Institute of Health & Wellbeing

Machine learning for the prediction of psychosocial outcomes in acquired brain injury

Emma Mawdsley, MSc, BSc Honours

Submitted in partial fulfilment of the requirements for the degree of Doctorate in Clinical Psychology

Institute of Health and Wellbeing

College of Medical, Veterinary and Life Sciences

University of Glasgow

September, 2020

Acknowledgements

I would like to give a special thanks to my supervisors, Breda Cullen, Brian O'Neill, and Samuel Leighton, for your never-ending support for the completion of this project. I felt very privileged to have such an engaged and enthusiastic team of supervisors, with everyone so quick to respond to queries and provide such helpful guidance throughout.

Thank you so much to the service users and staff at Graham Anderson House for enabling this project. It was a pleasure to meet you all over my visits to the wards. In particular, thank you so much to Brian for the long hours and late nights helping me get the data ready within minutes of the coronavirus lockdown. It was definitely a unique way to start a project.

To my colleague and friend, Bronagh Reynolds, thank you for your help as the second marker, the emotional support, and to always be available for a celebratory glass of wine after each of our hand in dates for our drafts. The doctorate was so much easier with a research buddy.

My whole class of 2017, it was a delight completing the doctorate with you. You will all hold such a special place in my heart.

And finally, thank you to my family for all of your support. A particular thanks to my partner, David, who not only was a huge support throughout this time, he also allowed me a break from laundry duty for a full four months. I will forever be grateful.

CONTENTS

List of Tables

List of Figures

Chapter One: Systematic Review

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] & Breda Cullen[1]

Chapter word count: 6,259

Prepared in accordance with the Journal of Neuropsychology (appendix 1.1, pg.47)

[1] **Mental Health and Wellbeing, Institute of Health and Wellbeing, University of Glasgow**
***Correspondence Address:**
Mental Health and Wellbeing
Administration Building
Gartnavel Royal Hospital
1055 Great Western Road Glasgow
Glasgow
G12 0XH
[2] **NHS Greater Glasgow and Clyde, Scotland**
[3] **NHS Ayrshire and Arran**, **Scotland**

ORCID ID: https://orcid.org/0000-0002-4061-5152

Declaration of Conflict of Interest: None

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

Abstract

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.

Word count: 207

Keywords

Machine learning; brain injury; stroke; predictive research; systematic review

Introduction

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. Being able to accurately predict psychosocial outcomes at a future time-point after ABI would serve timely resource allocation and risk management, as well as being able to adapt interventions for known risk factors to maximise 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. 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, Altman, Reitsma, 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 efficacy of using ML algorithms across ABIs and 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 1.2 on page 56. The search results were managed in the author's EndNote library (www.myendnoteweb.com). Duplicates were removed during database extraction, 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 1.3 (pg. 57). The form was piloted by the primary author for 5 studies, 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 3 studies giving consistency of 93.1%, 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. 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 both for 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 are reported. For the third review question, the rationale of the author's choice of methodology was summarised. 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

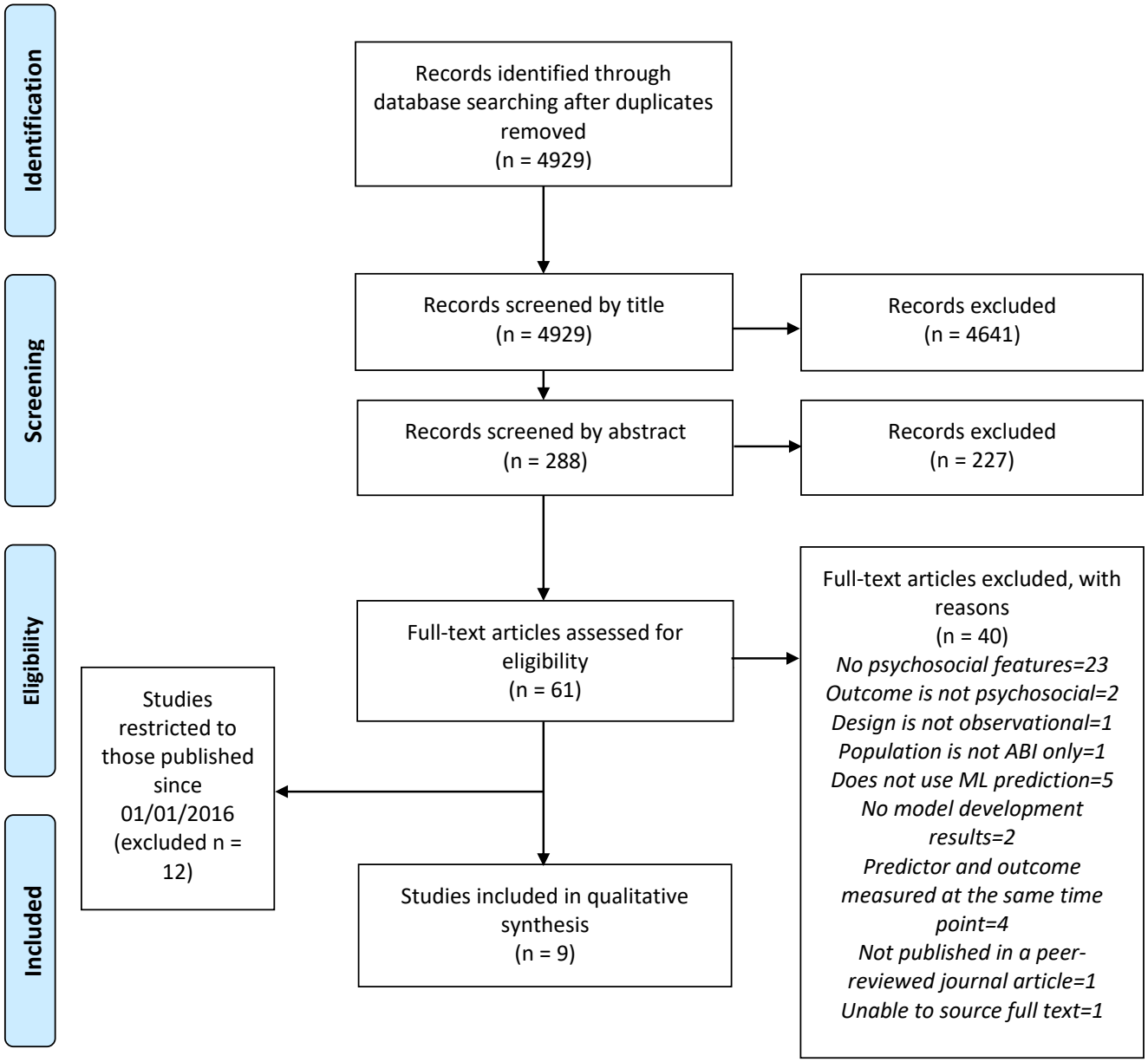
Study selection

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. Six were from the United States (Bergeron et al., 2019; Cnossen et al., 2017; Gupta et al., 2017; Hirata et al., 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.

Figure 1 PRISMA flow diagram of the study selection process



Abbreviations: ABI= acquired brain injury; ML= machine learning

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), 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).

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	1611 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 back-propagation 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	7867 (1 year) 6783 (2 year) 4927 (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

Abbreviations: ABI= Acquired brain injury; ANN= Artificial neural network; DT= Decision tree; GCS= Glasgow coma score; KNN= K-Nearest Neighbours; ML= Machine learning; NB= Naïve Bayes; RF= Random forest; RLR= Regularised logistic regression; SVM= Support vector machine; TBI= Traumatic brain injury

Across the nine studies there were a total of 11 types of ML: regularised 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, back propagation and radial basis function network), lasso regularisation with linear regression, and random forest used for feature selection with logistic regression. Algorithm descriptions can be found in Appendix 1.4 on pg. 59. 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 2), with the rationale for ROB scores in appendix 1.5 on pg. 61. 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 are included in Table 3. Models with an AUC of 0.80 or above are considered to show 'good' performance, between 0.70-0.79 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 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	Y	NI	High	High

PROBAST findings are aggregated by study since each model in each study had the same risk of bias ratings

Abbreviations: 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

Table 3 Summary of performance metrics and reliability of findings using machine learning to predict psychosocial outcomes in acquired brain injury

Machine learning algorithms			Performance metrics Results are area under the curve (AUC) unless otherwise stated			Overall risk of bias
			Model development	Internal validation	External validation	
Classification	Regularised logistic regression		1.) n/a 2.) Two models developed ranging from 0.74-0.76 6.) n/a	1.) Six models ranging from 0.63-0.69 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 0.63-0.69 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 0.70-0.77 9.) Three models developed ranging from 0.70-0.73	1.) Twelve models ranging from 0.59-0.64 for C4.5D algorithms and 0.60-0.67 for C4.5N algorithms 5.) n/a 8.) Three models ranging from 0.73-0.77 9.) Three models developed ranging from 0.69-0.73	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 0.66-0.74	1.) n/a	1.) High
	K-nearest neighbours		1.) n/a	1.) Six models ranging from 0.64-0.69	1.) n/a	1.) High
	Random forest		1.) n/a 4.) n/a 6.) n/a	1.) Twelve models ranging from 0.66-0.73 for 100-tree models, and 0.66-0.74 for 500-tree models 4.) Accuracy 69% (specificity 70% and sensitivity 64%) 6.) 0.85	1.) n/a 4.) n/a 6.) 0.87	1.) High 4.) High 6.) High
	Random Forest feature selection, used with logistic regression		3.) Two models developed ranging from 0.89-0.89	3.) n/a	3.) Two models developed ranging from 0.75-0.84	3.) High
	Artificial neural networks	Multilayer perceptron	1.) n/a	1.) Six models ranging from 0.63-0.67	1.) n/a	1.) High
		Back propagation	7.) n/a	7.) 86.9	7.) n/a	7.) High
		Radial basis function network	1.) n/a	1.) Six models ranging from 0.61-0.71	1.) n/a	1.) High
Regression	Least absolute shrinkage and selection operator regularisation 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); 9. Walker et al (2018)

Which ML algorithms are most commonly used?

DT 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, 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 are presented in Table 4. There was no reported information for NB, radial basis function network, multilayer perceptron, or *K*NN, as not all authors included a detailed rationale for their choices of ML algorithms (Bergeron, 2019, Huttunen et al., 2016). For example, Bergeron and colleagues (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 minimise 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 4. 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.

Table 4 Rationale and limitations of ML algorithms as provided by the authors of reviewed studies

Machine Learning algorithm	Rationale for author choice of algorithm	Limitations as stated by study authors
Regularisation with logistic or linear regression	Regularisation (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 regularisation 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 clinicans 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, Back propagation	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.

Discussion

The primary aim of this systematic review was to evaluate the efficacy 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 seem 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 shortcomings included improper model evaluation without assessment of calibration for nine out of nine studies, followed by six of nine with either inadequate reporting or improper handling of missing data, 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 primary review question of which algorithms are most effective for predicting psychosocial outcomes in ABI.

DT 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 and colleagues (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, and the results are averaged to minimise risk of overfitting and adjust for model optimism.

To reduce bias, internal validation procedures with numerous repeats of model development (e.g. 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 dataset 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 if this was a nested cross validation to fully minimise 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 prioritisation 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. Calibration assessment can inform of likely over- or underfitting to consider how the models will perform in new samples. If models are poorly calibrated, findings may be inaccurate for new predictions. 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 datasets better than traditional statistical modelling, such as with embedded feature selection, not every ML algorithm is suitable for every dataset. Just like traditional statistical modelling, ML algorithms cope differently with sample size to dimension ratio, and 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 are 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 minimised due to the high interrater concordance.

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 and type of data, such as whether their research question would benefit understanding of important predictors (such as with RLR or RF), or whether an easily interpretable method would be beneficial for translation to clinical practice (such as DTs). Greater a priori decisions for the suitability of the data for different algorithms (such as appropriate sample sizes and power calculations, analysis of missing data, and suitable validation methods for data size), as well as 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.

References

Bergeron, M. F., Landset, S., Maugans, T. A., Williams, V. B., Collins, C. L., Wasserman, E. B., et al. (2019). Machine Learning in Modeling High School Sport Concussion Symptom Resolve. *Medicine & Science in Sports & Exercise*, *51*(7), 1362-1371. DOI: 10.1249/mss.0000000000001903.

Briggs, J. (2017). Critical Appraisal Tools :Checklist for Cohort Studies. Joanna Briggs Institute. Avaiable from https://joannabriggs.org/ebp/critical_appraisal_tools. Accessed 17/August/2019.

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

Christodoulou, E., 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. doi:https://doi.org/10.1016/j.jclinepi.2019.02.004.

Cnossen, M. C., Winkler, E., Yue, J., Steyerberg, E. W., Lingsma, H., & Manley, G. (2017). Development of a prediction model for postconcussive symptoms following mild traumatic brain injury: A track-TBI pilot study. *Journal of Neurotrauma*, *34* (16), 2396-2409. https://doi.org/10.1089/neu.2016.4819

Cruz-Cunha, M. M. (Ed.). (2013). *Handbook of Research on Enterprise 2.0: Technological, Social, and Organizational Dimensions: Technological, Social, and Organizational Dimensions*. IGI Global.

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. doi: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., et al. (2017). Prognosticating Functional Outcome After Intracerebral Hemorrhage: The ICHOP Score. *World Neurosurgery*, *101,* 577-583. doi:10.1016/j.wneu.2017.02.082.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: data mining, inference, and prediction.* Springer Science & Business Media.

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*(5), 1090-1095. https://doi.org/10.1016/j.jstrokecerebrovasdis.2015.12.042

Huttunen, J., Lindgren, A., Kurki, M. I., Huttunen, T., Frosen, J., von und zu Fraunberg, M., et al. (2016). Antidepressant Use After Aneurysmal Subarachnoid Hemorrhage A Population-Based Case-Control Study. *Stroke*, *47*(9), 2242-2248. doi: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*(12), 2455-2465. doi:10.1017/S0033291716001367.

Liberati, A., Altman, D. G., Tetzlaff, J., Mulrow, C., Gøtzsche, P. C., Ioannidis, J. P. A., et al. (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. doi:10.1007/s12124-016-9357-3.

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

Moons, K. M., Altman, D. G., Reitsma, J. B., et al. (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. doi:10.7326/M14-0698.

Nishi, H., Oishi, N., Ishii, A., Ono, I., Ogura, T., Sunohara, T., et al. (2019). Predicting Clinical Outcomes of Large Vessel Occlusion Before Mechanical Thrombectomy Using Machine Learning. *Stroke*, *50*(9), 2379-2388. doi:10.1161/strokeaha.119.025411.

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.

Senders, J. T., Staples, P. C., Karhade, A. V., Zaki, M. M., Gormley, W. B., Broekman, M. L. D., et al. (2018). Machine Learning and Neurosurgical Outcome Prediction: A Systematic Review. *World Neurosurgery*, *109*, 476-486.e1. doi: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. doi : 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. doi: 10.1097/HTR.0000000000000438.

Teasdale, G., & Jennett, B. (1974). Assessment of coma and impaired consciousness. A practical scale. *Lancet*, *2*(7872), 81-4. doi:10.1016/s0140-6736(74)91639-0.

Walker, W. C., Stromberg, K. A., Marwitz, J. H., Sima, A. P., Agyemang, A. A., Graham, K. M., et al. (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., et al. (2019). PROBAST: A Tool to Assess the Risk of Bias and Applicability of Prediction Model Studies. *Annals of Internal Medicine*, *170*(1), 51-58. doi:10.7326/M18-1376.

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*(11), 216. doi:10.21037/atm.2018.05.32.

Chapter Two: Major Research Project

Predicting psychosocial outcomes at the time of discharge from inpatient neurorehabilitation for acquired brain injury: Development and internal validation of machine learning models

Emma Mawdsley [1,2], Samuel Leighton[1], Brian O'Neill[3] & Breda Cullen[1]

Chapter word count: 6,589

Prepared in accordance with the Journal of Neuropsychology (appendix 1.1, pg. 47)

[1] **Mental Health and Wellbeing, Institute of Health and Wellbeing, University of Glasgow**
***Correspondence Address:**
Mental Health and Wellbeing
Administration Building
Gartnavel Royal Hospital
1055 Great Western Road Glasgow
Glasgow
G12 0XH
[2] **NHS Greater Glasgow and Clyde, Glasgow, Scotland**
[3] **Brain Injury Rehabilitation Trust, Graham Anderson House, Glasgow**

ORCID ID: https://orcid.org/0000-0002-4061-5152

Declaration of Conflict of Interest: None

Plain English Summary of Major Research Project

Title: Predicting psychosocial outcomes at the time of discharge from inpatient neurorehabilitation for acquired brain injury: Development and internal validation of machine learning models

Background: Acquired brain injury (ABI) is an injury to the brain after birth, such as a stroke or a blow to the head. ABI can cause significant lifestyle changes for the person affected and their family members. Rehabilitation inpatient programmes have been developed for people with severe ABI, which aim to improve their functioning and help them adapt to living with a brain injury. After discharge from rehabilitation, the level of functioning can vary from person to person, with differing levels of support needed for living arrangements, work and recreational activities (known as psychosocial outcomes).

Machine learning (ML) is a type of data analysis for making predictions. ML methods overcome some of the limitations that traditional predictive statistics have. ABI rehabilitation centres generate large volumes of data, which ML might be able to use to make more accurate predictions than other methods. This could give us a better idea of likely psychosocial outcomes when people are later discharged, allowing us to plan care packages in advance.

Aims and questions: This study aimed to compare ML models with traditional statistical techniques to predict psychosocial outcomes after discharge from inpatient rehabilitation. In particular, the study questioned whether one of three ML methods, or a statistical method would perform better at predicting five different psychosocial outcomes.

Methods: A database was developed of people who had been admitted and discharged from Graham Anderson House, a Glasgow-based inpatient rehabilitation centre, between 2009 and March 2020. The data gathered from admission assessments were used as predictors in analyses using three types of ML and one type of traditional statistics to predict five outcomes. The main outcome was the likelihood a person could live independently after discharge. These models were evaluated to determine which method had the most superior predictive power.

Main Findings: The analyses showed that a type of ML, called random forest, had better performance than the traditional statistical method for predicting every outcome. For each of the four outcomes that could be analysed in this study, the random forest method had at least a 70% chance of being correct. One outcome of interest (quality of life) did not have enough available data and so was not analysed.

Conclusions: Although these prediction models require some further evaluation before they could be used in clinical practice, being able to predict what a person's likely outcomes will be from the time of admission will be helpful for clinicians and social care workers to make advance decisions about people's care plans. This would help to reduce any unnecessary delays to funding or living arrangements. Being able to understand more about what contributes to good outcomes could also mean that rehabilitation programmes might be able to be tailored better to support people with ABI to maximise their independence.

Word count: 483

Predicting psychosocial outcomes at the time of discharge from inpatient neurorehabilitation for acquired brain injury: Development and internal validation of machine learning models

Abstract

Acquired brain injury (ABI) can be a life changing condition, affecting housing, independence, and employment. Machine learning (ML) is increasingly used as a method to predict ABI outcomes, however improper model evaluation poses a potential bias to initially promising findings (Chapter One). This study aimed to evaluate, with transparent reporting, three common ML classification methods. Regularised logistic regression with elastic net, random forest and linear kernel support vector machine were compared with unregularised logistic regression to predict good psychosocial outcomes after discharge from ABI inpatient neurorehabilitation using routine cognitive, psychometric and clinical admission assessments. Outcomes were selected on the basis of decision making for care packages: accommodation status, functional participation, supervision needs, occupation and quality of life. The primary outcome was accommodation (n = 164), with models internally validated using repeated nested cross-validation. Random forest was statistically superior to logistic regression for every outcome with areas under the receiver operating characteristic curve (AUC) ranging from 0.81 (95% confidence interval 0.77-0.85) for the primary outcome of accommodation, to its lowest performance for predicting occupation status with an AUC of 0.72 (0.69-0.76). The worst performing ML algorithm was support vector machine, only having statistically superior performance to logistic regression for one outcome, supervision needs, with an AUC of 0.75 (0.71-0.80). Unregularised logistic regression models were poorly calibrated compared to ML indicating severe overfitting, unlikely to perform well in new samples. Overall, ML can predict psychosocial outcomes using routine psychosocial admission data better than other statistical methods typically used by psychologists.

Word count: 248

Keywords

Brain injury; stroke; cerebrovascular accident; hypoxia; neuroinfection; anoxic brain injury; machine learning; random forest; logistic regression; regularisation; elastic net; support vector machine; prediction; accommodation; employment; occupation; participation; quality of life; supervision; functional.

Introduction

The incidence of acquired brain injury (ABI, an injury to the brain sustained after birth) is increasing, contributing to approximately 1000 daily UK hospital admissions (Menon & Bryant, 2019), with outcomes varying greatly by the timing and intensity of interventions (Cullen, Chundamala, Bayley, & Jutai, 2007). Moderate to severe ABI often requires high-intensity care from inpatient rehabilitation centres. Even with individualised rehabilitation programmes, long term outcomes for people with ABI remain variable (Ponsford et al., 2014; Rassovsky et al., 2015).

Healthcare providers are challenged to provide care pathways so that housing, functional, and occupational needs are met at the time of discharge. Functional independence forms the basis of decision making for the required level of care, with average UK personal budgets after inpatient neurorehabilitation varying from averages of £306/week for low dependency needs to £1349/week for high dependency needs (Turner-Stokes, Williams, Bill, Bassett, & Sephton, 2016). Accurately predicting support needs would ensure resources are allocated efficiently and cost-effectively.

Previous research using traditional statistical methods indicates psychosocial data may be valuable for predicting outcomes after ABI inpatient neurorehabilitation. Neuropsychological measures have been shown to be strongly associated with functional productivity after rehabilitation using longitudinal methodology (Green et al., 2008), suggesting these may therefore have predictive value. As a contrast however, few demographic variables were found to be associated with long term functional outcomes after TBI (Ponsford, Draper & Schonberger, 2008). Using traditional logistic regression (LR), returning home after discharge from inpatient stroke neurorehabilitation was predicted using a range of psychosocial and neurological variables (Frank, Conzelmann, & Engelter, 2010) finding an area under the curve (AUC) of 0.86 (95% confidence interval [CI] 0.84-0.88). Without the use of model validation techniques however, these performance metrics are likely inflated. Together, these studies suggest variation in predictive value of psychosocial assessment data. Traditional regression algorithms are not capable of modelling high numbers of variables for risk of modelling sample noise, a phenomenon known as overfitting, posing problems to clinicians to know what assessment data will be valuable to use to inform of long-term outcomes.

Predictive models in clinical practice are often limited by statistical methods employed with findings unable to be replicated or generalised to clinical settings (Dwyer, Falkai, & Koutsouleris, 2018; Ioannidis, 2016). Routine clinical data often have high proportions of missingness and violate traditional regression assumptions, such as having collinear and skewed datasets. To overcome these challenges datasets are heavily cleaned for analysis, excluding participants with missing data or using inappropriate predictor selection methods resulting in biased models (Wolff et al., 2019).

Machine learning (ML), a branch of artificial intelligence, has previously predicted improvement after inpatient rehabilitation with model validation finding AUC values of 0.85-0.93 for different ML algorithms (Marcano-Cedeño et al., 2013). ML learns from the data how to best fit predictor variables to future or unknown events.

ML has better capacity to deal with multidimensional, missing and multicollinear data, typical of routine clinical data (Iniesta et al., 2016). With ML one can make use of pre-existing data sets more representative of true population characteristics, offering an alternative method for predicting outcomes with greater power and less concern of overfitting (Yarkoni & Westfall, 2017).

Neurorehabilitation centres undertake comprehensive assessments offering invaluable information for predictive modelling. Employing ML to model these data may offer greater accuracy and inform clinicians what clinical assessment data are useful for predicting probable outcome. Previous ML research in the field of ABI has however been limited by a lack of model evaluation (systematic review, Chapter One), showing that models need to be validated and evaluated by power, discrimination and calibration to ensure ML is meeting the research aims.

Aims

This study aimed to examine the effectiveness of ML methods for predicting psychosocial outcomes after ABI inpatient neurorehabilitation using routine psychometric, cognitive, demographic and medical history admission assessments. Given the expected variability for outcomes and algorithms (as summarised in Chapter One), this research aimed to evaluate the efficacy of regularised logistic modelling with elastic net (RLR), random forest (RF), linear kernel support vector machine (SVM) and traditional LR modelling for predicting accommodation status and other categorical psychosocial outcomes after discharge from inpatient neurorehabilitation. Models with an AUC of 0.8 or above (Safari et al., 2016), a calibration slope near 1 (Calster et al., 2016), and a sample size to indicate appropriate power by an a priori power analysis will be considered to show 'good' performance. For continuous outcomes, an R-squared value above 0.75 would be considered as substantial, and between 0.25-0.75 for moderate effect size (Cruz-Cunha, 2013).

Primary research questions

1. Is it possible to predict accommodation status at discharge better than chance using baseline demographic, clinical, cognitive and psychometric measures from admission?
2. Was ML (RLR, RF or SVM) or traditional (unregularised) LR more superior at predicting accommodation status?
3. Which features are superior predictors for accommodation status?

Additional research questions

4. Is it possible for ML to predict level of participation better than chance at discharge?
5. Is it possible for ML to predict level of supervision better than chance at discharge?
6. Is it possible for ML to predict occupational functioning better than chance at discharge?
7. Is it possible for ML to predict quality of life (QoL) better than chance 6-months after discharge?
8. Is it possible to predict length of admission with a moderate R-squared of 0.25-0.75 using regularised linear regression or traditional linear regression?

Methods

Design

This study used a retrospective single-centre cohort design.

Participants

Participant data were sourced from Graham Anderson House (GAH), a Brain Injury Rehabilitation Trust (BIRT) inpatient neurorehabilitation centre in Glasgow, United Kingdom. Inclusion criteria for the service are a diagnosis of moderate to severe ABI requiring inpatient rehabilitation (including ABI caused by cerebrovascular accident, traumatic injury, anoxic brain injury or infection), aged 16 or over, and for needs not better met by another service (e.g. alcohol-related brain damage service).

Database inclusion and exclusion criteria

A research database was constructed from routine clinical data at GAH. Participants were included in the database who were admitted and discharged from GAH between service opening (2009) and data extraction (12th March 2020). Participants were excluded from the database if they were prematurely discharged (e.g. self-discharge, death or transfer to another service). If participants were readmitted, data from their initial assessment and discharge was used and readmission data excluded. For each model participants were excluded if they were missing the outcome variable of interest.

Ethics

Ethical approval was provided by West of Scotland Research Ethics Committee 1 on 24/02/2020 (20/WS/0026; appendix 2.1 pg. 62). Management approval was provided by The Disabilities Trust on 24/10/2019 (appendix 2.2. pg. 66). At admission to GAH service users provide their consent for their anonymised data to be used for service evaluation projects. Access to patient records was required for on-site database development, with BIRT Caldicott approval granted on 10/01/2020 (appendix 2.3 pg. 67) and Greater Glasgow and Clyde NHS Caldicott approval on 06/01/2020 (appendix 2.4 on pg. 68). A research database was developed with participant data fully anonymised and stored in line with ethical approval and data protection regulations.

Measures

Participant data included routine clinical information collected and recorded prospectively from admission at GAH, to discharge assessments, and finally follow-up assessments 6-months after discharge. Assessment data were collected by trained members of the clinical team including clinical and assistant psychologists for neuropsychological assessments.

Outcome measures

Five psychosocial outcome measures were selected describing functioning in key areas following discharge (accommodation, participation, supervision, occupation, and service user-rated QOL) and a sixth outcome of length of admission. The primary outcome for the study was accommodation status since clinical opinion within the service believed this is service users', and families' primary concern at admission. The five

psychosocial outcomes were dichotomised by favourable or poor outcomes (Table 5) with favourable outcomes coded 'Yes' and poor coded 'No.' Decisions on dichotomisation of target variables were agreed by EM, BO'N and BC based on clinical judgement of meaningful categories before analysis. Models were developed to predict favourable clinical outcomes. Length of admission in weeks was kept as a continuous measure to be more clinically relevant.

Predictors

Variables were selected on the basis of a literature review for predictors with likely significant predictive power for psychosocial outcomes and availability within routine records. Candidate predictors included cognitive and psychometric measures recorded at admission, injury-related factors, demographics, and other medical history. A full list of candidate predictors and data processing are available in Appendix 2.5 on pg. 69, with a total of 30 candidate variables (represented in the models by 38 parameters after dummy coding). Where predictors from different measures needed to be converted into a common metric (e.g. memory scores from different batteries), decisions were made by EM, BO'N and BC based on clinical knowledge of neuropsychological instruments.

Analysis

Data analysis for model development and validation was performed with R Programming version 3.6.2 using the 'caret' package (Kuhn, 2019), "glmnet," "randomForest," and "e1071." A comparison of predictor variables between participants with and without available data on the primary outcome of accommodation was performed using IBM SPSS statistics version 26 to explore potential bias for exclusion by outcome. Categorical predictor variables between groups were compared with chi-squared tests, and continuous data compared with independent t-tests or Mann Whitney U test depending on data distribution. P-value corrections were employed using false discovery rate (FDR) adjustments using an online calculator [available from https://www.sdmproject.com/utilities/?show=FDR]. The complete R code and SPSS syntax is available at https://github.com/EmmaMawdsley/Predicting-brain-injury-outcomes-with-machine-learning.

An a priori power analysis was completed with the R package 'pROC' (Robin et al., 2019). Based on an estimated ratio of good:poor outcome of 4:1 for accommodation status (the primary outcome), from preliminary service data prior to conducting this research, to have 80% power to detect a significant effect at $p<0.05$ (two-sided), the minimum sample size required for analysis would be 27 for a superior AUC of 0.85, or 58 for a good AUC of 0.75.

Table 5 Outcome measures

Outcome measure	**Favourable outcome**	**Poor outcome**
Accommodation status Measured by either the Accommodation Scale within BIRT Independent Living Scale (BILS, Michael Oddy, Haye, & Goodson, 2018), or the Accommodation Rating Scale within Community Disposition Ratings (CDR-ARS, Eames, 1999) measured at patient's discharge. Scores range from 0-11 for the BILS and 0-10 for the CDR-ARS (with higher scores indicating greater accommodation support).	Scores ≤6: *'independent or community supported housing.'*	Scores ≥7: *'residential/hospital accommodation.'*
Participation in functional tasks The Participation subscale of the Mayo-Portland Adaptability Inventory (Bellon, Malec, & Kolakowsky-Hayner, 2012) measured at discharge, converted into standardised scores as per instrument manual (with higher scores indicating more severe functional disability).	T-scores ≤49: *'a good outcome or mild to moderate disability.'*	T-scores ≥50: *'moderate to severe disability.'*
Level of supervision The *Supervision Rating Scale* (SRS, Boake, 1996) measured at discharge with total scores ranging from 0-13 (with higher scores indicating greater supervision requirement).	Scores ≤7: *'part time or no supervision.'*	Scores ≥8: *'full time direct supervision.'*
Occupational functioning Measured by either the Occupational Participation Scale, within the BILS, or The Occupation Rating Scale within the CDR (CDR-ORS, Eames, 1999) at discharge with total scores ranging from 0-9 for the BILS and 0-8 for the CDR-ORS (with higher scores indicating less occupational activity).	Scores ≤6: *'productive occupational activity.'*	Scores of ≥7: *'recreational/non-occupational activity.'*
Quality of Life The *EuroQoL Instrument* (EQ5D, EuroQoL Group., 2010) a patient-rated measure for QoL from 0-100 (with higher scores indicating greater QOL), administered at the 6-month follow-up.	Dichotomised by the sample median with scores ≥75 representing a good outcome.	Scores ≤74 representing a poorer outcome.
<u>Outcome measure</u>	<u>Continuous outcome</u>	
Length of admission The number of weeks between admission and discharge.	Admission length was kept as a continuous outcome.	

Data processing

Pre-processing of data included removal of zero or near-zero variance predictors and highly correlated predictors (>70% correlation). For the primary (research questions 1-3) and secondary analyses (research questions 4-8), missing data on predictor variables were imputed using k=5 nearest neighbours (*K*NN) for variables with ≥80% complete data (variables with <80% complete data were omitted from the models). *K*NN is a non-parametric imputation method that involves matching a missing data point to its nearest *K* related cases based on other predictor variables (Beretta & Santaniello, 2016) with the outcome data removed. For

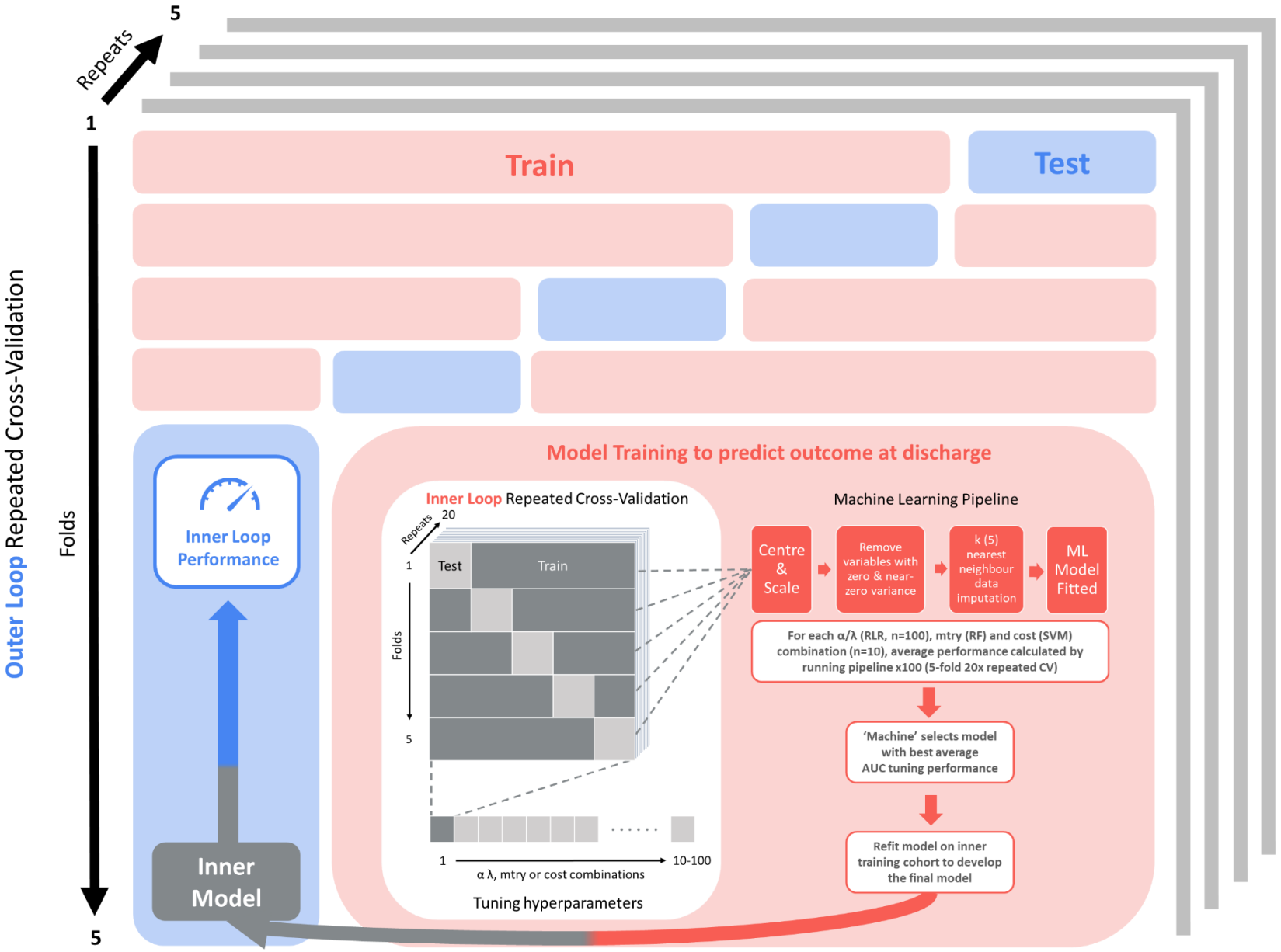
additional sensitivity analyses, missing data on variables with ≥50% complete data were imputed with *K*NN and included in the models (variables with <50% complete data were omitted).

Model development and validation

For dichotomous outcomes, RLR, RF and SVM algorithms (determined as suitable for predicting psychosocial outcomes in ABI, with the additional benefit of embedded feature selection for RLR and RF as described in Chapter One) were evaluated with the AUC and calibration slope being the primary performance metrics. For models with continuous outcomes, regularised linear regression with elastic net was evaluated by R-squared and root mean square error (RMSE) as primary performance metrics (with higher R-squared and lower RMSE indicating better performance). A description of each method can be found in Appendix 2.6 on pg. 73.

Models were initially developed using 5-fold cross-validation repeated ten times. Those with promising predictive ability by evaluation metrics and power were then internally validated using 5-fold repeated nested cross-validation (nested CV) repeated twenty times for the inner loop and five times for the outer loop whereby 80% of the data was used for training, with the remaining 20% of the data (chosen randomly each time) reserved for model validation (Figure 2). This was across 100 hyperparameter combinations for RLR, and ten hyperparameter combinations for RF and SVM (due to differences in the algorithms). The nested CV was performed on the best combination of hyperparameters chosen from the inner loop, with the AUC calculated by combining the results in sequence to reduce the likelihood of model optimism and overfitting. A permutation test was performed for each AUC to test the significance level of the obtained result, corrected with FDR. The final model was then assessed for calibration by the calibration slope (a plot of the observed outcomes and model predictions; only the slope is interpreted as the intercept is not relevant for internal validation). A perfectly calibrated model will have a slope of 1, with higher metrics indicating underfitting and lower metrics for overfitting (Calster et al., 2016).

Figure 2 Overview of analysis and internal validation procedure repeated for each method with each outcome



Adapted with permission from Samuel Leighton

Abbreviations and definitions:

α = alpha, elastic net penalty

AUC= area under the curve

cost = SVM tuning parameter

CV = cross validation

λ = lambda, elastic net tuning parameter

mtry = number of variables available at each split of the tree node in RF

RF=random forest

RLR=regularised logistic regression

SVM=support vector machine

Unregularised logistic and linear regression models were also constructed (for dichotomous and continuous outcomes, respectively) for comparison against the RLR, RF and SVM models. The unregularised regression models are similar to traditional regression models with the exception of nested cross-validation which tunes the parameters to give less optimistic performance during the nested CV procedure as described above. For the binary outcomes, the resulting AUCs for LR, RLR, RF and SVM were compared for statistical differences using the Delong test (DeLong, DeLong, & Clarke-Pearson, 1988), with p-values corrected for multiple analyses using FDR adjustment.

Predictor analysis

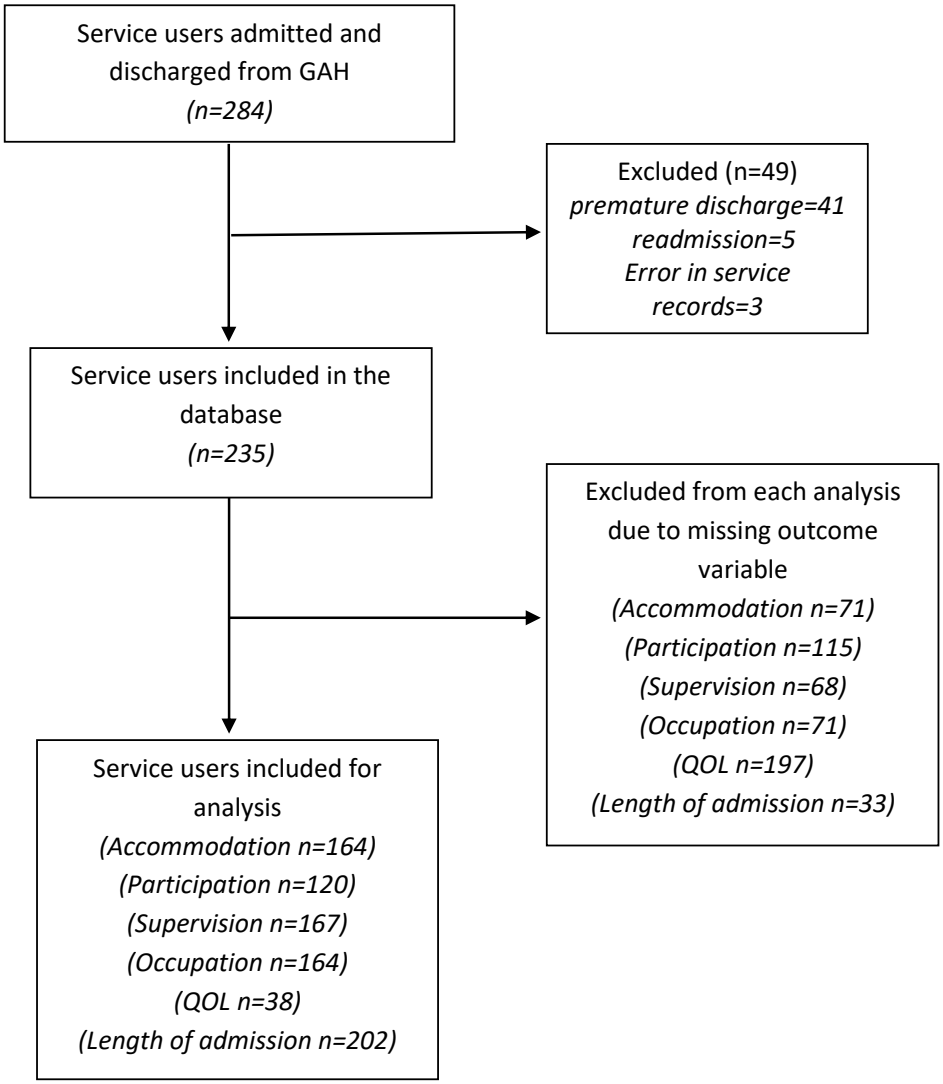
For RLR and RF, predictors identified through embedded feature selection were ranked by variable importance after internal validation. This analysis was not performed with SVM as it does not have embedded feature selection. For RLR, an additional analysis was performed for feature stability to show which features were selected across all of the models developed in the nested-cross validation stage (using the method reported in Nogueira, Sechidis, & Brown, 2017). Stability assessments were unable to be performed for RF and SVM due to different methods for feature selection for RF and no embedded feature selection for SVM.

Results

Sample characteristics

284 service users were discharged from GAH of whom 235 met the inclusion and exclusion criteria. Participants were excluded on a model by model basis if they were missing the outcome, resulting in 164 participants (69.8%) with the primary outcome of accommodation (Figure 3). A comparison of the distribution of predictor variables between those with and without the accommodation outcome is shown in Table 6, showing no significant between group differences on predictor variables.

Figure 3 Flow chart of included participants for analysis of each outcome



GAH= Graham Anderson House; QOL= quality of life

Table 6 Distribution of candidate predictors between patients with and without the primary outcome of accommodation

Candidate predictor	Statistic	Patients excluded by outcome (n = 71)	Patients included in analysis (n = 164)	P value (corrected with FDR)
Gender, Female	N (%)	17 (23.9)	26 (15.9)	0.29
Age at injury	Mean (SD)	44.3 (16.7)	44.0 (15.7)	0.91
Age at admission	Mean (SD)	46.3 (15.1)	46.9 (13.8)	0.88
Days between injury and admission	Mean (SD)	988.7 (1962.2)	952.6 (2321.5)	0.91
Diagnosis CVA	N (%)	11 (15.7)	18 (11.3)	0.11
Hypoxia		9 (12.9)	16 (10.1)	
Infection		6 (8.6)	4 (2.5)	
Neoplasm		2 (2.9)	0 (0)	
TBI		31 (44.3)	101 (63.5)	
Other		11 (15.7)	20 (12.6)	
Preinjury psychosis, Yes	N (%)	4 (5.6)	7 (4.3)	0.81
Drug dependence, Yes	N (%)	19 (27.1)	27 (19.6)	0.38
Alcohol abuse, Yes	N (%)	28 (40.6)	71 (51.4)	0.29
Multiple trauma, Yes	N (%)	10 (17.9)	31 (32.3)	0.17
Other medical condition, Yes	N (%)	31 (56.4)	52 (58.4)	0.88
SIMD Rank 2020	Median (IQR)	1407.50 (591-3589)	1746.5 (665-3795)	0.83
HADS Anxiety	Median (IQR)	6.78 (2-11)	7 (3-11)	0.68
Depression	Median (IQR)	5.6 (2-8)	6 (3-8)	0.68
Total MPAI	Mean (SD)	75.7 (13.0)	70.1 (16.4)	0.11
MPAI subscales Abilities	Median (IQR)	55 (50-59)	51 (46-57)	0.11
Adjustment	Median (IQR)	56.5 (54-59)	55 (52-58)	0.21
Participation	Median (IQR)	59 (53-65)	55 (49-62)	0.11
WAIS VCI	Mean (SD)	77.0 (13.4)	80.7 (15.0)	0.32
PRI	Mean (SD)	76.2 (10.0)	80.3 (13.1)	0.20
WMI	Median (IQR)	77.5 (66-87)	83 (74-92)	0.11
PSI	Median (IQR)	66.5 (56-71)	68 (59-76)	0.54
FSIQ	Mean (SD)	70.4 (11.9)	74.9 (11.7)	0.17

Candidate predictor		Statistic	Patients missing outcome (n = 71)	Patients included in analysis (n = 164)	P value (corrected with FDR)
Executive functioning	Impaired Borderline Low Average Average High Average	N (%)	27 (60) 6 (13.3) 1 (2.2) 10 (22.2) 1 (2.2)	74 (55.6) 20 (15.0) 15 (11.3) 22 (16.5) 2 (1.5)	0.61
Memory quotient score		Median (IQR)	67.4 (58-72)	66 (59-76)	0.61
Neuropsychological data availability	Full data Some data No data	N (%)	29 (40.8) 20 (28.2) 22 (31.0)	72 (43.9) 71 (43.3) 21 (12.8)	0.11

Abbreviations: CVA= Cerebrovascular accident; FDR= False discovery rate; HADS = Hospital Anxiety and Depression Scale; IQR= interquartile range; MPAI= Mayo Portland Adaptability Inventory; N= Number of participants; TBI= Traumatic brain injury; SD= Standard deviation; SIMD= Scottish Index of Multiple Deprivation; WAIS= Wechsler Adult Intelligence Scale (FSIQ= Full scale IQ; PRI= Perceptual reasoning index; PSI= Processing speed index; VCI= Verbal comprehension index; WMI= Working memory index).

The frequencies of favourable and poorer binary outcomes are reported in Table 7. For the continuous outcome of length of admission (n=202), the sample median was 27 weeks, interquartile range (IQR) between 13.8 and 51.0 weeks.

Table 7 Observed frequencies of favourable and poorer outcomes

	Accommodation	Participation	Supervision	Occupation	Quality of life
Favourable outcome	140 (85.4%)	90 (75%)	135 (80.8%)	68 (41.5%)	20 (52.6%)
Poorer outcome	24 (14.6%)	30 (25%)	32 (19.2%)	96 (58.5%)	18 (47.4%)

Predicting psychosocial outcomes

Comparisons between the different models' predictive ability after internal validation are shown in Table 8. ML results for sensitivity analyses with predictor imputation with ≥50% complete data are reported in appendix 2.7 on pg. 75. For primary and secondary analyses, LR performance varied between AUC values of 0.62-0.71 for the different outcomes, compared to RLR (0.65-0.79), RF (0.72-0.81) and SVM (0.61-0.77). Due to the low numbers of respondents for the QOL measure 6-months after discharge, QOL models were underpowered to detect a significant effect and therefore not further evaluated.

Table 8 Performance metrics including area under the curve, 95% confidence intervals, calibration slope, and p-value after permutation testing for predicting psychosocial outcomes

Outcome	Sample size	Method	AUC	95% CI	Calibration (slope)	Permutation p value (with FDR corrections)
Accommodation	164	LR	0.63	0.52-0.74	0.01	0.002
		RLR	0.79	0.74-0.83	0.99	<0.001
		RF	0.81	0.77-0.85	0.90	<0.001
		SVM	0.77	0.71-0.82	1.13	<0.001
Participation	120	LR	0.67	0.61-0.72	0.22	<0.001
		RLR	0.73	0.68-0.78	1.37	<0.001
		RF	0.73	0.68-0.78	1.04	<0.001
		SVM	0.72	0.67-0.77	0.84	<0.001
Supervision	167	LR	0.73	0.68-0.78	0.40	<0.001
		RLR	0.77	0.72-0.81	1.10	<0.001
		RF	0.81	0.77-0.85	0.89	<0.001
		SVM	0.75	0.71-0.80	1.01	<0.001
Occupation	164	LR	0.65	0.61-0.69	0.41	<0.001
		RLR	0.65	0.61-0.69	0.62	<0.001
		RF	0.72	0.69-0.76	1.14	<0.001
		SVM	0.61	0.58-0.65	0.80	<0.001
Quality of life	38	RLR	0.73	0.63-0.83	0.22	<0.001

AUC= Area under the receiver operator curve; CI= Confidence intervals; FDR= False discovery rate; LR= logistic regression; RLR= regularised logistic regression; RF= random forest; SVM= support vector machine

Psychosocial predictions using ML had fair to good performance by AUC and calibration, and LR had fair performance by AUC but poor calibration, severely overfitting. ML models, particularly RLR, indicated occasional under and over-fitting to the sample data although less extreme than LR. RF was the only ML algorithm showing a consistent pattern of superior performance than LR for each psychosocial outcome. Delong's test for a significant difference between AUCs is shown in Table 9, with RF performing significantly better than LR for all four binary outcomes that were internally validated, and with RLR statistically superior to LR for three of the four outcomes. SVM was statistically superior for LR for predicting supervision, although LR was statistically superior to SVM for predicting occupation.

Best performing predictors for psychosocial outcomes

The top five clinical data variables used in RLR and RF models (after internal validation) are reported in Table 10. Stability analyses after nested cross-validation were performed for RLR to inform which predictors were stable across nested cross-validation and the resulting means of coefficients are shown in Appendix 2.8 on pg. 76.

Table 9 Significance values using Delong's test to compare ROC curves adjusted with FDR corrections

Accommodation	**RLR**	**RF**	**SVM**
LR	0.03*	0.01*	0.06
RLR		0.1	0.5
RF			0.6
Participation	**RLR**	**RF**	**SVM**
LR	<0.001**	0.008**	<0.001**
RLR		1	0.84
RF			0.84
Supervision	**RLR**	**RF**	**SVM**
LR	<0.001**	<0.001**	0.05*
RLR		0.03*	0.1
RF			0.004**
Occupation	**RLR**	**RF**	**SVM**
LR	0.5	<0.001**	<0.001**
RLR		<0.001**	<0.001**
RF			<0.001**

**** p<0.05 ** p<0.01***

FDR= false discovery rate; LR= logistic regression; RLR= regularised logistic regression; ROC=receiver operating characteristic curve; RF= random forest; SVM= support vector machine

Predicting length of admission

Models were developed for predicting length of admission comparing regularised linear regression with elastic net (0.07 R^2, 49.65 RMSE) and unregularised linear regression (0.05 R^2, 52.76 RMSE) for ≥80% complete data imputation. These results suggest poor model performance therefore models were not further evaluated.

Table 10 Top five predictors for favourable psychosocial outcomes for RF and RLR identified from embedded feature selection.

Outcome	RF	RLR
Accommodation	1. Adjustment (MPAI) 2. Neuropsychological data availability 3. Days between injury and admission 4. Age at admission 5. SIMD rank	1. Neuropsychological data availability 2. Adjustment (MPAI) (-) 3. Diagnosis (other) (-) 4. Diagnosis (TBI) 5. Abilities (MPAI) (-)
Participation	1. Days between injury and admission 2. Adjustment (MPAI) 3. Abilities (MPAI) 4. Age at admission 5. Neuropsychological data availability	1. Neuropsychological data availability 2. Abilities (MPAI) (-) 3. Adjustment (MPAI) (-) 4. Gender (female) 5. Executive functioning
Supervision	1. Abilities (MPAI) 2. Adjustment (MPAI) 3. Days between injury and admission 4. Age at admission 5. SIMD rank	1. Neuropsychological data availability 2. Adjustment (MPAI) (-) 3. Abilities (MPAI) (-) 4. Diagnosis (Hypoxia) (-) 5. Days between injury and admission (-)
Occupation	1. Days between injury and admission 2. Age at admission 3. Abilities (MPAI) 4. SIMD rank 5. Adjustment (MPAI)	1. Gender (female) 2. Neuropsychological data availability 3. Abilities (MPAI) (-) 4. Diagnosis (other) (-) 5. Age at admission (-)

MPAI= Mayo Portland Adaptability Inventory; RF= random forest; RLR= regularised logistic regression; SIMD= Scottish Index of Multiple Deprivation; TBI=Traumatic brain injury
(-) under RLR show negative relationships; algorithms for RF do not show negative relationships as there are no coefficients

Discussion

The main aim of this study was to examine the effectiveness of using ML algorithms to predict favourable psychosocial outcomes after ABI inpatient neurorehabilitation using routinely collected psychometric, cognitive, clinical and demographic predictors. Three ML classification methods (RLR, RF, and SVM) were compared for four psychosocial outcomes (accommodation, participation, supervision and occupation at the time of discharge). Planned analyses for QOL 6-months after discharge were underpowered to detect a significant effect. Every evaluated model performed better than chance for predicting each outcome. ML models, particularly RF and RLR, had fair to good performance (Safari et al., 2016), and every ML model had better calibration than LR indicating ML models will perform better for prediction in new samples. This study also aimed to predict length of admission comparing linear regression with and without elastic net

regularisation, however both methods had a poor fit to the data suggesting psychosocial functioning may not be important determinants of length of admission for ABI.

Across all outcomes RF consistently showed superior performance than LR, as well as being superior to the other ML algorithms for predicting supervision and occupation. RF is an ensemble method of many weak learners, randomly generating a "forest" of decision trees which could increase RF's likelihood of fitting a good prediction. RF has evidenced capacity to handle complex data, such as skewed and unbalanced class distributions (Guo et al., 2010) and high dimensional noise in features (Guo & Balasubramanian, 2012) which could underlie its superiority in this study. Psychological assessments often include a degree of subjectivity or response bias, leading to greater variation and higher degrees of noise than many medical variables. RF's capacity to deal with these challenges may make it a superior method for complex psychosocial datasets.

Feature selection methods embedded in RF and RLR are main benefits of these ML algorithms over other ML methods which can be more complex to interpret, such as SVM or artificial neural networks (Nadkarni, 2016). RF and RLR frequently showed two subscales of the MPAI (adjustment, a measure of mood and interpersonal difficulties, and abilities, a measure of a person's cognitive and physical functioning), and neuropsychological data availability (an indicator of the extent to which a person underwent neuropsychological testing), as important predictors of a range of psychosocial outcomes at the time of discharge. As well as aiding our knowledge of strong predictors of outcomes, identifying these may help inform why people have poorer outcomes and may indicate other avenues for intervention. Further investigation of predictors is however required as embedded feature selection doesn't provide significance levels or confidence intervals, and identified features may be proxies for other important variables not otherwise measured (Shmueli, 2010).

Neuropsychological data availability interestingly had greater predictive power than any of the individual neuropsychological test results. Frequent reasons for a person not having a full neuropsychological battery are challenging behaviours or very severe impairment. A person's status as having less neuropsychological data could have been proxy for these difficulties not otherwise captured. Alternatively, more neuropsychological tests administered could better tailor the person's rehabilitation to their needs, leading to more favourable outcomes. Together with the strong predictors of better adjustment and abilities, the former interpretation might be likely that more neuropsychological tests are administered when a person has greater emotional stability, higher cognitive functioning and less interpersonal conflict, in turn contributing to more favourable outcomes. It makes clinical sense that these characteristics would make it more likely for a person to return home or to employment after moderate to severe brain injury.

Models built for predicting QOL 6-months after discharge had an inadequate sample size, resulting in lack of statistical power. As such, the internal validation procedures would have reduced the variance in predictors and biased results. Additionally, the QOL measure was likely subject to response bias given only around 16% of

discharged patients chose to complete the measure, providing an optimistic group median of 75/100 (with higher scores indicating greater self-rated quality of life). It is likely that representative samples would show a lower median QOL rating. Future research would benefit from a larger sample size to explore how ML performs at modelling psychosocial predictors for QOL after ABI. This would be of benefit as it may identify areas suitable for intervention.

Strengths and limitations

This study has a strength showing that with transparent reporting and robust methods, overcoming challenges described in Chapter One, ML methods RF and RLR have superior predictive power for psychosocial outcomes from a highly heterogeneous dataset with a range of ABIs. Models were properly internally validated to account for model optimism and evaluated by discrimination, calibration and power. Internally validated LR in this study demonstrated how statistical techniques typically used by psychologists may not be reliable for novel predictions. Calibration assessment indicated extreme overfitting for LR, likely leading to lower performing models in new samples. There was some indication of under- and overfitting for some ML models, however much less extreme than LR. External validation in a new dataset would be beneficial for generalisability. As GAH continues to discharge patients there is the possibility of a temporal validation cohort, or geographically with an alternative BIRT centre.

Limitations to the current dataset which, if improved, could further benefit models' predictive ability in future research. Firstly, this study was limited to features available from routine clinical assessment, meaning features previously identified as strong psychosocial predictors such as education history or length of post-traumatic amnesia (e.g. Stromberg et al., 2019), were unable to be modelled in this study due to inconsistent recording in service records. Including these predictors in future models would likely strengthen predictive performance. Secondly, whilst this study performed an a-prior power analysis, predictive models are at lower risk of bias when the number of participants to the number of candidate predictor parameters are at least twenty events per variable (Wolff et al., 2019). Unfortunately, this guidance wasn't published at the time of study design and power analysis, and so there is risk of models being underpowered due to the available sample size. Finally, routine clinical data has high degrees of missingness for certain clinical variables. Whilst our sensitivity analyses are promising for similar results to the primary and secondary analyses, the imputation strategy could have contributed to less accurate results. Multiple imputation may have been a more reliable strategy than *K*NN, which relies on single imputation, however due to the number of algorithms and complexity of the nested CV, multiple imputation would have been computationally intensive.

Conclusions

This study shows promising preliminary findings for predicting psychosocial outcomes after discharge from inpatient ABI neurorehabilitation based on routine admission data using ML. These datasets are typical of clinical data suggesting ML is a useful skill for clinicians. RF had superior performance for modelling

psychosocial data with better calibration than unregularised LR, although external validation in a novel dataset is required to increase reliability of the findings. Future use of ML modelling techniques could inform treatment planning and appropriate care pathways after discharge from ABI neurorehabilitation to be more efficient and cost-effective.

References

Bellon, K., Malec, J. F., & Kolakowsky-Hayner, S. A. (2012). Mayo-portland adaptability inventory-4. *The Journal of Head Trauma Rehabilitation, 27*(4), 314-316. doi: 10.1097/HTR.0b013e3182562f04.

Beretta, L., & Santaniello, A. (2016). Nearest neighbor imputation algorithms: a critical evaluation. *BMC Medical Informatics and Decision Making, 16*(3), 197-208. doi:doi:10.1186/s12911-016-0318-z

Boake, C. (1996). Supervision Rating Scale: a measure of functional outcome from brain injury. *Archives of Physical Medicine and Rehabilitation, 77*(8), 765-772. https://doi.org/10.1016/S0003-9993(96)90254-3.

Calster, B. V., Nieboer, D., Vergouwe, Y., Cock, B. D., Pencina, M. J., & Steyerberg, E. W. (2016). A calibration hierarchy for risk models was defined: from utopia to empirical data. *Journal of Clinical Epidemiology, 74*, 167-176. doi:10.1016/j.jclinepi.2015.12.005

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*. doi:https://doi.org/10.1016/j.jclinepi.2019.02.004

Cruz-Cunha, M. M. (2013). *Handbook of Research on Enterprise 2.0: Technological, Social, and Organizational Dimensions*: IGI Global.

DeLong, E. R., DeLong, D. M., & Clarke-Pearson, D. L. (1988). Comparing the Areas under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach. *Biometrics, 44*(3), 837-845. doi: 10.2307/2531595.

Dwyer, D. B., Falkai, P., & Koutsouleris, N. (2018). Machine learning approaches for clinical psychology and psychiatry. *Annual Review of Clinical Psychology, 14*, 91-118.https://doi.org/10.1146/annurev-clinpsy-032816-045037.

Feigin, V. L., Barker-Collo, S., Krishnamurthi, R., Theadom, A., & Starkey, N. (2010). Epidemiology of ischaemic stroke and traumatic brain injury. *Best Practice & Research Clinical Anaesthesiology, 24*(4), 485-494. doi:https://doi.org/10.1016/j.bpa.2010.10.006

Frank, M., Conzelmann, M., & Engelter, S. (2010). Prediction of discharge destination after neurological rehabilitation in stroke patients. *European Neurology, 63*(4), 227-233. https://doi.org/10.1159/000279491.

Eames P. (1999) Measuring Outcome: Development of an Approach. *Neuropsychological Rehabilitation*, *9*(3–4):363–371. https://doi.org/10.1080/096020199389437

Euroqol Group., 2010. EQ-5D. *In:* Preedy, V. R. & Watson, R. R. (eds.) *Handbook of Disease Burdens and Quality of Life Measures.* New York, NY: Springer New York.

Green, R. E., Colella, B., Hebert, D. A., Bayley, M., Kang, H. S., Till, C., & Monette, G. (2008). Prediction of return to productivity after severe traumatic brain injury: investigations of optimal neuropsychological tests and timing of assessment. *Archives of Physical Medicine and Rehabilitation*, *89*(12), S51-S60. https://doi.org/10.1016/j.apmr.2008.09.552

Guo, Y., & Balasubramanian, R. (2012). Comparative Evaluation of Classifiers in the Presence of Statistical Interactions Between Features in High Dimensional Data Settings. *The International Journal of Biostatistics, 8*(1). doi:10.1515/1557-4679.1373

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. doi:10.1186/1471-2105-11-447

Iniesta, R., Stahl, D., & McGuffin, P. (2016). Machine learning, statistical learning and the future of biological research in psychiatry. *Psychological Medicine, 46*(12), 2455-2465. doi:10.1017/S0033291716001367

Ioannidis, J. P. A. (2016). Why most clinical research is not useful. *PLoS medicine, 13*(6), e1002049. https://doi.org/10.1371/journal.pmed.1002049

Kuhn, M. (2019). The caret package. Retrieved from http://topepo.github.io/caret/index.html

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. *339*, b2700. doi:10.1136/bmj.b2700

Marcano-Cedeño, A., Chausa, P., García, A., Cáceres, C., Tormos, J. M., & Gómez, E. J. (2013). Artificial metaplasticity prediction model for cognitive rehabilitation outcome in acquired brain injury patients. *Artificial Intelligence in Medicine, 58*(2), 91-99. doi:https://doi.org/10.1016/j.artmed.2013.03.005

Menon, D. K., & Bryant, C. (2019). Time for change in acquired brain injury. *The Lancet Neurology, 18*(1), 28. https://doi.org/10.1016.S1474-4422(18)30463-0.

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. *Systematic Reviews, 4*(1), 1-9. doi:10.1186/2046-4053-4-1

Moons, K. M., Altman, D. G., Reitsma, J. B., & et al. (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. doi:10.7326/M14-0698

Nadkarni, P. (2016). Chapter 4 - Core Technologies: Machine Learning and Natural Language Processing. In *Clinical Research Computing* (pp. 85-114): Academic Press.

Nogueira, S., Sechidis, K., & Brown, G. (2017). On the stability of feature selection algorithms. *Journal of Machine Learning Research, 18*(1), 1-54. doi:3242031

Oddy, M., Haye, L., & Goodson, A. (2018). Preliminary investigation of the reliability and validity of the BIRT Independent Living Scale AU - Ramos, Sara D. S. *Disability and Rehabilitation, 40*(23), 2817-2823. doi:10.1080/09638288.2017.1362594

Ponsford, J. L., Downing, M. G., Olver, J., Ponsford, M., Acher, R., Carty, M., & Spitz, G. (2014). Longitudinal follow-up of patients with traumatic brain injury: outcome at two, five, and ten years post-injury. *Neurotrauma, 31*(1), 64-77. doi:10.1089/neu.2013.2997

Ponsford, J., Draper, K., & Schönberger, M. (2008). Functional outcome 10 years after traumatic brain injury: its relationship with demographic, injury severity, and cognitive and emotional status. *Journal of the International Neuropsychological Society*, *14*(2), 233-242. doi: 10.10170S1355617708080272

Rassovsky, Y., Levi, Y., Agranov, E., Sela-Kaufman, M., Sverdlik, A., & Vakil, E. (2015). Predicting long-term outcome following traumatic brain injury (TBI). *Journal of Clinical and Experimental Neuropsychology, 37*(4), 354-366. https://doi.org/10.1080/13803395.2015.1015498.

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C., . . . Doering, M. (2019). CRAN - Package pROC: Comprehensive R Archive Network (CRAN). Retrieved from https://cran.r-project.org/web/packages/pROC/index.html

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.

Shmueli, G. (2010). To explain or to predict? *Statistical Science, 25*(3), 289-310. doi:10.1214/10-STS330

Stromberg, K. A., Agyemang, A. A., Graham, K. M., Walker, W. C., Sima, A. P., Marwitz, J. H., . . . Merchant, R. (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. doi: 10.1097/HTR.0000000000000438.

Turner-Stokes, L., Williams, H., Bill, A., Bassett, P., & Sephton, K. (2016). Cost-efficiency of specialist inpatient rehabilitation for working-aged adults with complex neurological disabilities: a multicentre cohort analysis of a national clinical data set. *BMJ Open, 6*(2), e010238. doi:10.1136/bmjopen-2015-010238

Wechsler, D. (2008). *Wechsler Adult Intelligence Scale–Fourth Edition (WAIS–IV)*: San Antonio, TX: The Psychological Corporation.

Wolff, R. F., Moons, K. G. M., Riley, R. D., Whiting, P. F., Westwood, M., Collins, G. S., . . . for the, P. G. (2019). PROBAST: A Tool to Assess the Risk of Bias and Applicability of Prediction Model Studies. *Annals of Internal Medicine, 170*(1), 51-58. doi:10.7326/M18-1376

Yarkoni, T., & Westfall, J. (2017). Choosing Prediction Over Explanation in Psychology: Lessons From Machine Learning. *Perspectives on Psychological Science, 12*(6), 1100-1122. doi:10.1177/1745691617693393

Zigmond, A. S., & Snaith, R. P. (1983). The hospital anxiety and depression scale. *Acta Psychiatrica Scandinavica, 67*(6), 361-370. https://doi.org/10.1111/j.1600-0447.1983.tb09716.x.

Appendices

Appendix 1.1. Journal of Neuropsychology submission guidelines

1. SUBMISSION

Authors should kindly note that submission implies that the content has not been published or submitted for publication elsewhere except as a brief abstract in the proceedings of a scientific meeting or symposium.

Once the submission materials have been prepared in accordance with the Author Guidelines, manuscripts should be submitted online at http://www.editorialmanager.com/jnp

Click here for more details on how to use **Editorial Manager**.

All papers published in the *Journal of Neuropsychology* are eligible for Panel A: Psychology, Psychiatry and Neuroscience in the Research Excellence Framework (REF).

Data protection:

By submitting a manuscript to or reviewing for this publication, your name, email address, and affiliation, and other contact details the publication might require, will be used for the regular operations of the publication, including, when necessary, sharing with the publisher (Wiley) and partners for production and publication. The publication and the publisher recognize the importance of protecting the personal information collected from users in the operation of these services, and have practices in place to ensure that steps are taken to maintain the security, integrity, and privacy of the personal data collected and processed. You can learn more at **https://authorservices.wiley.com/statements/data-protection-policy.html**.

Preprint policy:

This journal will consider for review articles previously available as preprints. Authors may also post the submitted version of a manuscript to a preprint server at any time. Authors are requested to update any pre-publication versions with a link to the final published article.

2. AIMS AND SCOPE

The Journal of Neuropsychology publishes original contributions to scientific knowledge in neuropsychology including:

clinical and research studies with neurological, psychiatric and psychological patient populations in all age groups

behavioural or pharmacological treatment regimes

cognitive experimentation and neuroimaging

multidisciplinary approach embracing areas such as developmental psychology, neurology, psychiatry, physiology, endocrinology, pharmacology and imaging science

The following types of paper are invited:

papers reporting original empirical investigations

theoretical papers; provided that these are sufficiently related to empirical data

review articles, which need not be exhaustive, but which should give an interpretation of the state of research in a given field and, where appropriate, identify its clinical implications

brief reports and comments

case reports

fast-track papers (included in the issue following acceptation) reaction and rebuttals (short reactions to publications in JNP followed by an invited rebuttal of the original authors)

special issues.

3. MANUSCRIPT CATEGORIES AND REQUIREMENTS

Research papers should be no more than 6000 words (excluding the abstract, reference list, tables and figures). Multiple citations for a single point are usually duplicative and authors are urged to cite the best reference. In exceptional cases the Editor retains discretion to publish papers beyond this length where the clear and concise expression of the scientific content requires greater length (e.g., explanation of a new theory or a substantially new method). Authors must contact the Editor prior to submission in such a case.

Brief communications are short reports of original research or case reports. They are limited to a maximum of 1500 words (excluding the abstract, reference list, tables and figures) and have a total of up to three tables or figures, and no more than 10 references.

Theoretical or review articles are full-length reviews of, or opinion statements regarding, the literature in a specific scientific area. They should be no more than 4000 words (excluding the abstract, reference list, tables and figures) and have no more than 45 references. Multiple citations for a single point are usually duplicative and authors are urged to cite the best reference. In exceptional cases the Editor retains discretion to publish papers beyond this length where the clear and concise expression of the scientific content requires greater length (e.g., explanation of a new theory or a substantially new method). Authors must contact the Editor prior to submission in such a case.

Please refer to the separate guidelines for Registered Reports.

All systematic reviews must be pre-registered.

4. PREPARING THE SUBMISSION

Free Format Submission

Journal of Neuropsychology now offers free format submission for a simplified and streamlined submission process.

Before you submit, you will need:

Your manuscript: this can be a single file including text, figures, and tables, or separate files – whichever you prefer. All required sections should be contained in your manuscript, including abstract, introduction, methods, results, and conclusions. Figures and tables should have legends. References may be submitted in any style or format, as long as it is consistent throughout the manuscript. If the manuscript, figures or tables are difficult for you to read, they will also be difficult for the editors and reviewers. If your manuscript is difficult to read, the editorial office may send it back to you for revision.

The title page of the manuscript, including a data availability statement and your co-author details with affiliations. *(Why is this important? We need to keep all co-authors informed of the outcome of the peer review process.)* You may like to use this template for your title page.

Important: the journal operates a double-blind peer review policy. Please anonymise your manuscript and prepare a separate title page containing author details. *(Why is this important? We need to uphold rigorous ethical standards for the research we consider for publication.)*

An ORCID ID, freely available at https://orcid.org. *(Why is this important? Your article, if accepted and published, will be attached to your ORCID profile. Institutions and funders are increasingly requiring authors to have ORCID IDs.)*

To submit, login at **https://www.editorialmanager.com/jnp/default.aspx** and create a new submission. Follow the submission steps as required and submit the manuscript.

If you are invited to revise your manuscript after peer review, the journal will also request the revised manuscript to be formatted according to journal requirements as described below.

Revised Manuscript Submission

Contributions must be typed in double spacing. All sheets must be numbered.

Cover letters are not mandatory; however, they may be supplied at the author's discretion. They should be pasted into the 'Comments' box in Editorial Manager.

Parts of the Manuscript

The manuscript should be submitted in separate files: title page; main text file; figures/tables; supporting information.

Title Page

You may like to use **this template** for your title page. The title page should contain**:**

A short informative title containing the major key words. The title should not contain abbreviations (see Wiley's best practice SEO tips);

A short running title of less than 40 characters;

The full names of the authors;

The author's institutional affiliations where the work was conducted, with a footnote for the author's present address if different from where the work was conducted;

Abstract;

Keywords;

Data availability statement (see Data Sharing and Data Accessibility Policy);

Acknowledgments.

Authorship

Please refer to the journal's Authorship policy in the Editorial Policies and Ethical Considerations section for details on author listing eligibility. When entering the author names into Editorial Manager, the corresponding author will be asked to provide a CRediT contributor role to classify the role that each author played in creating the manuscript. Please see the **Project CRediT** website for a list of roles.

Abstract

Please provide an abstract which gives a concise statement of the intention, results or conclusions of the article. The abstract should not include any sub-headings.

Abstracts for Research Papers should not exceed 250 words.

Abstracts for theoretical or review articles should not exceed 250 words.

Abstracts for brief communications should not exceed 80 words.

Keywords

Please provide appropriate keywords.

Acknowledgments

Contributions from anyone who does not meet the criteria for authorship should be listed, with permission from the contributor, in an Acknowledgments section. Financial and material support should also be mentioned. Thanks to anonymous reviewers are not appropriate.

Main Text File

As papers are double-blind peer reviewed, the main text file should not include any information that might identify the authors.

The main text file should be presented in the following order:

Title

Main text

References

Tables and figures (each complete with title and footnotes)

Appendices (if relevant)

Supporting information should be supplied as separate files. Tables and figures can be included at the end of the main document or attached as separate files but they must be mentioned in the text.

As papers are double-blind peer reviewed, the main text file should not include any information that might identify the authors. Please do not mention the authors' names or affiliations and always refer to any previous work in the third person.

The journal uses British/US spelling; however, authors may submit using either option, as spelling of accepted papers is converted during the production process.

References

References should be prepared according to the *Publication Manual of the American Psychological Association* (6th edition). This means in text citations should follow the author-date method whereby the author's last name and the year of publication for the source should appear in the text, for example, (Jones, 1998). The complete reference list should appear alphabetically by name at the end of the paper. Please note that for journal articles, issue numbers are not included unless each issue in the volume begins with page 1, and a DOI should be provided for all references where available.

For more information about APA referencing style, please refer to the **APA FAQ**.

Reference examples follow:

Journal article

Beers, S. R. , & De Bellis, M. D. (2002). Neuropsychological function in children with maltreatment-related posttraumatic stress disorder. *The American Journal of Psychiatry*, *159*, 483–486. doi:**10.1176/appi.ajp.159.3.483**

Book

Bradley-Johnson, S. (1994). *Psychoeducational assessment of students who are visually impaired or blind: Infancy through high school* (2nd ed.). Austin, TX: Pro-ed.

Internet Document

Norton, R. (2006, November 4). How to train a cat to operate a light switch [Video file]. Retrieved from **http://www.youtube.com/watch?v=Vja83KLQXZs**

Tables

Tables should be self-contained and complement, not duplicate, information contained in the text. They should be supplied as editable files, not pasted as images. Legends should be concise but comprehensive – the table, legend, and footnotes must be understandable without reference to the text. All abbreviations must be defined in footnotes. Footnote symbols: †, ‡, §, ¶, should be used (in that order) and *, **, *** should be reserved for P-values. Statistical measures such as SD or SEM should be identified in the headings.

Figures

Although authors are encouraged to send the highest-quality figures possible, for peer-review purposes, a wide variety of formats, sizes, and resolutions are accepted.

Click here for the basic figure requirements for figures submitted with manuscripts for initial peer review, as well as the more detailed post-acceptance figure requirements.

Legends should be concise but comprehensive – the figure and its legend must be understandable without reference to the text. Include definitions of any symbols used and define/explain all abbreviations and units of measurement.

Colour figures. At the editors' discretion, colour figures can be provided for use in the journal. Good quality photographs will be considered where they add substantially to the argument, to a maximum of three per article. These can be supplied electronically as TIF files scanned to at least 300dpi. If they are printed in colour, then they can be reproduced in colour online and black and white in print.

Supporting Information

Supporting information is information that is not essential to the article, but provides greater depth and background. It is hosted online and appears without editing or typesetting. It may include tables, figures, videos, datasets, etc.

Click here for Wiley's FAQs on supporting information.

Note: if data, scripts, or other artefacts used to generate the analyses presented in the paper are available via a publicly available data repository, authors should include a reference to the location of the material within their paper.

General Style Points

For guidelines on editorial style, please consult the **APA Publication Manual** published by the American Psychological Association. The following points provide general advice on formatting and style.

Language: Authors must avoid the use of sexist or any other discriminatory language.

Abbreviations: In general, terms should not be abbreviated unless they are used repeatedly and the abbreviation is helpful to the reader. Initially, use the word in full, followed by the abbreviation in parentheses. Thereafter use the abbreviation only.

Units of measurement: Measurements should be given in SI or SI-derived units. Visit the Bureau International des Poids et Mesures (BIPM) website for more information about SI units.

Effect size: In normal circumstances, effect size should be incorporated.

Numbers: numbers under 10 are spelt out, except for: measurements with a unit (8mmol/l); age (6 weeks old), or lists with other numbers (11 dogs, 9 cats, 4 gerbils).

Wiley Author Resources

Manuscript Preparation Tips: Wiley has a range of resources for authors preparing manuscripts for submission available **here.** In particular, we encourage authors to consult Wiley's best practice tips on **Writing for Search Engine Optimization**.

Article Preparation Support: **Wiley Editing Services** offers expert help with English Language Editing, as well as translation, manuscript formatting, figure illustration, figure formatting, and graphical abstract design – so you can submit your manuscript with confidence.

Also, check out our resources for **Preparing Your Article** for general guidance and the **BPS Publish with Impact infographic** for advice on optimizing your article for search engines.

5. EDITORIAL POLICIES AND ETHICAL CONSIDERATIONS

Peer Review and Acceptance

Except where otherwise stated, the journal operates a policy of anonymous (double blind) peer review. Please ensure that any information which may reveal author identity is blinded in your submission, such as institutional affiliations, geographical location or references to unpublished research. We also operate a triage process in which submissions that are out of scope or otherwise inappropriate will be rejected by the editors without external peer review. Before submitting, please read **the terms and conditions of submission** and the **declaration of competing interests**.

The Journal receives a large volume of papers to review each year, and in order to make the process as efficient as possible for authors and editors alike, all papers are initially examined by the Editors to ascertain whether the article is suitable for full peer review. In order to qualify for full review, papers must meet the following criteria:
- the content of the paper falls within the scope of the Journal
- the methods and/or sample size are appropriate for the questions being addressed
- research with patient populations is appropriately defined
- the word count is within the stated limit for the Journal (i.e. 6000 words)

The *Journal of Neuropsychology* is committed to a fast and efficient turnaround of papers, aiming to complete the review process in under two months.

Further information about the process of peer review and production can be found in '**What happens to my paper?**' Appeals are handled according to the **procedure recommended by COPE**. Wiley's policy on the confidentiality of the review process is **available here**.

Research Reporting Guidelines

Accurate and complete reporting enables readers to fully appraise research, replicate it, and use it. Authors are encouraged to adhere to recognised research reporting standards. The EQUATOR Network collects more than 370 reporting guidelines for many study types, including for:

Randomised trials: CONSORT

Systematic reviews: PRISMA

Interventions: TIDieR

Clinical case reports: CARE

We encourage authors to adhere to the APA Style Journal Article Reporting Standards for:

Manuscripts that report primary qualitative research

Manuscripts that report the collection and integration of qualitative and quantitative data

Manuscripts that report new data collections regardless of research design

We also encourage authors to refer to and follow guidelines from the **FAIRsharing website**.

Conflict of Interest

The journal requires that all authors disclose any potential sources of conflict of interest. Any interest or relationship, financial or otherwise that might be perceived as influencing an author's objectivity is considered a potential source of conflict of interest. These must be disclosed when directly relevant or directly related to the work that the authors describe in their manuscript. Potential sources of conflict of interest include, but are not limited to: patent or stock ownership, membership of a company board of directors, membership of an advisory board or committee for a company, and consultancy for or receipt of speaker's fees from a company. The existence of a conflict of interest does not preclude publication. If the authors have no conflict of interest to declare, they must also state this at submission. It is the responsibility of the corresponding author to review this policy with all authors and collectively to disclose with the submission ALL pertinent commercial and other relationships.

Funding

Authors should list all funding sources in the Acknowledgments section. Authors are responsible for the accuracy of their funder designation. If in doubt, please check the Open Funder Registry for the correct nomenclature: **https://www.crossref.org/services/funder-registry/**

Authorship

All listed authors should have contributed to the manuscript substantially and have agreed to the final submitted version. Authorship is defined by the criteria set out in the APA Publication Manual:

"Individuals should only take authorship credit for work they have actually performed or to which they have substantially contributed (APA Ethics Code Standard 8.12a, Publication Credit). Authorship encompasses, therefore, not only those who do the actual writing but also those who have made substantial scientific contributions to a study. Substantial professional contributions may include formulating the problem or hypothesis, structuring the experimental design, organizing and conducting the statistical analysis, interpreting the results, or writing a major portion of the paper. Those who so contribute are listed in the byline." (p.18)

Data Sharing and Data Accessibility Policy

The *Journal of Neuropsychology* recognizes the many benefits of archiving data for scientific progress. Archived data provides an indispensable resource for the scientific community, making possible future replications and secondary analyses, in addition to the importance of verifying the dependability of published research findings.

The journal expects that where possible all data supporting the results in papers published are archived in an appropriate public archive offering open access and guaranteed preservation. The archived data must allow each result in the published paper to be recreated and the analyses reported in the paper to be replicated in full to support the conclusions made. Authors are welcome to archive more than this, but not less.

All papers need to be supported by a data archiving statement and the data set must be cited in the Methods section. The paper must include a link to the repository in order that the statement can be published.

It is not necessary to make data publicly available at the point of submission, but an active link must be included in the final accepted manuscript. For authors who have pre-registered studies, please use the Registered Report link in the Author Guidelines.

In some cases, despite the authors' best efforts, some or all data or materials cannot be shared for legal or ethical reasons, including issues of author consent, third party rights, institutional or national regulations or laws, or the nature of data gathered. In such cases, authors must inform the editors at the time of submission. It is understood that in some cases access will be provided under restrictions to protect confidential or proprietary information. Editors may grant exceptions to data access requirements provided authors explain the restrictions on the data set and how they preclude public access, and, if possible, describe the steps others should follow to gain access to the data.

If the authors cannot or do not intend to make the data publicly available, a statement to this effect, along with the reasons that the data is not shared, must be included in the manuscript.

Finally, if submitting authors have any questions about the data sharing policy, please access the **FAQs** for additional detail.

Publication Ethics

Authors are reminded that the *Journal of Neuropsychology* adheres to the ethics of scientific publication as detailed in the ***Ethical principles of psychologists and code of conduct*** (American Psychological Association, 2010). The Journal generally conforms to the Uniform Requirements for Manuscripts of the International Committee of Medical Journal Editors (**ICJME**) and is also a member and subscribes to the principles of the Committee on Publication Ethics (**COPE**). Authors must ensure that all research meets these ethical guidelines and affirm that the research has received permission from a stated Research Ethics Committee (REC) or Institutional Review Board (IRB), including adherence to the legal requirements of the study county.

Note this journal uses iThenticate's CrossCheck software to detect instances of overlapping and similar text in submitted manuscripts. Read Wiley's Top 10 Publishing Ethics Tips for Authors **here**. Wiley's Publication Ethics Guidelines can be found **here**.

ORCID

As part of the journal's commitment to supporting authors at every step of the publishing process, the journal requires the submitting author (only) to provide an ORCID iD when submitting a manuscript. This takes around 2 minutes to complete. **Find more information here.**

6. AUTHOR LICENSING

If a paper is accepted for publication, the author identified as the formal corresponding author will receive an email prompting them to log in to Author Services, where via the Wiley Author Licensing Service (WALS) they will be required to complete a copyright license agreement on behalf of all authors of the paper.

Authors may choose to publish under the terms of the journal's standard copyright agreement, or **OnlineOpen** under the terms of a Creative Commons License.

General information regarding licensing and copyright is available **here**. To review the Creative Commons License options offered under OnlineOpen, please **click here.** (Note that certain funders mandate a particular type of CC license be used; to check this please click **here**.)

BPS members:

Self-Archiving Definitions and Policies: Note that the journal's standard copyright agreement allows for self-archiving of different versions of the article under specific conditions. Please click **here** for more detailed information about self-archiving definitions and policies.

Open Access fees**:** Authors who choose to publish using OnlineOpen will be charged a fee. A list of Article Publication Charges for Wiley journals is available **here.**

Appendix 1.2. Search strategy for OVID interface

(machine*learning OR neural network* OR support vector machine OR multilayer perceptron OR random forest OR lasso OR ridge OR kernel OR Bayesian network OR classification tree OR regression tree OR relevance vector machine OR nearest neighbo*r OR probability estimation tree OR elastic net OR ensemble OR penali*ed OR regulari*ed OR bagging OR boosted OR boosting OR fuzzy OR na*ve bayes OR deep learning OR genetic algorithm*)

AND

(head injur* OR brain injur* OR stroke OR brain h*emorrhage OR head trauma OR brain trauma OR concussion OR TBI OR ABI OR HI OR mTBI OR cerebrovascular accident OR CVA OR subarachnoid h*emorrhage)

AND

(Educat* OR school* OR behavio*r* OR psychosocial* OR psychologi* OR neuropsychologi* OR problem solv* OR cogniti* OR executive OR memory OR attention* OR social OR stress OR (quality adj5 life) OR QoL OR hrqol OR depress* OR anxi* OR psychiatr* OR mental health OR well*being OR living OR accommodation OR independen* OR support* OR residen* OR placement OR destination OR domestic OR famil* OR relation* OR employ* OR work* OR occupation* OR job* OR affect* OR mood OR emotion* OR function* OR instrumental OR activ* OR ADL OR IADL)

Human/

Appendix 1.3. Data extraction template

- Bibliographic details
- Includes people with a diagnosis of ABI?
- Has a separate analysis has been included for ABI?
- Type of ABI
- If applicable, are comparator groups from the same population?
- Study design
- Were the exposures similarly measured between groups, or with all participants?
- Was the exposure measured in a valid and reliable way?
- Were confounding factors identified?
- Were strategies to deal with confounding factors stated?
- Were methods for missing data used and what?
- Were the outcomes measured in a valid and reliable way?
- Was the follow up time reported?
- Was follow up complete?
- If follow up was not complete, were the reasons to loss to follow up described and explored?
- Were strategies to address incomplete follow up utilized?
- What method of machine learning was used?
- Was the rationale for the type of ML algorithm described?
- Was an a priori power analysis performed?
- Were feature selection methods used?
- What performance metrics were used for model performance? E.g. AUC
- Were the reported metrics for model performance appropriate for the type of ML algorithm?
- What was the result of the performance metric for each algorithm/model reported? Record development and validation metrics.
- If more than one type of ML algorithm was used, which one had the most superior performance?
- Was the model validated in the study?
- If the model was validated, which validation methods were used?
- Were limitations of ML techniques discussed?
- If ML limitations were discussed, what were they?
- Relevant reported limitations to power of study?
- Sample size
- Sample demographics

- Country
- Conflicts of interest
- Funding source

Appendix 1.4. Machine learning algorithm definitions

Machine learning algorithms		Definition
Classification	Regularised 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 amongst 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 their 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 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) regularisation with linear regression	In the regression equation, lasso sets certain coefficients to 0, with the goal of increasing prediction accuracy while 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.

References

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5-32.

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

Cunningham, P., & Delany, S. J. (2020). k-Nearest Neighbour Classifiers--. *arXiv preprint arXiv:2004.04523*.

Nadkarni, P. (2016). Chapter 4 - Core Technologies: Machine Learning and Natural Language Processing. *Clinical Research Computing* (pp. 85-114). Academic Press. doi:https://doi.org/10.1016/B978-0-12-803130-8.00004-X.

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.

Rokach, L., & Maimon, O. Z. (2008). *Data mining with decision trees: theory and applications* (Vol. 69). World scientific.

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

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. doi:10.1109/5326.897072

Appendix 1.5. 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 minimising 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 minimising 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 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 4.9. Unclear if predictors in the final models correspond to results from analysis as training data presented only

Appendix 2.1: Ethics approval letter

WoSRES
West of Scotland Research Ethics Service

NHS
Greater Glasgow and Clyde

Dr Breda Cullen
Lecturer in Mental Health
University of Glasgow
University of Glasgow, 1st floor, Administration Building,
Clinical Psychology, Gartnavel Royal Hospital
1055 Great Western Road, Glasgow
G12 0XH

West of Scotland REC 1
Research Ethics
Clinical Research and Development
Ward 11
Dykebar Hospital
Grahamston Road
Paisley PA2 7DE

Date 24 February 2020
Direct line 0141 314 0212
E-mail WoSREC1@ggc.scot.nhs.uk

Please note: This is the favourable opinion of the REC only and does not allow you to start your study at NHS sites in England until you receive HRA Approval

Dear Dr Cullen

Study title: **Predicting psychosocial outcomes at the time of discharge from inpatient neurorehabilitation for acquired brain injury: Development and validation of a machine learning model.**
REC reference: **20/WS/0026**
Protocol number: **n/a**
IRAS project ID: **266548**

Thank you for your email of 18 February 2020, responding to the Committee's request for further information on the above research and submitting revised documentation.

The further information has been considered on behalf of the Committee by the Chair.

Confirmation of ethical opinion

On behalf of the Committee, I am pleased to confirm a favourable ethical opinion for the above research on the basis described in the application form, protocol and supporting documentation as revised, subject to the conditions specified below.

Conditions of the favourable opinion

The REC favourable opinion is subject to the following conditions being met prior to the start of the study.

Confirmation of Capacity and Capability (in England, Northern Ireland and Wales) or NHS management permission (in Scotland) should be sought from all NHS organisations involved in

the study in accordance with NHS research governance arrangements. Each NHS organisation must confirm through the signing of agreements and/or other documents that it has given permission for the research to proceed (except where explicitly specified otherwise).

Guidance on applying for HRA and HCRW Approval (England and Wales)/ NHS permission for research is available in the Integrated Research Application System.

For non-NHS sites, site management permission should be obtained in accordance with the procedures of the relevant host organisation.

Sponsors are not required to notify the Committee of management permissions from host organisations

Registration of Clinical Trials

It is a condition of the REC favourable opinion that **all clinical trials are registered** on a publicly accessible database. For this purpose, 'clinical trials' are defined as the first four project categories in IRAS project filter question 2. Registration is a legal requirement for clinical trials of investigational medicinal products (CTIMPs), except for phase I trials in healthy volunteers (these must still register as a condition of the REC favourable opinion).

Registration should take place as early as possible and within six weeks of recruiting the first research participant at the latest. Failure to register is a breach of these approval conditions, unless a deferral has been agreed by or on behalf of the Research Ethics Committee (see here for more information on requesting a deferral: https://www.hra.nhs.uk/planning-and-improving-research/research-planning/research-registration-research-project-identifiers/

As set out in the UK Policy Framework, research sponsors are responsible for making information about research publicly available before it starts e.g. by registering the research project on a publicly accessible register. Further guidance on registration is available at: https://www.hra.nhs.uk/planning-and-improving-research/research-planning/transparency-responsibilities/

You should notify the REC of the registration details. We will audit these as part of the annual progress reporting process.

It is the responsibility of the sponsor to ensure that all the conditions are complied with before the start of the study or its initiation at a particular site (as applicable).

After ethical review: Reporting requirements

The attached document "After ethical review – guidance for researchers" gives detailed guidance on reporting requirements for studies with a favourable opinion, including:

- Notifying substantial amendments
- Adding new sites and investigators
- Notification of serious breaches of the protocol
- Progress and safety reports
- Notifying the end of the study, including early termination of the study

- Final report

The latest guidance on these topics can be found at
https://www.hra.nhs.uk/approvals-amendments/managing-your-approval/.

Ethical review of research sites

NHS/HSC sites

The favourable opinion applies to all NHS/HSC sites listed in the application subject to confirmation of Capacity and Capability (in England, Northern Ireland and Wales) or management permission (in Scotland) being obtained from the NHS/HSC R&D office prior to the start of the study (see "Conditions of the favourable opinion").

Non-NHS/HSC sites (if applicable)

I am pleased to confirm that the favourable opinion applies to any non-NHS/HSC sites listed in the application, subject to site management permission being obtained prior to the start of the study at the site.

Approved documents

The final list of documents reviewed and approved by the Committee is as follows:

Document	*Version*	*Date*
Evidence of Sponsor insurance or indemnity (non NHS Sponsors only) [University of Glasgow insurance]		24 July 2019
Letter from sponsor [Sponsor letter]		24 October 2019
REC Application Form [REC_Form_14012020]		14 January 2020
Research protocol or project proposal [IRAS proposal version 4 (clean)]	4	09 February 2020
Research protocol or project proposal [Tracked proposal version 4]	4	09 February 2020
Response to Request for Further Information		18 February 2020
Summary CV for Chief Investigator (CI) [Breda Cullen CV]		31 July 2019
Summary CV for student [Emma Mawdsley CV]		03 October 2019

Statement of compliance

The Committee is constituted in accordance with the Governance Arrangements for Research Ethics Committees and complies fully with the Standard Operating Procedures for Research Ethics Committees in the UK.

User Feedback

The Health Research Authority is continually striving to provide a high quality service to all applicants and sponsors. You are invited to give your view of the service you have received and the application procedure. If you wish to make your views known please use the feedback form available on the HRA website:
http://www.hra.nhs.uk/about-the-hra/governance/quality-assurance/

HRA Learning

We are pleased to welcome researchers and research staff to our HRA Learning Events and online learning opportunities – see details at:
https://www.hra.nhs.uk/planning-and-improving-research/learning/

20/WS/0026	**Please quote this number on all correspondence**

With the Committee's best wishes for the success of this project.

Yours sincerely

On behalf of
Dr Malcolm Booth
Chair

Enclosures: "After ethical review – guidance for researchers"

Copy to: Miss Emma-Jane Gault, University of Glasgow
Ms Emma Mawdsley, NHS Greater Glasgow and Clyde

Appendix 2.2 The Disabilities Trust Management Approval

32 Market Place
Burgess Hill
West Sussex
RH15 9NP
Telephone: 01444 239123
Fax: 01444 244978
Email: foundation@thedtgroup.org

24 October 2019

To whom it may concern,

Re: Predicting psychosocial outcomes at the time of discharge from inpatient neurorehabilitation for acquired brain injury: Development and validation of a machine learning model

This letter is to confirm that The Disabilities Trust is aware of, and supports, the proposed analysis of anonymised service evaluation data to be carried out by Ms Emma Mawdsley as part of the requirements for her doctoral degree in Clinical Psychology at the University of Glasgow.

The Disabilities Trust is a leading national charity, providing innovative care, rehabilitation and support solutions for people with acquired brain injury. We welcome the opportunity to carry out this investigation in order to better understand and promote the factors that moderate the most positive outcomes following rehabilitation.

This letter is intended to give support in principle to the aforementioned project and is subject to the implementation of the recommendations of the Trust's Data Protection Officer and a valid Information Sharing Agreement between The Disabilities Trust and the University of Glasgow.

Please do not hesitate to contact me to discuss any aspect of this submission.

Yours faithfully

Dr Sara da Silva Ramos
Research Fellow
Research & Programme Evaluation Office
Kerwin Court
Brain injury Rehabilitation Trust
Five Oaks Road
Horsham
RH13 0TP

Cc: Dr Brian O'Neill, Consultant in Neuropsychology and Rehabilitation, Brain Injury Rehabilitation Trust
Mr Allan Darby, Data Protection Officer and Health and Safety Lead, Central Support Services, Quality Assurance

Appendix 2.3. The Disabilities Trust Caldicott Approval

Sue Copstick
Clinical Director / Caldicott Guardian
Disabilities Trust
sue.copstick@thedtgroup.org

BIRT
Brain Injury
Rehabilitation Trust
The Disabilities Trust

Emma-Jane Gault
Research Governance Officer
University of Glasgow
Wolfson Medical School Building
University Avenue
Glasgow
G12 8QQ

Email emmajane.gault@glasgow.ac.uk

10.01.2020

Dear Miss Gault,

Re: Research project at Graham Anderson House

We reviewed the Caldicott application for the proposed research project by Emma Mawdsley, Trainee Clinical Psychologist, predicting psychosocial outcomes at discharge from Graham Anderson House (IRAS version 3, 12.12.19).

I note that the researcher will need access to patient information to create the pseudo-anonymised database by gathering clinical information from different sources. The research database will have data identified by ID number which will be able to be traced back to the client via the clinical database. The research database will not be stored in the same location as the clinical database to minimise possibility of participant identification.

This letter is to signify Caldicott approval for access to necessary patient data for the project.

Yours sincerely,

Sue Copstick
Clinical Director

Appendix 2.4. NHS Greater Glasgow and Clyde Caldicott Approval

NHS
Greater Glasgow and Clyde

Emma Mawdsley
@student.gla.ac.uk

Data Protection Officer
Information Governance Department
NHS Greater Glasgow & Clyde
2nd Floor, 1 Smithhills Street
Paisley PA1 1EB

Date: 06/01/2020

Enquiries to: Isobel Brown
Tel: 0141 355 2020
Email: Isobel.Brown@ggc.scot.nhs.uk

Dear Emma,

Re: Predicting Psychosocial Outcomes from Inpatient Neurorehabillitation

Thank you for your Caldicott application received on 21/08/2016regarding your proposed Research Project.

I have reviewed this application and can confirm that I am happy to approve this application on behalf of the Caldicott Guardian.
Please note that this approval only covers access to NHSGGC patients.

Please find attached a signed copy of your application for your records.

Yours sincerely

Isobel Brown
Data Protection Officer
Information Governance

Appendix 2.5: Candidate baseline predictors and data processing

Candidate predictors		*Measure*	*Data type and processing if applicable*
Neuropsychological predictors *Assessed by a member of the psychological team at GAH following admission*	General cognitive ability	WAIS -IV (Wechsler, 2008)	5 domains including full scale IQ, verbal comprehension, perceptual reasoning, working memory, and processing speed, each kept as a continuous measure.
	Predicted premorbid intelligence	TOPF (Wechsler, 2011)	Predictions of premorbid functioning in 5 domains including full scale IQ, verbal comprehension, perceptual reasoning, working memory, and processing speed, each kept as a continuous measure.
	Executive functioning	BADS (Wilson, Evans, Alderman, Burgess, & Emslie, 1997) or the DKEFS (Delis, Kaplan, & Kramer, 2001) depending on administrator choice.	An ordinal measure categorised as either high average, average, low average, borderline or impaired as coded in the BADS. If the DKEFS was administered instead of the BADS, the total score was averaged across subtests if they had 3 or more administered. This mean scaled score was then converted into an ordinal scaled score as coded by the BADS or coded as impaired if administration had clearly been discontinued. If a service user had both DKEFS and BADS administered at baseline, the BADS score was chosen over DKEFS due to the established standardised method Data were coded 1=Impaired 2=Borderline 3=Low average 4=Average 5=High average
	Memory	Either RBMT (Wilson et al., 1999) or BMIPB (Oddy, Coughlan, & Crawford, 2007)	Scale quotient score as coded in the RBMT. If data was missing from the RBMT due to part administration, the total score was averaged across subtests if had 3 or more administered. For service users who had the BMIPB administered instead of the RBMT, the BMIPB score was converted into regression-based continuous norms using the Crawford equation to provide a quotient score equivalent of the RBMT [available from https://homepages.abdn.ac.uk/j.crawford/pages/dept/BMIPB_Programs.htm]. If a service user had both BMIPB and RBMT administered at baseline, the RBMT score was chosen over BMIPB due to the established standardised method

Candidate predictors		*Measure*	*Data type and processing if applicable*
	Neuropsychological data availability	An indicator of whether baseline neuropsychological assessments were administered	As neuropsychological test administration was expected not to be missing at random, an indicator for testing ability was developed as a categorical measure, labelled ordinally as "3" if service user had complete measures across WAIS, memory, executive and TOPF assessments, "2" if not all were administered or had WAIS subtests missing, and "1" if no neuropsychological tests were administered or had to be discontinued without enough to score a minimum of three subscales across all neuropsychological tests.
Psychometric predictors *Assessed by a member of the psychological team at GAH following admission*	Anxiety and Depression	HADS (Zigmond & Snaith, 1983) at admission responded to by the patient	Total scores of both the depression and anxiety subscales as continuous measures
	Physical, cognitive, emotional, behavioural, and social difficulties related to brain injury	MPAI 4th edition (Bellon et al., 2012) at admission. The MPAI-4 may be completed by the patient, healthcare professional or family member.	Total T-scores of the adjustment, participation, abilities and total subscales as continuous measures
Demographics		Age at injury	Continuous, in weeks
		Age at admission	Continuous, in weeks
		Gender	Categorised as male or female as per recording practices in the service
		SIMD	The SIMD rank score as a continuous measure with higher scores indicating lower levels of deprivation

Candidate predictors	*Measure*	*Data type and processing if applicable*
Medical history	Type of ABI	A categorical measure of either TBI, CVA, hypoxia, infection, neoplasm or other.
	Problematic drug use	Binary Yes or No
	Problematic alcohol use	Binary Yes or No
	Multiple trauma	Binary Yes or No
	Other medical history	Binary Yes or No
	Pre-injury psychosis	Binary Yes or No
	Days between injury and admission	In days as a continuous measure

Abbreviations: ABI= Acquired brain injury ; BADS= Behavioural Assessment of the Dysexecutive Syndrome ; BMIPB= BIRT Memory and Information Processing Battery ; CVA=Cerebrovascular accident ; DKEFS= Delis-Kaplan Executive Function System ; HADS = Hospital Anxiety and Depression Scale ; MPAI= Mayo Portland Adaptability Inventory ; RBMT= Rivermead Behavioural Memory Test ; SIMD= Scottish Index of Multiple Deprivation ; TBI= Traumatic brain injury ; TOPF= Test of Premorbid Functioning; WAIS= Wechsler Adult Intelligence Scale (FSIQ= Full scale IQ ; PRI= Perceptual reasoning index ; PSI= Processing speed index ; VCI= Verbal comprehension index ; WMI= Working memory index).

References

Bellon, K., Malec, J. F., & Kolakowsky-Hayner, S. A. (2012). Mayo-portland adaptability inventory-4. *The Journal of Head Trauma Rehabilitation, 27*(4), 314-316.

Delis, D. C., Kaplan, E., & Kramer, J. H. (2001). *Delis-Kaplan Executive Function System.* United Kingdom: Blackwell Publisher.

Oddy, M., Coughlan, A., & Crawford, J. (2007). BIRT memory and information processing battery. *Horsham, UK: Brain Injury Research Trust*.

Wechsler, D. (2008). *Wechsler Adult Intelligence Scale–Fourth Edition (WAIS–IV)*: San Antonio, TX: The Psychological Corporation.

Wechsler, D. (2011). Test of premorbid functioning. UK version (TOPF UK). *UK: Pearson Corporation*.

Wilson, B. A., Clare, L., Cockburn, J., Baddeley, A. D., Tate, R., & Watson, P. (1999). The rivermead behavioural memory test-extended version.

Wilson, B. A., Evans, J. J., Alderman, N., Burgess, P. W., & Emslie, H. (1997). Behavioural assessment of the dysexecutive syndrome. *Methodology of Frontal and Executive Function, 239*, 250.

Appendix 2.6: Machine learning algorithms

Regularised linear regression:

A regression algorithm from the field of statistics, with elastic net penalisation (as described below), to predict a continuous outcome where *Y* is predicted*:*

$$Y = \alpha + \beta 1x1 + \beta 2x2 + \beta 3x3 \ldots$$

Where α represents the intercept and β represents the coefficients to be learned and tuned from x, the predictors in the dataset using an iterative method. Unlike statistical regression, ML linear regression in this study used elastic net, a regularisation method which penalises some less contributing coefficients to 0 or near 0 to minimize overfitting, for embedded feature selection (Zou & Hastie, 2005). The goal with elastic net penalisation is to have the benefits from alternative penalisation methods, ridge and lasso, of reducing the variance between predicted and observed data points, to handle greater collinearity between variables and to use a tuning parameter that reduces the likelihood of overfitting. At the same time elastic net overcomes the challenges of ridge regularisation (not having feature selection), and lasso (less effective regularisation).

Regularised logistic regression:

A classification algorithm whereby coefficient weights are learned using an iterative method with adjustments within a linear algorithm (described above in regularised linear regression) before being transformed to predict a binary outcome using the sigmoid, or logistic function (Nadkarni, 2016). Regularised logistic regression in this study also used elastic net penalisation for regularisation, which is solved as below:

$$\min_{\beta_0,\beta} 1N\sum_{i=1N} w_i l(y_i,\beta_0+\beta_T x_i)+\lambda[(1-\alpha)||\beta||_{22}/2+\alpha||\beta||_1]$$

“over a grid of values of λ covering the entire range. Here l(y,η) is the negative log-likelihood contribution for observation ii; e.g. for the Gaussian case it is 12(y−η)2. The *elastic-net* penalty is controlled by α, and bridges the gap between lasso (α=1, the default) and ridge (α=0). The tuning parameter λ controls the overall strength of the penalty” (Hastie & Qian, 2014).

Random forest:

An ensemble method where a large number of decision trees are grown, each with a random split of training data from the original data with replacement (known as bootstrap samples). Each bootstrap sample is trained using random feature selection and node splits to create trees which are largely uncorrelated. Each tree then votes for the most popular class at input x (Breiman, 2001) combining the

results with a technique called “bagging.” The goal here is to produce a stronger, less biased model than single decision trees alone, which reduces the variance without increasing the bias. Random forest algorithms have “out the bag” error embedded during model development which adjusts the fit of the models based on the results of each bootstrap sample for validation during the model development process.

Linear kernel support vector machine

A classification algorithm whereby vectors are mapped into a high dimensional space to construct a linear decision surface with the goal of separating two decision categories with a maximum “margin” (the distance between the data points and the linear decision surface, or hyperplane) between them (Cortes & Vapnik, 1995). The “support vectors” are the data points closest to the margin, meaning data further from the decision margin are not used (Nadkarni, 2016).

References

Breiman, L. (2001). Random forests. *Machine learning*, 45(1), 5-32.

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

Hastie, T., & Qian, J. (2014). Glmnet Vignette. Retrieved from https://web.stanford.edu/~hastie/glmnet/glmnet_alpha.html

Nadkarni, P. (2016). Chapter 4 - Core Technologies: Machine Learning and Natural Language Processing. *Clinical Research Computing* (pp. 85-114). Academic Press. doi:https://doi.org/10.1016/B978-0-12-803130-8.00004-X.

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 67*(2), 301-320.

Appendix 2.7 Sensitivity analyses for ML algorithms with ≥50% complete data on predictor variables using *K*NN imputation

Outcome	Sample size	Method	AUC	95% CI	Calibration (slope)	Permutation p value
Accommodation	164	RLR	0.78	0.73-0.83	0.89	<0.001
		RF	0.82	0.78-0.86	1.09	<0.001
		SVM	0.78	0.73-0.83	0.96	<0.001
Participation	120	RLR	0.77	0.72-0.82	1.22	<0.001
		RF	0.79	0.75-0.86	1.47	<0.001
		SVM	0.72	0.66-0.77	0.82	<0.001
Supervision	167	RLR	0.77	0.73-0.82	1.04	<0.001
		RF	0.83	0.80-0.86	1.10	<0.001
		SVM	0.75	0.71-0.80	0.96	<0.001
Occupation	164	RLR	0.68	0.64-0.72	0.95	<0.001
		RF	0.66	0.63-0.70	0.93	<0.001
		SVM	0.62	0.58-0.66	0.75	<0.001
Quality of life	38	RLR	0.73	0.63-0.83	0.22	<0.001

AUC= Area under the curve; CI= Confidence intervals; LR= logistic regression; RF= random forest; RLR= regularised logistic regression; SVM= support vector machine.

Appendix 2.8. Stable predictors identified by 100% of developed models for RLR during nested cross-validation

Outcome	RLR (80% complete data) (coefficient mean in regression algorithm)	RLR (50% complete data) (coefficient mean in regression algorithm)
Accommodation	1. Neuropsychological data availability (0.506) 2. MPAI Adjustment (-0.379)	1. Neuropsychological data availability (0.514) 2. Diagnosis other (-0.387) 3. Total MPAI (-0.368)
Participation	1. Neuropsychological data availability (0.361)	1. Neuropsychological data availability (0.488)
Supervision	1. Neuropsychological data availability (0.349) 2. MPAI Adjustment (-0.355) 3. MPAI Abilities (-0.291) 4. Diagnosis Hypoxia (-0.237)	1. Neuropsychological data availability (0.358) 2. Processing speed WAIS (0.331) 3. MPAI Adjustment (-0.283) 4. Diagnosis Hypoxia (0.232)
Occupation	1. Gender male (-0.483) 2. Neuropsychological data availability (0.427) 3. MPAI Abilities (-0.366) 4. Diagnosis other (-0.191)	1. Gender male (-0.354) 2. Neuropsychological data availability (0.306) 3. Working memory WAIS (0.223) 4. HADS Anxiety (-0.223) 5. MPAI Abilities (-0.156)
QOL	1. HADS Anxiety (-1.053)	1. HADS Anxiety (-1.053)

HADS= Hospital anxiety and depression scale; MPAI= Mayo Portland adaptability inventory; RLR = regularised logistic regression; QOL= quality of life; WAIS= Wechsler adult intelligence scale

Appendix 2.9. Major research project proposal

REC Proposal –Version 4

Predicting psychosocial outcomes at the time of discharge from inpatient neurorehabilitation for acquired brain injury: Development and validation of a machine learning model

09/February/2020

Emma Mawdsley

Breda Cullen

Breda.Cullen@glasgow.ac.uk

Abstract

Background: Outcomes after inpatient neurorehabilitation for acquired brain injury (ABI) are variable, challenging the ability to make timely, cost-effective decisions for the level of required support for housing and functional needs at the time of discharge. Machine learning algorithms have potentially high predictive power for modelling high dimensional data with collinear predictors, providing a valuable tool for predicting clinical prognostic outcomes.

Aims: This project aims to build predictive models using neuropsychological, psychometric, medical and demographic variables to predict outcomes for service users' length of admission, independence of living, required level of supervision, occupational functioning, participation in functional tasks, and quality of life after discharge from inpatient neurorehabilitation.

Methods: This study will use a retrospective cohort study design of service users with ABI admitted and discharged from Graham Anderson House, neurorehabilitation centre. Data will be analysed using machine learning. Three different machine learning algorithms will be compared for their performance for predicting psychosocial outcomes after discharge. The algorithms will be selected based on the findings of the systematic review (also conducted by the researcher) for the most superior machine learning methods for predicting psychosocial outcomes in ABI research. These algorithms will be compared for their performance using a receiver operating characteristic (ROC) curve, with the area under the curve (AUC) being the primary metric used to determine model performance.

Applications: A strong predictive algorithm will aid timely, cost-effective decisions for appropriate accommodation and support needs at the time of discharge.

Introduction

The number of admissions for people with acquired brain injury (ABI) in the UK is increasing, raising concern for developing care pathways for this highly heterogeneous condition. Approximately 1000 daily hospital admissions in the UK are attributed to ABI (Menon & Bryant, 2019), with stroke and traumatic brain injury being in the most frequent causes of death and disability worldwide (Feigin, Barker-Collo, Krishnamurthi, Theadom, & Starkey, 2010). The burden following ABI can affect multiple domains of functioning and continue throughout the lifespan, with outcomes varying greatly with the timing and intensity of interventions (Cullen et al., 2007).

Brain Injury Rehabilitation Trust (BIRT) centres offer inpatient neurobehavioural rehabilitation for people with severe ABI. Rehabilitation programmes are individually tailored to improve functioning across cognitive, behavioural and motor domains. Even with the development of individualised rehabilitation programmes, long term outcomes for people with ABI remain variable (Ponsford et al., 2014; Rassovsky et

al., 2015). This challenges healthcare professionals ability to offer adequate guidance for service users (SUs), family members, and social services for the person's probable housing, functional, and occupation needs.

A person's functional independence forms the basis of decision making for the required level of care. UK multicentre outcomes between 2010 and 2015 showed mean costs of care following inpatient neurorehabilitation discharge differ from £306/week (95% Confidence Interval £271-£342) for low dependency needs, to £1349/week (95% CI £1315-£1384) for high dependency needs (Turner-Stokes et al., 2016). Accurately predicting support needs would ensure resources are allocated efficiently and cost-effectively.

Caregivers of people with ABI report main concerns of balancing their own emotional needs and the level of care they will need to provide (Powell et al., 2017). The sudden onset means caregivers have little time to prepare for these significant lifestyle changes. Caregiver burden is a particularly important consideration during early stages of the condition where carer distress is high (Qadeer et al., 2017). Predictive modelling can help prepare caregivers for discharge and provide timely support for those most in need.

The usefulness of predictive models in clinical practice are often limited by the statistical methods employed. Findings are often not replicated or are ungeneralisable to clinical settings (Dwyer et al., 2018; Ioannidis, 2005, 2016). Traditional methods of modelling, validated in the same dataset they are developed (Yarkoni & Westfall, 2017), may rely on unrepresentative data collected for the purpose of the research question (Agoston & Langford, 2017). With added model complexity, traditional regression models are strongly influenced by sample noise, a phenomenon known as overfitting, resulting in predictive models with inaccurate performance in new samples.

Machine learning (ML) approaches are growing in popularity for predicting healthcare outcomes with improved predictive power and less concern of overfitting (Yarkoni and Westfall, 2017). Rather than traditional statistical models, based on explanation or inference within a given dataset, ML-based predictive models make use of existing data to extrapolate to future or unknown events. ML's capacity to deal with multidimensional, missing and multicollinear data typical of clinical data (Iniesta et al., 2016) means one can make use of pre-existing data sets, more representative of true population characteristics.

With traditional regression models, neuropsychological measures predicted psychosocial outcomes after neurorehabilitation for people with ABI with linear regression accounting for 39-44% of the variance (Smith-Knapp et al., 1996) on functional measures. The limited explained variance of these models may reflect the omission of other potential predictors for these psychosocial outcomes, unable to be modelled for risk of overfitting. Traditional logistic regression predicted returning home after discharge from

inpatient Stroke neurorehabilitation using a wider range of socioeconomic, neurological and functional variables (Frank et al., 2010) with the area under the curve (AUC) of 0.86 (95% CI 0.84-0.88). Without the use of modern model validation techniques however, these models may not perform as well in new data sets. ML has previously evaluated inpatient rehabilitation efficacy with model validation, finding AUC values of 0.85-0.93 with different ML algorithms for accurately predicting patient improvement (Marcano-Cedeño et al., 2013). ML may therefore overcome these challenges, leading to more optimal model performance.

Neurorehabilitation centres generate large volumes of data during the assessment process. These data offer invaluable information for predictive modelling. Employing ML to model this data offers a way forward to making more accurate predictions of psychosocial outcomes following inpatient neurorehabilitation for people with ABI.

Aims and Hypotheses

Given the variability of psychosocial outcomes for SUs with severe ABI, this research aims to build predictive models using ML for psychosocial outcomes at discharge and quality of life (QoL) 6-months after discharge from inpatient neurorehabilitation. Our aim is to predict status at a given point in time, rather than to predict amount of change over time. Different ML algorithms will vary in their accuracy for predicting psychosocial outcomes in brain injury. This project aims to compare the performance of logistic modelling machine learning and two other machine learning methods (decided upon from the systematic review results for which algorithms have the most superior predictive power for psychosocial outcomes in ABI research).

In particular, the study aims to predict outcomes for length of admission, independence of living (accommodation status), level of supervision, occupational functioning, participation in functional tasks, and QoL. The primary outcome will be accommodation status, since clinical opinion within the service believe this is usually SUs primary concern at admission.

The primary outcome timepoint will be at discharge from the inpatient unit, with a secondary timepoint of six months post-discharge also analysed for SUs with available data.

The primary project hypothesis is that it is possible to predict, at better than chance level, independence of living at time of discharge, using baseline demographic, clinical, neuropsychological and psychometric measures from the time of admission. The null hypothesis is that the model's performance will not be reliably different from chance level. Models with better than chance performance and an AUC of 0.8 or above will be considered to show 'good' performance (Safari, Baratloo, Elfil, & Negida, 2016b)

Plan of investigation

Participants

SUs admitted and discharged with ABI from care of Graham Anderson House (GAH) between 2009 and 2018.

Inclusion and Exclusion Criteria

SUs admitted to GAH are aged between 16 and 84, with a diagnosis of ABI and complex needs in a stable condition by the time of referral. ABI may be caused by cerebrovascular accident, traumatic injury, anoxic brain injury or infection.

For SUs to be included in the study they need to have been discharged from GAH, with data available from baseline and discharge assessments. A further analysis will be performed for individuals with 6-month outcome data.

Where SUs have been readmitted, data from their initial assessment will be used, with readmission data excluded.

Participants will be excluded where routine neuropsychological assessments were unable to be administered at baseline.

Recruitment procedures

This is a retrospective cohort analysis of all SUs discharged from GAH using existing data. No new recruitment is necessary.

Measures

Baseline predictor measures to be entered into the model are displayed in Table 1.

Table 1: Baseline predictor measures:

Baseline Predictor		*Measure*
Neuropsychological predictors *Assessed by a member of the psychological team at GAH following admission*	General cognitive ability	Wechsler Adult Intelligence Scale IV (WAIS-IV, David Wechsler, 2008)
	Predicted premorbid intelligence	Test of Premorbid Functioning (TOPF, D. Wechsler, 2011)
	Executive functioning	Behavioural Assessment of the Dysexecutive Syndrome (BADS, Wilson et al., 1997)
	Memory	Either Rivermead Behavioural Memory Test (RBMT, (Wilson et al., 1999) or BIRT Memory and Information Processing Battery (BMIPB, M. Oddy et al., 2007)
	Attention	Test of Everyday Attention (TEA, Robertson, Ward, Ridgeway, & Nimmo-Smith, 1994)
Psychometric predictors *Assessed by a member of the psychological team at GAH following admission*	Anxiety and Depression	Hospital Anxiety and Depression Scale (HADS, Zigmond & Snaith, 1983) at admission responded to by the patient
	Physical, cognitive, emotional, behavioural, and social difficulties related to brain injury	Mayo Portland Adaptability Inventory 4th edition (MPAI-4, Bellon et al., 2012) at admission. The MPAI-4 may be completed by the patient, healthcare professional or family member.
Injury-related predictors		e.g. Time between injury and admission, total number of ABIs
Demographics		Age
		Gender
		Education level
		Scottish Index of Multiple Deprivation code (SIMD)
		Living arrangements prior to admission
Other medical history		Prior reported problematic drug/alcohol use
		The number of serious medical diagnoses in addition to ABI

N.B. Feasibility checks will be conducted to ascertain amount of missing data on each measure and only measures with at least 80% complete data will considered for the model.

Outcome Measures

For the purpose of the analysis, outcome measures (apart from length of admission) will be transformed into binary variables to be more practically relevant than a continuous score. Outcome data used for the purpose of the analysis will be gathered from routine outcome assessments at the time of the patient's discharge as assessed by a Clinical Psychologist within the service. Further models will be built for predicting QoL at 6-months post-discharge.

Primary Outcome Measure:

Accommodation status: Either the Accommodation Scale within the BIRT Independent Living Scale (BILS, Michael Oddy et al., 2018), or the Accommodation Rating Scale within the Community Disposition Ratings (CDR-ARS, *unpublished, appendix 1*) as rated by a member of the psychological team at GAH at the time of the patient's discharge. The scales will be converted into binary outcomes classifying the level of supported accommodation required, whereby scores ≤ 6 represent *"independent or community supported housing"* and scores ≥7 represent *"residential/hospital accommodation."*

Secondary outcome measures:

Length of admission: Length of admission will be a continuous outcome score in the number of weeks. The predicted length of admission will be used as a predictor measure in models for the other outcome measures if the model has good performance.

Participation in functional tasks: The Participation subscale of the MPAI (Bellon et al., 2012) as rated by a member of psychological team through interviewing the patient and family member where necessary. This measure is collected at the time of discharge, with binary outcome conversion of *t-scores* ≥50 representing *'moderate to severe disability',* and ≤49 representing *'a good outcome or mild to moderate disability.'*

Level of supervision: The *Supervision Rating Scale* (SRS, Boake, 1996) as rated by a member of the psychological team at the time of discharge, with binary conversions of scores ≥8 representing the person requires *'full time direct supervision,'* and ≤7 indicating *'part time or no supervision.'*

Occupational functioning: either the Occupational Participation Scale, within the BILS (Michael Oddy et al., 2018), or The Occupation Rating Scale within the CDR (CDR-ORS, *unpublished, appendix 1*), rated by a member of the psychological team at the time of discharge, with binary conversion scores of ≥7 representing *'recreational/non-occupational activity,'* and ≤ 6 indicating *'productive occupational activity.'*

QoL: The *EuroQoL Instrument* (EQ5D, EuroQoL Group., 2010) a patient-rated measure for QoL, converted as binary outcome measures. This patient-rated measure is administered at the 6-month follow-up rather than discharge. There are two commonly accepted ways to dichotomise favourable and poor EQ5D

outcomes: by dichotomising at the sample median to predict a favourable or poor outcome relative to the sample, or by dichotomising at the 0.5 point to predict a favourable outcome related to clinical populations more generally (Parkin, Devlin, & Feng, 2016). For the purpose of this study, two models will be developed to analyse outcomes of both approaches.

Design

The project will use a retrospective single-centre cohort design.

Research procedures

Ethical approval will be applied for via NHS ethics, because the majority of BIRT service users are NHS patients admitted under service-level agreements.

At admission to GAH, SUs are involved in a comprehensive assessment during which the baseline measures described above are conducted by Clinical and Assistant Psychologists within the team. When SUs are due to be discharged they undergo routine assessment of their outcomes administered by a member of the psychology team, including the outcome variables described above (apart from the EQ5D). At the 6-month follow-up, the EQ5D is administered, which is voluntarily provided by SUs by either telephone interview or postal survey. For this reason, the number for whom this data is available is considerably lower.

SU data from baseline and outcome measures will be collated in a database from paper files with the assistance of two honorary research assistants between October 2018 and April 2019. Data will be checked for accuracy prior to analyses. A research database will be collated with only anonymised data, identifiable by a participant ID number with no way of being traced back to the client in the clinical database. Patient name and other identifiable information will not be transferred to the research database. Clinical and research databases will be stored in separate locations. The research database will be stored on the University encrypted laptop and backed up regularly on University encrypted networks, given that GAH computers are not efficient for ML algorithms. The clinical database will be stored only at GAH. No patient identifiable information will be transferred to the research database.

Data analysis

Data analyses will be performed with R Programming using the Caret package. Missing data on predictor variables will be imputed using a non-parametric method (k nearest neighbours), as implemented within the Caret package. Logistic Model ML analyses, and two further machine learning algorithms (determined to be the most superior for predicting psychosocial outcomes in brain injury research from the findings of the systematic review also being conducted by the researcher), will be compared for their performance accuracy. Logistic modelling ML analysis will use elastic net (a regularization method to minimize

overfitting) for embedded feature selection. It is likely that a five-fold nested cross-validation approach will be used whereby 80% of the data will be used for training, with the remaining 20% of the data (chosen randomly each time) reserved for model validation. The predicted outcomes from the models will be evaluated using a receiver operating characteristic (ROC) curve, with AUC being the primary metric used to determine model performance.

Justification of sample size

The analysis will include approximately 250 people who have been discharged from GAH and meet the criteria for the cohort analysis, of whom around 100 people have 6 month follow-up data available (for the QoL analysis).

Power analyses were performed on the 'pROC' package in R, based on a typical ratio of poor:good outcome of 4:1 (based on preliminary service data at GAH) for accommodation status at the time of discharge (the primary outcome). To have 80% power to detect a significant effect at $p<.05$ (two-sided), the minimum sample size required for analysis would be 29 for an AUC of 0.85 (similar to Frank et al.'s 2010 model for predicting return home after neurorehabilitation), or 61 for a weaker AUC of 0.75.

Settings and Equipment

Data cleaning and organisation will be based at GAH. The researcher and clinical team will have access to the clinical database including identifiable information during this stage. The research database will be managed in an Excel spreadsheet with anonymised data. A data protection impact assessment has been completed. The research database will be exported to a University encrypted laptop and backed up regularly on a secure University network folder in line with GDPR. Appropriate statistical software will be installed on the laptop.

Anonymised data will be held securely at the University for 10 years.

The clinical database which includes identifiable information for clinical purposes will be stored in a different location, located at GAH.

Health and Safety Issues

None identified given this is using retrospective data.

Ethical Issues

Service users at GAH provide consent at admission for their anonymised data to be used for evaluating the rehabilitation service. The researcher will require access to records within the service to organise the anonymised database of assessment and outcome data. This has been approved by the Brain Injury

Rehabilitation Trust, and BIRT Caldicott approval has been provided. The researcher will require access to patient information in the neuropsychological and outcomes databases. These are not anonymised as they are used clinically. No member of the clinical team would have capacity to organise and collate the databases to anonymise them prior to research. To gather information for the analysis, the researcher may need to extract data from clinical letters and clinical databases into a anonymised database. The neuropsychological database does not have the patient number included, therefore the collation of data from different databases is currently only identifiable by name. A separate anonymised database will be collated prior to analysis. The anonymised database will be identifiable only by participant ID number, which will not be able to be traced back to clinical information stored in the clinical database at GAH. NHS and BIRT Caldicott approval has been approved. A PVG check has been approved for working within GAH. An NHS ethics application will be required for research analysis of previously gathered patient information. A separate BIRT ethics application is not required in addition to NHS ethics in accordance to BIRT research procedures, however BIRT management/governance approval has been approved. Anonymised data will be held securely at the University for 10 years.

Results will be disseminated via publication in peer reviewed journals, presentations at research conferences, and either the BIRT newsletter or website.

Financial Issues

None identified.

Timetable

There will be assistance from honorary research assistants from October 2018 to April 2019 for data entry. Planning meetings will be regularly scheduled to monitor progress with data entry.

The researcher is currently undertaking training courses in ML and R Programming, managed in her own time.

An ethics application will be made once the proposal has been accepted, likely June 2019.

BIRT Caldicott approval was applied for in April 2019 and NHS Caldicott approval applied for in July 2019. Once this is received, I can continue with data cleaning (but no data analysis), because this is part of routine data curation in the service, while awaiting ethical approval. I aim to have data cleaning completed by September 2019.

Data analysis will commence following ethical approval and the findings from the systematic review which will inform choice of ML algorithms (likely September to October 2019). Data analysis will be completed by

April 2020, with write up and supervisor review by July 2020. Regular supervision meetings will be scheduled, including supervisor Samuel Leighton for support for ML coding using R.

Practical applications

The results of this study will help service users, their families, and social services make timely, cost-effective decisions for appropriate accommodation and support needs to be ready for discharge.

References

Agoston, D. V. & Langford, D. 2017. Big Data in traumatic brain injury; promise and challenges. *Concussion,* 2**,** CNC45.

Bellon, K., Malec, J. F. & Kolakowsky-Hayner, S. A. 2012. Mayo-portland adaptability inventory-4. *The Journal of head trauma rehabilitation,* 27**,** 314-316.

Boake, C. 1996. Supervision Rating Scale: a measure of functional outcome from brain injury. *Archives of Physical Medicine and Rehabilitation,* 77**,** 765-772.

Cullen, N., Chundamala, J., Bayley, M. & Jutai, J. 2007. The efficacy of acquired brain injury rehabilitation. *Brain Injury,* 21**,** 113-132.

Dwyer, D. B., Falkai, P. & Koutsouleris, N. 2018. Machine learning approaches for clinical psychology and psychiatry. *Annual review of clinical psychology,* 14**,** 91-118.

Euroqol Group., 2010. EQ-5D. *In:* Preedy, V. R. & Watson, R. R. (eds.) *Handbook of Disease Burdens and Quality of Life Measures.* New York, NY: Springer New York.

Feigin, V. L., Barker-Collo, S., Krishnamurthi, R., Theadom, A. & Starkey, N. 2010. Epidemiology of ischaemic stroke and traumatic brain injury. *Best Practice & Research Clinical Anaesthesiology,* 24**,** 485-494.

Frank, M., Conzelmann, M. & Engelter, S. 2010. Prediction of discharge destination after neurological rehabilitation in stroke patients. *European Neurology,* 63**,** 227-233.

Iniesta, R., Stahl, D. & Mcguffin, P. 2016. Machine learning, statistical learning and the future of biological research in psychiatry. *Psychological Medicine,* 46**,** 2455-2465.

Ioannidis, J. P. A. 2016. Why most clinical research is not useful. *PLoS medicine,* 13**,** e1002049.

Oddy, M., Coughlan, A. & Crawford, J. 2007. BIRT memory and information processing battery. *Horsham, UK: Brain Injury Research Trust*.

Oddy, M., Haye, L. & Goodson, A. 2018. Preliminary investigation of the reliability and validity of the BIRT Independent Living Scale AU - Ramos, Sara D. S. *Disability and Rehabilitation,* 40**,** 2817-2823.

Marcano-Cedeño, A., Chausa, P., García, A., Cáceres, C., Tormos, J. M. & Gómez, E. J. 2013. Artificial metaplasticity prediction model for cognitive rehabilitation outcome in acquired brain injury patients. *Artificial Intelligence in Medicine,* 58**,** 91-99.

National Institute for Health and Care Excellence (2014). Head injury: assessment and early management. Clinical guideline [CG176].

Parkin, D., Devlin, N. & Feng, Y. 2016. What Determines the Shape of an EQ-5D Index Distribution? *Medical Decision Making,* 36**,** 941-51.

Ponsford, J. L., Downing, M. G., Olver, J., Ponsford, M., Acher, R., Carty, M. & Spitz, G. 2014. Longitudinal follow-up of patients with traumatic brain injury: outcome at two, five, and ten years post-injury. *Journal of Neurotrauma,* 31**,** 64-77.

Powell, J. M., Wise, E. K., Brockway, J. A., Fraser, R., Temkin, N. & Bell, K. R. 2017. Characteristics and Concerns of Caregivers of Adults With Traumatic Brain Injury. *Journal of Head Trauma Rehabilitation,* 32**,** E33-E41.

Qadeer, A., Khalid, U., Amin, M., Murtaza, S., Khaliq, M. F. & Shoaib, M. 2017. Caregiver's Burden of the Patients With Traumatic Brain Injury. *Cureus,* 9.

Rassovsky, Y., Levi, Y., Agranov, E., Sela-Kaufman, M., Sverdlik, A. & Vakil, E. 2015. Predicting long-term outcome following traumatic brain injury (TBI). *Journal of clinical and experimental neuropsychology,* 37**,** 354-366.

Robertson, I. H., Ward, T., Ridgeway, V. & Nimmo-Smith, I. 1994. The test of everyday attention (TEA). *San Antonio, TX: Psychological Corporation*.

Safari, S., Baratloo, A., Elfil, M. & Negida, A. 2016. Part 5: Receiver Operating Characteristic Curve and Area under the Curve. *Emergency,* 4**,** 111-113.

Smith-Knapp, K. I. P., Corrigan, J. D. & Arnett, J. A. 1996. Predicting functional independence from neuropsychological tests following traumatic brain injury. *Brain Injury,* 10**,** 651-662.

Turner-Stokes, L., Williams, H., Bill, A., Bassett, P. & Sephton, K. 2016. Cost-efficiency of specialist inpatient rehabilitation for working-aged adults with complex neurological disabilities: a multicentre cohort analysis of a national clinical data set. *British Medical Journal Open,* 6**,** e010238.

Wechsler, D. 2008. *Wechsler Adult Intelligence Scale–Fourth Edition (WAIS–IV)*, San Antonio, TX: The Psychological Corporation.

Wechsler, D. 2011. *Test of premorbid functioning.* UK version (TOPF UK). UK: Pearson Corporation.

Wilson, B. A., Clare, L., Cockburn, J., Baddeley, A. D., Tate, R. & Watson, P. 1999. *The rivermead behavioural memory test-extended version.* Bury St Edmunds: Thames Valley Test Company.

Wilson, B. A., Evans, J. J., Alderman, N., Burgess, P. W. & Emslie, H. 1997. *Behavioural assessment of the dysexecutive syndrome.* Methodology of frontal and executive function*,* 239**,** 250.

Yarkoni, T. & Westfall, J. 2017. Choosing Prediction Over Explanation in Psychology: Lessons From Machine Learning. *Perspecives in Psychological Science,* 12**,** 1100-1122.

Zigmond, A. S. & Snaith, R. P. 1983. The hospital anxiety and depression scale. *Acta psychiatrica scandinavica,* 67**,** 361-370.