Machine learning versus regression for prognostication in traumatic brain injury: between cohort differences dominated

Corresponding author:
Benjamin Y. Gravesteijn BSc
Departments of Public Health
Erasmus MC – University Medical Centre Rotterdam
Postbus 2040
3000 CA, Rotterdam
The Netherlands
b.gravesteijn@erasmusmc.nl
+316-83448055

Daan Nieboer MSc
Departments of Public Health
Erasmus MC – University Medical Centre Rotterdam
The Netherlands

Ari Ercole PhD
Division of Anaesthesia
University of Cambridge
United Kingdom

Hester F. Lingsma PhD
Departments of Public Health
Erasmus MC – University Medical Centre Rotterdam
The Netherlands

David Nelson PhD
Department of Physiology and Pharmacology, Section of Perioperative Medicine and Intensive Care
Karolinska Institutet
Stockholm, Sweden

Ben van Calster PhD
Department of Development and Regeneration
KU Leuven
Belgium
Department of Biomedical Data Sciences

Leiden university medical centre
The Netherlands

Ewout W. Steyerberg PhD
Department of Biomedical Data Sciences
Leiden University Medical Centre
Leiden, The Netherlands
&
Department of Public Health
Erasmus MC – University Medical Centre Rotterdam
Rotterdam, The Netherlands

and the CENTER-TBI collaborators

Running title
Machine learning for prediction in traumatic brain injury

Take home message
Flexible machine learning algorithms may not perform better than traditional regression approaches in a low-dimensional setting for outcome prediction after moderate or severe TBI. Similar to regression-based prediction models, ML algorithms should be rigorously validated to ensure applicability to new populations.

Word count of the paper: 3313

Word count of the abstract: 190

Number of tables: 3

Number of figures: 2

Number of references: 55

Number of additional files: 2

Objective
We aimed to explore the added value of common machine learning (ML) algorithms for prediction of outcome for moderate and severe traumatic brain injury.

Study Design and Setting
We performed logistic (LR), lasso, and ridge regression with key baseline predictors in the IMPACT-II database (15 studies, n=11,022). ML algorithms included support vector machines, random forests, gradient boosting machines, and artificial neural networks, and were trained using the same predictors. To assess generalizability of predictions, we performed internal, internal-external, and external validation on the recent CENTER-TBI study (patients with GCS<13, n = 1,554). Both calibration (calibration slope/intercept) and discrimination (AUC) was quantified.

Results

In the IMPACT-II database, 3,332/11,022(30%) died and 5,233(48%) had unfavorable outcome (Glasgow Outcome Scale below 4). In the CENTER-TBI study, 348/1,554(29%) died and 651(54%) had unfavorable outcome. Discrimination and calibration varied widely between the studies, and less so between the studied algorithms. The mean AUC was 0.82 for mortality and 0.77 for unfavorable outcome in CENTER-TBI.

Conclusion

ML algorithms may not outperform traditional regression approaches in a low-dimensional setting for outcome prediction after moderate or severe TBI. Similar to regression-based prediction models, ML algorithms should be rigorously validated to ensure applicability to new populations.

Keywords

Machine learning; Prognosis; Traumatic brain injury; Prediction; Data science; Cohort study

1. Introduction

Traumatic brain injury (TBI) is a common disease, with a significant societal burden[1]: TBI is estimated to be responsible for around 300 hospital admissions and 12 deaths per 100,000 persons per year in Europe[2]. TBI is a heterogeneous disease in terms of phenotype and prognosis [3]. Therefore, prognostic models, which predict outcome for a patient given a particular combination of baseline characteristics, are important: they may give us insight in mechanisms of disease that lead to poor outcome, and allow for risk-based stratification of patients for logistic, research, and clinical reasons.

A large number of prediction models have been developed to predict outcome for TBI patients, mostly using traditional regression techniques [4]. However, these models have not yet been widely implemented in clinical practice. In recent years, more flexible machine learning (ML) algorithms have enjoyed enthusiasm as a potentially promising techniques to improve outcome prognostication [5]. Frequently used methods are support vector machines (SVM)[6], deep neural networks (NN) [7], random forests (RF) [8], and gradient boosting machine (GBM) [9]. Some of these algorithms have been used to develop prediction models on small datasets (<200 events) [10–12]. Since ML algorithms are more prone to overfitting [13], it remains unclear what the impact on prognostication is of these novel techniques.

Although the incremental value of flexible ML methods has been previously assessed, these comparisons were potentially subject to bias [14]. The incremental value of ML algorithms is potentially overrated, because studies up to this point mainly focused on the ability of the methods to discriminate between patients with good and poor outcome [15–19]. Performance of prediction models is however commonly measured across at least two dimensions: calibration and discrimination [20,21]. Calibration refers to the agreement of predicted probabilities of a model and observed outcomes (e.g. “if the risk of death is x%, do x% of the patients with this prediction actually die?”). Poor calibration of prediction models may lead to harmful decision making when applying these models [22–24].

One of the more thoroughly validated prediction models with good performance exists in the field of traumatic brain injury (TBI): the IMPACT model [25]. This model comprises of baseline clinical characteristics, presence of secondary insults, imaging findings, and lab characteristics. Using the variables of this model, the current study aims to fairly assess the potential incremental value of flexible ML methods beyond classical regression approaches.

2. Methods

This study was reported conform the TRIPOD guidelines [23].

2.1. Study population

We included 15 studies from the IMPACT-II database. These include four observational studies and eleven randomized controlled trials on moderate to severe TBI (Glasgow Coma Scale [GCS] ≤ 12), which were conducted between 1984 and 2004 [26]. Furthermore, we validated models in the moderate to severe TBI patients (GCS ≤ 12) from the CENTER-TBI Core study. This is a recent prospective study, which included patients from 2014 to 2018 [27]. Data for the CENTER-TBI study has been collected through the Quesgen e-CRF (Quesgen Systems Inc, USA), hosted on the INCF platform and extracted via the INCF Neurobot tool (INCF, Sweden). Version 1.0 of the CENTER-TBI data was used for this analysis.

2.2. Model specification

The outcomes which were predicted were 6 months mortality and unfavourable outcome (Glasgow Outcome Scale < 3, or Glasgow Outcome Scale - Extended <5). The predictors included in the models were 11 predictors of the IMPACT laboratory model [25]. Continuous variables were included as continuous variables in the model (no categorization). An overview of the included variables, and their specifications, is shown in Table 1. The baseline GCS score was defined as the last GCS in the emergency department ("post-stabilization"). If this score was missing, the nearest GCS at an earlier moment was used. In total, eleven predictors were included, representing 19 parameters (or degrees of freedom [df]). In the case of mortality, 3491 events (or 184 events per parameter) were on average present in our database for each training. The variables were normalized or one-hot encoded, because this is standard practice for training algorithms which use gradient descent optimization.

Table 1, model specification: 11 predictors, with 19 degrees of freedom (df)

Variable in the model	Characteristics
Age	Continuous
Motor GCS score	Categorical, 1-6
Pupils	Categorical, 3 levels: - Both reactive - One reactive - Two reactive
CT class	Categorical, 5 levels: - No visible pathology - Diffuse injury - Diffuse injury with swelling - Diffuse injury with shift - Mass
Traumatic Subarachnoid Hemorrhage	Binary
Epidural hematoma	Binary
Hypoxia	Binary
Hypotension	Binary
Glucose, first measured	Continuous
Sodium, first measured	Continuous
Hemoglobin, first measured	Continuous

GCS = Glasgow Coma Scale; CT = computed tomography

2.3. Regression techniques

The regression techniques which were compared to the ML algorithms included standard logistic regression, but also penalized regression: lasso and ridge regression [28]. These algorithms were developed to improve the performance of logistic regression models by shrinking the coefficients during estimation [29,30]. The objective is to obtain models that are less prone to making too extreme predictions (overfitting). The *glmnet* function from the *glmnet* package was used (alpha=0 for ridge, and alpha=1 for lasso). No non-linear or interaction terms were included in the regression models.

2.4. Machine learning algorithms

All analyses were performed using R (R Core Team (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria). The script can be found on https://github.com/bgravesteijn/ML_baseline_pred_code.

The flexible ML algorithms that were compared to logistic regression were support vector machine, neural network, random forest, and gradient boosting machine. All these algorithms have so called "hyperparameters", that need to be optimized for the algorithms to work optimally. To select the optimal hyperparameters, the framework of the *caret* package was used. The best combination of hyperparameters of the algorithms were chosen based on the highest log-likelihood. The average log-likelihood over 10 repetitions of tenfold cross-validation was used to select the optimal parameters (figure 1). For a detailed description of what algorithms were used, and what hyperparameters were considered, see appendix B.

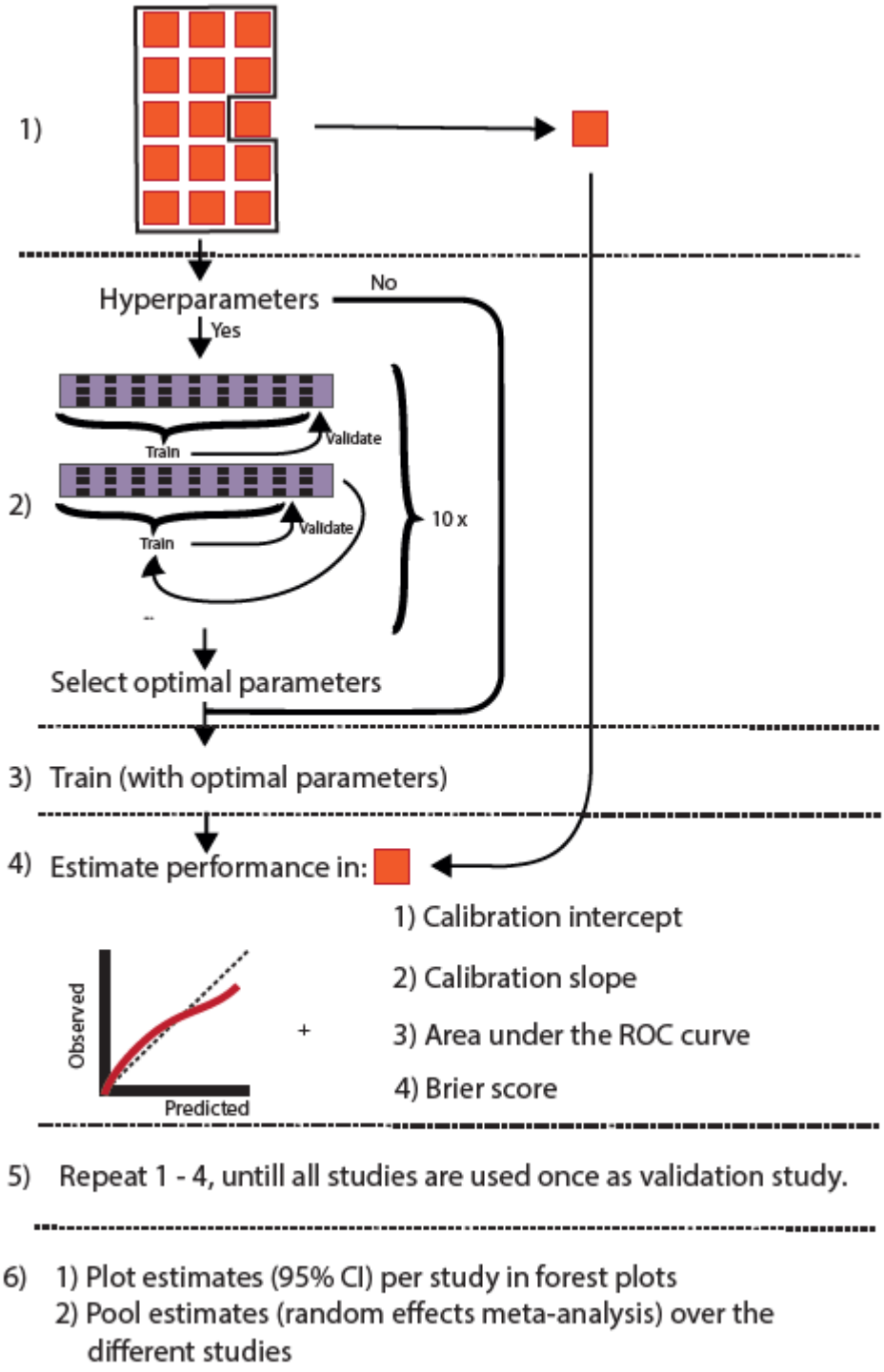


Fig. 1 Overview of the experimental setup. Step 1 is selecting a study as a validation study. Step 2 is selecting the optimal hyperparameters through 10 times 10-fold cross validation. If the algorithm did not require hyperparameters, this step was skipped. Step 3 is the training of the final model with optimal hyperparameters on the full training data. The model of step 3) was validated in step 4 with the study that was left out of the training set. Step 5 is repeating 1 – 4 until all studies are used once as validation study. Finally step 6 is the presentation of the results, and pooling the results over the different studies.

The included flexible ML methods, just like regression, do not allow for missing values. Unlike regression, however, they are not readily compatible with multiple imputation: not every algorithm uses weights as core operators. Moreover, for the algorithms that use weights, there is no implementation of pooling these weights over multiple datasets using Rubin's rules[31]. Therefore, multiple imputation using the *mice* package was performed[32], but only one imputed dataset was used to train the models. The outcome and all predictors were included in the imputation model. To check for stability of results, a sensitivity analysis was performed with a different imputed dataset.

2.5. Cross-validation

The models were validated using three different strategies. First, they were cross-validated per study: the algorithms were trained on all but one study, and calibration and discrimination were assessed by applying the models to the study not used at model development. This procedure has been referred to as 'Internal-external cross-validation' [33,34]. For an overview of the analytical steps of internal-external cross validation, see figure 1. Second, internal validation was performed in the IMPACT-II database using 10 times 10-fold random cross validation (10x10 CV). For this method, the data were randomly divided by deciles. The model was developed on 9/10, and validated on 1/10 of the data. This process was repeated until all patients were used once as validation sample. Finally, a fully external validation was performed, with training of the models in the IMPACT-II database, and validating in CENTER-TBI.

The performance was assessed in three domains. First, calibration was examined graphically and quantified using a calibration slope and the calibration intercept: the calibration test proposed by Cox [35]. Second, discrimination was quantified using the *c*-statistic, also known as area under the ROC curve. The confidence intervals of the c-statistic were obtained using the DeLong et al method [36], using the *ci.auc* function from the *pROC* package. Third, as a measure of overall performance, the Brier score was calculated [37]. More extensive descriptions of these metrics can be found in appendix B.

The estimates and 95% confidence intervals were plotted in forest plots, to visually inspect the variation. To obtain estimates per model and outcome, the estimates (and standard errors) in every validation were pooled using a random effects meta-analysis, using the DerSimonian and Laird estimator for τ^2 [38]. Since the CENTER-TBI database is a recent study, unlike the IMPACT-II studies, the estimates obtained from validating in this study were presented separately.

To compare whether observed variation of the performance measures can be attributed to differences in performance across study population or type of model used we used mixed effects linear regression. This was performed in the internal-external validation framework. The performance measure was used as dependent variable, and two random intercepts were included in the model: one for what algorithm was used and one for what study the models were validated in. These random intercepts were assumed to follow a normal distribution with mean 0 and variance τ^2. The percentage variation in performance attributable to in which study the model was validated was calculated by dividing the τ^2 of study by the total variance (the sum of the variance of the random intercepts of study and algorithm, and the residuals): $\tau^2_{study}/(\tau^2_{study}+\tau^2_{algorithm}+\tau^2_{residuals})$. Similarly, the percentage variation in performance attributable to what algorithm was trained, was calculated.

3. Results

3.1. Patient characteristics

The baseline characteristics differed substantially between the IMPACT-II and the CENTER-TBI data. In the IMPACT-II database, patients were younger (35 versus 47.4 years), had less traumatic subarachnoid hemorrhages (4016 [45%] versus 759 [74%]), and presented less often with a motor GCS of one (1565 [16%] versus 615 [45%]). However, the patients showed similar Glasgow Outcome Scale in the two studies: In the IMPACT-II database, 3332 (30%) died and 5233 (48%) had an unfavourable outcome, and in the CENTER-TBI study 348 (29%) died and 651 (54%) had unfavourable outcome (table 2). For an overview of the patient characteristics per study in IMPACT-II and CENTER-TBI, see table A1.

Table 2, baseline characteristics of the CENTER-TBI and IMPACT-II databases.

		IMPACT-II	CENTER-TBI	Missing data, total %
N		11022	1375	
Age (median [IQR])		31 [22, 46]	48 [28, 65]	0.0
Hypoxia (%)		1707 (22)	217 (16.8)	26.3
Hypotension (%)		1518 (17.2)	205 (15.9)	18.3
Marshall CT class (%)				40.6
	1	379 (5.9)	81 (8.3)	
	2	2281 (36)	428 (43.9)	
	3	1259 (20)	86 (8.8)	
	4	248 (3.9)	19 (2.0)	
	5	2223 (35)	360 (37.0)	
Traumatic subarachnoid hemorrhage (%)		4016 (44.6)	759 (73.6)	19.1
Epidural hematoma (%)		1275 (13.4)	172 (16.7)	14.8
Glucose (median mmol/l (SD))		8.84 (3.46)	8.18 (2.95)	44.5
Hemoglobin (mean g/dl (SD))		12.46 (2.42)	7.96 (2.36)	52.2
				7.4
GCS motor (%)				
	1	1565 (15.5)	615 (44.7)	
	2	1285 (12.7)	77 (5.6)	
	3	1362 (13.5)	80 (5.8)	
	4	2438 (24.1)	136 (9.9)	
	5	2791 (27.6)	357 (26.0)	
	6	658 (6.5)	110 (8.0)	
Pupil (%)				12.8
	Both reactive	6292 (66.3)	973 (73.7)	
	One reactive	1192 (12.6)	110 (8.3)	
	None reactive	2010 (21.2)	238 (18.0)	
Glasgow outcome scale (%)				1.4
	2	3322 (30.1)	348 (29.0)	
	3	1911 (17.3)	303 (25.2)	
	4	2262 (20.5)	246 (20.5)	

5	3527 (32.0)	303 (25.2)	

CT = computed tomography; GCS = Glasgow Coma Scale; SD = standard deviation; IQR = interquartile range

3.2. Discrimination

At internal-external validation, the difference between maximum and minimum c-statistic of the algorithms was only 0.02 for mortality and unfavourable outcome. The discriminatory performance of the implementation of random forest was suboptimal: the median and IQR of c-statistic of the random forest were 0.79 (0.77 – 0.82) for mortality (the overall average was 0.81) and 0.79 (0.76 – 0.81) for unfavourable outcome (the overall average was 0.80). The discriminative performances varied substantially per study (figure A2 and table 3). At internal validation in IMPACT-II, a similar pattern was seen, but the c-statistics were somewhat higher. For example the gradient boosting machine showed a c-statistic of 0.81 (0.79 – 0.83) at internal-external validation, and 0.83 (0.82 – 0.84) at internal validation. When performing external validation in CENTER-TBI, this pattern was also seen: The random forest showed a median and 95% CI for the c-statistic of 0.81 (0.78 - 0.84) for mortality (overall average was 0.82) and 0.76 (0.74 - 0.79) for unfavourable outcome (overall average was 0.77). Similar results were observed over a different imputed set, see table A5.

Table 3, results for discriminative performance of all algorithms, in all three validation strategies: Internal-external (per-study CV), internal (10-fold CV) and external (CENTER-TBI) validation. Estimates and 95% CI are shown.

Algorithm	Outcome	Internal-external	Internal	External
Logistic regression	Mortality	0.81 (0.79 - 0.84)	0.82 (0.81 - 0.83)	0.82 (0.79 - 0.84)
Support vector machine		0.81 (0.78 - 0.83)	0.82 (0.82 - 0.83)	0.81 (0.79 - 0.84)
Random forest		0.79 (0.77 - 0.82)	0.79 (0.78 - 0.81)	0.81 (0.78 - 0.84)
Neural network		0.81 (0.79 - 0.84)	0.82 (0.81 - 0.83)	0.82 (0.79 - 0.84)
Gradient boosting machine		0.81 (0.79 - 0.84)	0.83 (0.82 - 0.84)	0.83 (0.81 - 0.86)
Lasso regression		0.81 (0.79 - 0.84)	0.82 (0.82 - 0.83)	0.82 (0.79 - 0.84)
Ridge regression		0.81 (0.79 - 0.84)	0.82 (0.82 - 0.83)	0.82 (0.79 - 0.84)
Logistic regression	Unfavourable outcome	0.81 (0.79 - 0.83)	0.82 (0.81 - 0.82)	0.77 (0.75 - 0.80)
Support vector machine		0.80 (0.79 - 0.82)	0.81 (0.81 - 0.82)	0.78 (0.75 - 0.80)
Random forest		0.79 (0.76 - 0.81)	0.79 (0.78 - 0.80)	0.76 (0.74 - 0.79)
neural network		0.80 (0.79 - 0.82)	0.81 (0.81 - 0.82)	0.78 (0.76 - 0.80)
Gradient boosting machine		0.80 (0.78 - 0.82)	0.81 (0.80 - 0.82)	0.78 (0.76 - 0.80)
Lasso regression		0.81 (0.79 - 0.83)	0.81 (0.80 - 0.82)	0.77 (0.75 - 0.80)
Ridge regression		0.81 (0.79 - 0.83)	0.81 (0.80 - 0.82)	0.77 (0.75 - 0.80)

3.3. Calibration

At internal-external validation, the average calibration intercepts across the algorithms did not vary substantially: the range of calibration intercepts was -0.08 - -0.02 for mortality, and for unfavourable outcome, the calibration intercepts were 0.02 (figure 2B and table A2). The range of calibration slopes was larger: 0.85 - 1.05 for mortality and 0.89 - 1.06 for unfavourable outcome (figure 2C and table A3). The random forest made too extreme predictions, with a median (95% CI) calibration slope

of 0.85 (0.77 - 0.93) for mortality, while the overall mean was 0.97; and 0.89 (0.82 - 0.96) for unfavourable outcome, while the overall mean was 0.99. At internal validation in IMPACT-II, calibration slopes and intercepts were similar. In external validation in CENTER-TBI, the random forest had again a too low calibration slope (0.88, 95% CI: 0.77 – 0.99 for mortality).

The calibration intercept for mortality was generally low in CENTER-TBI: the overall mean was -0.58, indicating that the 6-month mortality was lower than expected in CENTER-TBI.

3.4. Overall predictive ability

The Brier score was very similar at internal-external validation, internal, and external validation for both outcomes (table A4). The brier score was somewhat higher at external validation, but consistent for all methods (e.g.: 0.19 versus 0.18 for logistic regression to predict unfavourable outcome).

3.5. Explained heterogeneity

At internal-external validation, the variation in c-statistic, calibration intercept, and Brier score was mainly attributable to the study in which the algorithm was validated (table 4): for mortality, the variation in c-statistic was for 97% attributable to the study in which the algorithm was validated (versus 2.0% to what algorithm was used); while the variation in calibration intercept was for 98% attributable to the study in which the algorithm was validated (versus 0.3% to what algorithm was used); and variation in Brier score was for 96% attributable to the study in which the algorithm was validated (versus 2.0% to what algorithm was used). Variation in calibration slope was slightly more attributable to what algorithm was used, compared to the other metrics (Figure A1). For mortality, the variation in calibration slope was for 11% attributable to the algorithm used, and 86% attributable to the study in which the algorithm was validated. This was mostly caused by the low calibration slope of the random forest algorithm. This algorithm displayed the worst calibration slope, as indicated in figure 2C. For unfavourable outcome, the results were similar.

Table 4, percentage of variation in performance attributable to what study the algorithms were validated in. An example is shown in the supplemental material (figure 1).

Outcome	C-statistic	Calibration intercept	Calibration slope	Brier score
Mortality				
Algorithm	2.0	0.3	11	2.0
Study	97	98	86	96
Unfavourable				
Algorithm	2.9	0.0	12	2.5
Study	96	99	85	97

3.6. Non-additivity and non-linearity

To explore whether non-additive and non-linear effects were frequently appropriate to assume in our data, we performed a post-hoc analysis. Per study, logistic regression models allowing for non-additivity and non-linearity were tested with likelihood ratio tests (omnibus tests) to the model which did not allow for relaxation of those assumptions [20]. It was observed that the model predicting mortality had a better fit when non-linearity was allowed for in 7 (44%) studies. Less often, the assumption of non-additivity improved the model fit (Table A6).

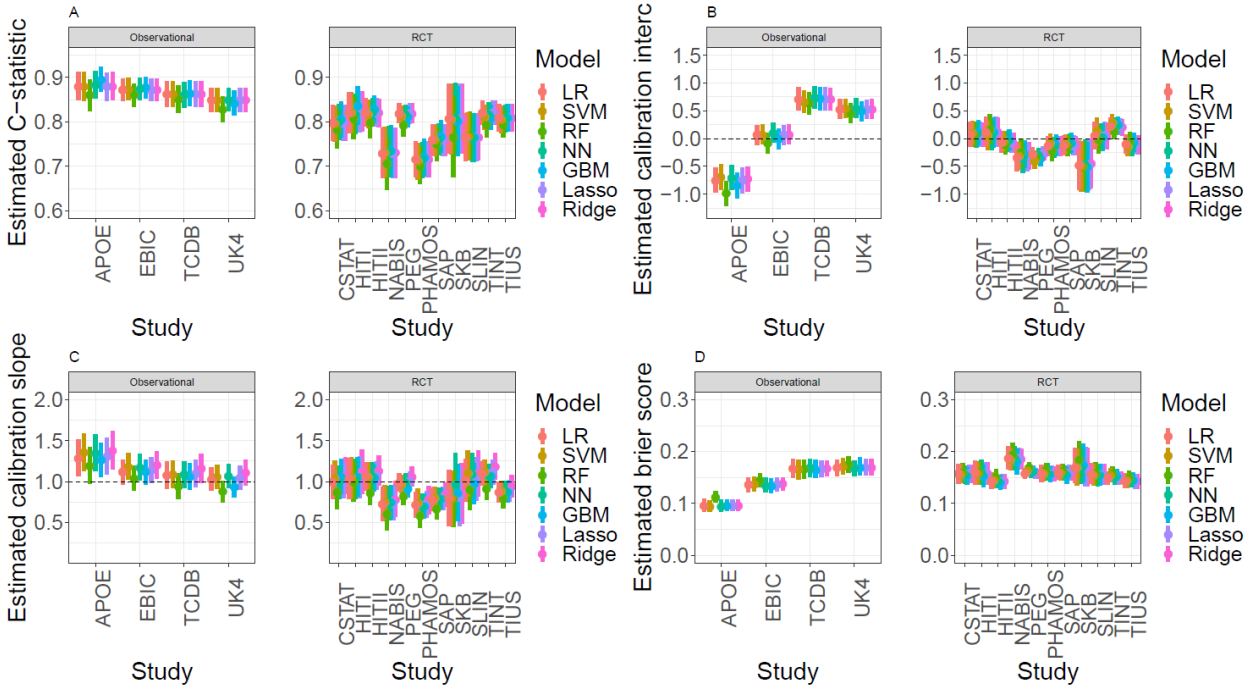


Fig. 2 Results the internal-external cross-validation for mortality. Panel A shows the results of the c-statistic/area under the ROC curve, panel B shows the calibration intercept, and panel C shows the calibration slope. The validation results are displayed per study (left: observational, right: randomized controlled trials), and per algorithm.

4. Discussion

This study aimed to compare flexible ML algorithms to more traditional logistic regression in contemporary patient data. We trained the algorithms to obtain a model with both high discrimination and good calibration. This was achieved by optimizing the log-likelihood for both regression and ML algorithms. All models and algorithms were developed and validated in large datasets, including the recent prospective cohort study CENTER-TBI [27]. Performance was assessed in terms of both discrimination and calibration, which are both important characteristics to be assessed in algorithm validation [22,24,39]. Similar performance of most methods was found across a large number of studies from different time periods.

The algorithm that relatively underperformed was the random forest: the discrimination was somewhat lower, but it clearly underperformed in terms of calibration. In particular, the random forest showed a calibration slope that was far below one. This indicates overfitting, a problem often arising in small datasets [37]. According to theoretical arguments, the RF algorithm should not overfit [40]. The discrepancy between the theory and the empirical evidence of our study should be explored further. There could be a role for the selection of hyperparameters, in particular the number of random variables at the split, and the fraction of observations in the training sample [41]. Since the random forest shows signs of overfitting, even in large datasets, the discriminative performance should be interpreted with caution: due to optimism, the discrimination in new datasets can be lower [21]. As a contrast, this method was one of the better performing methods in other studies [15,42], which however did not assess calibration. Since calibration is a crucial step

before implementation of a prediction model in clinical practice [20,39,43], our study encourages the use of other modeling techniques than random forests for outcome prediction.

The variation in observed performance was more explained by the cohorts where the algorithms were validated, than by which algorithms was used. This implies that prediction models need continuous updating and validation, because their performance is often worse in new cohorts[44]. This is a limitation which needs to be addressed, in order to effectively use these models in clinical practice [45]. This finding does raise concerns about the validity of individual patient data meta-analysis in the context of prediction modeling.

A recent systematic review compared flexible ML methods to traditional statistical techniques in relatively small datasets (median sample size was 1250), and did not find incremental value [14]. This was perhaps to be expected, since modern ML methods are known to be data hungry compared to classical statistical techniques [13,46]. However, due to the increased sharing of data, international collaborations, and the availability of data from electronical health records and other datasets with routinely collected data, datasets are becoming increasingly large [47–49]. Our study shows that in this situation, flexible ML methods are not improving outcome prognostication as well.

A limitation of our study is that we only used a linear kernel function of support vector machine. Other kernels could have increased the performance of the algorithm, since the performance of the algorithm is substantially dependent on its hyperparameters [41]. Unfortunately, the computation time increased drastically when this kernel was implemented (the expected running time for one series of cross-validation was 21 days). Since the first six iterations did not show substantial increase in discriminative performance, we decided to use the linear basis function instead.

Second, we only considered a relatively small number of predictors (11 predictors, with 19 df). The reason for not including more predictors is that there were no other common data elements between all databases. This potentially limits the performance of ML techniques, since it has been suggested that flexible ML techniques perform better than traditional regression techniques when a large number of predictors are being considered, i.e. high-dimensional data [50,51]. A reason for such presumed superiority is the flexibility of these algorithms, enabling them to capture complex non-linear and interaction effects. It should be noted that regression-based techniques can also be extended by non-linear and interaction effects [20]. Given that ML algorithms did not outperform regression, these effects are not likely to be essential in the field of outcome prediction in TBI patients. Our study was not able to fully utilize the potential benefit of multidimensional data, because of a phenomenon that is expected in big data research: larger volumes of data for better models may come at the price of less detailed or lower quality data.

We do believe that although we could perhaps not utilize the full potential performance of ML algorithms, our comparison is just as relevant. Published machine learning based prediction algorithms often include a large number of predictors, sometimes with the suggestion to result in high discriminative performance [52,53]. We note that external validation of these high-dimensional prediction algorithms is challenging, since availability of predictors may differ from one setting to the other. For prediction with genomics data, this may be feasible if sufficient standardization and harmonization was performed [54]. However, clinical variables often have different definitions, notations, or units, which complicate the validation procedure with a large number (say n>50) of predictors. External validation remains an essential step before implementing prediction algorithms

in clinical practice. To train and validate high-dimensional data, a sophisticated IT environment is necessary [55]. Therefore, we believe that the low-dimensional setting, such as our study, might be more relevant for clinical practice, also for the near future. Powerful predictions for outcome after TBI can apparently be made with linear effects which are captured with simple algorithms.

Finally, this study should be replicated in other fields than TBI to ensure the generalizability of our findings, again from a largely neutral perspective [54]. Preferably, a wide range of studies should be used, representing different settings in terms of study design (RCTs vs observational), geography (different countries), types of centers (level I trauma centers vs other), etc. Most studies that compared algorithms used only one or a limited number of study populations [15–19]. Since the performance heavily relies on the study population, comparing the methods in multiple populations is recommended.

4.1. Conclusion

In a low-dimensional setting, flexible machine learning algorithms do not perform better than more traditional regression models in outcome prediction after moderate or severe TBI. This is potentially explained by the most important prognostic effects acting as independent, linear effects. Predictive performance is more dependent on the population in which the model is applied, than the type of algorithm used. This finding has strong implications: continuous validation and updating of prediction models is necessary to ensure applicability to new populations of both machine learning algorithms and regression-based models. To improve prognostication for TBI, future studies should extend current prognostic models with new predictors (biomarkers, imaging, genomics) with strong incremental value, for the reliable identification of patients with poor versus good prognosis.

List of abbreviations

TBI	Traumatic brain injury
ML	Machine learning
SVM	Support vector machine
GBM	Gradient boosting machine
NN	Neural network
RF	Random Forest
LR	Logistic regression
GCS	Glasgow coma scale

Declarations

Ethics approval and consent to participate

The authors declare that all participants signed informed consent to be included in the study. Ethical approval was obtained for each recruiting sites.

Consent for publication

The authors declare that approval for publication was obtained.

Availability of data and material

As a EU funded project, CENTER-TBI is an open-access database. Access can be obtained as collaborator, after declaring to adhere to the CENTER-TBI data use agreement. For more information, see https://www.center-tbi.eu/data.

Competing interests

The authors declare to have no competing interests.

Funding

Data used in preparation of this manuscript were obtained in the context of CENTER-TBI, a large collaborative project with the support of the European Union 7th Framework program (EC grant 602150).

The funder had no role in the study design, enrolment, collection of data, writing or publication decisions.

Trial registration

ClinicalTrials.gov Identifier: NCT02210221

Authors' contributions

	Benjamin Y. Gravesteijn	Daan Nieboer	Ari Ercole	Hester F. Lingsma	David Nelson	Ben van Calster	Ewout W. Steyerberg
Conceptualization	X	x				x	x
Data curation	X	X					
Formal analysis	x	x					
Funding acquisition			x	x			X
Investigation				x			X
Methodology	x	x					X
Project administration				X			x
Resources				X			
Software							
Supervision		X		X			x
Validation	X						
Visualization	X						
Writing – original draft	x						
Writing – review & editing	x	x	x	x	x	x	x

Acknowledgements

The CENTER-TBI participants and investigators:

Cecilia Åkerlund[1], Krisztina Amrein[2], Nada Andelic[3], Lasse Andreassen[4], Audny Anke[5], Anna Antoni[6], Gérard Audibert[7], Philippe Azouvi[8], Maria Luisa Azzolini[9], Ronald Bartels[10], Pál Barzó[11], Romuald Beauvais[12], Ronny Beer[13], Bo-Michael Bellander[14], Antonio Belli[15], Habib Benali[16], Maurizio Berardino[17], Luigi Beretta[9], Morten Blaabjerg[18], Peter Bragge[19], Alexandra Brazinova[20], Vibeke Brinck[21], Joanne Brooker[22], Camilla Brorsson[23], Andras Buki[24], Monika Bullinger[25], Manuel Cabeleira[26], Alessio Caccioppola[27], Emiliana Calappi [27], Maria Rosa Calvi[9], Peter Cameron[28], Guillermo Carbayo Lozano[29], Marco Carbonara[27], Giorgio Chevallard[30], Arturo Chieregato[30], Giuseppe Citerio[31, 32], Maryse Cnossen[33], Mark Coburn[34], Jonathan Coles[35], D. Jamie Cooper[36], Marta Correia[37], Amra Čović [38], Nicola Curry[39], Endre Czeiter[24], Marek Czosnyka[26], Claire Dahyot-Fizelier[40], Helen Dawes[41], Véronique De Keyser[42], Vincent Degos[16], Francesco Della Corte[43], Hugo den Boogert[10], Bart Depreitere[44], Đula Đilvesi [45], Abhishek Dixit[46], Emma Donoghue[22], Jens Dreier[47], Guy-Loup Dulière[48], Ari Ercole[46], Patrick Esser[41], Erzsébet Ezer[49], Martin Fabricius[50], Valery L. Feigin[51], Kelly Foks[52], Shirin Frisvold[53], Alex Furmanov[54], Pablo Gagliardo[55], Damien Galanaud[16], Dashiell Gantner[28], Guoyi Gao[56], Pradeep George[57], Alexandre Ghuysen[58], Lelde Giga[59], Ben Glocker[60], Jagoš Golubovic[45], Pedro A. Gomez [61], Johannes Gratz[62], Benjamin Gravesteijn[33], Francesca Grossi[43], Russell L. Gruen[63], Deepak Gupta[64], Juanita A. Haagsma[33], Iain Haitsma[65], Raimund Helbok[13], Eirik Helseth[66], Lindsay Horton [67], Jilske Huijben[33], Peter J. Hutchinson[68], Bram Jacobs[69], Stefan Jankowski[70], Mike Jarrett[21], Ji-yao Jiang[56], Kelly Jones[51], Mladen Karan[47], Angelos G. Kolias[68], Erwin Kompanje[71], Daniel Kondziella[50], Evgenios Koraropoulos[46], Lars-Owe Koskinen[72], Noémi Kovács[73], Alfonso Lagares[61], Linda Lanyon[57], Steven Laureys[74], Fiona Lecky[75], Rolf Lefering[76], Valerie Legrand[77], Aurelie Lejeune[78], Leon Levi[79], Roger Lightfoot[80], Hester Lingsma[33], Andrew I.R. Maas[42], Ana M. Castaño-León[61], Marc Maegele[81], Marek Majdan[20], Alex Manara[82], Geoffrey Manley[83], Costanza Martino[84], Hugues Maréchal[48], Julia Mattern[85], Catherine McMahon[86], Béla Melegh[87], David Menon[46], Tomas Menovsky[42], Davide Mulazzi[27], Visakh Muraleedharan[57], Lynnette Murray[28], Nandesh Nair[42], Ancuta Negru[88], David Nelson[1], Virginia Newcombe[46], Daan Nieboer[33], Quentin Noirhomme[74], József Nyirádi[2], Otesile Olubukola[75], Matej Oresic[89], Fabrizio Ortolano[27], Aarno Palotie[90, 91, 92], Paul M. Parizel[93], Jean-François Payen[94], Natascha Perera[12], Vincent Perlbarg[16], Paolo Persona[95], Wilco Peul[96], Anna Piippo-Karjalainen[97], Matti Pirinen[90], Horia Ples[88], Suzanne Polinder[33], Inigo Pomposo[29], Jussi P. Posti [98], Louis Puybasset[99], Andreea Radoi [100], Arminas Ragauskas[101], Rahul Raj[97], Malinka Rambadagalla[102], Ruben Real[38], Jonathan Rhodes[103], Sylvia Richardson[104], Sophie Richter[46], Samuli Ripatti[90], Saulius Rocka[101], Cecilie Roe[105], Olav Roise[106 140], Jonathan Rosand[107], Jeffrey V. Rosenfeld[108], Christina Rosenlund[109], Guy Rosenthal[54], Rolf Rossaint[34], Sandra Rossi[95], Daniel Rueckert[60], Martin Rusnák[110], Juan Sahuquillo[100], Oliver Sakowitz[85, 111], Renan Sanchez-Porras[111], Janos Sandor[112], Nadine Schäfer[76], Silke Schmidt[113], Herbert Schoechl[114], Guus Schoonman[115], Rico Frederik Schou[116], Elisabeth Schwendenwein[6], Charlie Sewalt[33], Toril Skandsen[117, 118] , Peter Smielewski[26],

Abayomi Sorinola[119], Emmanuel Stamatakis[46], Simon Stanworth[39], Ana Kowark[34], Robert Stevens[120], William Stewart[121], Ewout W. Steyerberg[33, 122], Nino Stocchetti[123], Nina Sundström[124], Anneliese Synnot[22, 125], Riikka Takala[126], Viktória Tamás[119], Tomas Tamosuitis[127], Mark Steven Taylor[20], Braden Te Ao[51], Olli Tenovuo[98], Alice Theadom[51], Matt Thomas[82], Dick Tibboel[128], Marjolein Timmers[71], Christos Tolias[129], Tony Trapani[28], Cristina Maria Tudora[88], Peter Vajkoczy [130], Shirley Vallance[28], Egils Valeinis[59], Zoltán Vámos[49], Gregory Van der Steen[42], Joukje van der Naalt[69], Jeroen T.J.M. van Dijck [96], Thomas A. van Essen[96], Wim Van Hecke[131], Caroline van Heugten[132], Dominique Van Praag[133], Thijs Vande Vyvere[131], Audrey Vanhaudenhuyse[16, 74], Roel P. J. van Wijk[97], Alessia Vargiolu[32], Emmanuel Vega[79], Kimberley Velt[33], Jan Verheyden[131], Paul M. Vespa[134], Anne Vik[117, 135], Rimantas Vilcinis[127], Victor Volovici[65], Nicole von Steinbüchel[38], Daphne Voormolen[33], Petar Vulekovic[45], Kevin K.W. Wang[136], Eveline Wiegers[33], Guy Williams[46], Lindsay Wilson[67], Stefan Winzeck[46], Stefan Wolf[137], Zhihui Yang[136], Peter Ylén[138], Alexander Younsi[85], Frederik A. Zeiler[46,139], Veronika Zelinkova[20], Agate Ziverte[59] , Tommaso Zoerle[27]

[1] Department of Physiology and Pharmacology, Section of Perioperative Medicine and Intensive Care, Karolinska Institutet, Stockholm, Sweden
[2] János Szentágothai Research Centre, University of Pécs, Pécs, Hungary
[3] Division of Surgery and Clinical Neuroscience, Department of Physical Medicine and Rehabilitation, Oslo University Hospital and University of Oslo, Oslo, Norway
[4] Department of Neurosurgery, University Hospital Northern Norway, Tromso, Norway
[5] Department of Physical Medicine and Rehabilitation, University Hospital Northern Norway, Tromso, Norway
[6] Trauma Surgery, Medical University Vienna, Vienna, Austria
[7] Department of Anesthesiology & Intensive Care, University Hospital Nancy, Nancy, France
[8] Raymond Poincare hospital, Assistance Publique – Hopitaux de Paris, Paris, France
[9] Department of Anesthesiology & Intensive Care, S Raffaele University Hospital, Milan, Italy
[10] Department of Neurosurgery, Radboud University Medical Center, Nijmegen, The Netherlands
[11] Department of Neurosurgery, University of Szeged, Szeged, Hungary
[12] International Projects Management, ARTTIC, Munchen, Germany
[13] Department of Neurology, Neurological Intensive Care Unit, Medical University of Innsbruck, Innsbruck, Austria
[14] Department of Neurosurgery & Anesthesia & intensive care medicine, Karolinska University Hospital, Stockholm, Sweden
[15] NIHR Surgical Reconstruction and Microbiology Research Centre, Birmingham, UK
[16] Anesthesie-Réanimation, Assistance Publique – Hopitaux de Paris, Paris, France
[17] Department of Anesthesia & ICU, AOU Città della Salute e della Scienza di Torino - Orthopedic and Trauma Center, Torino, Italy
[18] Department of Neurology, Odense University Hospital, Odense, Denmark
[19] BehaviourWorks Australia, Monash Sustainability Institute, Monash University, Victoria, Australia
[20] Department of Public Health, Faculty of Health Sciences and Social Work, Trnava University, Trnava, Slovakia
[21] Quesgen Systems Inc., Burlingame, California, USA
[22] Australian & New Zealand Intensive Care Research Centre, Department of Epidemiology and Preventive Medicine, School of Public Health and Preventive Medicine, Monash University, Melbourne, Australia
[23] Department of Surgery and Perioperative Science, Umeå University, Umeå, Sweden
[24] Department of Neurosurgery, Medical School, University of Pécs, Hungary and Neurotrauma Research Group, János Szentágothai Research Centre, University of Pécs, Hungary
[25] Department of Medical Psychology, Universitätsklinikum Hamburg-Eppendorf, Hamburg, Germany
[26] Brain Physics Lab, Division of Neurosurgery, Dept of Clinical Neurosciences, University of Cambridge, Addenbrooke's Hospital, Cambridge, UK
[27] Neuro ICU, Fondazione IRCCS Cà Granda Ospedale Maggiore Policlinico, Milan, Italy
[28] ANZIC Research Centre, Monash University, Department of Epidemiology and Preventive Medicine, Melbourne, Victoria, Australia
[29] Department of Neurosurgery, Hospital of Cruces, Bilbao, Spain
[30] NeuroIntensive Care, Niguarda Hospital, Milan, Italy
[31] School of Medicine and Surgery, Università Milano Bicocca, Milano, Italy

[32] NeuroIntensive Care, ASST di Monza, Monza, Italy
[33] Department of Public Health, Erasmus Medical Center-University Medical Center, Rotterdam, The Netherlands
[34] Department of Anaesthesiology, University Hospital of Aachen, Aachen, Germany
[35] Department of Anesthesia & Neurointensive Care, Cambridge University Hospital NHS Foundation Trust, Cambridge, UK
[36] School of Public Health & PM, Monash University and The Alfred Hospital, Melbourne, Victoria, Australia
[37] Radiology/MRI department, MRC Cognition and Brain Sciences Unit, Cambridge, UK
[38] Institute of Medical Psychology and Medical Sociology, Universitätsmedizin Göttingen, Göttingen, Germany
[39] Oxford University Hospitals NHS Trust, Oxford, UK
[40] Intensive Care Unit, CHU Poitiers, Potiers, France
[41] Movement Science Group, Faculty of Health and Life Sciences, Oxford Brookes University, Oxford, UK
[42] Department of Neurosurgery, Antwerp University Hospital and University of Antwerp, Edegem, Belgium
[43] Department of Anesthesia & Intensive Care, Maggiore Della Carità Hospital, Novara, Italy
[44] Department of Neurosurgery, University Hospitals Leuven, Leuven, Belgium
[45] Department of Neurosurgery, Clinical centre of Vojvodina, Faculty of Medicine, University of Novi Sad, Novi Sad, Serbia
[46] Division of Anaesthesia, University of Cambridge, Addenbrooke's Hospital, Cambridge, UK
[47] Center for Stroke Research Berlin, Charité – Universitätsmedizin Berlin, corporate member of Freie Universität Berlin, Humboldt-Universität zu Berlin, and Berlin Institute of Health, Berlin, Germany
[48] Intensive Care Unit, CHR Citadelle, Liège, Belgium
[49] Department of Anaesthesiology and Intensive Therapy, University of Pécs, Pécs, Hungary
[50] Departments of Neurology, Clinical Neurophysiology and Neuroanesthesiology, Region Hovedstaden Rigshospitalet, Copenhagen, Denmark
[51] National Institute for Stroke and Applied Neurosciences, Faculty of Health and Environmental Studies, Auckland University of Technology, Auckland, New Zealand
[52] Department of Neurology, Erasmus MC, Rotterdam, the Netherlands
[53] Department of Anesthesiology and Intensive care, University Hospital Northern Norway, Tromso, Norway
[54] Department of Neurosurgery, Hadassah-hebrew University Medical center, Jerusalem, Israel
[55] Fundación Instituto Valenciano de Neurorrehabilitación (FIVAN), Valencia, Spain
[56] Department of Neurosurgery, Shanghai Renji hospital, Shanghai Jiaotong University/school of medicine, Shanghai, China
[57] Karolinska Institutet, INCF International Neuroinformatics Coordinating Facility, Stockholm, Sweden
[58] Emergency Department, CHU, Liège, Belgium
[59] Neurosurgery clinic, Pauls Stradins Clinical University Hospital, Riga, Latvia
[60] Department of Computing, Imperial College London, London, UK
[61] Department of Neurosurgery, Hospital Universitario 12 de Octubre, Madrid, Spain
[62] Department of Anesthesia, Critical Care and Pain Medicine, Medical University of Vienna, Austria

[63] College of Health and Medicine, Australian National University, Canberra, Australia
[64] Department of Neurosurgery, Neurosciences Centre & JPN Apex trauma centre, All India Institute of Medical Sciences, New Delhi-110029, India
[65] Department of Neurosurgery, Erasmus MC, Rotterdam, the Netherlands
[66] Department of Neurosurgery, Oslo University Hospital, Oslo, Norway
[67] Division of Psychology, University of Stirling, Stirling, UK
[68] Division of Neurosurgery, Department of Clinical Neurosciences, Addenbrooke's Hospital & University of Cambridge, Cambridge, UK
[69] Department of Neurology, University of Groningen, University Medical Center Groningen, Groningen, Netherlands
[70] Neurointensive Care , Sheffield Teaching Hospitals NHS Foundation Trust, Sheffield, UK
[71] Department of Intensive Care and Department of Ethics and Philosophy of Medicine, Erasmus Medical Center, Rotterdam, The Netherlands
[72] Department of Clinical Neuroscience, Neurosurgery, Umeå University, Umeå, Sweden
[73] Hungarian Brain Research Program - Grant No. KTIA_13_NAP-A-II/8, University of Pécs, Pécs, Hungary
[74] Cyclotron Research Center , University of Liège, Liège, Belgium
[75] Emergency Medicine Research in Sheffield, Health Services Research Section, School of Health and Related Research (ScHARR), University of Sheffield, Sheffield, UK
[76] Institute of Research in Operative Medicine (IFOM), Witten/Herdecke University, Cologne, Germany
[77] VP Global Project Management CNS, ICON, Paris, France
[78] Department of Anesthesiology-Intensive Care, Lille University Hospital, Lille, France
[79] Department of Neurosurgery, Rambam Medical Center, Haifa, Israel
[80] Department of Anesthesiology & Intensive Care, University Hospitals Southhampton NHS Trust, Southhampton, UK
[81] Cologne-Merheim Medical Center (CMMC), Department of Traumatology, Orthopedic Surgery and Sportmedicine, Witten/Herdecke University, Cologne, Germany
[82] Intensive Care Unit, Southmead Hospital, Bristol, Bristol, UK
[83] Department of Neurological Surgery, University of California, San Francisco, California, USA
[84] Department of Anesthesia & Intensive Care,M. Bufalini Hospital, Cesena, Italy
[85] Department of Neurosurgery, University Hospital Heidelberg, Heidelberg, Germany
[86] Department of Neurosurgery, The Walton centre NHS Foundation Trust, Liverpool, UK
[87] Department of Medical Genetics, University of Pécs, Pécs, Hungary
[88] Department of Neurosurgery, Emergency County Hospital Timisoara , Timisoara, Romania
[89] School of Medical Sciences, Örebro University, Örebro, Sweden
[90] Institute for Molecular Medicine Finland, University of Helsinki, Helsinki, Finland
[91] Analytic and Translational Genetics Unit, Department of Medicine; Psychiatric & Neurodevelopmental Genetics Unit, Department of Psychiatry; Department of Neurology, Massachusetts General Hospital, Boston, MA, USA
[92] Program in Medical and Population Genetics; The Stanley Center for Psychiatric Research, The Broad Institute of MIT and Harvard, Cambridge, MA, USA
[93] Department of Radiology, Antwerp University Hospital and University of Antwerp, Edegem, Belgium
[94] Department of Anesthesiology & Intensive Care, University Hospital of Grenoble, Grenoble, France

[95] Department of Anesthesia & Intensive Care, Azienda Ospedaliera Università di Padova, Padova, Italy
[96] Dept. of Neurosurgery, Leiden University Medical Center, Leiden, The Netherlands and Dept. of Neurosurgery, Medical Center Haaglanden, The Hague, The Netherlands
[97] Department of Neurosurgery, Helsinki University Central Hospital
[98] Division of Clinical Neurosciences, Department of Neurosurgery and Turku Brain Injury Centre, Turku University Hospital and University of Turku, Turku, Finland
[99] Department of Anesthesiology and Critical Care, Pitié -Salpêtrière Teaching Hospital, Assistance Publique, Hôpitaux de Paris and University Pierre et Marie Curie, Paris, France
[100] Neurotraumatology and Neurosurgery Research Unit (UNINN), Vall d'Hebron Research Institute, Barcelona, Spain
[101] Department of Neurosurgery, Kaunas University of technology and Vilnius University, Vilnius, Lithuania
[102] Department of Neurosurgery, Rezekne Hospital, Latvia
[103] Department of Anaesthesia, Critical Care & Pain Medicine NHS Lothian & University of Edinburg, Edinburgh, UK
[104] Director, MRC Biostatistics Unit, Cambridge Institute of Public Health, Cambridge, UK
[105] Department of Physical Medicine and Rehabilitation, Oslo University Hospital/University of Oslo, Oslo, Norway
[106] Division of Orthopedics, Oslo University Hospital
[107] Broad Institute, Cambridge MA Harvard Medical School, Boston MA, Massachusetts General Hospital, Boston MA, USA
[108] National Trauma Research Institute, The Alfred Hospital, Monash University, Melbourne, Victoria, Australia
[109] Department of Neurosurgery, Odense University Hospital, Odense, Denmark
[110] International Neurotrauma Research Organisation, Vienna, Austria
[111] Klinik für Neurochirurgie, Klinikum Ludwigsburg, Ludwigsburg, Germany
[112] Division of Biostatistics and Epidemiology, Department of Preventive Medicine, University of Debrecen, Debrecen, Hungary
[113] Department Health and Prevention, University Greifswald, Greifswald, Germany
[114] Department of Anaesthesiology and Intensive Care, AUVA Trauma Hospital, Salzburg, Austria
[115] Department of Neurology, Elisabeth-TweeSteden Ziekenhuis, Tilburg, the Netherlands
[116] Department of Neuroanesthesia and Neurointensive Care, Odense University Hospital, Odense, Denmark
[117] Department of Neuromedicine and Movement Science, Norwegian University of Science and Technology, NTNU, Trondheim, Norway
[118] Department of Physical Medicine and Rehabilitation, St.Olavs Hospital, Trondheim University Hospital, Trondheim, Norway
[119] Department of Neurosurgery, University of Pécs, Pécs, Hungary
[120] Division of Neuroscience Critical Care, John Hopkins University School of Medicine, Baltimore, USA
[121] Department of Neuropathology, Queen Elizabeth University Hospital and University of Glasgow, Glasgow, UK
[122] Dept. of Department of Biomedical Data Sciences, Leiden University Medical Center, Leiden, The Netherlands

[123] Department of Pathophysiology and Transplantation, Milan University, and Neuroscience ICU, Fondazione IRCCS Cà Granda Ospedale Maggiore Policlinico, Milano, Italy
[124] Department of Radiation Sciences, Biomedical Engineering, Umeå University, Umeå, Sweden
[125] Cochrane Consumers and Communication Review Group, Centre for Health Communication and Participation, School of Psychology and Public Health, La Trobe University, Melbourne, Australia
[126] Perioperative Services, Intensive Care Medicine and Pain Management, Turku University Hospital and University of Turku, Turku, Finland
[127] Department of Neurosurgery, Kaunas University of Health Sciences, Kaunas, Lithuania
[128] Intensive Care and Department of Pediatric Surgery, Erasmus Medical Center, Sophia Children's Hospital, Rotterdam, The Netherlands
[129] Department of Neurosurgery, Kings college London, London, UK
[130] Neurologie, Neurochirurgie und Psychiatrie, Charité – Universitätsmedizin Berlin, Berlin, Germany
[131] icoMetrix NV, Leuven, Belgium
[132] Movement Science Group, Faculty of Health and Life Sciences, Oxford Brookes University, Oxford, UK
[133] Psychology Department, Antwerp University Hospital, Edegem, Belgium
[134] Director of Neurocritical Care, University of California, Los Angeles, USA
[135] Department of Neurosurgery, St.Olavs Hospital, Trondheim University Hospital, Trondheim, Norway
[136] Department of Emergency Medicine, University of Florida, Gainesville, Florida, USA
[137] Department of Neurosurgery, Charité – Universitätsmedizin Berlin, corporate member of Freie Universität Berlin, Humboldt-Universität zu Berlin, and Berlin Institute of Health, Berlin, Germany
[138] VTT Technical Research Centre, Tampere, Finland
[139] Section of Neurosurgery, Department of Surgery, Rady Faculty of Health Sciences, University of Manitoba, Winnipeg, MB, Canada
[140] Institute of Clinical Medicine, Faculty of Medicine, University of Oslo

References

[1] Maas AIR, Menon DK, Adelson PD, Andelic N, Bell MJ, Belli A, et al. Traumatic brain injury: integrated approaches to improve prevention, clinical care, and research. Lancet Neurol 2017;4422. doi:10.1016/S1474-4422(17)30371-X.

[2] Majdan M, Plancikova D, Brazinova A, Rusnak M, Nieboer D, Feigin V, et al. Epidemiology of traumatic brain injuries in Europe: a cross-sectional analysis. Lancet Public Heal 2016;1:e76–83. doi:10.1016/S2468-2667(16)30017-2.

[3] Saatman KE, Duhaime A, Bullock R, Maas AI, Valadka A, Manley GT, et al. Classification of Traumatic Brain Injury for Targeted Therapies 2008. doi:10.1089/neu.2008.0586.

[4] Lingsma HF, Roozenbeek B, Steyerberg EW, Murray GD, Maas AI. Early prognosis in traumatic brain injury: from prophecies to predictions. Lancet Neurol 2010;9:543–54. doi:10.1016/S1474-4422(10)70065-X.

[5] Liu NT, Salinas J. Machine Learning for Predicting Outcomes in Trauma. SHOCK 2017;48:504–10. doi:10.1097/SHK.0000000000000898.

[6] Burges CJC. A Tutorial on Support Vector Machines for Pattern Recognition. vol. 2. 1998.

[7] Jain AK, Jianchang Mao, Mohiuddin KM. Artificial neural networks: a tutorial. Computer (Long Beach Calif) 1996;29:31–44. doi:10.1109/2.485891.

[8] Afanador NL, Smolinska A, Tran TN, Blanchet L. Unsupervised random forest: a tutorial with case studies. J Chemom 2016;30:232–41. doi:10.1002/cem.2790.

[9] Natekin A, Knoll A. Gradient boosting machines, a tutorial. Front Neurorobot 2013;7:21. doi:10.3389/fnbot.2013.00021.

[10] Rau C-S, Kuo P-J, Chien P-C, Huang C-Y, Hsieh H-Y, Hsieh C-H. Mortality prediction in patients with isolated moderate and severe traumatic brain injury using machine learning models. PLoS One 2018;13:e0207192. doi:10.1371/journal.pone.0207192.

[11] Matsuo K, Aihara H, Nakai T, Morishita A, Tohma Y, Kohmura E. Machine Learning to Predict In-Hospital Morbidity and Mortality after Traumatic Brain Injury. J Neurotrauma 2019:neu.2018.6276. doi:10.1089/neu.2018.6276.

[12] Feng J, Wang Y, Peng J, Sun M, Zeng J, Jiang H. Comparison between logistic regression and machine learning algorithms on survival prediction of traumatic brain injuries. J Crit Care 2019;54:110–6. doi:10.1016/j.jcrc.2019.08.010.

[13] van der Ploeg T, Austin PC, Steyerberg EW. Modern modelling techniques are data hungry: a simulation study for predicting dichotomous endpoints. BMC Med Res Methodol 2014;14:137. doi:10.1186/1471-2288-14-137.

[14] Evangelia christodoulou, Jie MA, Collins GS, Steyerberg EW, Verbakel JY, van Calster B. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. J Clin Epidemiol 2019;0. doi:10.1016/j.jclinepi.2019.02.004.

[15] van Os HJA, Ramos LA, Hilbert A, van Leeuwen M, van Walderveen MAA, Kruyt ND, et al. Predicting Outcome of Endovascular Treatment for Acute Ischemic Stroke: Potential Value of Machine Learning Algorithms. Front Neurol 2018;9:784. doi:10.3389/fneur.2018.00784.

[16] Churpek MM, Yuen TC, Winslow C, Meltzer DO, Kattan MW, Edelson DP. Multicenter Comparison of Machine Learning Methods and Conventional Regression for Predicting Clinical Deterioration on the Wards. Crit Care Med 2016;44:368–74. doi:10.1097/CCM.0000000000001571.

[17] Lee H-C, Yoon S, Yang S-M, Kim W, Ryu H-G, Jung C-W, et al. Prediction of Acute Kidney Injury after Liver Transplantation: Machine Learning Approaches vs. Logistic Regression Model. J Clin Med 2018;7:428. doi:10.3390/jcm7110428.

[18] Bisaso KR, Karungi SA, Kiragga A, Mukonzo JK, Castelnuovo B. A comparative study of logistic regression based machine learning techniques for prediction of early virological suppression in antiretroviral initiating HIV patients. BMC Med Inform Decis Mak 2018;18:77. doi:10.1186/s12911-018-0659-x.

[19] Decruyenaere A, Decruyenaere P, Peeters P, Vermassen F, Dhaene T, Couckuyt I. Prediction of delayed graft function after kidney transplantation: comparison between logistic regression and machine learning methods. BMC Med Inform Decis Mak 2015;15:83. doi:10.1186/s12911-015-0206-y.

[20] Harrell FE. Regression Modeling Strategies. New York, NY: Springer New York; 2001.

doi:10.1007/978-1-4757-3462-1.

[21] Steyerberg EW. Clinical Prediction Models. New York, NY: Springer New York; 2009. doi:10.1007/978-0-387-77244-8.

[22] Van Calster B, Nieboer D, Vergouwe Y, De Cock B, Pencina MJ, Steyerberg EW. A calibration hierarchy for risk models was defined: from utopia to empirical data. J Clin Epidemiol 2016;74:167–76. doi:10.1016/J.JCLINEPI.2015.12.005.

[23] Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): The TRIPOD Statement. Ann Intern Med 2015;162:55. doi:10.7326/M14-0697.

[24] Moons KGM, Kengne AP, Grobbee DE, Royston P, Vergouwe Y, Altman DG, et al. Risk prediction models: II. External validation, model updating, and impact assessment. Heart 2012;98:691–8. doi:10.1136/heartjnl-2011-301247.

[25] Steyerberg EW, Mushkudiani N, Perel P, Butcher I, Lu J, McHugh GS, et al. Predicting Outcome after Traumatic Brain Injury: Development and International Validation of Prognostic Scores Based on Admission Characteristics. PLoS Med 2008;5:e165. doi:10.1371/journal.pmed.0050165.

[26] Marmarou A, Lu J, Butcher I, McHugh GS, Mushkudiani NA, Murray GD, et al. IMPACT Database of Traumatic Brain Injury: Design And Description. J Neurotrauma 2007;24:239–50. doi:10.1089/neu.2006.0036.

[27] Maas AIR, Menon DK, Steyerberg EW, Citerio G, Lecky F, Manley GT, et al. Collaborative European neurotrauma effectiveness research in traumatic brain injury (CENTER-TBI): A prospective longitudinal observational study. Neurosurgery 2015;76:67–80. doi:10.1227/NEU.0000000000000575.

[28] Steyerberg EW. Clinical Prediction Models: A Practical Approach to Development, Validation and Updating. New York: Springer; 2009.

[29] Tibshirani R. Regression Shrinkage and Selection Via the Lasso. J R Stat Soc Ser B 1996;58:267–88. doi:10.1111/j.2517-6161.1996.tb02080.x.

[30] FIRTH D. Bias reduction of maximum likelihood estimates. Biometrika 1993;80:27–38. doi:10.1093/biomet/80.1.27.

[31] Rubin DB. Multiple imputation for nonresponse in surveys. Wiley-Interscience; 2004.

[32] Buuren S van. Flexible imputation of missing data. CRC Press; 2018.

[33] Royston P, Parmar MKB, Sylvester R. Construction and validation of a prognostic model across several studies, with an application in superficial bladder cancer. Stat Med 2004;23:907–26. doi:10.1002/sim.1691.

[34] Steyerberg EW, Harrell FE. Prediction models need appropriate internal, internal-external, and external validation HHS Public Access. J Clin Epidemiol 2016;69:245–7. doi:10.1016/j.jclinepi.2015.04.005.

[35] Cox D. Two further applications of a model for binary regression. Biometrika 1958;45:562–5.

[36] DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics

1988;44:837–45.

[37] Steyerberg EW, Vickers AJ, Cook NR, Gerds T, Gonen M, Obuchowski N, et al. Assessing the performance of prediction models: a framework for traditional and novel measures. Epidemiology 2010;21:128–38. doi:10.1097/EDE.0b013e3181c30fb2.

[38] DerSimonian R, Laird N. Meta-analysis in clinical trials. Control Clin Trials 1986;7:177–88. doi:10.1016/0197-2456(86)90046-2.

[39] Steyerberg EW, Vergouwe Y. Towards better clinical prediction models: Seven steps for development and an ABCD for validation. Eur Heart J 2014;35:1925–31. doi:10.1093/eurheartj/ehu207.

[40] Breiman L. Random Forests. vol. 45. 2001.

[41] Probst P, Boulesteix A-L, Bischl B. Tunability: Importance of Hyperparameters of Machine Learning Algorithms. 2018.

[42] Sakr S, Elshawi R, Ahmed AM, Qureshi WT, Brawner CA, Keteyian SJ, et al. Comparison of machine learning techniques to predict all-cause mortality using fitness data: the Henry ford exercIse testing (FIT) project. BMC Med Inform Decis Mak 2017;17:174. doi:10.1186/s12911-017-0566-6.

[43] König IR, Malley JD, Weimar C, Diener H-C, Ziegler A, German Stroke Study Collaboration. Practical experiences on the necessity of external validation. Stat Med 2007;26:5499–511. doi:10.1002/sim.3069.

[44] Thelin EP, Nelson DW, Vehvilä Inen J, Nyström H, Kivisaari R, Siironen J, et al. Evaluation of novel computerized tomography scoring systems in human traumatic brain injury: An observational, multicenter study 2017. doi:10.1371/journal.pmed.1002368.

[45] Ngiam KY, Khor IW. Big data and machine learning algorithms for health-care delivery. Lancet Oncol 2019;20:e262–73. doi:10.1016/S1470-2045(19)30149-4.

[46] van der Ploeg T, Nieboer D, Steyerberg EW. Modern modeling techniques had limited external validity in predicting mortality from traumatic brain injury. J Clin Epidemiol 2016;78:83–9. doi:10.1016/J.JCLINEPI.2016.03.002.

[47] Poldrack RA, Gorgolewski KJ. Making big data open: data sharing in neuroimaging. Nat Neurosci 2014;17:1510–7. doi:10.1038/nn.3818.

[48] Neurology TL. The changing landscape of traumatic brain injury research. Lancet Neurol 2012;11:651. doi:10.1016/S1474-4422(12)70166-7.

[49] Charles D, Gabriel M, Searcy T. Adoption of Electronic Health Record Systems among U.S. Non-Federal Acute Care Hospitals: 2008-2014. 2015.

[50] Rajkomar A, Oren E, Chen K, Dai AM, Hajaj N, Hardt M, et al. ARTICLE Scalable and accurate deep learning with electronic health records 2018;1:18. doi:10.1038/s41746-018-0029-1.

[51] Beam AL, Kohane IS. Big Data and Machine Learning in Health Care. JAMA 2018;319:1317. doi:10.1001/jama.2017.18391.

[52] Delahanty RJ, Kaufman D, Jones SS. Development and Evaluation of an Automated Machine Learning Algorithm for In-Hospital Mortality Risk Adjustment Among Critical Care Patients. Crit Care Med 2018. doi:10.1097/CCM.0000000000003011.

[53] Desautels T, Das R, Calvert J, Trivedi M, Summers C, Wales DJ, et al. Prediction of early unplanned intensive care unit readmission in a UK tertiary care hospital: A cross-sectional machine learning approach. BMJ Open 2017. doi:10.1136/bmjopen-2017-017199.

[54] Klau S, Jurinovic V, Hornung R, Herold T, Boulesteix A-L. Priority-Lasso: a simple hierarchical approach to the prediction of clinical outcome using multi-omics data. BMC Bioinformatics 2018;19:322. doi:10.1186/s12859-018-2344-6.

[55] Reps JM, Schuemie MJ, Suchard MA, Ryan PB, Rijnbeek PR. Design and implementation of a standardized framework to generate and evaluate patient-level prediction models using observational healthcare data n.d. doi:10.1093/jamia/ocy032.

Appendix A

Table 1 Baseline characteristics the 15 IMPACT studies (n=…) and selected patients from the CENTER-TBI Core study (n=1554)

	TINT	TIUS	SLIN	SAP	BYR	HITI	UK4	TCDB	SKB	EBIC	HITII	NABIS	CSTAT	PHAMOS	APOE	CENTER-TBI	NA
N	**1118**	**1041**	**409**	**919**	**1510**	**350**	**791**	**603**	**126**	**822**	**819**	**385**	**517**	**856**	**756**	**1554**	
Age (mean (sd))	33·61 (14·52)	32·78 (12·48)	32·35 (13·4)	35·63 (15·43)	30·44 (13·26)	35·49 (15·38)	39·64 (19·2)	32·97 (16·74)	30·61 (13·28)	41·79 (20·3)	36·30 (15·6)	31·82 (12·47)	31·79 (13·2)	35·01 (13·78)	41·14 (18·8)	47·42 (20·98)	1·4
Hypoxia (%)	149 (15·1)	265 (28·7)	24 (5·9)	110 (12·9)	-	-	197 (25·5)	109 (18·1)	29 (34·5)	233 (28·5)	-	124 (33·4)	62 (13·0)	212 (24·8)	193 (28·1)	217 (16·8)	27·4
Hypotension (%)	153 (14·1)	224 (22·1)	-	128 (14·9)	-	17 (5·0)	205 (26·2)	143 (23·7)	21 (20·2)	199 (24·4)	81 (9·9)	56 (15·1)	85 (16·8)	132 (15·4)	74 (10·8)	205 (15·9)	19·4
Marshall CT class (%)																	41·4
1	51 (4·6)	98 (9·5)	2 (0·5)	39 (4·3)	-	-	-	-	3 (2·4)	98 (12·2)	69 (8·5)	4 (1·1)	-	15 (1·8)	-	81 (8·3)	
2	421 (38·1)	360 (35·0)	152 (37·2)	358 (39·4)	-	-	-	-	51 (40·5)	226 (28·0)	270 (33·4)	31 (8·9)	-	412 (48·5)	-	428 (43·9)	
3	218 (19·7)	196 (19·1)	94 (23·0)	145 (16·0)	-	-	-	-	40 (31·7)	81 (10·0)	89 (11·0)	193 (55·3)	-	203 (23·9)	-	86 (8·8)	
4	46 (4·2)	39 (3·8)	26 (6·4)	22 (2·4)	-	-	-	-	2 (1·6)	21 (2·6)	31 (3·8)	5 (1·4)	-	56 (6·6)	-	19 (2·0)	
5	370 (33·5)	335 (32·6)	135 (33·0)	344 (37·9)	-	-	-	-	30 (23·8)	380 (47·1)	350 (43·3)	116 (33·2)	-	163 (19·2)	-	360 (37·0)	
Traumatic subarachnoid hemorrhage (%)	566 (52·5)	420 (42·7)	317 (78·5)	399 (43·9)	619 (41·0)	71 (27·6)	-	227 (43·2)	99 (78·6)	327 (41·9)	268 (32·7)	-	-	511 (59·7)	192 (25·7)	759 (73·6)	20·3
Epidural hematoma (%)	181 (16·5)	88 (8·6)	-	186 (20·5)	147 (9·7)	63 (18·0)	110 (14·6)	-	14 (11·1)	76 (9·2)	134 (16·4)	-	60 (11·7)	166 (19·4)	50 (6·6)	172 (16·7)	16·0
Glucose (mean (sd))	8·60 (3·58)	9·22 (3·43)	-	7·74 (2·78)	9·70 (3·66)	8·78 (3·30)	-	-	9·99 (5·95)	-	-	9·51 (2·89)	-	8·02 (2·68)	-	8·18 (2·95)	45·3

Table 1, continued

	TINT	TIUS	SLIN	SAP	BYR	HITI	UK4	TCDB	SKB	EBIC	HITII	NABIS	CSTAT	PHAMOS	APOE	CENTER-TBI	NA
N	**1118**	**1041**	**409**	**919**	**1510**	**350**	**791**	**603**	**126**	**822**	**819**	**385**	**517**	**856**	**756**	**1554**	
GCS motor (%)																	8·8
1	5 (0·4)	9 (0·9)	0 (0·0)	141 (15·4)	475 (31·5)	122 (35·8)	113 (18·9)	136 (23·8)	34 (31·2)	150 (22·1)	210 (27·6)	82 (21·9)	40 (7·7)	43 (5·9)	5 (1·2)	615 (44·7)	
2	136 (12·2)	143 (13·7)	55 (13·4)	123 (13·4)	180 (11·9)	41 (12·0)	85 (14·2)	107 (18·7)	22 (20·2)	80 (11·8)	70 (9·2)	62 (16·5)	88 (17·0)	91 (12·5)	2 (0·5)	77 (5·6)	
3	237 (21·2)	132 (12·7)	91 (22·2)	143 (15·6)	165 (10·9)	45 (13·2)	37 (6·2)	74 (12·9)	14 (12·8)	55 (8·1)	92 (12·1)	55 (14·7)	86 (16·6)	136 (18·6)	0 (0·0)	80 (5·8)	
4	327 (29·2)	300 (28·8)	127 (31·1)	223 (24·3)	334 (22·1)	56 (16·4)	141 (23·6)	121 (21·2)	16 (14·7)	113 (16·6)	181 (23·8)	76 (20·3)	180 (34·8)	225 (30·8)	18 (4·2)	136 (9·9)	
5	384 (34·3)	406 (39·0)	134 (32·8)	286 (31·2)	309 (20·5)	77 (22·6)	191 (32·0)	113 (19·8)	21 (19·3)	182 (26·8)	199 (26·2)	97 (25·9)	122 (23·6)	235 (32·2)	35 (8·2)	357 (26·0)	
6	29 (2·6)	51 (4·9)	2 (0·5)	0 (0·0)	47 (3·1)	0 (0·0)	30 (5·0)	21 (3·7)	2 (1·8)	99 (14·6)	8 (1·1)	3 (0·8)	1 (0·2)	0 (0·0)	365 (85·9)	110 (8·0)	
Pupil (%)																	14·0
Both reactive	757 (72·4)	673 (67·9)	308 (77·4)	-	758 (52·0)	230 (66·9)	387 (54·3)	299 (49·6)	-	503 (65·1)	570 (71·3)	229 (61·9)	302 (70·9)	642 (78·1)	634 (84·4)	973 (73·7)	
One reactive	165 (15·8)	114 (11·5)	77 (19·3)	-	153 (10·5)	50 (14·5)	103 (14·4)	55 (9·1)	-	73 (9·4)	96 (12·0)	48 (13·0)	72 (16·9)	147 (17·9)	39 (5·2)	110 (8·3)	
None reactive	123 (11·8)	204 (20·6)	13 (3·3)	-	548 (37·6)	64 (18·6)	223 (31·3)	249 (41·3)	-	197 (25·5)	133 (16·6)	93 (25·1)	52 (12·2)	33 (4·0)	78 (10·4)	238 (18·0)	
Glasgow outcome scale (%)																	2·8
2	322 (28·8)	267 (25·6)	108 (26·4)	236 (25·7)	476 (31·5)	109 (31·1)	372 (47·0)	298 (49·4)	40 (31·7)	299 (36·4)	220 (26·9)	119 (30·9)	142 (27·5)	191 (22·3)	123 (16·3)	348 (29·0)	
3	134 (12·0)	128 (12·3)	69 (16·9)	142 (15·5)	298 (19·7)	62 (17·7)	146 (18·5)	95 (15·8)	30 (23·8)	123 (15·0)	108 (13·2)	98 (25·5)	69 (13·3)	246 (28·7)	163 (21·6)	303 (25·2)	
4	171 (15·3)	180 (17·3)	84 (20·5)	174 (18·9)	374 (24·8)	64 (18·3)	130 (16·4)	103 (17·1)	27 (21·4)	159 (19·3)	199 (24·3)	91 (23·6)	92 (17·8)	203 (23·7)	211 (27·9)	246 (20·5)	
5	491 (43·9)	466 (44·8)	148 (36·2)	367 (39·9)	362 (24·0)	115 (32·9)	143 (18·1)	107 (17·7)	29 (23·0)	241 (29·3)	292 (35·7)	77 (20·0)	214 (41·4)	216 (25·2)	259 (34·3)	303 (25·2)	
HB (mean (sd))	11·90 (2·75)	13·06 (2·16)	-	11·44 (1·96)	13·22 (2·10)	13·21 (2·09)	-	-	12·10 (2·33)	-	-	12·88 (2·12)	-	12·72 (2·50)	-	7·96 (2·36)	52·2

CT = computed tomography; GCS = Glasgow Coma Scale; HB= hemoglobin. For the study descriptions, see http://www.tbi-impact.org/?p=impact/datasets.

Table 2 Results of the calibration intercept at internal-external (per study CV), internal (10-fold cross validation), and external (CENTER-TBI) validation

Algorithm	Outcome	Internal-external	Internal	External
Logistic regression	Mortality	-0.02 (-0.20 - 0.15)	-0.01 (-0.04 - 0.03)	-0.61 (-0.75 - -0.47)
Support vector machine		-0.02 (-0.20 - 0.16)	0.02 (-0.01 - 0.05)	-0.44 (-0.57 - -0.30)
Random forest		-0.08 (-0.26 - 0.11)	-0.04 (-0.10 - 0.03)	-0.83 (-0.97 - -0.69)
neural network		-0.03 (-0.20 - 0.15)	0.01 (-0.05 - 0.07)	-0.49 (-0.63 - -0.35)
Gradient boosting machine		-0.03 (-0.21 - 0.15)	-0.00 (-0.05 - 0.05)	-0.57 (-0.71 - -0.42)
Lasso regression		-0.02 (-0.20 - 0.15)	0.00 (-0.07 - 0.08)	-0.60 (-0.74 - -0.46)
Ridge regression		-0.02 (-0.19 - 0.15)	0.00 (-0.07 - 0.07)	-0.55 (-0.69 - -0.42)
Logistic regression	Unfavourable outcome	0.02 (-0.15 - 0.20)	-0.01 (-0.05 - 0.03)	-0.18 (-0.30 - -0.05)
Support vector machine		0.02 (-0.15 - 0.19)	-0.02 (-0.05 - 0.01)	-0.04 (-0.17 - 0.08)
Random forest		0.02 (-0.15 - 0.19)	0.02 (-0.02 - 0.06)	-0.32 (-0.45 - -0.20)
neural network		0.02 (-0.15 - 0.20)	-0.03 (-0.07 - 0.02)	-0.08 (-0.21 - 0.05)
Gradient boosting machine		0.02 (-0.15 - 0.19)	0.01 (-0.03 - 0.05)	-0.08 (-0.21 - 0.04)
Lasso regression		0.02 (-0.15 - 0.20)	0.00 (-0.07 - 0.07)	-0.17 (-0.29 - -0.04)
Ridge regression		0.02 (-0.15 - 0.19)	0.00 (-0.06 - 0.07)	-0.14 (-0.26 - -0.02)

Table 3 Results of the calibration slope at internal-external (per study CV), internal (10-fold cross validation), and external (CENTER-TBI) validation

Algorithm	Outcome	Internal-external	Internal	External
Logistic regression	Mortality	0.98 (0.90 - 1.05)	0.98 (0.94 - 1.03)	0.96 (0.84 - 1.07)
Support vector machine		1.00 (0.91 - 1.08)	1.00 (0.98 - 1.02)	0.97 (0.85 - 1.09)
Random forest		0.85 (0.77 - 0.93)	0.78 (0.72 - 0.83)	0.88 (0.77 - 0.99)
neural network		0.99 (0.91 - 1.08)	0.99 (0.95 - 1.03)	0.98 (0.86 - 1.09)
Gradient boosting machine		0.96 (0.89 - 1.04)	0.98 (0.93 - 1.03)	0.92 (0.81 - 1.03)
Lasso regression		0.99 (0.91 - 1.07)	1.01 (0.96 - 1.06)	0.97 (0.85 - 1.09)
Ridge regression		1.05 (0.96 - 1.13)	1.07 (1.02 - 1.12)	1.03 (0.90 - 1.15)
Logistic regression	Unfavourable outcome	0.99 (0.93 - 1.05)	1.00 (0.97 - 1.04)	0.82 (0.71 - 0.92)
Support vector machine		1.00 (0.94 - 1.06)	1.01 (0.97 - 1.05)	0.82 (0.72 - 0.92)
Random forest		0.89 (0.82 - 0.96)	0.81 (0.77 - 0.85)	0.76 (0.66 - 0.86)
neural network		0.98 (0.92 - 1.03)	0.99 (0.94 - 1.05)	0.81 (0.71 - 0.91)
Gradient boosting machine		1.01 (0.95 - 1.07)	1.02 (0.97 - 1.06)	0.82 (0.72 - 0.92)
Lasso regression		1.00 (0.94 - 1.06)	1.01 (0.96 - 1.06)	0.83 (0.73 - 0.93)
Ridge regression		1.06 (1.00 - 1.13)	1.07 (1.02 - 1.12)	0.88 (0.77 - 0.99)

Table 4 Results of the Brier score at internal-external (per study CV), internal (10-fold cross validation), and external (CENTER-TBI) validation

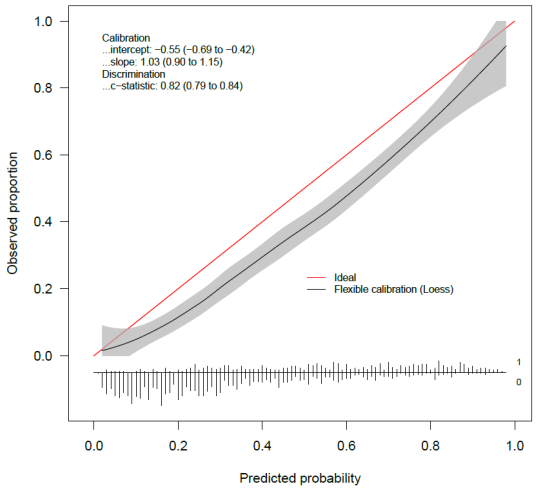
Algorithm	Outcome	Internal-external	Internal	External
Logistic regression	Mortality	0.15 (0.14 - 0.16)	0.15 (0.15 - 0.15)	0.16 (0.15 - 0.17)
Support vector machine		0.15 (0.14 - 0.17)	0.15 (0.15 - 0.15)	0.16 (0.15 - 0.17)
Random forest		0.16 (0.15 - 0.17)	0.16 (0.16 - 0.16)	0.17 (0.16 - 0.18)
neural network		0.15 (0.14 - 0.16)	0.15 (0.15 - 0.15)	0.16 (0.15 - 0.17)
Gradient boosting machine		0.15 (0.14 - 0.16)	0.15 (0.14 - 0.15)	0.15 (0.14 - 0.16)
Lasso regression		0.15 (0.14 - 0.16)	0.15 (0.14 - 0.15)	0.16 (0.15 - 0.17)
Ridge regression		0.15 (0.14 - 0.16)	0.15 (0.14 - 0.15)	0.16 (0.15 - 0.17)
Logistic regression	Unfavourable outcome	0.18 (0.17 - 0.19)	0.17 (0.17 - 0.18)	0.19 (0.18 - 0.21)
Support vector machine		0.18 (0.17 - 0.19)	0.18 (0.17 - 0.18)	0.19 (0.18 - 0.20)
Random forest		0.19 (0.18 - 0.20)	0.19 (0.18 - 0.19)	0.20 (0.19 - 0.21)
neural network		0.18 (0.17 - 0.19)	0.18 (0.17 - 0.18)	0.19 (0.18 - 0.20)
Gradient boosting machine		0.18 (0.17 - 0.19)	0.18 (0.17 - 0.18)	0.19 (0.18 - 0.21)
Lasso regression		0.18 (0.17 - 0.19)	0.18 (0.17 - 0.18)	0.19 (0.18 - 0.21)
Ridge regression		0.18 (0.17 - 0.19)	0.18 (0.17 - 0.18)	0.19 (0.18 - 0.21)

Table 5 Results of the c-statistic at internal-external (per study CV), internal (10-fold cross validation), and external (CENTER-TBI) validation, in a different imputed dataset.

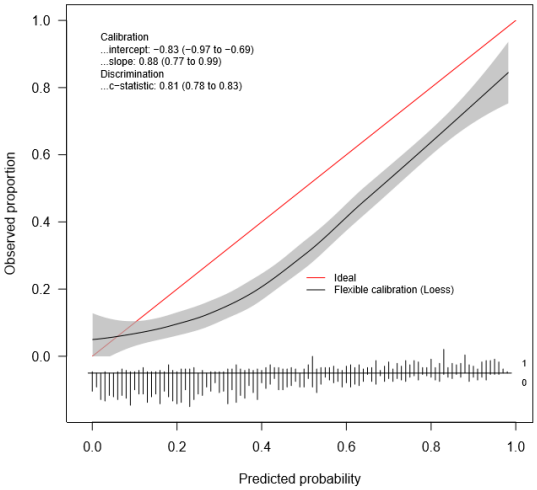
Algorithm	Outcome	Internal-external	Internal	External
Logistic regression	Mortality	0.81 (0.79 - 0.84)	0.83 (0.82 - 0.84)	0.82 (0.79 - 0.84)
Support vector machine		0.81 (0.78 - 0.83)	0.83 (0.82 - 0.83)	0.81 (0.79 - 0.84)
Random forest		0.79 (0.77 - 0.82)	0.81 (0.80 - 0.81)	0.81 (0.78 - 0.84)
neural network		0.81 (0.79 - 0.84)	0.83 (0.82 - 0.84)	0.82 (0.79 - 0.84)
Gradient boosting machine		0.81 (0.79 - 0.84)	0.83 (0.82 - 0.84)	0.83 (0.81 - 0.86)
Lasso regression		0.81 (0.79 - 0.84)	0.83 (0.82 - 0.84)	0.82 (0.79 - 0.84)
Ridge regression		0.81 (0.79 - 0.84)	0.83 (0.82 - 0.84)	0.82 (0.79 - 0.84)
Logistic regression	Unfavourable outcome	0.81 (0.79 - 0.83)	0.81 (0.80 - 0.82)	0.77 (0.75 - 0.80)
Support vector machine		0.80 (0.79 - 0.82)	0.81 (0.80 - 0.82)	0.78 (0.75 - 0.80)
Random forest		0.79 (0.76 - 0.81)	0.78 (0.78 - 0.79)	0.76 (0.74 - 0.79)
neural network		0.80 (0.79 - 0.82)	0.81 (0.80 - 0.81)	0.78 (0.76 - 0.80)
Gradient boosting machine		0.80 (0.78 - 0.82)	0.81 (0.80 - 0.82)	0.78 (0.76 - 0.80)
Lasso regression		0.81 (0.79 - 0.83)	0.81 (0.80 - 0.82)	0.77 (0.75 - 0.80)
Ridge regression		0.81 (0.79 - 0.83)	0.81 (0.80 - 0.82)	0.77 (0.75 - 0.80)

Figure 1 Two algorithms are compared in two studies: A+B is CENTER-TBI, C+D is UK4, A+C is ridge regression, and B+D random forest. The c-statistic is primarily determined by study: the c-statistics between studies (A-C and B-D) differ more than within studies (A-B and C-D). This is also found with respect to the calibration intercept, and calibration slope.

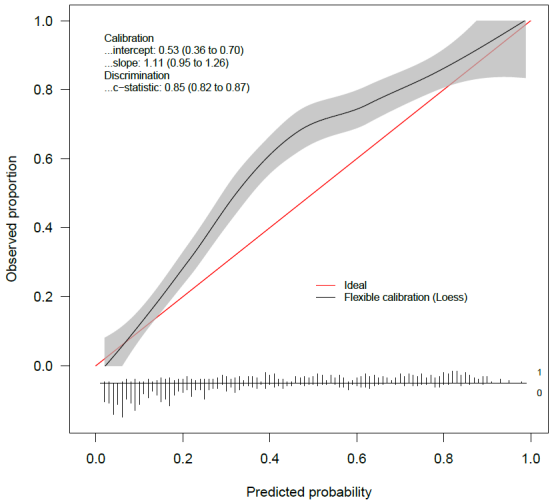
A) CENTER-TBI Ridge regression



B) CENTER-TBI Random Forest



C) UK4 Ridge regression



D) UK4 Random Forest

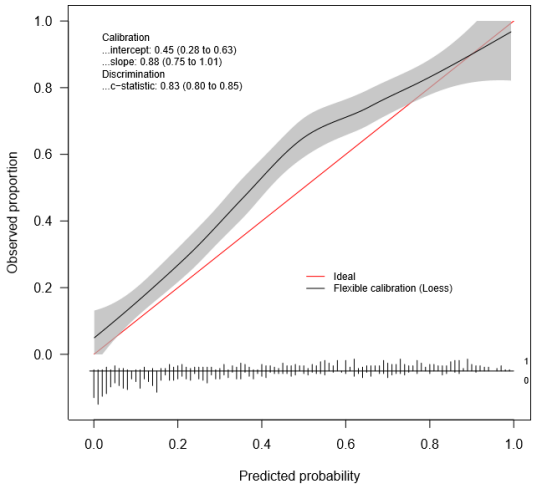


Table 6 Results for empirical testing of the assumption of non-linearity and non-additivity in the datasets. The p-value, both crude and adjusted by the false discovery rate, are shown.

Study	Assumption	P-value	Adjusted p-value
TINT	Non-linearity	0.207	0.476
TIUS		0.567	0.798
SLIN		0.662	0.820
SAP		**0.003**	**0.026**
BYR		0.473	0.798
HITI		0.101	0.312
UK4		0.643	0.820
TCDB		**0.021**	0.108
SKB		0.821	0.877
EBIC		**0.001**	**0.012**
HIT II		**<0.001**	**0.005**
NABIS		0.158	0.446
CSTAT		0.073	0.251
PHAMOS		0.882	0.911
APOE		**0.005**	0.031
CENTER-TBI		**<0.001**	**0.005**
TINT	Non-additivity (interaction)	0.782	0.865
TIUS		0.966	0.966
SLIN		0.527	0.798
SAP		0.246	0.509
BYR		0.543	0.798
HITI		0.215	0.476
UK4		0.549	0.798
TCDB		0.759	0.865
SKB		NA*	NA*
EBIC		0.739	0.865
HIT II		0.556	0.798
NABIS		0.595	0.802
CSTAT		0.417	0.798
PHAMOS		**0.048**	0.211
APOE		0.071	0.251
CENTER-TBI		0.195	0.476

* Too few observations to fit the model with all interaction terms

Appendix B

Calibration

Calibration refers to the agreement between the predicted probability of the outcome versus the observed. Calibration was examined graphically and quantified using a calibration slope and the calibration intercept: the calibration test proposed by Cox [1]. The calibration slope was calculated by fitting a logistic regression model with the observed outcome as dependent variable and the log odds of

the predicted probability of the model. The calibration slope is given by the coefficient of the log odds of the predicted probability, and should ideally be equal to one. To assess calibration in the large, or calibration intercept, a similar model was fitted, with the log-odds of the predictions as an offset variable. The intercept of this model is the calibration intercept, which reflects calibration-in-the-large, or difference between observed and predicted risks. Ideally, this measure should be equal to zero [2,3].

Discrimination

Discrimination refers to the ability of the model to distinguish between patients with and without the event of interest. It was quantified using the *c*-statistic, also known as area under the ROC curve. The c-statistic ranges from 0.5 to 1 where 0.5 indicates that model predictions are no better than a coin toss and 1 indicates perfect discrimination. The confidence intervals of the c-statistic were obtained using the DeLong et al method [4], using the *ci.auc* function from the *pROC* package.

Brier score

The Brier score is both influenced by discrimination and calibration, and can therefore be seen as a metric of overall performance [5]. It is calculated by summarizing the distance of the calculated probability of an event to its realization (1 in case the event occurred, and 0 if not). The brier score of flipping a fair coin is 0.25, and perfect prediction translates into a brier score of 0: a probability of 1 if the event will occur, and a probability of 0 if not.

Machine learning algorithms

Extended tuning grids of possible combinations of hyperparameters in was used, see the table below.

Support vector machine constructs hyperplanes in a multidimensional space, defined by both the kernel function and the input variables [6]. The hyperplane then discerns classes. The most commonly used kernel functions are radial, polynomial, and linear. The linear kernel function was used in this study, because using other kernel functions lead to convergence problems. Using the distance to this hyperplane, the probability of belonging to a class can be calculated. The svmLinear function from the e1071 package was used.

Standard neural networks are 'multilayer perceptrons'. These start with an input layer that contains the predictors (input neurons), and output layer to generate predictions (output neurons), and then one or more hidden layers in between (consisting of hidden neurons). In each layer, neurons are activated by input from neurons of the previous layer [7]. This activation is similar to regression to the extent that the neurons evaluates all input signals with weights to come to an output signal to the next layer. The networks can be built with different gradients of complexity. We considered neural networks with 1 hidden layer only. The hyperparameters of this algorithm were size (1, 3, 5, 10, or 15 middle neurons), and decay (0, 0.1, 0.01, and 0.001). By using one binary output neuron, the network functions as a logistic regression model, in the sense that it trains to predict probabilities. The nnet function from the nnet package was used.

Finally, random forest and gradient boosting machine are both decision-tree based methods. They classify by partitioning the classes based on input variables. Both methods consist of multiple created

decision trees. The random forest algorithm uses a repeated bootstrap sample procedure to randomly create decision trees [8]. The hyperparameters of the algorithm are the number of trees (500) and the number of random variables used for partitioning the tree (2, 5, 10, or 18). The ranger function from the ranger package was used. This function creates as a default 500 trees in the forest. The size of the terminal nodes is used as a criterion for the maximum tree depth: if 5 or less patients are left at the end of the creating of a tree, the algorithm terminates.

The gradient boosting machine uses information from prior created trees to optimize the creation or branching of consecutive trees[9]. The proportion of trees that voted in favor of the outcome was used as the predicted probability of the outcome. The hyperparameters of the algorithm are depth of the trees (1, 2, or 3 layers), number of trees (50, 100, 150, 200, 300), shrinkage (0, 0.1, 0.01), and the minimal number of patients per node (50, 100, or 150). The gbm function from the gbm package was used.

[1] Cox D. Two further applications of a model for binary regression. Biometrika 1958;45:562–5.

[2] Steyerberg EW, Vergouwe Y. Towards better clinical prediction models: seven steps for development and an ABCD for validation. Eur Heart J 2014;35:1925–31. doi:10.1093/eurheartj/ehu207.

[3] Steyerberg EW. Clinical Prediction Models. New York, NY: Springer New York; 2009. doi:10.1007/978-0-387-77244-8.

[4] DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics 1988;44:837–45.

[5] Steyerberg EW, Vickers AJ, Cook NR, Gerds T, Gonen M, Obuchowski N, et al. Assessing the performance of prediction models: a framework for traditional and novel measures. Epidemiology 2010;21:128–38. doi:10.1097/EDE.0b013e3181c30fb2.

[6] Burges CJC. A Tutorial on Support Vector Machines for Pattern Recognition. vol. 2. 1998.

[7] Jain AK, Jianchang Mao, Mohiuddin KM. Artificial neural networks: a tutorial. Computer (Long Beach Calif) 1996;29:31–44. doi:10.1109/2.485891.

[8] Breiman L. Random Forests. vol. 45. 2001.

[9] Natekin A, Knoll A. Gradient boosting machines, a tutorial. Front Neurorobot 2013;7:21. doi:10.3389/fnbot.2013.00021.

The grid of hyperparemeters used to optimize the machine learning algorithms.

Algorithm	**Hyperparameter**	**Values to select from**
Lasso/Ridge	Lambda	log(-6) - log(2)
SVM	Cost	0.5, 0.6, 0.7, 0.8, 0.9, 1.0
NNet	Size	1, 3, 5, 10, 15
	Decay	0, 0.1,0.01, 0.001

RF	N trees	500
	N random variables	2, 5, 10, 18
GBM	Tree depth	1, 2, 3,5
	N trees	50, 100, 150, 200, 300
	Shrinkage	0, 0.01, 0.1
	Min N observed per node	50, 100, 150