission features | XGBoost
120h | NICU only (static/dynamic) | Admission features | Ordinal
120h | NICU only (static/dynamic) | Admission + NICU features | XGBoost
120h | NICU only (static/dynamic) | Admission + NICU features | Ordinal
96h | NICU only (static/dynamic) | Admission features | XGBoost
96h | NICU only (static/dynamic) | Admission features | Ordinal
96h | NICU only (static/dynamic) | Admission + NICU features | XGBoost
96h | NICU only (static/dynamic) | Admission + NICU features | Ordinal
72h | NICU only (static/dynamic) | Admission features | XGBoost
72h | NICU only (static/dynamic) | Admission features | Ordinal
72h | NICU only (static/dynamic) | Admission + NICU features | XGBoost
72h | NICU only (static/dynamic) | Admission + NICU features | Ordinal
48h | NICU only (static/dynamic) | Admission features | XGBoost
48h | NICU only (static/dynamic) | Admission features | Ordinal
48h | NICU only (static/dynamic) | Admission + NICU features | XGBoost
48h | NICU only (static/dynamic) | Admission + NICU features | Ordinal
24h | NICU only (static/dynamic) | Admission features | XGBoost
24h | NICU only (static/dynamic) | Admission features | Ordinal
24h | NICU only (static/dynamic) | Admission + NICU features | XGBoost
24h | NICU only (static/dynamic) | Admission + NICU features | Ordinal
24h | All-admitted | Admission features | XGBoost
24h | All-admitted | Admission features | Ordinal
0.1
0.2
0.3
0.4
0.5
Test QWK (Quadratic Weighted Kappa)

Figure 17. Barbell plots illustrating the two primary findings of the study. (a) Admission vs admission and physiological data from the neurointensive care unit (NICU) showing improved performance with the addition of physiological variables in all static/dynamic cohorts. Blue fill colour indicates models trained on admission features only and orange colour indicates models trained on admission and NICU features. (b) Showing temporal versus stratified split showing absence of the improved performance of the added physiological variables in the XGBoost models but in the ordinal logistic regression models in the 72-, 96-, 120-, 144- and 168-hour cohorts. Circles represent models trained on static/dynamic cohorts while the stars represent the models trained on the full 'static' cohort. Green fill colour indicates stratified split cross validation, and pink colour indicates temporal split validation. Ordinal/ordinal logistic = ordinal logistic regression, NICU = neurointensive care unit.

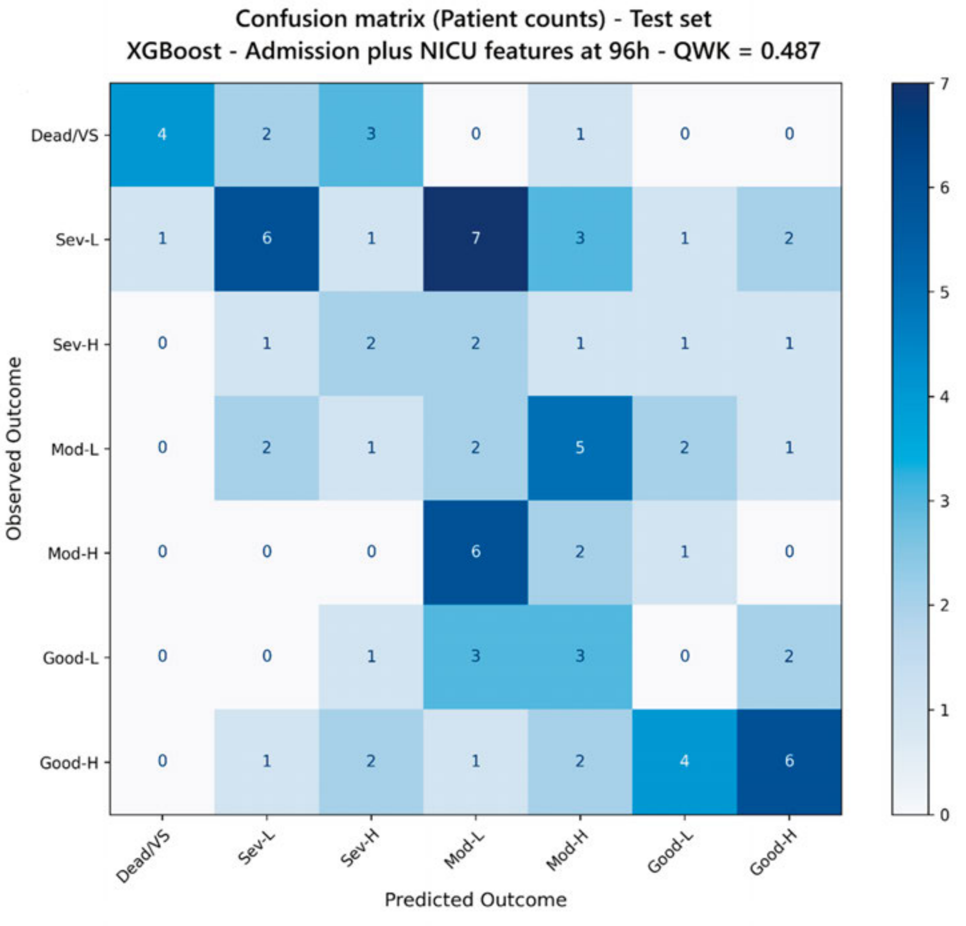


Figure 18. Confusion matrix of actual counts of predictions of the best performing 96-hour XGBoost model with admission and physiological features. While all categories are represented there are still misclassifications of all classes. Blue colour represents more patients. VS = vegetative state. Sev-L = lower severe disability. Sev-H = Upper severe disability, Mod-L=lower moderate disability, Mod-H=upper moderate disability, Good-L=lower good recovery and Good-H=upper good recovery. NICU = neurointensive care unit.

Table 11. Performance summary of XGBoost and Ordinal Logistic Regression models using the stratified split strategy. Results are shown for each daily timeframe up to one week post-trauma comparing the static to the static/dynamic cohorts in which models were trained either with only admission data or admission and physiological data. CV = cross validation. QWK = Quadratic weighted kappa. AW1 = accuracy within one category. AW2 = accuracy within two categories. AUROC-m = area under the receiver operator curve for mortality. AUROC-f = area under the receiver operator curve for favourable outcome. ECE = expected calibration error. XGB = XGBoost. OLR = Ordinal logistic regression.

Timepoint	Cohort	Feature set	Model	CV score	Test QWK	Accuracy	AW1	AW2	AUROC-m	AUROC-f	ECE-m	Calibration slope-m	ECE-m	Calibration slope-f	Train n=	Test n=	Features n=
Admission	All-admitted	Admission only	XGB	0.35	0.53	0.22	0.55	0.82	0.92	0.79	0.05	1.49	0.07	1.25	393	99	43
Admission	All-admitted	Admission only	OLR	0.41	0.51	0.16	0.54	0.82	0.91	0.8	0.09	2.08	0.09	1.38	393	99	43
24	ICU Head-to-head	Admission only	XGB	0.35	0.25	0.08	0.44	0.75	0.84	0.79	0.06	1.27	0.18	2.51	263	71	43
24	ICU Head-to-head	Admission only	OLR	0.4	0.39	0.15	0.49	0.82	0.84	0.73	0.08	1.47	0.09	1.19	263	71	43
24	ICU Head-to-head	Admission and ICU	XGB	0.34	0.27	0.2	0.49	0.69	0.84	0.7	0.1	0.68	0.12	0.5	263	71	95
24	ICU Head-to-head	Admission and ICU	OLR	0.43	0.34	0.15	0.52	0.69	0.81	0.67	0.07	1.02	0.11	0.86	263	71	95
48	ICU Head-to-head	Admission only	XGB	0.35	0.24	0.1	0.44	0.74	0.83	0.78	0.06	1.2	0.17	2.41	308	77	43
48	ICU Head-to-head	Admission only	OLR	0.4	0.38	0.14	0.48	0.82	0.85	0.71	0.07	1.56	0.07	1.14	308	77	43
48	ICU Head-to-head	Admission and ICU	XGB	0.36	0.39	0.21	0.49	0.77	0.84	0.75	0.1	0.7	0.11	0.82	308	77	95
48	ICU Head-to-head	Admission and ICU	OLR	0.37	0.43	0.16	0.51	0.82	0.87	0.7	0.05	0.98	0.14	0.88	308	77	95
72	ICU Head-to-head	Admission only	XGB	0.35	0.24	0.1	0.43	0.72	0.85	0.79	0.05	1.31	0.13	2.49	330	82	43
72	ICU Head-to-head	Admission only	OLR	0.4	0.37	0.13	0.46	0.8	0.81	0.73	0.07	1.37	0.09	1.22	330	82	43
72	ICU Head-to-head	Admission and ICU	XGB	0.35	0.39	0.16	0.49	0.79	0.83	0.83	0.05	1.34	0.15	3.07	330	82	95
72	ICU Head-to-head	Admission and ICU	OLR	0.36	0.45	0.17	0.44	0.78	0.88	0.77	0.07	1.17	0.11	1.27	330	82	95
96	ICU Head-to-head	Admission only	XGB	0.35	0.24	0.12	0.45	0.73	0.86	0.78	0.06	1.33	0.14	2.41	338	86	43
96	ICU Head-to-head	Admission only	OLR	0.4	0.36	0.15	0.48	0.8	0.81	0.72	0.07	1.39	0.11	1.2	338	86	43
96	ICU Head-to-head	Admission and ICU	XGB	0.44	0.49	0.26	0.59	0.83	0.8	0.78	0.08	0.72	0.1	1.05	338	86	95
96	ICU Head-to-head	Admission and ICU	OLR	0.39	0.4	0.21	0.48	0.72	0.75	0.74	0.08	0.48	0.14	0.54	338	86	95
120	ICU Head-to-head	Admission only	XGB	0.35	0.24	0.12	0.45	0.73	0.86	0.78	0.06	1.33	0.14	2.41	343	86	43
120	ICU Head-to-head	Admission only	OLR	0.4	0.36	0.15	0.48	0.8	0.81	0.72	0.07	1.39	0.11	1.2	343	86	43
120	ICU Head-to-head	Admission and ICU	XGB	0.45	0.43	0.17	0.52	0.85	0.81	0.78	0.08	0.7	0.11	1.24	343	86	95
120	ICU Head-to-head	Admission and ICU	OLR	0.4	0.4	0.28	0.51	0.7	0.76	0.72	0.06	0.49	0.15	0.42	343	86	95
144	ICU Head-to-head	Admission only	XGB	0.35	0.24	0.11	0.45	0.74	0.86	0.78	0.06	1.34	0.13	2.34	344	88	43
144	ICU Head-to-head	Admission only	OLR	0.4	0.35	0.15	0.48	0.8	0.8	0.71	0.07	1.3	0.08	1.07	344	88	43
144	ICU Head-to-head	Admission and ICU	XGB	0.46	0.42	0.22	0.57	0.76	0.84	0.83	0.06	0.85	0.15	1.64	344	88	95
144	ICU Head-to-head	Admission and ICU	OLR	0.36	0.47	0.16	0.55	0.78	0.85	0.81	0.04	1.03	0.1	1.23	344	88	95
168	ICU Head-to-head	Admission only	XGB	0.35	0.24	0.11	0.45	0.74	0.86	0.78	0.06	1.34	0.13	2.34	346	88	43
168	ICU Head-to-head	Admission only	OLR	0.4	0.35	0.15	0.48	0.8	0.8	0.71	0.07	1.3	0.08	1.07	346	88	43
168	ICU Head-to-head	Admission and ICU	XGB	0.43	0.41	0.26	0.58	0.81	0.82	0.8	0.09	0.75	0.12	1.15	346	88	95
168	ICU Head-to-head	Admission and ICU	OLR	0.34	0.46	0.17	0.55	0.8	0.85	0.78	0.1	0.96	0.11	0.99	346	88	95

Table 12. Performance summary of XGBoost and Ordinal Logistic Regression models using the temporal split strategy. Results are shown for each daily timeframe up to one week post-trauma comparing the static to the static/dynamic cohorts in which models were trained either with only admission data or admission and physiological data. CV = cross validation. QWK = Quadratic weighted kappa. AW1 = accuracy within one category. AW2 = accuracy within two categories. AUROC-m = area under the receiver operator curve for mortality. AUROC-f = area under the receiver operator curve for favourable outcome. ECE = expected calibration error. XGB = XGBoost. OLR = Ordinal logistic regression.

Timepoint	Cohort	Feature set	Model	CV score	Test QWK	Accuracy	AW1	AW2	AUROC-m	AUROC-f	ECE-m	Calibration slope-m	ECE-m	Calibration slope-f	Train n=	Test n=	Features n=
Admission	All-admitted	Admission only	XGB	0.42	0.3	0.27	0.52	0.68	0.88	0.71	0.14	0.72	0.18	0.51	393	99	43
Admission	All-admitted	Admission only	OLR	0.44	0.34	0.23	0.55	0.76	0.88	0.7	0.14	1.33	0.19	0.7	393	99	43
24	ICU Head-to-head	Admission only	XGB	0.35	0.26	0.17	0.51	0.72	0.87	0.68	0.17	0.7	0.16	0.57	256	78	42
24	ICU Head-to-head	Admission only	OLR	0.42	0.27	0.22	0.49	0.76	0.97	0.68	0.17	1.59	0.16	0.46	256	78	42
24	ICU Head-to-head	Admission and ICU	XGB	0.41	0.07	0.12	0.4	0.63	0.93	0.54	0.15	1.01	0.25	0.1	256	78	94
24	ICU Head-to-head	Admission and ICU	OLR	0.38	0.17	0.21	0.44	0.68	0.99	0.59	0.17	2.08	0.27	0.15	256	78	94
48	ICU Head-to-head	Admission only	XGB	0.35	0.25	0.17	0.5	0.71	0.88	0.67	0.17	0.72	0.17	0.57	303	82	43
48	ICU Head-to-head	Admission only	OLR	0.42	0.26	0.21	0.49	0.76	0.98	0.68	0.17	1.61	0.13	0.48	303	82	43
48	ICU Head-to-head	Admission and ICU	XGB	0.36	0.18	0.17	0.52	0.74	0.76	0.68	0.14	0.68	0.15	0.56	303	82	95
48	ICU Head-to-head	Admission and ICU	OLR	0.32	0.28	0.22	0.55	0.79	0.93	0.67	0.17	2.25	0.11	0.55	303	82	95
72	ICU Head-to-head	Admission only	XGB	0.35	0.25	0.18	0.5	0.71	0.88	0.67	0.17	0.72	0.14	0.58	328	84	43
72	ICU Head-to-head	Admission only	OLR	0.42	0.26	0.2	0.48	0.75	0.98	0.67	0.17	1.63	0.14	0.45	328	84	43
72	ICU Head-to-head	Admission and ICU	XGB	0.39	0.18	0.21	0.51	0.7	0.86	0.67	0.1	0.97	0.15	0.62	328	84	95
72	ICU Head-to-head	Admission and ICU	OLR	0.36	0.35	0.23	0.51	0.7	0.93	0.71	0.17	0.83	0.17	0.41	328	84	95
96	ICU Head-to-head	Admission only	XGB	0.35	0.25	0.19	0.5	0.72	0.88	0.66	0.16	0.73	0.17	0.56	338	86	43
96	ICU Head-to-head	Admission only	OLR	0.42	0.25	0.2	0.48	0.74	0.98	0.67	0.17	1.64	0.15	0.44	338	86	43
96	ICU Head-to-head	Admission and ICU	XGB	0.31	0.26	0.21	0.52	0.74	0.8	0.72	0.11	0.69	0.14	0.81	338	86	95
96	ICU Head-to-head	Admission and ICU	OLR	0.35	0.38	0.2	0.53	0.74	0.97	0.72	0.16	1.16	0.18	0.38	338	86	95
120	ICU Head-to-head	Admission only	XGB	0.35	0.25	0.19	0.5	0.72	0.88	0.66	0.16	0.73	0.17	0.56	343	86	43
120	ICU Head-to-head	Admission only	OLR	0.42	0.25	0.2	0.48	0.74	0.98	0.67	0.17	1.64	0.15	0.44	343	86	43
120	ICU Head-to-head	Admission and ICU	XGB	0.34	0.42	0.29	0.58	0.8	0.7	0.76	0.12	0.83	0.15	1.08	343	86	95
120	ICU Head-to-head	Admission and ICU	OLR	0.35	0.42	0.23	0.56	0.76	0.9	0.73	0.19	1	0.17	0.39	343	86	95
144	ICU Head-to-head	Admission only	XGB	0.35	0.25	0.19	0.5	0.72	0.88	0.66	0.16	0.73	0.17	0.56	346	86	43
144	ICU Head-to-head	Admission only	OLR	0.42	0.25	0.2	0.48	0.74	0.98	0.67	0.17	1.64	0.15	0.44	346	86	43
144	ICU Head-to-head	Admission and ICU	XGB	0.37	0.29	0.27	0.51	0.76	0.62	0.67	0.13	0.5	0.13	0.56	346	86	95
144	ICU Head-to-head	Admission and ICU	OLR	0.31	0.39	0.21	0.52	0.78	0.95	0.73	0.18	1.13	0.18	0.46	346	86	95
168	ICU Head-to-head	Admission only	XGB	0.35	0.25	0.19	0.5	0.72	0.88	0.66	0.16	0.73	0.17	0.56	348	86	43
168	ICU Head-to-head	Admission only	OLR	0.42	0.25	0.2	0.48	0.74	0.98	0.67	0.17	1.64	0.15	0.44	348	86	43
168	ICU Head-to-head	Admission and ICU	XGB	0.35	0.31	0.23	0.59	0.79	0.61	0.68	0.13	0.68	0.07	0.8	348	86	95
168	ICU Head-to-head	Admission and ICU	OLR	0.39	0.42	0.21	0.56	0.83	0.95	0.75	0.14	1.11	0.12	0.84	348	86	95

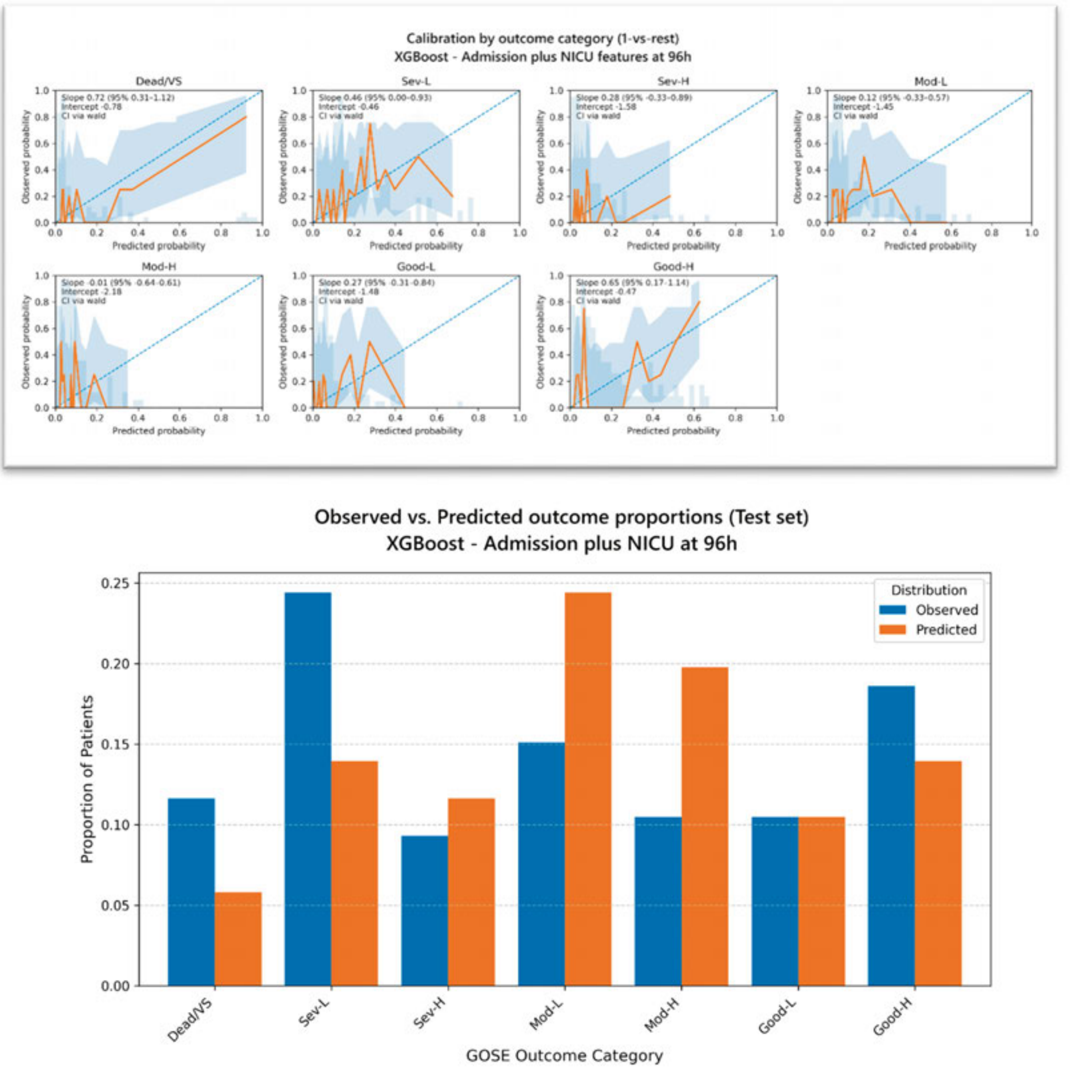


Figure 19. a) Calibration slopes and b) category proportions plot of the best performing model (XGBoost at 96 hours using both admission and physiological variables). Calibration was best for the edge categories (Dead/VS and Good-upper), while the middle categories showed poorer calibration. Furthermore, while all categories are predicted, there is clear preference of the middle categories of the scale and underprediction of the edge categories. The ideal calibration slope value is 1, a value of <1 means that the model is over-confident, and >1 that the model is under-confident. VS = vegetative state. Sev-L = lower severe disability. Sev-H = Upper severe disability, Mod-L=lower moderate disability, Mod-H=upper moderate disability, Good-L=lower good recovery and Good-H=upper good recovery. NICU = neurointensive care unit.

Discussion

Study I

In study I we validated the IMPACT prognostic model in a Swedish cohort in order to understand the applicability of the IMPACT model in Swedish patients with severe TBI. The IMPACT model is a well-established model for predicting outcome after severe-to-moderate traumatic brain injury and in our cohort all three components of the model performed in accordance with previous publications in predicting mortality and unfavorable outcome. Interestingly, there was no apparent difference between the core, CT and laboratory models and the best AUC was 0.89. In order to assess the full spectrum of GOSE we investigated the relationship between actual outcome in the full-scale GOSE compared to predicted risk. The predicted mortality was overestimated for GOSE groups 2-8, but in GOSE 1 group the model produced a greater range of predicted risk with more chance of survival. There could be several possible reasons for this. Firstly, there are significant differences between the original IMPACT cohort and ours. Sweden as a country is relatively sparsely populated and Uppsala University Hospital has a geographically large uptake region. The IMPACT model excludes all patients with an initial GCS of 12 and better, and in our cohort the patients were included depending on GCS upon arrival at UUH NICU as opposed to GCS at arrival to the initial local clinic, which could have been hours to days earlier. As most of the patients in our data came from relocated patients, there is an expected recurring delay in GCS evaluation. However, there is no certain information as to whether the patients generally deteriorated or improved during this delay, and whether this potential change in evaluated GCS significantly affected the predicted risk of the patients. Nevertheless, the laboratory data from these relocated patients was also obtained upon arrival to Uppsala University hospital and since the AUC value for these patients did not differ in the core and CT module, this could hint that this delay did not cause significant disturbance in IMPACT parameters or prognostic calculation. A second reason for the predicted underestimation of mortality in GOSE 1 group could spring from the fact that the IMPACT models were partially developed using patient data from drug studies in which the exclusion criteria could differ from ours. It is conceivable that patients that were included in this study would have been deemed non-treatable and excluded from the drug studies that IMPACT was based upon. Finally, it is worth mentioning that 6-month mortality depends on several care-giving instances including pre-hospital care, treatment in the intensive care unit, as well as rehabilitation care, all of which could differ from the original IMPACT cohort.

Looking at the predicted unfavorable outcome, it was correctly predicted to be low in GOSE groups 5-8 but the model underestimates risk in the GOSE 3-4 groups, and more correctly predicts an unfavorable outcome in GOSE groups 1-2. This means that the included parameters of the model more correctly predict if the outcome is death or a vegetative state but has poorer discriminative ability into placement into the severe and moderate disability groups. However, it should be noted that a large group of patients end up in the range of 40-60% risk of mortality or unfavorable outcome in particular in the core model, which in reality don't contribute to any outcome prediction in real practice since it is almost like "tossing a coin" for these patients.

It would be highly valuable to identify these patients where IMPACT calculator shows poor performance and if they share some common elements such as poor autoregulation and/or genetic vulnerability. Interestingly all three modules produced similar results in predicting both mortality and unfavorable outcome. In line with previous studies the CT and laboratory models slightly increased AUC but this was not the case for unfavorable outcome.

One must be cautious in applying prediction models intended for large cohorts in an individual patient setting. The goal of prediction models might be evaluating the outcome of different cohorts and the care of different centers, but they could also provide insight into whether individualized interventions could improve outcome. A prediction model could also identify patients with unexplainable outlier outcomes that could be of interest for further examination, for example to find out if there is an unknown reason some patients estimated to have a poor outcome fare much better than predicted.

Furthermore, prognostic information could advance our understanding of the pathophysiology of TBI. By adding novel parameters to the prognostic calculators, such as genetics, and analyzing whether this information impacts the calibration and discrimination of the calculator it could be possible to analyze the relevance of other pathological and physiological findings and their impact on the prognosis of TBI. This would also mean that the parameters of a prognostic calculator are dynamic, and that future novelties and discoveries of the pathophysiology of TBI could impact the calculator and vice versa. In the future, this could lead to more individualized healthcare, where novel parameters could impact the clinical management and treatment of patients.

Limitations

The primary limitations of this study were data completeness and cohort timing. A significant portion of laboratory data was lost due to a change in the medical records system. Furthermore, because 80% of the cohort consisted of patients transferred from other hospitals, the Glasgow Coma Scale (GCS) scores were recorded upon arrival at the tertiary center rather than at the initial time of injury; this delay allows for clinical fluctuation that may conflict with the original IMPACT model's design.

Study II

In study II we have explored different machine learning models' ability to predict full spectrum 8-scale GOSE outcome using the admission IMPACT variables from a Swedish TBI cohort. We were also able to perform external validation in two different TBI cohorts, one from Leuven, Belgium, and the other from the ProTECT III study which is a US based multicenter study. We found that a proportional ordinal logistic regression, a random forest and a neural network model perform at a similar accuracy value around 0.3-0.35 in internal data, and that all models actually perform better in external validation in the Leuven cohort but worse in the ProTECT III cohort.

TBI is a burdensome disease with oftentimes severe and lifelong implications for the patient, the family, and for society. Better prognostication is vital to further characterize the disease and to provide better care for these patients. [24] Although there are previous studies with GOSE as the dependent variable, for example by Cremer et al. [67], the outcome have always been collapsed into subsets of 2-3 ordinal levels for statistical reasons. An advanced prognostic model could assist in both evaluating results from clinical trials as well as in providing exclusion criteria for patients that are either deemed so severely injured that they are beyond rescue or rather on track towards a full recovery with current standard of care. In both cases it makes sense to exclude these patients from a clinical trial as they otherwise potentially obscure treatment effects. Furthermore, as TBI is a very heterogeneous disease there is substantial variability in patient outcomes which makes it difficult to detect minor improvements from drug treatment which then requires very large study populations. In fact, adjusting for baseline characteristics using prognostic models have been found to increase statistical power and reduce the required sample size by 25-30%. [13]

Outcome prediction models after TBI has been extensively studied, and we have attempted to set up the evaluation of our models in accordance with what has been previously recommended. These recommendations include well defined and justified variables, the model should have clear evaluation of discrimination and validation, and that the models should be externally validated on cohorts that differ in time or place as to exclude over-fitting. The models should preferably be represented in a user-friendly fashion to be understandable and clinically available. [18]

Variables identified in the IMPACT study have been analyzed for their prognostic power and validated in many different cohorts. [24] Age, injury severity and computer tomography findings were found to have the highest R^2 value. Furthermore, Maas et al. found high univariate odds ratios for age, low GCS-motor, pupil reactivity and CT-classification which is the main reason they were chosen for this study. [24] In our previous study we did not find that the laboratory variables of IMPACT (blood glucose and hemoglobin) increased the prognostic predictive power and thus excluded these from current study. [68]

Evaluation of our models

In study II, random forest and neural network performing slightly better accuracy in both internal and external data. All models also produce similar AUROC values in the range of 0.75-0.9 in internal and Leuven data which is comparable to previous studies, where several external validations studies of IMPACT have previously been made resulting in an AUROC around 0.6-0.9[21-23]. The ProTECT III data proved harder to predict for these models developed in a single Swedish center which is clear by the lower accuracy and lower AUROC values of the models.

Specifically, comparing the accuracy in the Leuven cohort to the ProTECT III cohort we found that the accuracy of the models was in the range of 0.44-0.52 and 0.18-0.21 respectively. Basically, the models found it easier to predict that the Leuven patients would fare well but grossly underestimated the outcome of the ProTECT III patients. An explanation for this is the fact that the Leuven cohort had a large majority of milder TBI compared to the ProTECT III cohort. The ProTECT III cohort had a higher number of abnormal pupillary reactions and generally worse GCS-motor scores. It could also be because of different trauma mechanisms or treatment regimens between the cohorts and the training data.

Further examination of the models reveals that they tend to favor the most common category in the training data, especially the POLR model, which results in them basically only predicting GOSE 1, 3 and 8. This results in them also having low TCPD and MCPD. Interestingly, the more complex models RF and NN made some more finely granulated predictions meaning that they might have picked up some interactions in the data that the POLR model could not. Although the low accuracy and overprediction of some categories is suboptimal, it is perhaps understandable as other factors after admission should contribute to specifying patient outcome. It could be that the problem of predicting full spectrum GOSE outcome is too much to ask from these models when using only admission variables. The fact that the training dataset was imbalanced might also affect the outcome of model fitting, and it might be that larger patient cohorts would even out the categories.

Although the calibration plots (Figure 12) clearly display a lack of predicted probability in several GOSE categories for all models, the error distribution graphs and the relatively low MD and SSD show that they are, however, overall fairly balanced at least in the internal and Leuven datasets.

Defining the optimal prediction model

TBI is a very heterogenous disease, and high accuracy in predicting the exactly correct category might be difficult using only admission data. Our models perform at best 35% accuracy in internal validation, and that is probably due to the complexity of the problem. Since high accuracy on the GOSE scale might be too great of a demand of the prediction models, it could be necessary to evaluate models trained on admission data using AW1 and AW2 instead.

The interpretability of machine learning models is important and higher interpretability facilitates understanding why and how the prediction has been made. Logistic regression generally has high interpretability, while neural networks and random forests generally have low interpretability. An outcome prediction model or calculator should also be user-friendly which will facilitate the clinical implementation.

To summarize, the optimal model depends on the purpose of the model. For the evaluation of the care of TBI patients and the improvement of clinical studies, the model should have high accuracy and be well-calibrated. For practical reasons, the independent variables should also be relatively easily available and not require too much data imputation.

Strengths and limitations

A strong advantage of our study is that we could perform external validation in a TBI cohort from Leuven and use the results from the multicenter ProTECT III study, thus including result from many different centers.

The major limitation was the fact that the models were developed in a single center and while they had adequate results in external validation in Leuven which could suggest potential generalizability, they had poor results in the ProTECT III data. Furthermore, both model training and validation was conducted in cohorts from high-income environments and further studies are needed especially in developing countries.

Furthermore, since the IMPACT variables are generally easily available, we had no issues with data availability and chose not to impute any data. This however means that the models presented here are unable to handle any missing data. In future settings, data selection and imputation might be necessary for complex models to be practical and usable. It is important to note that data imputation could lead to important data being missed, and that choosing variables that are easily available are important when developing a prediction model.

In this setting, a simulation of the accuracy of POLR models with an increasing number in the sample size increased accuracy until it plateaued at around 600 patients. This could indicate that a larger number of patients might not increase accuracy of the POLR models further (Figure 13). It is however probable that adding more variables to prediction models would require a higher number of patients, and cohorts with more evenly distributed GOSE outcome should be relevant in improving prediction of the middle categories. Adding more variables should also lead to improved models. Previous studies have shown that the IMPACT variables only provide an R^2 approaching 0.35 meaning that most contributing factors are not included. [24] In this study we chose to include the IMPACT admission variables as a starting point, but other variables could easily be added depending on the clinical setting wanted to be evaluated.

Great care should be taken when applying prognostic models that were developed from large patient cohorts into the clinical settings and when making decisions in the care of individual patients. Nevertheless, the fact is that clinicians apply their own prognostic evaluation in everyday care and tools are needed to standardize this prediction.

Study III

Study III is a real-world study of the actual clinical benefit from using an AI application for ICH detection in CT scans of the brain from 2306 patients. The application alerted the radiologist on positive findings by a visual cue in the work list, facilitating rapid assessment and potential time benefit. Clinical benefit was examined using data extraction from medical charts. There are three main findings and several secondary findings.

The first main finding was that the positive predictive value in AI-annotated scans was 0.881 or 0.823 – with or without considering tumors as relevant findings – in an authentic clinical environment off-site from the development and testing of the software. This suggests that clinical implementation of AI-annotation is feasible and that further evaluation and clinical implementation of these software could be worthwhile.

The second main finding was that among the individual patients with time benefit from the AI detection – i.e. patients with potential clinical benefit from the AI support tool – an absolute majority had a time benefit of less than 30 minutes (the majority less than 15 minutes), compared to AI negative patients from the same prioritization category. From the point of view of triage assistance of neurosurgical patients, this benefit is of limited value for the patients. On the group level, there was no statistically significant time benefit within any category.

The third main finding was that there was no significant time benefit in patients who required elevation of care or acute surgery compared to those who did not. Furthermore, there was no significant correlation between time to surgery and time benefit from AI-annotation. Among the limited number of patients (n=17) with more than 30 minutes of time benefit, there was no signs of clinical benefit as defined as elevated level of care or emergency surgery.

The second and third main findings combined suggests that in our clinical setting there was no measurable outcome benefit from triage assistance by the AI support system. This corresponds to the primary endpoint of the study, which was to evaluate actual clinical benefit for patients with intracranial hemorrhage generated by the supporting AI application.

The lack of tangible clinical benefit in this study could have several explanations. Time benefit was defined as shorter time from scan to report in annotated cases compared to the normal median waiting time within that priority category. Time from scan to evaluation were already prioritized by the referring physician based on the contents of the referral, and “acute” CT scans would already be prioritized and have lower median waiting time compared to the “subacute” and “routine” scans. In our setting there was limited benefit from AI annotations in “routine” scans, which is partly dependent on the examined population and local traditions (such as CT technicians alerting the radiologist on suspected findings). It is possible that in another clinical setting, AI-positive “routine” scans could have larger benefit from being prioritized for human evaluation.

Furthermore, in our clinical context, patients who require elevation of the level of care or acute surgery are usually identified prior to radiology depending on their clinical status. In other words, if a patient presents with decreased level of consciousness the CT scan would probably be referred as “acute”, the decision to elevate level of care would already be made before the CT scan and they would also be prioritized for surgery. On the other hand, patients with favorable clinical status would be more likely to have “subacute” or “routine” scan referral, and the patients would more likely have smaller hemorrhages that might not require elevation of care or surgery at all. It is also important to note that the number of patients who needed emergency surgery in the AI-positive group was only 7, and that it is not really possible to draw certain conclusions from this. In this study, time benefit from AI-annotation was higher in patients in patients that had *longer* time to acute surgery. This was primarily associated with patients with non-acute conditions, discovered in non-acute scans.

During office hours, median times for acute and subacute cases were slightly *longer* for AI-positive cases (Table 8). This is counter intuitive but probably reflects, to some extent, patients that are well known for the staff and the radiologist, and where the report is not rushed due to the AI annotation.

In a recent systematic review and meta-analysis, Jørgensen et al (2022)[54] conclude that detection of intracranial hemorrhage remains a promising prospect of conventional neural networks, and that sensitivity and specificity often is on par with radiologists. Furthermore, that implementation of AI for detection of ICH could improve workflow and workload of radiologists and hopefully reduce time to diagnosis and improve clinical outcome. However, they conclude that there is need for additional studies, especially including external validation. While the field of image recognition from AI remains promising, clinical implementation of AI tools in a variety of fields remains to be realized. In fact, in a systematic review Lam et al (2022)[69], the authors conclude that there is a very limited number of randomized clinical trials for the implementation of clinical AI tools and that while a majority of these report improved primary end points compared to standard care, very few show clinically relevant improvement. This study addresses some aspects of these concerns.

There is a potential *increase* of the time it takes for the radiologist to confidently reject a false positive AI evaluation, due to fear of missing something that the AI detected. Any potential time-savings from the true positives can be partially mitigated by false positive cases. In the present study, 17.6% of the AI-positive findings were false (not ICH), either due to obvious reasons such as postoperative status or calcifications, or without an obvious explanation. In addition, any false negatives can *also* increase the time to diagnosis since the radiologist can feel safe and ignore such cases as not requiring immediate attention. These factors were not measured in the present study but should ideally be accounted for in the evaluation of potential benefits from implementing artificially intelligent decision support.

There was no case in this study where a false positive classification by the AI misled the interpreting radiologist to a false positive report. When the program was implemented in the clinic, the possibility of false positives was (extensively) discussed among the colleagues, which probably was an important contextual factor. Regarding the radiologist's confidence in working in conjunction with AI, the availability of reliable saliency maps ("heatmaps"), would probably be of high importance. [70] More advanced models providing saliency maps, anatomical location and volume of lesions could also provide other interesting information in the clinical setting and for future research.

Limitations

The primary limitation was that this was a single center open-label study. Additionally, there was a limited number of patients who needed acute surgery, and no certain conclusions should be made regarding the correlation between time benefit from AI-assisted prioritization and time to surgery. Nevertheless, prioritization for acute surgery is a multivariable decision and the clinical status of the patient is equally or more important as the radiological image.

A second limitation was that the time benefit from AI-annotation was retrospectively defined as the difference from the median of normal scans in the same priority category. While this enables identification of a group with potential benefit from AI-prioritizing, it is possible that there is loss of data from the context of individual cases. In future studies, it should be preferable to prospectively randomize groups for AI evaluation and compare the groups from each priority category.

The effects and potential benefits of an intervention in a real-world clinical setting is heavily dependent on the composition of the patient population. The number of patients with, for example, chronic subdural hematomas and acute parenchymal hemorrhage depends on the demographics, and the number of "neurosurgical" patients is associated with the size of the neurosurgical clinic and the selection process in referring patients etc. Thus, the potential benefit will be different in different settings and results are not universally generalizable. The present study represents a medium sized tertiary university hospital with a complete neurosurgical organization.

Study IV

In study IV we continued the development of the models by including dynamic physiological data. Our primary finding is that dynamic physiological data from the NICU at later time points, can significantly improve prognostic accuracy for 7-grade GOSE outcomes. This result, best exemplified by the 96-hour XGBoost model, confirms the long-held clinical hypothesis that the evolving trajectory of a patient's physiology contains valuable prognostic information beyond admission variables. However, a second key finding is that this performance gain is highly sensitive to the validation strategy; the improvement was attenuated and less consistent when evaluated on a temporal split, highlighting the critical challenge of dataset shift for clinical implementation. Interestingly, while the more complex machine learning model had higher performance in the stratified split the classical ordinal logistic regression showed more robustness in the temporal split setting further highlighting the delicate balance between performance and overfitting. Furthermore, there was a clear drop in prognostic performance between the static and static/dynamic NICU cohorts. The static/dynamic cohort is, by definition, composed of patients who survived long enough to have NICU monitoring and whose data was successfully recorded which might select for a more challenging-to-predict group. The static cohort includes patients which were not monitored in the NICU, which could be because they were deemed either too healthy or too sick to be candidates for intensive care and whose prognosis might be easier to separate based on admission data alone.

In our previous work we found that accurate prediction of the full GOSE scale proved difficult using only the admission variables from the IMPACT model, at least when the models are optimized for accuracy. The models over-predicted the most common categories (GOSE 1, 3 and 8), while prediction of the middle GOSE categories proved more elusive[71]. Using admission data, we also found similar performance of logistic regression models compared to more advanced machine learning models such as tree-based random forest models and neural network models. Comparisons of accuracy between this study and the previous study are not beneficial because of differences in inclusion and exclusion criteria and since the models in this study were optimized for QWK. Nevertheless, the calibration plots indicate that the models in this study also had difficulties correctly stratifying the risk of the middle GOSE categories. In essence, we found that prediction of the full GOSE scale continues to prove challenging in the moderate-severe TBI patient group.

In the current clinical setting, initial treatment of TBI is reliant on clinical snapshots – initial GCS and CT scan[72,73]. While powerful predictors, the clinical trajectory of initially similar patients may be very different and lead to dramatically different outcomes. Treatment of secondary injuries are often reactive and put into place only when the secondary insult is already happening[73]. Improved understanding and prognostication of clinical trajectories using dynamic physiological data could open a window for targeted interventions to provide timely interventions to the correct patient and avoiding over-treating the wrong patient[73,74].

However, dynamic physiological data is noisy, incomplete and artifact rich and finding true insults is difficult. To combat this, we attempted to implement modern handling of these variables to optimize their prognostic power. Several recent studies have addressed similar challenges. Bhattacharyay et al, predicted 6-month outcomes on the full GOSE scale using information from the first 24 hours after NICU admission[35]. They found that the biggest driver of performance was a wider set of highly predictive predictors and not deeper models. They also found that the models tended to overfit to higher thresholds GOSE >5, but that consistent prediction of full recovery (GOSE 8) remains the hardest using 24-hour data. They suggest that future models should not necessarily be "deeper" but "wider" using meaningful predictors from advanced monitoring, moving to dynamic ordinal predictions. Several previous models predicting outcome after TBI using NICU data have incorporated statistical features derived from ICP, MAP, and CPP. Raj et al developed dynamic prediction models of 30-day mortality using rolling time windows of mean, median, and slopes across time windows to detect trends[25]. They also calculated NICU burdens based on the fractions of time physiological variables exceeded clinically relevant thresholds (e.g. ICP >20, ICP<10, MAP >120), as well as 10^{th} and 90^{th} percentiles of the data. They found that the slopes and trends of the data, as well as the most recent time-windows, captured the best predictive signals.

In an analysis of the correlation between ICP and CPP insult burden and GOS outcome at 6 months, Güiza et al. found that worse neurological outcome was associated with an ICP above the thresholds of 15, 20, 25 and 30 mmHg for 223, 37, 12 and 8 minutes, respectively. They also found that effect on outcome was dependent on the CPP, and that patients had increased tolerance to ICP insults when CPP thresholds increased from 50 to 60 to 70 mm Hg. The relationship worked both ways: when CPP was under 50, outcome was worse regardless of ICP insult burden, and when ICP was above 25 mmHg, a higher CPP had no effect on outcome[36].

Sheth et al. also looked at the "dose" of intracranial hypertension, quantified as the area under the curve or pressure-time dose above the threshold 20 mmHg. They found that high-resolution pressure-time dose over 24–72h was associated with higher 30-day mortality and poorer GOSE at 3 and 6 months after adjustment for age and admission GCS, whereas the number of 5-minute ICP episodes showed no association[37]. Furthermore, Vik et al. found that higher ICP AUC independently predicted 6-month poor outcome and death[38]. Likewise, Donnelly et al. demonstrated that intracranial pressure (ICP) elevations exceeding 20 mmHg became strongly associated with poor neurological outcome when sustained for longer than 13 minutes. Critically, multivariable analysis revealed that the cumulative "pressure-time burden" spent in these high-risk zones was a powerful independent predictor of both mortality and unfavourable outcome. These findings move beyond static treatment thresholds, quantifying the dose-dependent nature of secondary brain injury from elevated ICP[40]. In a large, high-resolution TBI cohort, Åkerlund et al. identified that the crossing an intracranial pressure (ICP) threshold of 18 ± 4 mmHg was associated with worse functional outcome, refining previous estimates. Crucially, this tolerance was highly dependent on cerebrovascular autoregulation; during periods of impaired autoregulation (PRx > 0.3), no safe ICP threshold could be identified, suggesting any elevation is potentially harmful. Conversely, with intact autoregulation, insults above 19 mmHg for as little as five minutes were associated with injury. These findings strongly suggest that future ICP management should move beyond fixed values towards dynamic, autoregulation-guided thresholds to minimize secondary injury[41]. Furthermore, Svedung Wettervik et al. investigated the correlation between double hits and outcome and found that elevated ICP >20 together with PRx over 0 as well as CPP <70 together with PRx above 0.2–0.4 correlated with unfavourable outcome[39]. Finally, in a narrative review, Cuccuiolini et al also describes that the functional outcome after TBI is strongly determined not by isolated ICP peaks but by the cumulative burden of intracranial hypertension, or the "ICP dose". They describe that pathological ICP patterns, particularly sustained plateau waves (A waves) lasting over 30 minutes, signify a depleted compensatory reserve and are directly associated with increased mortality. Advanced computational techniques, such as morphological clustering of the ICP pulse waveform, are being developed to automatically analyse these dynamic changes, offering deeper prognostic insights beyond a single pressure value[42].

Despite the demonstrated prognostic value, extracting robust signal from high-frequency data remains challenging, and several factors likely hampered the full potential of our models. Foremost is the role of unmeasured confounding by treatment intensity. High-frequency physiological data are not passive observations, rather, they serve as triggers for clinical interventions. An episode of high ICP, depending on the magnitude and duration, prompts interventions such as increased sedation, hyperventilation, osmotic therapy or surgery[72,75,76].

These treatments are explicitly designed to mitigate the harm of the secondary insult. A machine learning model, seeing only the physiological data and not treatment intensity, cannot distinguish between a patient with refractory ICP despite maximal therapy and a patient whose ICP is easily controlled with minimal intervention. Even though we have attempted mitigation by including information on the use of barbiturates and decompressive hemicraniectomy, this confounding likely masks the true prognostic signal. Our models may be learning that 'high ICP' is not a strong predictor because, in many cases, it is successfully treated. Another way we attempted to mitigate this was by including ICP<10, which would indicate a stable patient which likely receives less interventions than a patient fluctuating between 15-20.

The significant performance drop observed on the temporal split has profound implications for the clinical translation of predictive models. This "dataset shift" suggests that the patient population, care protocols, or data recording practices evolved over the training and testing periods[77]. The phenomenon particularly affected the more complex XGBoost models, while the simpler Ordinal Logistic Regression models showed greater robustness, suggesting that simpler models may be less susceptible to overfitting. During the data collection period (2008-2024) the treatment protocol and treatment targets have remained largely the same, so the dataset shift might have other causes such as unmeasured demographics changes, trauma-related differences such as safer vehicles, or other diagnostic changes such as improved radiology. This finding is a strong reminder that a model's performance can be substantially overestimated by standard cross-validation, and it underscores the absolute necessity of continuous monitoring, regular recalibration, and prospective validation before any such tool can be safely deployed.

When dichotomized for favourable versus unfavourable outcomes, most of our models achieved fair AUROC values of 0.8-0.9 for mortality which is comparable or even improved compared to those reported in the IMPACT trial and recent external validations[20,68,78].

The AUROC values of favourable outcome had a slightly wider range of 0.67-0.83 depending on the model. A central challenge in this work was addressing the significant class imbalance of the 8-point GOSE scale. We systematically tuned both class weighting schemes during training and ordinal thresholding methods during inference. We observed a critical trade-off: models aggressively optimized for the highest QWK score tended to bias their predictions toward the more central GOSE categories. This represents a form of model “hedging” to minimize overall error by avoiding the higher risk of misclassifying rare outcomes at the extremes of the scale. Interestingly, on some models this led to a scenario where the model has high AUROC for mortality but never actually predicted the category. This highlights that the optimal model configuration is application-dependent, and future implementation designed to maximize sensitivity for specific outcomes (e.g., mortality or severe disability) may require different parameters than one seeking the highest QWK.

At present, this study adds to the literature the feasibility of ordinal GOSE prediction, but further refinement and validation is needed before clinical implementation. It is possible that variables that better reflect the extent of the brain injury and its anatomical location, such as biomarkers and improved radiological classification, could improve the prediction accuracy of these middle GOSE categories. A recent study by Hibi et al. showed that a multimodal neural network model that incorporated raw CT data had improved predictions of both dichotomized and categorical GOSE compared to models that included only the IMPACT variables and manual CT scoring systems[52].

Nevertheless, the potential of ordinal outcome prediction models to increase the power in clinical trials, and to provide refined prognostic information to clinical practitioner, the patient and their family, remain compelling.

Strength and limitations

The primary strength of this study is the use of a large well-characterized cohort with high-frequency physiological data, allowing for consistent data quality and care protocols. Being a single-center study has both advantages and limitations: although it allows detailed and consistent data collection, our results require external validation to ensure generalizability across other settings. Our rigorous methodology, including the static/dynamic design on the same patient cohort and the comparison of splitting strategies, adds robustness to our conclusions.

The most significant limitation, as discussed, is the lack of detailed treatment data. Incorporating variables such as sedative infusions, volumes of cerebrospinal fluid drained from ventricular drainage, doses of osmotic agents, and ventilator settings will be the next step to separate the prognostic signal from treatment effects. Finally, missing data required imputation, which may have influenced model performance.

Conclusions

Study I

The IMPACT model produces similar prediction values of mortality at 6 months in our TBI cohort as has been previously reported. However, the model shows poorer discriminative ability for the rest of the GOSE spectrum. Future studies using more complex machine learning models or other variables such as intensive care unit data and genetics that can be incorporated into new prognostic models in order to predict outcome following TBI in the full GOSE spectrum.

Study II

This is the first study predicting full scale GOSE outcome after TBI. We demonstrate that both the development and evaluation of such models is feasible. We conclude that all models perform well and that neural networks, random forests, and ordinal regression models could be considered for future studies of TBI outcome prediction and for the evaluation of specific TBI interventions. Furthermore, both simpler and more complex machine learning prediction models perform similarly with the IMPACT admission variables and to produce more specific outcome prediction more complex patient data is probably required.

Study III

The examined AI tool, Zebra Medical Vision – HealthICH+, had adequate positive predictive value in finding patients with intracranial hemorrhage. However, the central conclusion from this this study is that although a number of patients with AI-positive findings had a shortened time delay between examination and report, this time gain did not infer any measurable clinical benefit for the patients, such as prompt neurosurgical attention, elevated level of care, or earlier surgery. While future use of machine learning tools in health care is promising, its implementation must be adapted to meet real-world needs in the clinical setting, and potential benefit should be systematically evaluated.

Study IV

Dynamic physiological data contains a clear prognostic signal for long-term recovery after TBI, particularly when analysed 2-4 days after injury. Our study demonstrates that machine learning models can harness this information to improve prediction beyond what is possible with admission data alone.

Furthermore, we show that cohort selection and validation strategy have high impact on the model's performance. Future studies in TBI prognostication should include temporal validation to prove robustness to temporal dataset shift. For these tools to be clinically useful, future research must focus not only on incorporating richer data streams, like treatment variables, but also on developing robust strategies for continuous model monitoring and updating to ensure they remain accurate and reliable over time.

Overall conclusions

This thesis explores the development, evaluation, and clinical translation of machine learning systems for traumatic brain injury (TBI) and intracranial hemorrhage. Collectively, these studies demonstrate that while the shift from binary to ordinal outcome prediction is both feasible and statistically sound, the path to clinical utility requires more than just sophisticated machine learning models.

We identified that current "gold standard" models (IMPACT) are limited by their focus on dichotomized outcomes and admission data. In response, we demonstrated that predicting the full spectrum of recovery (GOSE) is achievable, though we found that advanced machine learning methods perform similarly to standard regression when restricted to baseline variables. To improve predictions further, advanced algorithms require richer, more detailed data. Our work confirms that incorporating dynamic physiological data captures prognostic signals missed by static admission models, significantly enhancing accuracy.

However, technical accuracy does not guarantee clinical efficacy. Our evaluation of an AI diagnostic tool revealed a critical gap between algorithm performance and tangible patient benefit, highlighting that implementation strategies are just as vital as the technology itself. Finally, we emphasize that TBI models are not static tools; they are susceptible to dataset shift over time. Therefore, future research must prioritize temporal validation and strategies for continuous model updating. This thesis supports a shift toward ordinal, dynamic, and continuously validated prognostic models.

Future directions

Ordinal outcome prediction represents a promising step forward for TBI research, offering enhanced statistical power and more granular prognostic information. Future models should incorporate predictors that more accurately capture injury severity and pathology, specifically through biomarkers and advanced neuroimaging classifications such as volumetric analysis. Additionally, the application of Bayesian models offers the advantage of quantified uncertainty, in other words how sure the model is in its predictions, which could prove to be a vital feature for clinical utility. However, prior to clinical implementation, rigorous external validation and continued recalibration would be essential to ensure robustness within an ever-changing healthcare environment.

Summary in Swedish – Sammanfattning på svenska

Traumatisk hjärnskada är en vanlig åkomma och en stor orsak till död och handikapp. Skadan drabbar ofta äldre som ramlar i hemmet, samt yngre som är med om högenergivåld som bilolyckor eller misshandel. Vården av dessa patienter förbättrades stort med introduktionen av neurointensivvården på 80- och 90-talet, men utvecklingen har stagnerat de senaste årtiondena. Flera läkemedel som visade potential i djurstudier kunde inte uppvisa någon effekt då de testades kliniskt.

En av anledningarna till detta är att sjukdomsförloppet är så heterogent. Man har sett att ålder spelar stor roll för återhämtningen, men också skadans utbredning, vilken sorts hjärnblödning den ger upphov till, vilka underliggande sjukdomar och vilken pågående medicinering man har osv. Dessutom har vi tillgång till stor mängd information om vad som händer i huvudet efter skadan under vårdtiden som också kan påverka utfallet, till exempel trycket i skallen och blodtryck.

Genom att använda maskininlärning kan man utveckla modeller som kan förutspå hur det kommer att gå för de här patienterna, vilket kan minska heterogeniteten i kliniska studier, och potentiellt guida sjukvårdspersonal att identifiera riskpatienter.

Den senaste tiden har det varit mycket uppståndelse kring maskininlärning och artificiell intelligens (AI). Detta beror dels på grund av de stora framgångar som språkmodeller och chatbottar har uppnått och fortsätter att uppnå, dels på grund av den fortsatta utvecklingen av hårdvara och grafikkort som möjliggör träningen av allt större och mer avancerade modeller. Det innebär att man kan utveckla relativt avancerade maskininlärningsmodeller genom att använda sin vanliga dator, även om de mest avancerade modellerna fortfarande kräver superdatorer/datakluster.

Maskininlärningsmodeller som neurala nätverk har funnits sedan mitten på 1900-talet, men dess användning i medicinsk forskning har varit relativt begränsad. De senaste åren har dock algoritmer för bildigenkänning och EKG-tolkning tagit stora kliv framåt. De nuvarande modellerna som predikterar utfallet för patienter med traumatisk hjärnskada använder ankomstdata, är byggda med klassiska statistiska modeller och ger en binär (tvådelad) prediktion: exempelvis död/överlevnad. Vårt mål med den här forskningen var att utveckla maskininlärningsmodeller som förutspår mer ett fingraderat utfall efter traumatisk hjärnskada genom att använda både inskrivningsdata och data från intensivvårdsperioden.

Vi fann att avancerade maskininlärningsmetoder presterar på liknande nivå som standard statistiska modeller när de begränsas till inskrivningsvariabler. För att förbättra prognostiseringen ytterligare krävs mer detaljerade data. Vi fann att dynamiska fysiologiska data fångar prognostiska signaler som statiska intagningsmodeller missar, vilket förbättrade prediktionerna.

Vi utvärderade även ett AI-diagnostikverktyg som automatiskt identifierade hjärnblödningar på datortomografiröntgen. Det visade sig att även om verktyget lyckades hitta de flesta blödningarna så såg vi ingen större konkret klinisk nytta för patienten. Det understryker att implementeringsstrategin är minst lika viktig som teknologin i sig.

Slutligen måste det betonas att de här modellerna inte är statiska verktyg. De är känsliga för förändringar över tid och därför måste framtida implementering prioritera kontinuerlig uppdatering av liknande modeller.

Acknowledgements

First and foremost, I would like to acknowledge all the patients and their families who participated in the research presented in this thesis. It is important to remember and respect that every data point in every graph represents a human being who suffered injury to the brain.

This thesis would not be possible without the help and support from many people whom I would like to express my gratitude and gratefulness towards.

To Elham Rostami, my main supervisor, I would like to extend my sincerest thanks for the extensive support and guidance through this thesis. I could always reach out to you for help with the small or large obstacles through this journey. You are a true scientific visionary, who taught me that the sky is the limit.

To Anders Hånell, my co-supervisor, whose guidance made this work possible. Your practical counsel helped me immensely and will forever be appreciated, and your continued support and input was invaluable.

To Per Enblad, my co-supervisor, thank you for supporting me as a PhD student at the department. Your vast clinical and scientific knowledge is incredibly inspiring. Your input in the papers was vital, and your leadership of the department laid the foundation for the thesis.

To Anders Lewén, my co-supervisor, who is responsible for the TBI registry which enabled this thesis. Thank you for your invaluable input in the papers, and I aspire to have your attention to details and quality in both research and clinical practice.

To David Fällmar, my co-supervisor, for enabling and supervising paper III. Your excellence in neuroradiology is paramount in both research and clinical practice, and I consider the potential in continued multidisciplinary collaboration a fundament for the future development of healthcare.

To Ludvig Hult and Thomas Schön, for the incredibly rewarding and friendly interdisciplinary meetings, and invaluable input in practical machine learning problem solving.
To Samuel Lenell thank you for the data collection of radiological data and your excellent input in the papers. It is always a pleasure to interact in the clinic.

To Timothy Howells, thank you for the input in paper I and for the creation of the Odin browser which enabled the collection of high-frequency physiological data.

To Teodor Svedung-Wettervik, thank you for your invaluable input in paper IV. I always appreciate our discussions and your knowledge in the field. Your drive is an inspiration in both the research and clinical arena.

To Bart Depreitere and David W. Wright, thank you for the opportunity to collaborate on paper II. The future of TBI research must include continued international collaboration.

To Julia Basu, Dimitrios Toumpanakis, Johan Burwick Nyberg, Tomas Bjerner for their essential input and contributions to the data collection in paper III.

To Sami Abu Hamdeh, for your excellent clinical mentorship and guidance throughout the years.

To Hans Ericson and Mats Ryttlefors, who hired me in the neurosurgery department, which made all this possible. Thank you, as well to Andreas Fahlström, for giving me time for research.

To all my colleagues at the department of neurosurgery, who all provide the best possible care for this challenging patient group. A special thank you to all senior colleagues who I rely on daily for clinical guidance and thank you to my fellow junior colleagues who always make the days a little brighter.

To my parents, Lise-Lotte and Per and my bonus parent Ylva for their unconditional love and support.

To my sister Frida, for your love and support through the years.

To my parents-in-law, Marianne and HG, as well as siblings-in-law Johan, Sophia, and Erik with families, for all your support and hospitality.

To Lovisa. My soulmate, wife and guiding star in the night. My deepest love to you and our children, August and Alice, who make life worthwhile.

References

1. Dewan MC, Rattani A, Gupta S, et al. Estimating the global incidence of traumatic brain injury. J Neurosurg 2018:1–18. DOI: 10.3171/2017.10.JNS17352.
2. Hyder AA, Wunderlich CA, Puvanachandra P, Gururaj G, Kobusingye OC. The impact of traumatic brain injuries: a global perspective. NeuroRehabilitation 2007;22(5):341–53. (https://www.ncbi.nlm.nih.gov/pubmed/18162698).
3. Miller GF, DePadilla L, Xu L. Costs of Nonfatal Traumatic Brain Injury in the United States, 2016. Med Care 2021;59(5):451–455. DOI: 10.1097/MLR.0000000000001511.
4. Jain S, Iverson LM. Glasgow Coma Scale. StatPearls. Treasure Island (FL)2023.
5. Permenter CM F-dTR, Sherman Al. Postconcussive Syndrome: Treasure Island (FL): StatPearls Publishing, 2022.
6. Marshall LF, Marshall SB, Klauber MR, et al. A new classification of head injury based on computerized tomography. Journal of Neurosurgery 1991;75(Supplement):S14–S20. (In English). DOI: https://doi.org/10.3171/sup.1991.75.1s.0s14.
7. Manley GT, Dams-O'Connor K, Alosco ML, et al. A new characterisation of acute traumatic brain injury: the NIH-NINDS TBI Classification and Nomenclature Initiative. Lancet Neurol 2025;24(6):512–523. DOI: 10.1016/S1474-4422(25)00154-1.
8. RIKSSTROKE Quality of the Swedish stroke care 2021. A breif summary of data for the full year 2021. (https://www.riksstroke.org/wp-content/uploads/2022/11/Arsrapport-2021-engelsk-sammanfattning_final.pdf).
9. van Asch CJ, Luitse MJ, Rinkel GJ, van der Tweel I, Algra A, Klijn CJ. Incidence, case fatality, and functional outcome of intracerebral haemorrhage over time, according to age, sex, and ethnic origin: a systematic review and meta-analysis. Lancet Neurol 2010;9(2):167–76. DOI: 10.1016/S1474-4422(09)70340-0.
10. Elliott J, Smith M. The acute management of intracerebral hemorrhage: a clinical review. Anesth Analg 2010;110(5):1419–27. DOI: 10.1213/ANE.0b013e3181d568c8.
11. Head Injury: Triage, Assessment, Investigation and Early Management of Head Injury in Children, Young People and Adults. London2014.
12. Loane DJ, Faden AI. Neuroprotection for traumatic brain injury: translational challenges and emerging therapeutic strategies. Trends Pharmacol Sci 2010;31(12):596–604. DOI: 10.1016/j.tips.2010.09.005.
13. Hernandez AV, Steyerberg EW, Taylor GS, Marmarou A, Habbema JD, Maas AI. Subgroup analysis and covariate adjustment in randomized clinical trials of traumatic brain injury: a systematic review. Neurosurgery 2005;57(6):1244–53; discussion 1244–53. DOI: 10.1227/01.neu.0000186039.57548.96.

14. Perel P, Wasserberg J, Ravi RR, Shakur H, Edwards P, Roberts I. Prognosis following head injury: a survey of doctors from developing and developed countries. J Eval Clin Pract 2007;13(3):464–5. DOI: 10.1111/j.1365-2753.2006.00713.x.
15. Keegan MT, Soares M. What every intensivist should know about prognostic scoring systems and risk-adjusted mortality. Rev Bras Ter Intensiva 2016;28(3):264–269. DOI: 10.5935/0103-507X.20160052.
16. 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):477–85. DOI: 10.1038/nrneurol.2016.89.
17. 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):573–85. DOI: 10.1089/neu.1998.15.573.
18. Perel P, Edwards P, Wentz R, Roberts I. Systematic review of prognostic models in traumatic brain injury. BMC Med Inform Decis Mak 2006;6:38. DOI: 10.1186/1472-6947-6-38.
19. Collaborators MCT, Perel P, Arango M, et al. Predicting outcome after traumatic brain injury: practical prognostic models based on large cohort of international patients. BMJ 2008;336(7641):425–9. DOI: 10.1136/bmj.39461.643438.25.
20. Steyerberg EW, Mushkudiani N, Perel P, 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; discussion e165. DOI: 10.1371/journal.pmed.0050165.
21. Roozenbeek B, Chiu YL, Lingsma HF, et al. Predicting 14-day mortality after severe traumatic brain injury: application of the IMPACT models in the brain trauma foundation TBI-trac(R) New York State database. J Neurotrauma 2012;29(7):1306–12. DOI: 10.1089/neu.2011.1988.
22. Roozenbeek B, Lingsma HF, Lecky FE, 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):1609–17. DOI: 10.1097/CCM.0b013e31824519ce.
23. Sun H, Lingsma HF, Steyerberg EW, Maas AI. External Validation of the International Mission for Prognosis and Analysis of Clinical Trials in Traumatic Brain Injury: Prognostic Models for Traumatic Brain Injury on the Study of the Neuroprotective Activity of Progesterone in Severe Traumatic Brain Injuries Trial. J Neurotrauma 2016;33(16):1535–43. DOI: 10.1089/neu.2015.4164.
24. 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):543–54. DOI: 10.1016/S1474-4422(10)70065-X.
25. Raj R, Wennervirta JM, Tjerkaski J, et al. Dynamic prediction of mortality after traumatic brain injury using a machine learning algorithm. NPJ Digit Med 2022;5(1):96. DOI: 10.1038/s41746-022-00652-3.
26. Zeiler FA, McFadyen C, Newcombe VFJ, et al. Genetic Influences on Patient-Oriented Outcomes in Traumatic Brain Injury: A Living Systematic Review of Non-Apolipoprotein E Single-Nucleotide Polymorphisms. J Neurotrauma 2021;38(8):1107–1123. DOI: 10.1089/neu.2017.5583.
27. Thelin E, Al Nimer F, Frostell A, et al. A Serum Protein Biomarker Panel Improves Outcome Prediction in Human Traumatic Brain Injury. J Neurotrauma 2019;36(20):2850–2862. DOI: 10.1089/neu.2019.6375.

28. O'Neill TJ, Xi Y, Stehel E, et al. Active Reprioritization of the Reading Worklist Using Artificial Intelligence Has a Beneficial Effect on the Turnaround Time for Interpretation of Head CT with Intracranial Hemorrhage. Radiol Artif Intell 2021;3(2):e200024. DOI: 10.1148/ryai.2020200024.
29. Kinoshita K. Traumatic brain injury: pathophysiology for neurocritical care. J Intensive Care 2016;4:29. DOI: 10.1186/s40560-016-0138-3.
30. Kaloostian P, Robertson C, Gopinath SP, et al. Outcome prediction within twelve hours after severe traumatic brain injury by quantitative cerebral blood flow. J Neurotrauma 2012;29(5):727–34. DOI: 10.1089/neu.2011.2147.
31. Sorrentino E, Diedler J, Kasprowicz M, et al. Critical thresholds for cerebrovascular reactivity after traumatic brain injury. Neurocrit Care 2012;16(2):258–66. DOI: 10.1007/s12028-011-9630-8.
32. Czosnyka M, Smielewski P, Kirkpatrick P, Laing RJ, Menon D, Pickard JD. Continuous assessment of the cerebral vasomotor reactivity in head injury. Neurosurgery 1997;41(1):11–7; discussion 17–9. DOI: 10.1097/00006123-199707000-00005.
33. Gritti P, Bonfanti M, Zangari R. Cerebral autoregulation in traumatic brain injury: looking upwards. Crit Care 2025;29(1):263. DOI: 10.1186/s13054-025-05489-2.
34. BTF 4th-ed. Guidelines (Guidelines for the Management of Severe TBI, 4th Edition — Brain Trauma Foundation). (https://braintrauma.org/coma/guidelines/severe-tbi).
35. Bhattacharyay S, Milosevic I, Wilson L, et al. The leap to ordinal: Detailed functional prognosis after traumatic brain injury with a flexible modelling approach. PLoS One 2022;17(7):e0270973. DOI: 10.1371/journal.pone.0270973.
36. Guiza F, Depreitere B, Piper I, et al. Visualizing the pressure and time burden of intracranial hypertension in adult and paediatric traumatic brain injury. Intensive Care Med 2015;41(6):1067–76. DOI: 10.1007/s00134-015-3806-1.
37. Sheth KN, Stein DM, Aarabi B, et al. Intracranial pressure dose and outcome in traumatic brain injury. Neurocrit Care 2013;18(1):26–32. DOI: 10.1007/s12028-012-9780-3.
38. Vik A, Nag T, Fredriksli OA, et al. Relationship of "dose" of intracranial hypertension to outcome in severe traumatic brain injury. J Neurosurg 2008;109(4):678–84. DOI: 10.3171/JNS/2008/109/10/0678.
39. Svedung Wettervik T, Beqiri E, Hanell A, et al. Visualization of Cerebral Pressure Autoregulatory Insults in Traumatic Brain Injury. Crit Care Med 2024;52(8):1228–1238. DOI: 10.1097/CCM.0000000000006287.
40. Donnelly J, Guiza F, Depreitere B, Meyfroidt G, Czosnyka M, Smielewski P. Visualising the pressure-time burden of elevated intracranial pressure after severe traumatic brain injury: a retrospective confirmatory study. Br J Anaesth 2021;126(1):e15–e17. DOI: 10.1016/j.bja.2020.09.018.
41. Akerlund CA, Donnelly J, Zeiler FA, et al. Impact of duration and magnitude of raised intracranial pressure on outcome after severe traumatic brain injury: A CENTER-TBI high-resolution group study. PLoS One 2020;15(12):e0243427. DOI: 10.1371/journal.pone.0243427.
42. Cucciolini G, Motroni V, Czosnyka M. Intracranial pressure for clinicians: it is not just a number. J Anesth Analg Crit Care 2023;3(1):31. DOI: 10.1186/s44158-023-00115-5.
43. Muehlematter UJ, Daniore P, Vokinger KN. Approval of artificial intelligence and machine learning-based medical devices in the USA and Europe (2015-20):

a comparative analysis. Lancet Digit Health 2021;3(3):e195–e203. DOI: 10.1016/S2589-7500(20)30292-2.
44. Ye G, Balasubramanian V, Li JK, Kaya M. Machine Learning-Based Continuous Intracranial Pressure Prediction for Traumatic Injury Patients. IEEE J Transl Eng Health Med 2022;10:4901008. DOI: 10.1109/JTEHM.2022.3179874.
45. Krenn M, Pollice R, Guo SY, et al. On scientific understanding with artificial intelligence. Nat Rev Phys 2022;4(12):761–769. DOI: 10.1038/s42254-022-00518-3.
46. Trevor Hastie RT, Jerome Friedman. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer Series in Statistics: Springer, 2017.
47. Breiman L. Random Forests. Machine Learning 45, 5–32 (2001)October 2001.
48. Tianqi Chen CG. XGBoost: A Scalable Tree Boosting System. ArXiv 2016. DOI: https://doi.org/10.48550/arXiv.1603.02754.
49. Goodfellow I, Bengio Y, Courville A. Deep learning. Cambridge, Massachusetts: The MIT Press, 2016.
50. Nielsen MA. Neural Networks and Deep learning: Determination Press, 2015.
51. (FDA) UFaDA. Artificial Intelligence-Enabled Medical Devices. (https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-enabled-medical-devices).
52. Hibi A, Cusimano MD, Bilbily A, Krishnan RG, Tyrrell PN. Development of a Multimodal Machine Learning-Based Prognostication Model for Traumatic Brain Injury Using Clinical Data and Computed Tomography Scans: A CENTER-TBI and CINTER-TBI Study. J Neurotrauma 2024;41(11-12):1323–1336. DOI: 10.1089/neu.2023.0446.
53. Graham P, Jackson R. The analysis of ordinal agreement data: beyond weighted kappa. J Clin Epidemiol 1993;46(9):1055–62. DOI: 10.1016/0895-4356(93)90173-x.
54. Daugaard Jorgensen M, Antulov R, Hess S, Lysdahlgaard S. Convolutional neural network performance compared to radiologists in detecting intracranial hemorrhage from brain computed tomography: A systematic review and meta-analysis. Eur J Radiol 2022;146:110073. DOI: 10.1016/j.ejrad.2021.110073.
55. Flanders AE, Prevedello LM, Shih G, et al. Construction of a Machine Learning Dataset through Collaboration: The RSNA 2019 Brain CT Hemorrhage Challenge. Radiol Artif Intell 2020;2(3):e190211. DOI: 10.1148/ryai.2020190211.
56. Kuo W, Häne C, Mukherjee P, Malik J, Yuh EL. Expert-level detection of acute intracranial hemorrhage on head computed tomography using deep learning. Proc Natl Acad Sci U S A 2019;116(45):22737–22745. DOI: 10.1073/pnas.1908021116.
57. Chilamkurthy S, Ghosh R, Tanamala S, et al. Deep learning algorithms for detection of critical findings in head CT scans: a retrospective study. Lancet 2018;392(10162):2388–2396. DOI: 10.1016/S0140-6736(18)31645-3.
58. Gibson E, Georgescu B, Ceccaldi P, et al. Artificial Intelligence with Statistical Confidence Scores for Detection of Acute or Subacute Hemorrhage on Noncontrast CT Head Scans. Radiol Artif Intell 2022;4(3):e210115. DOI: 10.1148/ryai.210115.
59. Ginat DT. Analysis of head CT scans flagged by deep learning software for acute intracranial hemorrhage. Neuroradiology 2020;62(3):335–340. DOI: 10.1007/s00234-019-02330-w.

60. O'Neill TJ, Xi Y, Stehel E, et al. Active Reprioritization of the Reading Worklist Using Artificial Intelligence Has a Beneficial Effect on the Turnaround Time for Interpretation of Head CT with Intracranial Hemorrhage. Radiology: Artificial Intelligence 2021;3(2):e200024. DOI: 10.1148/ryai.2020200024.
61. Yu AC, Eng J. One Algorithm May Not Fit All: How Selection Bias Affects Machine Learning Performance. Radiographics 2020;40(7):1932–1937. DOI: 10.1148/rg.2020200040.
62. Nyholm L, Howells T, Enblad P, Lewen A. Introduction of the Uppsala Traumatic Brain Injury register for regular surveillance of patient characteristics and neurointensive care management including secondary insult quantification and clinical outcome. Ups J Med Sci 2013;118(3):169–80. DOI: 10.3109/03009734.2013.806616.
63. Robin X, Turck N, Hainard A, et al. pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics 2011;12:77. DOI: 10.1186/1471-2105-12-77.
64. Wilson L, Boase K, Nelson LD, et al. A Manual for the Glasgow Outcome Scale-Extended Interview. J Neurotrauma 2021;38(17):2435–2446. DOI: 10.1089/neu.2020.7527.
65. H. W. ggplot2: Elegant Graphics for Data Analysis: Springer-Verlag New York., 2016.
66. Wright DW, Yeatts SD, Silbergleit R, et al. Very early administration of progesterone for acute traumatic brain injury. N Engl J Med 2014;371(26):2457–66. DOI: 10.1056/NEJMoa1404304.
67. Cremer OL, Moons KG, van Dijk GW, van Balen P, Kalkman CJ. Prognosis following severe head injury: Development and validation of a model for prediction of death, disability, and functional recovery. J Trauma 2006;61(6):1484–91. DOI: 10.1097/01.ta.0000195981.63776.ba.
68. Rostami E, Gustafsson D, Hanell A, et al. Prognosis in moderate-severe traumatic brain injury in a Swedish cohort and external validation of the IMPACT models. Acta Neurochir (Wien) 2022;164(3):615–624. DOI: 10.1007/s00701-021-05040-6.
69. Lam TYT, Cheung MFK, Munro YL, Lim KM, Shung D, Sung JJY. Randomized Controlled Trials of Artificial Intelligence in Clinical Practice: Systematic Review. J Med Internet Res 2022;24(8):e37188. DOI: 10.2196/37188.
70. Yeo M, Tahayori B, Kok HK, et al. Evaluation of techniques to improve a deep learning algorithm for the automatic detection of intracranial haemorrhage on CT head imaging. Eur Radiol Exp 2023;7(1):17. DOI: 10.1186/s41747-023-00330-3.
71. Bark D, Boman M, Depreitere B, et al. Refining outcome prediction after traumatic brain injury with machine learning algorithms. Sci Rep 2024;14(1):8036. DOI: 10.1038/s41598-024-58527-4.
72. Maas AIR, Menon DK, Manley GT, et al. Traumatic brain injury: progress and challenges in prevention, clinical care, and research. Lancet Neurol 2022;21(11):1004–1060. DOI: 10.1016/S1474-4422(22)00309-X.
73. Lazaridis C, Foreman B. Management Strategies Based on Multi-Modality Neuromonitoring in Severe Traumatic Brain Injury. Neurotherapeutics 2023;20(6):1457–1471. DOI: 10.1007/s13311-023-01411-2.
74. Lindblad C, Raj R, Zeiler FA, Thelin EP. Current state of high-fidelity multimodal monitoring in traumatic brain injury. Acta Neurochir (Wien) 2022;164(12):3091–3100. DOI: 10.1007/s00701-022-05383-8.

75. Hutchinson PJ, Kolias AG, Timofeev IS, et al. Trial of Decompressive Craniectomy for Traumatic Intracranial Hypertension. N Engl J Med 2016;375(12):1119–30. DOI: 10.1056/NEJMoa1605215.
76. Asehnoune K, Lasocki S, Seguin P, et al. Association between continuous hyperosmolar therapy and survival in patients with traumatic brain injury - a multicentre prospective cohort study and systematic review. Crit Care 2017;21(1):328. DOI: 10.1186/s13054-017-1918-4.
77. Guo LL, Pfohl SR, Fries J, et al. Systematic Review of Approaches to Preserve Machine Learning Performance in the Presence of Temporal Dataset Shift in Clinical Medicine. Appl Clin Inform 2021;12(4):808–815. DOI: 10.1055/s-0041-1735184.
78. Yue JK, Lee YM, Sun X, et al. Performance of the IMPACT and CRASH prognostic models for traumatic brain injury in a contemporary multicenter cohort: a TRACK-TBI study. J Neurosurg 2024;141(2):417–429. DOI: 10.3171/2023.11.JNS231425.

Acta Universitatis Upsaliensis

Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Medicine 2217

Editor: The Dean of the Faculty of Medicine

A doctoral dissertation from the Faculty of Medicine, Uppsala University, is usually a summary of a number of papers. A few copies of the complete dissertation are kept at major Swedish research libraries, while the summary alone is distributed internationally through the series Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Medicine. (Prior to January, 2005, the series was published under the title "Comprehensive Summaries of Uppsala Dissertations from the Faculty of Medicine".)

Distribution: publications.uu.se
urn:nbn:se:uu:diva-571980

ACTA UNIVERSITATIS
UPSALIENSIS
2025