the british psychological society
promoting excellence in psychology

Journal of Neuropsychology (2021), 15, 319–339

www.wileyonlinelibrary.com

Theoretical Or Review Article

A systematic review of the effectiveness of machine learning for predicting psychosocial outcomes in acquired brain injury: Which algorithms are used and why?

Emma Mawdsley*[1,2], Bronagh Reynolds[1,3] and Breda Cullen[1]

[1]Mental Health and Wellbeing, Institute of Health and Wellbeing, University of Glasgow, UK
[2]NHS Greater Glasgow and Clyde, UK
[3]NHS Ayrshire and Arran, UK

Clinicians working in the field of acquired brain injury (ABI, an injury to the brain sustained after birth) are challenged to develop suitable care pathways for an individual client's needs. Being able to predict psychosocial outcomes after ABI would enable clinicians and service providers to make advance decisions and better tailor care plans. Machine learning (ML, a predictive method from the field of artificial intelligence) is increasingly used for predicting ABI outcomes. This review aimed to examine the efficacy of using ML to make psychosocial predictions in ABI, evaluate the methodological quality of studies, and understand researchers' rationale for their choice of ML algorithms. Nine studies were reviewed from five databases, predicting a range of psychosocial outcomes from stroke, traumatic brain injury, and concussion. Eleven types of ML were employed with a total of 75 ML models. Every model was evaluated as having high risk of bias, unable to provide adequate evidence for predictive performance due to poor methodological quality. Overall, there was limited rationale for the choice of ML algorithms and poor evaluation of the methodological limitations by study authors. Considerations for overcoming methodological shortcomings are discussed, along with suggestions for assessing the suitability of data and suitability of ML algorithms for different ABI research questions.

The variation in psychosocial outcomes after an acquired brain injury (ABI, an injury to the brain sustained after birth including stroke and traumatic brain injury [TBI]) challenges health and social care services to provide advice and guidance to the person, their family, and for socioeconomic implications. Currently, 'evidence-based practice' relies almost exclusively on the results of parametric analyses of group-level central tendency derived from randomized clinical trials, which offers very little guidance for individualized care. The study of clinical prediction rules to accurately predict an individual's psychosocial outcome at a future time point after ABI would serve timely resource allocation and risk

**Correspondence should be addressed to Emma Mawdsley, Mental Health and Wellbeing, Administration Building, Gartnavel Royal Hospital, 1055 Great Western Road Glasgow, Glasgow G12 0XH, UK (email: efmawdsley@gmail.com).*

DOI:10.1111/jnp.12244

management, as well as being able to adapt interventions for known risk factors to maximize the likelihood of more favourable outcomes.

Machine learning (ML) is an evolving methodology in clinical research, offering a possible solution to limitations with traditional methods of modelling and potentially providing better applicability of research findings to individualized clinical decisions through developing clinical prediction rules. Supervised ML learns from the data how to best predict the outcome in question (Hastie, Tibshirani, & Friedman, 2009; Ch 2). Whilst ML was predominantly employed by data scientists and statisticians, it is becoming an increasingly popular approach for clinicians and clinical researchers to consider its use for tackling the large and complex data sets typical of routine clinical data.

The clinical applications of ML have expanded from medical and genetic research, to psychological research questions. Predicting psychosocial outcomes, such as the likelihood of developing mood disorders or being able to return to work after an ABI, typically have a higher degree of subjectivity than medical outcomes, and the measurement around such variables can include higher proportions of noise (Mascolo, 2016). Despite growing popularity, how well ML performs at predicting such outcomes in ABI is unknown.

To date, there has been no review or guidance for using ML to predict psychosocial outcomes in ABI; however, a previous systematic review has shown superior power for ML methodologies to predict neurosurgical outcomes (Senders et al., 2018). Unfortunately, as no risk of bias (ROB) assessment was completed for the review it greatly limits the applicability of their findings. In recent years, guidance has been developed for prediction research (e.g., Moons et al., 2015; Wolff et al., 2019), allowing thorough evaluation of prediction models. Without such guidance, common data mistakes can lead to biased results. By evaluating psychosocial ABI research, clinicians will benefit from being able to understand the effectiveness of using ML algorithms across ABIs, consider the suitability of ML for data sets commonly available within services, and work towards developing accurate prediction tools to assist clinical decision-making.

Objectives

This systematic review aimed to evaluate research employing ML to develop models for the prediction of psychological, social, and/or functional outcomes after ABI.

In particular, this review set out to answer:

1. How effective is ML for making psychosocial predictions for people with ABI?
2. Which ML algorithms are most commonly used?
3. What is the rationale for the choice of ML algorithms, as stated by the study authors?

Method

Protocol and registration

The protocol of this systematic review was written in accordance with PRISMA-P (Moher et al., 2015) and registered on PROSPERO on 15/July/2019, registration number CRD42019140546 [available from: https://www.crd.york.ac.uk/PROSPERO/display_record.php?RecordID=140546]. This review has been written in accordance with PRISMA (Liberati et al., 2009).

Eligibility criteria

Research reports were included with an English language version available in a peer-reviewed journal. All reports up until the search date of 22/July/2019 were initially considered for the review. Due to the large number of eligible studies identified, studies were then limited to those published between 1st January 2016 and 22nd July 2019 to cover articles published after the Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD) guidance (Moons et al., 2015).

Participants

Studies included participants with a diagnosis of ABI, such as TBI (mild, moderate, or severe) or stroke. This review included people of any age, gender, or geographical location. Studies which included conditions other than ABI (e.g., other types of physical trauma or neurodegenerative conditions) in the same analysis with people with ABI were excluded.

Exposures and comparators

Studies were included with at least one psychosocial predictor in the final model. Psychosocial was defined as a measure of psychological or behavioural factors (e.g., cognition, mental health, challenging behaviours) or social factors (e.g., participation, accommodation status, employment). Studies were excluded where predictors were all biological (e.g., physical measurements, vital signs, or neuroimaging) or primarily all impairment-based (e.g., Glasgow Coma Scale [GCS], Teasdale & Jennett, 1974). The comparator was the absence of the exposure (predictor) or lower levels of the exposure where measured on a dimensional scale.

Outcomes of interest

Studies predicting a psychosocial outcome were included, with psychosocial defined as above. Studies were excluded where predictors and outcomes were measured at the same time point (e.g., questionnaire items predicting questionnaire outcome). This review excluded outcomes designed specifically for disciplines other than psychology (e.g., speech and language therapy measures, physiotherapy measures), measures which are primarily impairment-based (e.g., GCS) or neurological (e.g., neuroimaging, cerebrospinal fluid).

Study design

Studies were required to be observational designs which reported the development of a supervised ML model. ML was defined as 'algorithms [which search] through a large space of candidate programs, guided by training experience, to find a program that optimizes the performance metric.' (Bzdok, Krzywinski, & Altman, 2017 p. 1119). An ML technique is 'supervised' if it uses known outcome data as part of model learning. Studies reporting the application of a previously developed model and which did not include model development results were excluded.

Search and study selection

Published literature was reviewed from MEDLINE (PubMed), Web of Science, EMBASE (OVID interface, 1990 onwards), CINAHL, and PsycINFO (EBSCOhost interface, 1990 onwards), up until the date of 22/July/2019. The full search strategy is presented in Appendix S1. The search results were managed in the author's EndNote library (www.myendnoteweb.com). Duplicates were removed during database extraction, and then, titles were screened to remove papers that were not eligible. This screening process was repeated for abstracts and lastly full texts. A second reviewer independently repeated this process for 50 records at the title/abstract stage, and 10 records at the full text stage to check for consistency, showing 100% concordance.

Data collection process

A data extraction template was developed to extract relevant data from eligible studies combined from the Joanna Briggs Institute critical appraisal checklist for cohort studies (Briggs, 2017), TRIPOD (Moons et al., 2015), and additional items specific to the review questions. A full list of extracted data items is available in Appendix S2. The data extraction template was piloted by the primary author for five studies and then amended with two additional items. The final data extraction template was used by the primary author for all studies, and the second reviewer independently for three studies giving an inter-rater agreement of 93.1% (calculated as the percentage of agreement between raters on items), with discrepancies resolved by discussion.

Risk of bias in individual studies

The Prediction model Risk Of Bias ASsessment Tool (PROBAST, Wolff et al., 2019) was used at study level to evaluate bias for each presented ML model in each article, completed by the first author for all included articles and by the second reviewer independently for 3 records to check for consistency. The PROBAST assesses risk of bias across four areas in prediction studies (participants, predictors, outcomes, and analysis), rated by 20 items for ROB and 3 items for applicability. Examples of PROBAST items include the appropriateness of inclusion and exclusion criteria, or whether overfitting, underfitting, and model optimism have been considered in the performance of the model. Inter-rater agreement was 91.7%, indicating high consistency. Differences in opinion were discussed until consensus was reached.

Summary measures and synthesis of results

A narrative synthesis was performed, presented in text and tables. To address the first review question, performance metrics are reported for both the internal validation models and, if applicable, the external validation model, with the area under the receiver operating characteristic curve (AUC, also known as the c-index) being the primary metric of choice. Alternative metrics are reported for some studies. Performance metrics of models were then evaluated as being reliable or unreliable dependent on the ROB ratings of the models. To address the second review question, the frequency of the algorithms used by researchers is reported. For the third review question, the rationale of the author's choice of methodology was summarized. The findings of these three questions are then used to provide considerations for designing an ML study for predicting psychosocial outcomes in ABI for future researchers.

Results

Study selection

Figure 1 shows the flow diagram of the search procedure and the results.

Study characteristics

A total of nine studies were included for the systematic review, with brief abstracts available in Appendix S3. Six were from the United States (Bergeron et al., 2019; Cnossen et al., 2017; Gupta et al., 2017; Hirata, Ovbiagele, Markovic, & Towfighi, 2016; Stromberg et al., 2019; Walker et al., 2018), one from Finland (Huttunen et al., 2016), one from Japan (Nishi et al., 2019), and one from Iran (Shafiei et al., 2017). A brief review of study design and analysis by study is included in Table 1.

One study predicted outcomes after concussive incidents (1611 incidents with multiple concussions per person, Bergeron et al., 2019), and the remaining eight predicted outcomes from 64,325 people with ABI in total, including cerebrovascular accident (Gupta et al., 2017; Hirata et al., 2016; Huttunen et al., 2016; Nishi et al., 2019),

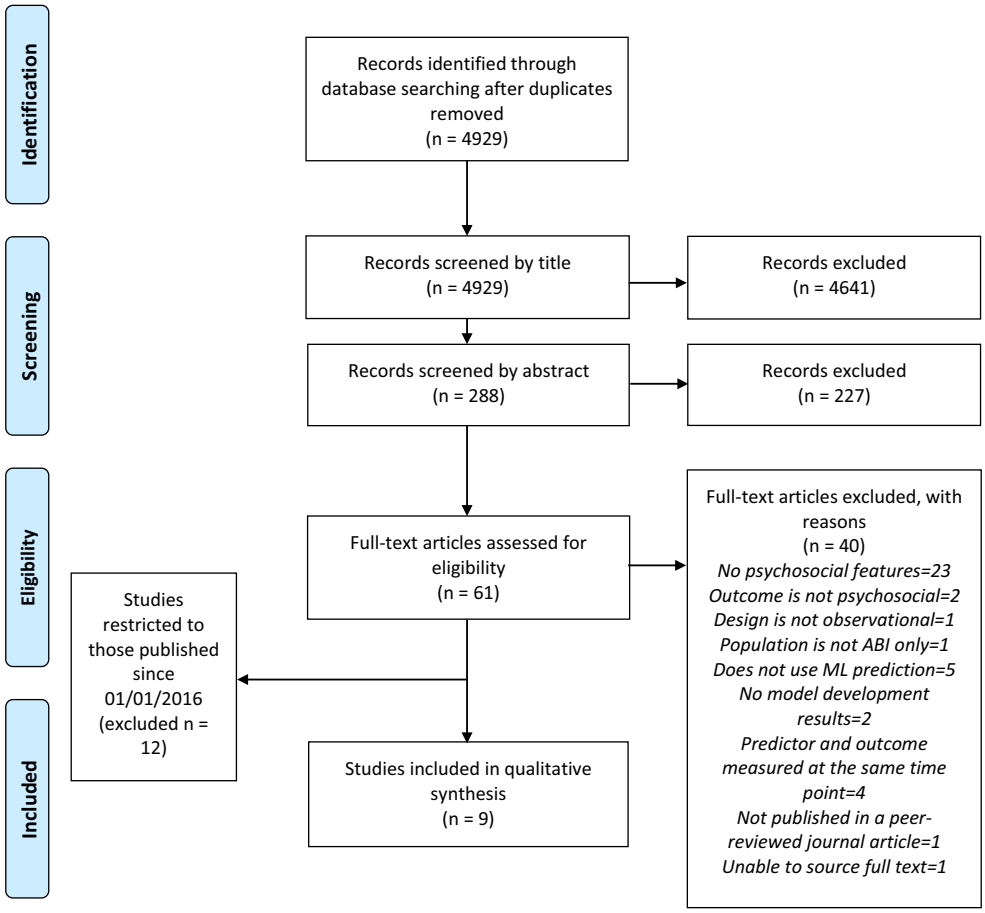

Figure 1. PRISMA flow diagram of the study selection process. Abbreviations: ABI = acquired brain injury; ML = machine learning.

Table 1. Characteristics of studies included in systematic review

Study	ABI population	Outcome	Sample size	Analysis design	ML methodology	Validation procedures
1. Bergeron et al. (2019)	Concussion	Time to symptom resolve	1,611 concussive incidents	Classification	NB, SVM, KNN, DTs (C4.5D and C4.5N), RF (with 100 and 500 trees), ANNs (multilayer perceptron and radial basis function network)	10-fold cross-validation, 1 segment reserved for internal validation
2. Cnossen et al. (2017)	Mild TBI GCS 13–15	Post-concussive symptoms (cognitive, somatic and psychological subscales, and severity)	277	Regression	RLR (lasso)	Bootstrap with 100 samples
3. Gupta et al. (2017)	Intracerebral haemorrhage	Functional outcome at 3 and 12 months	365 (3 months) 321 (12 months)	Classification and regression	RF for feature selection and then traditional linear and logistic regression	External validation
4. Hirata et al. (2016)	Stroke	Depression	17,132	Classification	RF	Within random forest uses 'out the bag', an embedded validation procedure, but no cross-validation
5. Huttunen et al. (2016)	Aneurysmal subarachnoid haemorrhage	Antidepressant use	940	Classification	DT	None
6. Nishi et al. (2019)	Acute stroke from large vessel occlusion who received mechanical thrombectomy	Good clinical outcome	387 development, 115 external validation	Classification	RLR, SVM and RF	10-fold nested cross-validation and external validation
7. Shafiei et al. (2017)	Mild TBI GCS 13-15	Psychological symptoms	100	Classification	ANN backpropagation algorithm	50/50 train test cross-validation repeated 300 times
8. Stromberg et al (2018)	TBI (moderate to severe)	Current competitive employment at 1, 2 and 5 years	7,867 (1 year) 6,783 (2 year) 4,927 (5 year)	Classification	DT	85/15 training test split with no cross-validation
9. Walker et al. (2018)	Non-penetrating TBI (moderate to severe)	Global outcome at 1, 2 and 5 years	10,125 (1 year) 8,821 (2 year) 6,165 (5 year)	Classification	DT	85/15 training test split with no cross-validation

ABI = acquired brain injury; ANN = artificial neural network; DT = decision tree; GCS = Glasgow Coma Scale; KNN = K-nearest neighbours; ML = machine learning; NB = naïve Bayes; RF = random forest; RLR = regularized logistic regression; SVM = support vector machine; TBI = traumatic brain injury.

mild TBI (Cnossen et al., 2017; Shafiei et al., 2017), and moderate to severe TBI (Stromberg et al., 2019; Walker et al., 2018). Two studies used the same database (Stromberg et al., 2019; Walker et al., 2018), and therefore, the same participants were likely in both studies. Outcomes included post-concussive symptoms (Bergeron et al., 2019; Cnossen et al., 2017), functional outcome (Gupta et al., 2017; Nishi et al., 2019; Walker et al., 2018), indicators of mood and psychological symptoms (Hirata et al., 2016; Huttunen et al., 2016; Shafiei et al., 2017), and employment (Stromberg et al., 2019).

Across the nine studies, there were a total of 11 types of ML: regularized logistic regression (RLR), support vector machine (SVM), decision trees (DT), naïve Bayes (NB), *K*-nearest neighbours (*K*NN), random forest (RF), artificial neural networks (ANNs, including multilayer perceptron, backpropagation, and radial basis function network), lasso regularization with linear regression, and random forest used for feature selection with logistic regression. Algorithm descriptions can be found in Table 2. Two studies compared more than one type of ML algorithm (Bergeron et al., 2019; Nishi et al., 2019), and five studies examined more than one time point or outcome (Bergeron et al., 2019; Cnossen et al., 2017; Gupta et al., 2017; Stromberg et al., 2019; Walker et al., 2018), giving a total of 75 ML models analysed.

Quality of the evidence

Quality ratings of the 75 models were aggregated by study since each model received the same score within each study (reported in Table 3), with the rationale for ROB scores in Table 4. Across the studies reviewed, each of the 75 ML models scored as being high ROB, with the main source of bias being the analysis. Every study failed to appropriately evaluate the developed models with use of calibration metrics, meaning the model's performance for individual probabilities is unknown. One study reported no model evaluation statistics for performance, discrimination, or calibration (Huttunen et al., 2016). Other common causes for high ROB were improper handling of missing data, not using appropriate techniques to account for model optimism and overfitting (such as internal nested cross-validation or bootstrapping), and poor reporting for how models performed after post-hoc refinement.

Only one study was high ROB for predictors and outcome (Bergeron et al., 2019), and three studies did not provide enough information to make a conclusion for either participant selection or variable handling (Shafiei et al., 2017; Stromberg et al., 2019; Walker et al., 2018). The other studies were well designed with regard to participant sources and measures to answer their research questions but failed to support their conclusions due to introducing bias from either the conduct or reporting of their analysis.

How effective is ML for making psychosocial predictions for people with ABI?

A summary of the performance metrics of the models along with the related ROB reliability ratings of the findings is included in Table 5. Models with an AUC of 0.80 or above are considered to show 'good' performance, between 0.70 and 0.79 as fair, and below 0.70 as poor (Safari, Baratloo, Elfil, & Negida, 2016). For linear algorithms, whilst it is a heavily disputed subject, an approximate rule for interpretation of R^2 is 0.75 for a substantial effect, 0.5 for moderate, and 0.25 for weak (Cruz-Cunha, 2013). However, due to the unreliability of each model from the ROB ratings, this review was unable to conclude which ML algorithm was most effective for predicting psychosocial outcomes. Considerations for choosing an ML algorithm are presented in the discussion.

Table 2. Machine learning algorithm definitions

Machine learning algorithms	Definition
Classification	
Regularized logistic regression	A classification algorithm whereby coefficient weights are learned using an iterative method with adjustments within a linear algorithm before being transformed to predict a binary outcome using the sigmoid or logistic function (Nadkarni, 2016)
Support vector machine	Most commonly used as a classification algorithm whereby vectors are mapped into a high-dimensional space to construct a linear decision surface (Cortes & Vapnik, 1995), with the goal of separating two decision categories
Decision trees	Decision trees classify predictors by their values among a series of decision branches, until ending with a fairly homogenous class of the target variable (Rokach & Maimon, 2008)
Naïve Bayes	A probability model based on Bayesian theory, where features are naïve in the sense that they assume independence from other features in a given class (Rish, 2001)
K-nearest neighbours (5NN)	Commonly used as a classification algorithm where new values are predicted based on the results of other, similar instances (or neighbours). It is common to take the results of more than one neighbour (*k*) for class determination (Cunningham & Delany, 2020)
Random forest	An ensemble algorithm where a large number of decision trees are grown, each with a random split of training data from the original data with replacement, using random feature selection/node splits. After which each tree votes for the most popular class at input *x* (Breiman, 2001). The goal here is to produce a stronger model than single decision trees alone
Artificial neural networks	Non-linear classification methods which make no underlying assumptions to limit their fit to the data (Zhang, 2000). A series of interconnected nodes are linked between predictors and output in a similar way as a neural network in the human brain
Regression	
Least absolute shrinkage and selection operator (lasso) regularization with linear regression	In the regression equation, lasso sets certain coefficients to 0, with the goal of increasing prediction accuracy whilst maintaining interpretability (Tibshirani, 1996)
Random forest feature selection, used with linear regression	Features identified by random forest (as described previously) are used to enhance performance of statistical regression algorithms

Table 3. Summary of aggregated risk of bias ratings using PROBAST (Wolff et al., 2019) by study (n = 75 total risk of bias ratings)

Study	Number of models evaluated with PROBAST	Participants			Predictors				Outcome							Analysis										ROB conclusion for overall assessment
		1.1	1.2	Overall	2.1	2.2	2.3	Overall	3.1	3.2	3.3	3.4	3.5	3.6	Overall	4.1	4.2	4.3	4.4	4.5	4.6	4.7	4.8	4.9	Overall	
1. Bergeron et al. (2019)	N = 60	Y	PY	Low	N	NI	Y	High	PN	PY	N	NI	PN	PY	High	Y	NI	NI	NI	Y	N/A	N	Y	N/A	High	High
2. Cnossen et al. (2017)	N = 1	Y	Y	Low	Y	Y	Y	Low	Y	Y	Y	Y	Y	Y	Low	PY	Y	N	Y	Y	N/A	N	Y	PY	High	High
3. Gupta et al. (2017)	N = 2	Y	Y	Low	Y	Y	Y	Low	Y	Y	Y	Y	Y	Y	Low	PY	Y	N	N	Y	Y	N	N	PY	High	High
4. Hirata et al. (2016)	N = 1	Y	PY	Low	Y	NI	Y	Low	Y	Y	Y	Y	PY	Y	Low	Y	PY	Y	N	Y	N/A	N	N	N/A	High	High
5. Huttunen et al. (2016)	N = 1	Y	Y	Low	PY	PY	Y	Low	Y	Y	Y	Y	Y	Y	Low	Y	NI	Y	PY	Y	N/A	N	N	PY	High	High
6. Nishi et al. (2019)	N = 3	Y	Y	Low	PY	Y	Y	Low	Y	Y	Y	Y	PY	Y	Low	PY	Y	Y	N	Y	N/A	N	Y	NI	High	High
7. Shafiei et al. (2017)	N = 1	Y	Y	Low	PY	Y	Y	Low	Y	Y	Y	Y	NI	Y	Unclear	PN	NI	PY	PY	Y	N/A	N	PN	N/A	High	High
8. Stromberg et al (2018)	N = 3	Y	Y	Low	Y	NI	Y	Unclear	PY	Y	Y	PY	PY	Y	Low	Y	Y	PY	N	Y	N/A	N	N	NI	High	High
9. Walker et al. (2018)	N = 3	Y	Y	Low	Y	NI	Y	Unclear	Y	Y	Y	Y	PY	Y	Low	Y	Y	N	N	Y	N/A	N	N	NI	High	High

N = information sufficient to conclude high ROB; NI = no information to assess ROB; PN = information provided is not sufficient to confirm high ROB, but due to other important information high ROB can be inferred; PY = sufficient information has not been provided to conclude low ROB but due to design or other important information low ROB can be inferred ;ROB = risk of bias; Y = sufficient information provided to conclude low ROB for the item.
PROBAST findings are aggregated by study since each model in each study had the same risk of bias ratings.

Table 4. Rationale for risk of bias ratings by study from an aggregated synthesis of each prediction model

Study	Rationale for ROB conclusion
1. Bergeron et al. (2019)	2.1. Symptoms are measured inconsistently by either verbal disclosure or a self-report checklist 3.1. Outcome likely to include measurement error 3.3. Predictors were not excluded from outcome which was time until absence of predictors 3.4. No information on how time until symptom resolution was measured 3.5. Predictor information likely to be known due to outcome definition 4.2. Pre-processing of predictor information not adequately described 4.3. Not adequately described 4.4. Not adequately described 4.7. Improper model evaluation, not assessing calibration
2. Cnossen et al. (2017)	4.3. Although participants were excluded with missing outcomes, between-group differences were explored for missing outcomes, showing no difference in baseline characteristics for lost to follow-up, thus minimizing this bias 4.7. Improper model evaluation, not assessing calibration
3. Gupta et al. (2017)	4.1. No reporting of events per candidate to fully assess dimensionality of data when sample size is small 4.3. Participants were excluded with missing predictors and outcomes; between-group differences were explored for missing outcomes, showing no difference in baseline characteristics for lost to follow-up, thus minimizing this bias 4.4. As with 4.3 4.7. Improper model evaluation, not assessing calibration 4.8. Internal cross-validation was not used to account for overfitting
4. Hirata et al. (2016)	4.4. Participants were excluded for missing the outcome variable. No information is provided on handling of missing predictor information 4.7. Improper model evaluation, not assessing calibration 4.8. No use of internal or external validation
5. Huttunen et al. (2016)	4.2. Data handling not adequately described 4.7. No model evaluation 4.8. No internal or external validation to account for overfitting
6. Nishi et al. (2019)	4.1. No reporting of events per candidate to fully assess dimensionality of data when sample size is small 4.3. Inappropriate exclusion for people with missing predictor and outcome data with no imputation 4.7. Improper model evaluation, not assessing calibration

Continued

Table 4. (*Continued*)

Study	Rationale for ROB conclusion
	4.9. Final predictive algorithms and coefficients are not reported
7. Shafiei et al. (2017)	3.5. Prospective design and no information on blinding to predictor variables during outcome determination
	4.1. Small sample size with a complex model architecture
	4.2. No information on handling of predictor variables
	4.7. Improper model evaluation, not assessing calibration
	4.8. Likely overfitting due to the 50/50 training test split for internal validation without external validation to accommodate, meaning parameter estimates have less variance
8. Stromberg et al (2018)	2.2. No information on whether predictor assessments were made without knowledge of outcome data
	4.4. Missing outcome excluded without exploration for impact on ROB
	4.7. Improper model evaluation, not assessing calibration
	4.8 A single split 85/15 validation was used increasing likelihood of overfitting and model optimism
	4.9. No information on whether the model was refitted after pruning
9. Walker et al. (2018)	4.3. Removal of participant data beyond those stated by exclusion criteria
	4.4. Missing outcome and missing covariate excluded without exploration for ROB
	4.7. Improper model evaluation, not assessing calibration
	4.8 A single split 85/15 validation was used increasing likelihood of overfitting and model optimism
	4.9. Unclear if predictors in the final models correspond to results from analysis as training data presented only

Table 5. Summary of performance metrics and reliability of findings using machine learning to predict psychosocial outcomes in acquired brain injury

	Performance metrics Results are area under the curve (AUC) unless otherwise stated			
Machine learning algorithms	Model development	Internal validation	External validation	Overall risk of bias
Classification				
Regularized logistic regression	(1) n/a (2) Two models developed ranging from 0.74 to 0.76 (6) n/a	(1) Six models ranging from 0.63 to 0.69 (2) n/a (6) 0.86	(1) n/a (2) n/a (6) 0.90	(1) High (2) High (6) High
Support vector machine	(1) n/a (6) n/a	(1) Six models ranging from 0.63 to 0.69 (6) 0.86	(1) n/a (6) 0.89	(1) High (6) High
Decision trees	(1) n/a (5) n/a (8) Three models developed ranging from 0.70 to 0.77 (9) Three models developed ranging from 0.70 to 0.73	(1) Twelve models ranging from 0.59 to 0.64 for C4.5D algorithms and 0.60-0.67 for C4.5N algorithms (5) n/a (8) Three models ranging from 0.73 to 0.77 (9) Three models developed ranging from 0.69 to 0.73	(1) n/a (5) n/a (8) n/a (9) n/a	(1) High (5) High (8) High (9) High
Naïve Bayes	(1) n/a	(1) Six models ranging from 0.66 to 0.74	1.) n/a	(1) High
K-nearest neighbours	(1) n/a	(1) Six models ranging from 0.64 to 0.69	1.) n/a	(1) High
Random forest	(1) n/a (4) n/a (6) n/a	(1) Twelve models ranging from 0.66 to 0.73 for 100-tree models, and 0.66-0.74 for 500-tree models	(1) n/a (4) n/a (6) 0.87	(1) High (4) High (6) High

Continued

Table 5. (*Continued*)

Machine learning algorithms		Performance metrics Results are area under the curve (AUC) unless otherwise stated Model development	Internal validation	External validation	Overall risk of bias
			(4) Accuracy 69% (specificity 70% and sensitivity 64%) (6) 0.85		
Random forest feature selection, used with logistic regression		(3) Two models developed ranging from 0.89 to 0.89	(3) n/a	3.) Two models developed ranging from 0.75 to 0.84	(3) High
Artificial neural networks	Multilayer perceptron	(1) n/a	(1) Six models ranging from 0.63 to 0.67	(1) n/a	(1) High
	Backpropagation	(7) n/a	(7) 86.9	(7) n/a	(7) High
	Radial basis function network	(1) n/a	(1) Six models ranging from 0.61 to 0.71	(1) n/a	(1) High
Regression					
Least absolute shrinkage and selection operator regularization with linear regression		(2) 21% of the variance	(2) 14% of the variance	(2) n/a	(2) High

1. Bergeron et al. (2019); 2. Cnossen et al. (2017); 3. Gupta et al. (2017); 4. Hirata et al. (2016); 5. Huttunen et al. (2016); 6. Nishi et al. (2019); 7. Shafiei et al. (2017); 8. Stromberg et al. (2018); and 9. Walker et al. (2018).

Which ML algorithms are most commonly used?

Decision trees methodology was most commonly used for predicting psychosocial outcomes in the field of ABI over recent years with four studies using the technique (Bergeron et al., 2019; Huttunen et al., 2016; Stromberg et al., 2019; Walker et al., 2018), followed by RF (Bergeron et al., 2019; Hirata et al., 2016; Nishi et al., 2019) and RLR (Bergeron et al., 2019; Cnossen et al., 2017; Nishi et al., 2019) with three studies each and then SVM (Bergeron et al., 2019; Nishi et al., 2019) and ANNs (Bergeron et al., 2019; Shafiei et al., 2017) with two studies each.

What is the rationale for the choice of ML algorithms, as stated by the study authors?

The rationale for the authors' choices in ML algorithms is presented in Table 6. There was no reported information for NB, radial bias function network, multilayer perceptron, or *K*NN, as not all authors included a detailed rationale for their choices of ML algorithms (Bergeron et al., 2019; Huttunen et al., 2016). For example, Bergeron et al. (2019) opted to compare ten different algorithms due to the absence of published guidance for suitability of different algorithms, and Nishi et al. (2019) chose three commonly used algorithms, although with the further rationale that they benefited from ranking of features.

Of the nine studies, only one (Cnossen et al., 2017) provided an a priori consideration for whether the type of analysis was suitable for their data (whether sample size was appropriate for the algorithm to minimize risk of overfitting). One study (Gupta et al., 2017) conducted a post-hoc power analysis; however since the findings scored at high ROB, the power analysis would also be unreliable. A further four did consider the possible implications of sample size in their limitations (Cnossen et al., 2017; Nishi et al., 2019; Stromberg et al., 2019; Walker et al., 2018). Only four of the nine studies critically evaluated the ML methodology in their limitations, as reported in Table 6. Some of these reported limitations are considered in the discussion of this review as to how these could have been overcome by more suitable study design, analysis, and model evaluation.

Discussion

The primary aim of this systematic review was to evaluate the effectiveness of using ML to predict psychosocial outcomes after ABI; however, no study reviewed had reliable findings when assessed for ROB to allow a conclusion. Whilst this might make ML seems like a daunting method for clinicians, bias tended to be introduced from improper analysis design relevant for ML and traditional predictive methods alike. The most common data and analysis shortcoming was improper model evaluation without assessment of calibration for nine out of nine studies. Calibration assessment can inform of likely over- or underfitting to consider how the models will perform in new samples. This is commonly quantified by the calibration slope (based on a plot of the observed outcomes and model predictions), with values near 1 representing better calibration. If models are poorly calibrated, findings may be inaccurate for new predictions, limiting the applicability of the models for future clinical cases (i.e., the external validity). Further data and analysis shortcomings included either inadequate reporting or improper handling of missing data in six of the nine studies, five studies not fully accounting for model optimism or overfitting, and four studies having excluded people inappropriately from the analysis. The resulting high ROB meant that this review was unable to answer the

Table 6. Rationale and limitations of machine learning algorithms as provided by the authors of reviewed studies

Machine learning algorithm	Rationale for author choice of algorithm	Limitations as stated by study authors
Regularization with logistic or linear regression	Regularization (lasso) gives less extreme ß values which improves external validity (Cnossen et al., 2017). Coefficient ranking allows for understanding the contribution of each feature, and deals with feature selection, multicollinear variables and overfitting better than statistical regression models (Nishi et al., 2019)	Lasso regularization as used by Cnossen et al. (2017) focussed on overall fit of the predictors, meaning poorly contributing predictors could still be included in their model
Support vector machine	Allows for understanding the contribution of each feature (Nishi et al., 2019)	None reported
Decision trees	Easily interpreted by clinicians due to similar decision-making process allowing greater clinical utility than ensemble methods (Stromberg et al., 2019). Predictors are identified by branching logic allowing flexible predictions (Walker et al., 2018)	Decision tree methodology may have limited predictive power compared to statistical regression (Stromberg et al., 2019; Walker et al., 2018). Branching is limited by sample size in terminal nodes, and its data-driven nature means different models may not be consistent (Stromberg et al., 2019)
Random forest	Feature selection is a strength with less decision-making error than traditional statistical methods (Gupta et al., 2017; Hirata et al., 2016). Allows for understanding the contribution of each feature (Nishi et al., 2019)	None reported
Artificial neural networks and backpropagation	Are not limited by parametric formulas allowing greater flexibility and more complexity (Shafiei et al., 2017)	Increasing hidden layer nodes can contribute to overfitting to the training data. Also does not benefit from feature ranking, is interpretationally complex, and computationally time-consuming (Shafiei et al., 2017)

Limitations and strengths reported in this table are from information presented in the original articles. Where limitations can be overcome by study design, this is mentioned in the discussion of this review.

primary review question of which algorithms are most effective for predicting psychosocial outcomes in ABI.

Decision trees methodology was the most popular choice for psychosocial ABI research over the review dates, being easy to interpret and lending well to clinical decision-making. As noted above, the application of the technique was unfortunately too poor to allow conclusions to be drawn regarding its efficacy. Stromberg et al. (2019) note as a limitation to DTs that when models are repeated, they are prone to modelling the data differently. This is actually true for all ML techniques (each time learning from the data). In order to overcome this limitation, models should be thoroughly internally validated, a process where multiple models are developed by dividing the data set into 'training' and 'testing' segments, where commonly, the model is trained using the data in one section, and then tested in the reserved section of data, adjusting its algorithm based on the accuracy of each tested prediction. The aim here is to minimize risk of overfitting and adjust for model optimism; thus, the more times this process is repeated, the more the model learns from its error to tune its performance. External validation then assesses the generalizability of a given model by testing its performance in a novel data set.

To reduce bias, internal validation procedures with numerous repeats of model development (e.g., nested cross-validation or bootstrapping) give a more stable and reliable fit to the training data (Wolff et al., 2019). Three of the four DT studies reviewed here employed improper techniques to internally validate their models (such as splitting the data set once where 85% of the data was used for model development and the remaining 15% reserved for validation, without repeating the process), leading to models which are likely overly optimistic and without reliable predictor branching (Huttunen et al., 2016; Stromberg et al., 2019; Walker et al., 2018). The other DT study did employ a 10-fold cross-validation procedure (Bergeron et al., 2019); however, it is unclear whether this was a nested cross-validation to fully minimize risk of overfitting. The unfortunate result means the produced models are unreliable for clinicians to be able to apply the DT to clinical cases (the ultimate goal of clinical predictive modelling), being unable to make use of this easily interpretable and time-efficient method for clinical decisions.

As well as DT methodology, RF, RLR, and SVM were commonly used approaches for psychosocial ABI research, which collectively allow for prioritization of predictors in order of importance (with RLR and RF having embedded feature selection). Feature ranking serves obvious benefits for clinicians working with ABI, allowing easy identification of risk factors for poor outcomes and, after further investigation, possibly even serving as targets for intervention. ANNs were also used more frequently for predicting psychosocial outcomes (Bergeron et al., 2019; Shafiei et al., 2017). ANNs however are often described as being a 'black box' when it comes to interpretation, informing little regarding predictors of value (Zhang et al., 2018). Methods with embedded feature selection may therefore be preferable for many of the research questions ABI clinicians have, inspecting a wider range of features for predictive power than is possible with traditional statistical methods.

Further common sources of ROB came from excluding people for missing the outcome of interest in predictive models which can introduce bias if missing not at random (Wolff et al., 2019). Two studies addressed this ROB by exploring differences between those with and without outcome data, showing no significant differences (Cnossen et al., 2017; Gupta et al., 2017). This benefits readers' understanding, knowing how response bias could impact on results and therefore how reliable the algorithm might be for new clinical cases.

Additionally, every study reviewed here failed to evaluate ML models by calibration. This omission in predictive modelling is not unique to ABI research: a previous prediction systematic review found that around 80% of studies did not assess calibration (Christodoulou et al., 2019). Together, these limitations of poor calibration assessment, inadequate validation procedures, and infrequent exploration around outcomes not missing at random mean these models provide little evidence for their benefit for future clinical decision-making.

Finally, authors often provided minimal information for their choice of ML algorithms. This may be because guidance around ML for psychosocial predictions in ABI has previously been limited. Among all studies reviewed, only one study reported an a priori decision about the suitability of their data for the algorithm (Cnossen et al., 2017). Although some ML algorithms handle high-dimensional data sets better than traditional statistical modelling, such as with embedded feature selection, not every ML algorithm is suitable for every data set. Just like traditional statistical modelling, ML algorithms cope differently with the number of predictor variables in relation to number of patient cases, as well as the noise in predictor variables (Guo, Graber, McBurney, & Balasubramanian, 2010). Whilst ML is often put forward as being a methodology with less concern of overfitting and better capability for dealing with multicollinear and multidimensional data than traditional statistical techniques (Iniesta, Stahl, & McGuffin, 2016), ML is not immune to these problems. Consideration of appropriateness of the analysis for the data, as well as thorough model evaluation, is still required as part of study design to determine efficacy.

Limitations of the review

Whilst this review benefits from being the first to systematically review ML for making psychosocial predictions in ABI, there are several limitations. Firstly, papers in this review were restricted to those published from 2016. This was because the TRIPOD statement (Moons et al., 2015) was not released until 2015 so it is likely there was a change in publication quality in articles published after. Additionally, for using PROBAST (Wolff et al., 2019) it is advised that a statistical expert fully reviews the articles; however, this was not possible within the scope of this work. Finally, our screening and rating method was completed for only a percentage of total articles by both raters. There is the possibility of some differing opinions, but this should mostly be minimized due to the high inter-rater concordance.

Future directions

This systematic review has identified a number of common omissions in ABI research using ML which limit the applicability of the produced models for future clinical decision-making. In addition to the more general guidance published in PROBAST (Wolff et al., 2019) and TRIPOD (Moons et al., 2015), researchers in this field may benefit from the following considerations when designing an ML study for predicting psychosocial outcomes in ABI:

Data handling, pre-processing, and algorithm selection

1. Inspect and/or clean the data for issues that may affect algorithm performance (e.g., highly correlated predictor variables, predictors with little variance, patterns of missing data, the ratio of predictor variables to patient cases). Consider either

cleaning the data to remove these variables if applicable or to select an algorithm that is less affected by the issues of a particular data set.
2. Calculate an a priori power analysis (e.g., events per variable) to ensure the model is sufficiently powered to minimize risk of error.
3. Algorithm selection: Researchers should keep both the research question and appropriateness for data in mind when choosing which ML algorithm to use (e.g., RF or RLR for research questions aiming to understand more about important predictors, DT (with proper validation methods) for studies aiming for easy translation to clinical practice, or opting for simpler models for smaller sample sizes (e.g., linear models over non-parametric models)).
4. Handling of missing data:

   a. Outcome data: Whilst whole sample analyses are preferable for the external validity of the model, these are not always possible with clinical data sets. With specific methods, the outcome variable can be imputed, or otherwise if those with missing outcome data are excluded, bias will be minimized through exploration of whether data are missing at random (e.g., significance testing of differences in predictor variables between those with and without the outcome of interest).
   b. Predictor data: Where possible, missing data should be imputed rather than excluded when appropriate quantities of complete data are available.

Model development and evaluation

1. *Validation*: Certain methods of internal validation commonly used in studies reviewed are often prone to bias by not repeating the procedure multiple times to reduce risk of overfitting or model optimism (e.g., cross-validation, or single split train/test validation methods). Nested cross-validation (which also optimizes hyperparameters) and bootstrapping are superior methods for internal validation. External and/or temporal validation are important for assessing model accuracy for clinical applicability, but these should be used in conjunction with, not instead of, thorough internal validation procedures.
2. *Model evaluation*: Binary models are frequently evaluated by the AUC only; however, this informs little for applying the model to new clinical cases. Researchers should evaluate models by discrimination, calibration, and power, and evaluate limitations for transparent reporting.

Conclusions

Overall, this review was unable to provide a conclusion as to which ML algorithm was most suitable for psychosocial ABI research; however, it has demonstrated current poor methodological quality and a lack of rationale for use of ML algorithms by clinical researchers. Researchers should consider which ML algorithms will be most suitable for the purpose of the research question, as well as the suitability of their data for different algorithms (such as appropriate sample sizes, power calculations, analysis of missing data, and suitable validation methods for data size). More thorough post-hoc model evaluation by calibration, discrimination, and where possible external validation will greatly increase the quality and reliability for the application of ML for new clinical predictions. Clearly, moving to a more systematically planned application of ML rather than a 'try it and see'

approach is needed to ensure the method and study design are able to answer the research questions for future applications.

Conflicts of interest

All authors declare no conflict of interest.

Author contributions

Emma Mawdsley, DClinPsy (Conceptualization; Formal analysis; Investigation; Project administration; Writing – original draft; Writing – review & editing) Bronagh Reynolds (Investigation; Validation; Writing – review & editing) Breda Cullen (Conceptualization; Supervision; Writing – review & editing).

Data Availability Statement

Data sharing is not applicable to this article as no new data were created or analysed in this study.

References

Bergeron, M. F., Landset, S., Maugans, T. A., Williams, V. B., Collins, C. L., Wasserman, E. B., & Khoshgoftaar, T. M. (2019). Machine learning in modeling high school sport concussion symptom resolve. *Medicine and Science in Sports & Exercise*, *51*, 1362–1371. https://doi.org/10.1249/mss.0000000000001903

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Briggs, J. (2017). *Critical appraisal tools:Checklist for cohort studies*. Adelaide, Australia: Joanna Briggs Institute.

Bzdok, D., Krzywinski, M., & Altman, N. (2017). Machine learning: A primer. *Nature Methods*, *14*, 1119–1120. https://doi.org/10.1038/nmeth.4526

Christodoulou, E., Ma, J., Collins, G. S., Steyerberg, E. W., Verbakel, J. Y., & Van Calster, B. (2019). A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. *Journal of Clinical Epidemiology*, *110*, 12–22. https://doi.org/10.1016/j.jclinepi.2019.02.004

Cnossen, M. C., Winkler, E. A., Yue, J. K., Okonkwo, D. O., Valadka, A. B., Steyerberg, E. W., . . . the TRACK-TBI Investigators, (2017). Development of a prediction model for postconcussive symptoms following mild traumatic brain injury: A track-TBI pilot study. *Journal of Neurotrauma*, *34*, 2396–2409. https://doi.org/10.1089/neu.2016.4819

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20*(3), 273–297. https://doi.org/10.1007/BF00994018

Cruz-Cunha, M. M. (Ed.). (2013). *Handbook of research on enterprise 2.0: Technological Social, and organizational dimensions: technological, social, and organizational dimensions*. Pennsylvania, PA: IGI Global. https://doi.org/10.4018/978-1-4666-4373-4

Cunningham, P., & Delany, S. J. (2020). k-Nearest Neighbour Classifiers: 2nd Edition (with Python examples). ArXiv, abs/2004.04523. https://arxiv.org/pdf/2004.04523.pdf

Guo, Y., Graber, A., McBurney, R. N., & Balasubramanian, R. (2010). Sample size and statistical power considerations in high-dimensionality data settings: A comparative study of classification algorithms. *BMC Bioinformatics*, *11*, 447. https://doi.org/10.1186/1471-2105-11-447

Gupta, V. P., Garton, A. L. A., Sisti, J. A., Christophe, B. R., Lord, A. S., Lewis, A. K., . . . Connolly, E. S. (2017). Prognosticating functional outcome after intracerebral hemorrhage: The ICHOP Score. *World Neurosurgery*, *101*, 577–583. https://doi.org/10.1016/j.wneu.2017.02.082

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction*. Berlin, Germany: Springer Science & Business Media. https://web.stanford.edu/~hastie/Papers/ESLII.pdf

Hirata, S., Ovbiagele, B., Markovic, D., & Towfighi, A. (2016). Key factors associated with major depression in a national sample of stroke survivors. *Journal of Stroke and Cerebrovascular Diseases*, *25*, 1090–1095. https://doi.org/10.1016/j.jstrokecerebrovasdis.2015.12.042

Huttunen, J., Lindgren, A., Kurki, M. I., Huttunen, T., Frösen, J., von und zu Fraunberg, M., . . . Immonen, A. (2016). Antidepressant use after aneurysmal subarachnoid hemorrhage a population-based case-control study. *Stroke*, *47*, 2242–2248. https://doi.org/10.1161/strokeaha.116.014327

Iniesta, R., Stahl, D., & McGuffin, P. (2016). Machine learning, statistical learning and the future of biological research in psychiatry. *Psychological Medicine*, *46*, 2455–2465. https://doi.org/10.1017/S0033291716001367

Liberati, A., Altman, D. G., Tetzlaff, J., Mulrow, C., Gøtzsche, P. C., Ioannidis, J. P. A., . . . Moher, D. (2009). The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate healthcare interventions: Explanation and elaboration. *Journal of Clinical Epidemiology*, *62*(10), e1–e34. https://doi.org/10.1016/j.jclinepi.2009.06.006

Mascolo, M. F. (2016). Beyond objectivity and subjectivity: The intersubjective foundations of psychological science. *Integrative Psychological and Behavioral Science*, *50*(4), 543–554. https://doi.org/10.1007/s12124-016-9357-3

Moher, D., Shamseer, L., Clarke, M., Ghersi, D., Liberati, A., Petticrew, M., . . . Stewart, L. A. (2015). Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015 statement (OriginalPaper). *Systematic Reviews*, *4*(1), 1–9. https://doi.org/10.1186/2046-4053-4-1

Moons, K. G. M., Altman, D. G., Reitsma, J. B., Ioannidis, J. P. A., Macaskill, P., Steyerberg, E. W., . . . Collins, G. S. (2015). Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (tripod): Explanation and elaboration. *Annals of Internal Medicine*, *162*(1), W1–W73. https://doi.org/10.7326/M14-0698

Nadkarni, P. (2016). Chapter 4 - core technologies: Machine learning and natural language processing. *Clinical research computing* (pp. 85-114). Cambridge, MA: Academic Press.

Nishi, H., Oishi, N., Ishii, A., Ono, I., Ogura, T., Sunohara, T., . . . Miyamoto, S. (2019). Predicting clinical outcomes of large vessel occlusion before mechanical thrombectomy using machine learning. *Stroke*, *50*, 2379–2388. https://doi.org/10.1161/strokeaha.119.025411

Rish, I. (2001, August). *An empirical study of the naive Bayes classifier*. In *IJCAI 2001 workshop on Empirical Methods in Artificial intelligence*, 3(22), 41–46. https://doi.org/10.1.1.330.2788

Rokach, L., & Maimon, O. Z. (2008). *Data mining with decision trees: Theory and applications* (Vol. *69*). Singapore City, Singapore: World Scientific. https://doi.org/10.1142/6604

Safari, S., Baratloo, A., Elfil, M., & Negida, A. (2016). Evidence based emergency medicine; Part 5 receiver operating curve and area under the curve. *Emergency*, *4*(2), 111–113. https://doi.org/10.22037/AAEM.V4I2.232.G232

Senders, J. T., Staples, P. C., Karhade, A. V., Zaki, M. M., Gormley, W. B., Broekman, M. L. D., . . . Arnaout, O. (2018). Machine learning and neurosurgical outcome prediction: A systematic review. *World Neurosurgery*, *109*, 476–486.e1. https://doi.org/10.1016/j.wneu.2017.09.149

Shafiei, E., Fakharian, E., Omidi, A., Akbari, H., Delpisheh, A., & Nademi, A. (2017). Comparison of artificial neural network and logistic regression models for prediction of psychological symptom six months after mild traumatic brain injury. *Iranian Journal of Psychiatry and Behavioral Sciences*, *11*(3), e5849. https://doi.org/10.17795/ijpbs-5849

Stromberg, K. A., Agyemang, A. A., Graham, K. M., Walker, W. C., Sima, A. P., Marwitz, J. H. et al (2019). Using decision tree methodology to predict employment after moderate to severe

traumatic brain injury. *The Journal of Head Trauma Rehabilitation*, *34*(3), E64–E74. https://doi.org/10.1097/HTR.0000000000000438

Teasdale, G., & Jennett, B. (1974). Assessment of coma and impaired consciousness. *A Practical Scale. Lancet*, *2*, 81–84. https://doi.org/10.1016/s0140-6736(74)91639-0

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, *58*(1), 267–288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

Walker, W. C., Stromberg, K. A., Marwitz, J. H., Sima, A. P., Agyemang, A. A., Graham, K. M., . . . Merchant, R. (2018). Predicting long-term global outcome after traumatic brain injury: Development of a practical prognostic tool using the traumatic brain injury model systems national database. *Journal of Neurotrauma*, *35*(14), 1587–1595. https://doi.org/10.1089/neu.2017.5359

Wolff, R. F., Moons, K. G. M., Riley, R. D., Whiting, P. F., Westwood, M., Collins, G. S., . . . Mallett, S. (2019). PROBAST: A tool to assess the risk of bias and applicability of prediction model studies. *Annals of Internal Medicine*, *170*(1), 51–58. https://doi.org/10.7326/M18-1376

Zhang, G. P. (2000). Neural networks for classification: A survey. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, *30*(4), 451–462. https://doi.org/10.1109/5326.897072

Zhang, Z., Beck, M. W., Winkler, D. A., Huang, B., Sibanda, W., & Goyal, H. (2018). Opening the black box of neural networks: Methods for interpreting neural network models in clinical applications. *Annals of Translational Medicine*, *6*, 216. https://doi.org/10.21037/atm.2018.05.32

Received 14 November 2020; revised version received 1 March 2021

Supporting Information

The following supporting information may be found in the online edition of the article:

Appendix S1. Search strategy for OVID interface.
Appendix S2. Data extraction template.
Appendix S3. Abstract summaries of included studies.