applied sciences

MDPI

Systematic Review

Machine Learning Approaches to Intracranial Pressure Prediction in Patients with Traumatic Brain Injury: A Systematic Review

George R. E. Bradley *, **María Roldán** and **Panayiotis A. Kyriacou**

Research Centre for Biomedical Engineering, City University of London, London EC1V 0HB, UK
* Correspondence: george.bradley@city.ac.uk

Abstract: Purpose: *Intracranial pressure* (ICP) monitoring is a "gold standard" monitoring modality for severe traumatic brain injury (TBI) patients. The capacity to predict ICP crises could further minimise the rate of secondary brain injury and improve the outcomes of TBI patients by facilitating timely intervention prior to a potential crisis. This systematic review sought (i) to identify the most efficacious approaches to the prediction of ICP crises within TBI patients, (ii) to access the clinical suitability of existing predictive models and (iii) to suggest potential areas for future research. **Methods:** Peer-reviewed primary diagnostic accuracy studies, assessing the performance of ICP crisis prediction methods within TBI patients, were included. The QUADAS-2 tool was used to evaluate the quality of the studies. **Results:** Three optimal solutions to predicting the ICP crisis were identified: a *long short-term memory* (LSTM) model, a *Gaussian processes* (GP) approach and a logistic regression model. These approaches performed with an *area under the receiver operating characteristics curve* (AUC-ROC) ranging from 0.86 to 0.95. **Conclusions:** The review highlights the existing disparity of the definition of an ICP crisis and what prediction horizon is the most clinically relevant. Moreover, this review draws attention to the existing lack of focus on the clinical intelligibility of algorithms, the measure of how algorithms improve patient care and how algorithms may raise ethical, legal or social concerns. The review was registered with the International Prospective Register of Systematic Reviews (PROSPERO) (ID: CRD42022314278).

Keywords: intracranial pressure; traumatic brain injury; brain injury; machine learning

check for updates

Citation: Bradley, G.R.E.; Roldán, M.; Kyriacou, P.A. Machine Learning Approaches to Intracranial Pressure Prediction in Patients with Traumatic Brain Injury: A Systematic Review. *Appl. Sci.* **2023**, *13*, 8015. https://doi.org/10.3390/app13148015

Academic Editors: Mohammed Abdelsamea and Mohamed Medhat Gaber

Received: 6 June 2023
Revised: 30 June 2023
Accepted: 7 July 2023
Published: 9 July 2023

1. Introduction

Traumatic brain injury (TBI) is defined as an alteration in brain function pathology by a sudden trauma causing damage to the brain. Symptoms can be mild, moderate or severe, depending on the extent of the damage to the brain [1]. The global incidence rate of TBI is estimated to be 69 million cases per annum, of which 5.48 million are estimated to be severe cases [2]. TBI causes a significant burden on individuals and their families, through health loss and disability. TBI has also created a large and growing burden on healthcare systems and nations, due to the complex and expensive medical care that the condition necessitates, and the consequent loss to productivity. The incidence rate of TBI is increasing over time, which is likely due to increases in population density, population ageing and the increasing use of motor vehicles, motorcycles and bicycles [3]. The total annual burden of TBI has been estimated at USD 400 billion [4]

Global modelling suggests that the incidence of TBI in low to middle income countries (LMICs) is significantly higher than in high income countries (HICs), the main cause of TBI in LMICs being road traffic collisions [3]. By contrast, the CENTER-TBI, a largely HIC-focused registry, collected patient demographics, injury, care pathway and acute care outcome data in 56 acute trauma receiving hospitals across 17 countries in Europe and Israel. The study reported that 56% of the 21,681 patients with TBI acquired their injury through falls, of which the majority, 71%, were ground-level (low-energy) falls [5].

The incidence rate of TBI is increasing. Kureshi et al. conducted a retrospective cohort study of all patients in Nova Scotia who presented with severe TBI between 2002 and 2018. Over the duration of the study, there were 5590 severe TBI patients, and the rate of severe TBI increased by 39%. The study's results mirrored the findings of the CENTER-TBI registry, that the mechanism of injury amongst patients was predominantly falls (45%). The rate of fall-related TBIs more than doubled between 2002 and 2017 [6]. Hsia et al.'s study analysed non-public patient-level data from California's Office of Statewide Health Planning and Development between the years 2005 and 2014. In line with the findings of Kuershi et al, the study found a 57.7% increase in the number of TBI emergency department visits, which represented a 40.5% increase in TBI related hospital visit rates over the 10-year period [7].

In this context, intracranial pressure (ICP) monitoring has become a standard practice in neurocritical care, for the timely identification of intracranial hypertension, as elevated ICP is associated with poor neuropsychological performance and functional outcomes in TBI patients. ICP is the pressure inside the cranial vault. The Monro–Kellie hypothesis states that under normal conditions the intracranial compartment space, cerebral blood volume and volume inside the cranium are fixed [8].

Changes in ICP are caused by changes in brain volume, cerebral blood volume and the production and/or clearance of cerebrospinal fluid (CSF) [8,9]. These changes can be caused by a variety of pathological processes, many of which may be caused by TBI, such as localised mass lesions, obstruction to major venous sinuses and cerebral edema or swelling. If any of these changes cause volumetric increases within the cranial vault, compensatory mechanisms occur, in an effort to maintain ICP within the normal ranges, which are generally from 10 to 15 millimetres of mercury (mmHg) for adults and from 3 to 7 mmHg for young children [10]. A critical threshold is reached when space-occupying lesions can no longer expand: without efficacious monitoring and subsequent intervention, this can lead to secondary injury to the brain through neuronal injury, herniation and brain death [11–13].

ICP monitoring is used either as a guide to treatment or as a diagnostic modality in a number of pathological conditions resulting in neurological injury. The American Brain Trauma Foundation (BTF) guidelines set indications for ICP monitoring in patients with severe TBI with a normal CT scan. ICP monitoring is indicated if two or more of the following features are noted on admission: (i) age over 40 years; (ii) unilateral or bilateral motor posturing; (iii) systolic blood pressure $<$90 mmHg. The BTF guidelines recommend treating ICP above 22 mmHg [14]. Elevated levels of ICP are referred to as intracranial hypertension (ICH).

The gold standard ICP monitoring modality is intraventricular pressure monitoring using an extraventricular drain (EVD) [15,16]. In patients with raised ICP, the desired outcome of ICP management is to (i) maintain ICP less than 22 mmHg, as per the guidelines set out by the BTF [14], and (ii) maintain adequate cerebral perfusion pressure (CPP), which, in healthy adults, is between 50 and 150 mmHg [17].

The ability to predict the onset of ICP crises levels in neurocritical care is valuable since it could enable early intervention and treatment, allowing healthcare providers to take proactive measures to prevent or minimise secondary injury to the brain and/or death caused by elevated ICP. The capacity to predict ICP crises could also help in resource optimisation, by allocating healthcare resources to at-risk patients. Ultimately, the capacity to predict ICP crises could improve patient outcomes, optimise resource allocation and enhance overall patient care in neurocritical settings.

The three main objectives of this review are (i) to identify the most efficacious machine learning approaches to the prediction of ICP crises within TBI patients, compared to invasive ICP measurements, (ii) to access the clinical suitability of existing predictive models and (iii) to suggest potential areas for future research.

2. Methods

2.1. Protocol and Registration

The review was performed in accordance with the *Preferred Reporting Items for Systematic Reviews and Meta-Analysis* (PRISMA) guidelines [18]. The study was funded by City University of London, and there were no competing interests. The review's protocol was designed with reference to the Cochrane Handbook for Systematic Reviews of Interventions [19]. The review was registered with the *International Prospective Register of Systematic Reviews* (PROSPERO) (ID: CRD42022314278).

2.2. Information Sources and Search Strategy

The reviewers (G.R.E.B, M.R) carried out a systematic search of PubMed, Scopus, Web of Science and ArXiv between 21 March 2022 and 25 March 2022. The systematic search strategy was orientated around keywords relating to traumatic brain injury, intracranial pressure, artificial intelligence and prediction. Variations of these keywords and indexing terms were used, to include all potentially eligible studies (Table A1). The results were filtered, for studies published in the last 10 years.

2.3. Selection Process and Eligibility Criteria

Once duplicate studies had been removed from the search results, the two independent reviewers (G.R.E.B, M.R) reviewed the titles and abstracts of the potentially eligible studies. Studies that fulfilled the inclusion criteria were retrieved for full-text assessment. Inclusion criteria: (i) studies of patients diagnosed with TBI; (ii) studies comparing predicted ICP and invasive ICP values; (iii) studies with accessible full text; (iv) diagnostic test accuracy studies; (v) studies published in English. Exclusion criteria: (i) reviews, conference proceedings, case reports, white papers, letters, editorials, animal and in vitro studies, case control studies, summaries, expert opinions and comments; (ii) studies without sufficient data; (iii) duplicate publications with the same dataset or non-independent studies; (iv) studies misreporting data. An independent arbiter (P.A.K) was used to resolve any disagreements that arose between the main reviewers during the full text review of the selected studies.

2.4. Data Collection Process and Data Items

The main reviewers (G.R.E.B, M.R) independently completed the data extraction. The data extraction was carried out manually, with the use of Microsoft Excel®. The following data items were extracted from the selected studies: (i) title; (ii) authors; (iii) publication year; (iv) country; (v) journal; (vi) dataset(s); (vii) ICP monitoring technique; (viii) ICP prediction model; (ix) size of training dataset; (x) size of testing dataset; (xi) definition of ICP crises; (xii) prediction horizon; (xiii) patient inclusion criteria; (xiv) patient exclusion criteria; (xv) model evaluation approach; (xvi) findings and model performance.

2.5. Study Risk of Bias Assessment

Two risk evaluation tools are being developed: *QUADAS-AI* and *PROBAST-AI* [20,21]. These tools may have been more suitable to apply to the studies in this review; however, as they were not yet released, the QUADAS-2 tool was used instead, to access and evaluate the possible risk of bias and applicability within the primary diagnostic accuracy studies selected for the full text review. The two main reviewers (G.R.E.B, M.R) independently completed the QUADAS-2 assessment [22]. The QUADAS-2 tool consists of four key domains of assessment: (i) patient selection; (ii) index test; (iii) reference standard; (iv) flow and timing. There are a series of questions pertaining to each domain, which are used to access the risk of bias and applicability. These questions are answered by one of three qualifiers: *"yes"*, *"no"* or *"unclear"*. The risk of bias and the applicability of each domain is rated as either *"low"*, *"high"* or *"unclear"*. Only studies with low risk of bias and high applicability were selected for synthesis.

2.6. Data Items and Synthesis of Results

Aggregated data were used to conduct a qualitative synthesis. The data items used for the qualitative synthesis included (i) author, (ii) publication year, (iii) invasive monitoring technique, (iv) definition of ICP crisis, (v) prediction model, (vi) prediction horizon, (vii) data window size, (viii) training dataset size, (ix) test dataset size, (x) median age, (xi) sex (m/f), (xii) median *Glasgow coma scale* (GCS) score, (xiii) median *length of stay* (LOS) and (xiv) outcome. The data items of each study were synthesised into a tabulated format with the use of Microsoft Excel®, which facilitated the comparison of study characteristics and outcomes between studies. Using the aggregated data, the study population and the size of the datasets were the main factors used to access the heterogeneity of the studies.

3. Results

3.1. Study Selection

The systematic search identified 228 studies, of which, 28 (12.28%) were duplicates and were removed. Of the remaining 200 studies, a further 9 (4.50%) were removed for being non-original or non-independent. A further 183 studies (91.50%) were removed as they did not meet the inclusion criteria. Of the remaining 8 studies selected for full text review, 5 (62.50%) were removed as they either did not meet the selection criteria or they failed the quality assessment. Figure 1 illustrates the process of identification, screening, eligibility and inclusion of papers in the review.

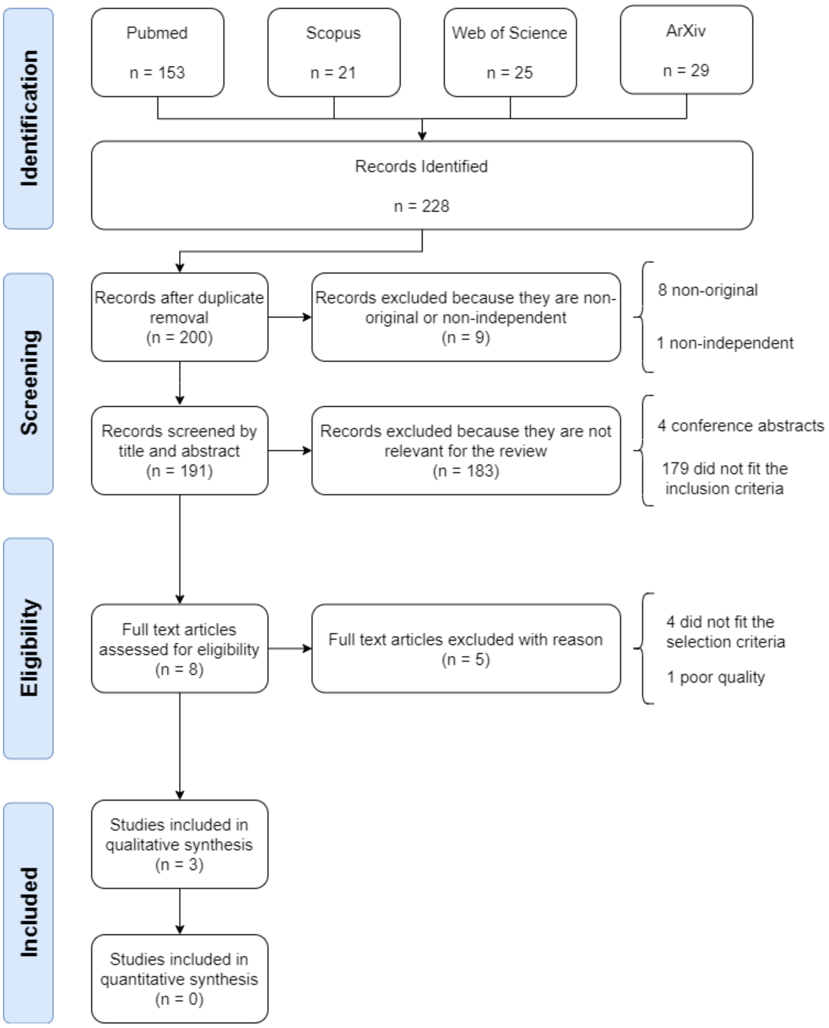
Identification
Pubmed
n = 153
Scopus
n = 21
Web of Science
n = 25
ArXiv
n = 29
Records Identified
n = 228
Screening
Records after duplicate removal
(n = 200)
Records excluded because they are non-original or non-independent
(n = 9)
8 non-original
1 non-independent
Records screened by title and abstract
(n = 191)
Records excluded because they are not relevant for the review
(n = 183)
4 conference abstracts
179 did not fit the inclusion criteria
Eligibility
Full text articles assessed for eligibility
(n = 8)
Full text articles excluded with reason
(n = 5)
4 did not fit the selection criteria
1 poor quality
Included
Studies included in qualitative synthesis
(n = 3)
Studies included in quantitative synthesis
(n = 0)

Figure 1. Flow diagram illustrating the systematic review process, including identification, screening, eligibility and inclusion stages.

3.2. Risk of Bias within Studies

Of the original 228 studies, the title and abstract of 191 papers were reviewed, and 8 were kept for full text assessment of quality and eligibility: of these, 4 studies were rejected as they did not fulfil the selection criteria, and 1 study was removed for being *"low quality"*. Both main reviewers (G.R.E.B, M.R) agreed on the quality assessments of the studies. The *"low quality"* study was judged to have a high risk of bias because (i) a case-control study was not avoided, (ii) the index test was interpreted with knowledge of the reference standard and (iii) the reference standard was interpreted with knowledge of the index test. Figure 2 depicts the results of the QUADAS-2 quality assessments.

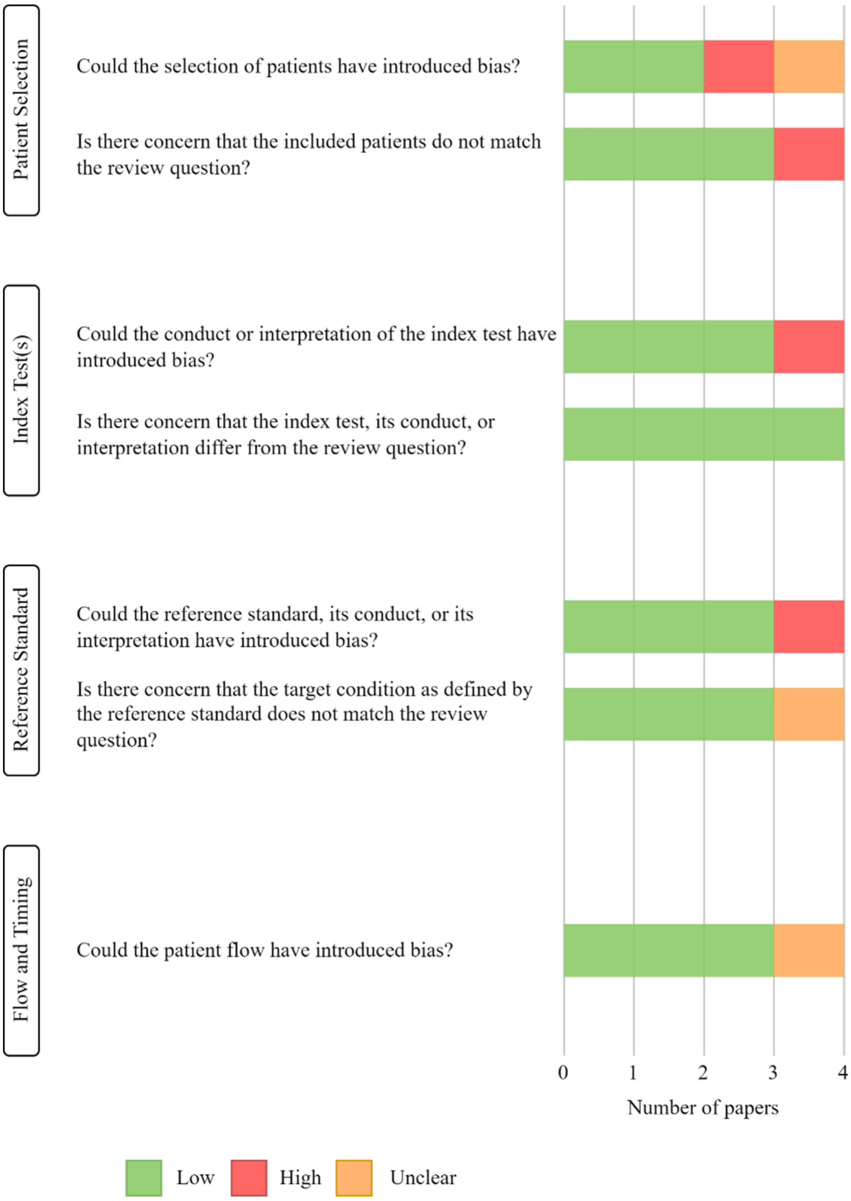

Figure 2. Results of QUADAS-2 tool assessment, displaying the quality and the risk of bias of the included studies.

3.3. Study Characteristics

A total of 2402 participants, with median ages ranging from 30 to 55, received invasive intracranial pressure monitoring [23–25]. The weighted average percentage of the participants was 69% male, 31% female. The majority had a moderate-to-severe TBI diagnosis. Table 1 depicts the bibliometric, demographic and technological characteristics of each study.

Table 1. Summary of bibliometric data of articles included in the systematic review.

Author	*Year*	*Invasive Monitoring Technique*	*Definition of ICP Crisis*	*Prediction Model*	*Prediction Horizon (min)*	*Data Window Size (min)*	*Dataset Window Size (min)*	*Dataset Size (Patients)*	*Test Dataset Size (Patients)*	*Median Age*	*Sex M/F (%)*	*Median GCS*	*GCS Value*	*Median LOS (days)*	*LOS Value*	*Outcome*	*Value*
Schweingruber et al. [23]	2022	Intraparenchymal ICP probe (Codman Microsensor, Integra LifeSciences)	Hour defined as a critical phase if at least one ICP measurement was ≥22 mmHg	LSTM	120	Arbitary	1346	1077	269	55	59.9/40.1	NR	NR	LOS	13	AUC-ROC AUC-PR AUC-ROC (Mimic dataset) AUC-ROC (eICU dataset)	0.95 0.71 0.948 0.903
Guiza et al. [24]	2013	NR	ICP measurement ≥30 mmHg for ≥10 min	Gaussian processes	30	240	264	178	61	31	80.9/19.1	GCS eye GCS motor GCS verbal GCS total	1 4 1 7	NR	NR	AUC-ROC Hosmer Lemeshow p value Calibration in the large Calibration slope Brier score Brier score scaled (%) Accuracy (%) Sensitivity (%) Specificity (%)	0.872 0.12 −0.019 1.02 0.137 39.4 77.4 81.6 75.2
Myers et al. [25]	2016	Extraventricular drain or an intraparenchymal fiberoptic probe	ICP measurement ≥20 mmHg for ≥15 min	Logistic regression	30	30	874	368	261	30	79.5/20.5	GCS eye GCS motor	7 5	ICU LOS Hospital LOS	16.8 24	AUC-ROC	0.86

Abbreviations: long short-term memory (LSTM); not reported (NR); area under the receiver operating characteristics (AUC-ROC); area under the precision–recall curve (AUC-PR); Glasgow coma scale (GCS); length of stay (LOS).

3.4. Results of Individual Studies

Schweingruber et al. [23] used a *long short-term memory* (LSTM) network architecture [26] to predict the onset of critical ICP *"phases"*. The authors defined critical ICP "phases" as durations above an ICP threshold. An hour was defined as a critical phase if at least one ICP measurement was $\geq$22 mmHg. A duration of two sequential critical hours was defined as a short critical phase. A duration of more than two sequential critical hours was defined as a long critical phase. A primary dataset of 1346 patients with invasive ICP monitoring was used for the training and testing of the predictive model. The median patient age in this dataset was 54.8. The dataset comprised 59.9% and 40.1% male and female patients, respectively. The main diagnosis was TBI. The LSTM model was trained using 80% of the data, and was tested on the additional 20%. In addition to this primary dataset, the authors used the eICU and MIMIC datasets [27,28] for external validation of their model. Schweingruber et al. focused their analysis on the two-hour prediction horizon, as it was deemed the most clinically relevant. The model performed with an *area under the receiver operating characteristics curve* (AUC-ROC) of 0.95 (±0.0009) on the testing dataset when predicting long (>2 h) and short ($\leq$2 h) critical phases. The model returned an *area under the precision–recall curve* (AUC-PR) of 0.71 (±0.0067) for the prediction of long critical phases. Similar results were reported in the external validation datasets, with an AUC-ROC of 0.948 (±0.0025) in the MIMIC dataset, and an AUC-ROC of 0.903 (±0.0033) in the eICU dataset. The authors suggested that the most important features for predicting long critical phases were ICP, mean arterial pressure (MAP) and *cerebral perfusion pressure* (CPP). CPP was calculated using ICP and MAP, using the equation CPP = MAP − ICP. Consequently, there may have been a collinearity issue in the interpretation of the three most significant features presented by the authors; however, given that the core aim of the developed model was prediction rather than interpretation, having correlated prediction variables did not inhibit prediction performance.

Guiza et al. [24] developed a *Gaussian processes* (GP) model to predict ICP crises. The authors defined an ICP crisis as an event where ICP $>$ 30 mmHg lasted for $\geq$10 min. The BrainIT dataset [29] was used in the development of this predictive model, which contained 264 TBI patients admitted to 22 neuro-ICUs in 11 European countries. The median age of the patients was 31. The dataset comprised 80% male patients and 20% female patients. The median total GCS score of patients within the dataset was 7. A brain injury was defined as severe if the GCS score was within the range 3–8. Minute-by-minute ICP and MAP of 239 patients was recorded for a minimum of four consecutive hours. The training dataset comprised data from 178 patients, and the remaining 61 patients made up the testing dataset. Güiza et al. judged a 30 min prediction horizon to be clinically appropriate, suggesting that 30 min offers sufficient time to allow for therapeutic intervention to reduce ICP. The authors reported a GP with an AUC-ROC of 0.872 on the testing dataset, with an accuracy of 77.4%, sensitivity of 81.6% and a specificity of 75.2% as their optimal prediction model. The study used 4 h of data to make predictions. Features used as input to the prediction model were derived from the analysis of the 4 h data window of ICP and MAP data along with ICU LOS at the time of prediction. Güiza et al. reported that the features with the most predictive power were within the ICP signal, with the most recent measurements being more relevant.

Myers et al. [25] developed and studied three models for the prediction of ICP crises: (i) GP; (ii) logistic regression and (iii) *autoregressive ordinal regression* (AR-OR). The authors defined ICP crises as an event where ICP was $\geq$20 mmHg for $\geq$15 min. They constructed and used a dataset of 817 patients with severe TBI diagnoses. The patients of the dataset had a median age of 30; 85% were male, and 15% were female. The patients constituting the dataset had a median GCS eye score and a GCS motor score of 7 and 5, respectively. Patient ICP, MAP, *end-tidal* CO2 (ETco2), *oxygen saturation and brain tissue oxygenation* (PbtO2) levels were recorded every 36 s. The study's dataset was collected over a period of 24 years, between 1989 and 2000 and between 2006 and 2013. The dataset was divided into three cohorts: (i) a *"study cohort"* (collected between 1989 and 1996), used to develop candidate predictive models; (ii) a *"selection cohort"* (collected between 1996 and 2000), used to evaluate the performance of the candidate predictive models; (iii) a *"validation cohort"* (collected between 2006 and 2013), used to test the performance of the best-performing models. A 30 min window of physiological data was used along with the time since the last crisis to predict impending crisis events, with prediction horizons ranging from 15 to 360 min. Myers et al. suggested that a shorter prediction horizon of 30 min was clinically more appropriate. The predictive model that returned the maximal AUC-ROC reported was a five-feature logistic regression using ICP measurements and the time since the last crisis, which was able to predict an ICP crisis with a 30 min prediction horizon, with an AUC-ROC of 0.83 on the *"validation cohort"*. The authors reported that the two physiological signals most associated with precursors to ICP crises were ICP and change in ICP.

There is significant methodological heterogeneity between the three studies, which renders a meta-analysis inappropriate. The studies presented three different solutions for predicting ICP crises. There was significant disparity in the sizes of the datasets used to train and test the models described in the studies. Although the main diagnosis in the Schweingruber et al. study was TBI (33.9%), the remaining diagnoses were not [*intracerebral haemorrhage, 21.0%; stroke, 16.1%; MISC, 13.5%; subarachnoid haemorrhage, 11.1%; tumour, 4.4%*], unlike the two other studies.

3.5. Assessing the Clinical Suitability of ICP-Predictive Models

An objective of this review was to assess the clinical suitability of the three models used in the respective studies. The two main reviewers (G.R.E.B, M.R) employed the use of the checklist developed by Scott et al. [30] for assessing the suitability of machine learning applications in healthcare and applied it to the three studies reviewed. Table 2 details the results of Scott et al.'s clinical suitability checklist.

The results of the clinical suitability assessment assert areas of possible shortcomings within the existing literature, which include (i) lack of pre-specification and rationale of the amount of data needed to train a model, (ii) lack of external validation of developed models, (iii) lack of findings or focus on clinical intelligibility and (iv) lack of measurements describing possible ways in which algorithms can potentially harm patients or raise ethical, legal or social concerns.

3.6. Risk of Bias Across Studies

There was significant heterogeneity introduced into the studies by the different sizes of the datasets used to train and test the models described above. In Schweingruber et al., the training data were obtained from 1077 patients—143.27% larger than the training dataset used in Güiza et al.'s study (178 patients) and 98.13% larger than the training dataset used by Myers et al. (368 patients). The Schweingruber et al. study returned an AUC-ROC of 0.95, which was 8.56% and 9.95% greater than the respective approaches presented by Güiza et al. and Myers et al.

The performance of Schweingruber et al.'s approach may have been influenced by data bias. It seems reasonable to suggest that the significantly larger training dataset used by Schweingruber et al., compared to the other two studies, allowed for the greater generalisation of their developed model, which may have contributed to the higher AUC-ROC value. An additional factor that may have affected the results was the diagnoses of the patients who comprised Schweingruber et al.'s dataset. Unlike the other two studies, whose datasets were solely TBI patients, in the Schweingruber et al. study, TBI was the main diagnosis (33.9%) [*intracerebral haemorrhage, 21.0%; stroke, 16.1%; MISC, 13.5%; subarachnoid haemorrhage, 11.1%; tumour, 4.4%*].

Table 2. Clinical checklist for assessing suitability of machine learning applications in healthcare.

Item	Response
1. What is the purpose of the algorithm?	The objective and context of the algorithm were adequately stated in the included studies.
2a. How good were the data used to train the algorithm? 2b. To what extent were the data accurate and free of bias? 2c. Were the data standardised and interoperable?	All three studies (100%) reported inclusion criteria. One study (33.3%) reported exclusion criteria. All three studies (100%) reported age. All three studies (100%) reported sex. The Glasgow coma scale (GCS) score for patients was reported in two of the studies (66.7%). Two of the studies (66.7%) used prospectively collected data. Although all three studies (100%) reported the extent of missing data, how missing data was handled was poorly reported in all three studies. All the data used in two of the studies (66.7%) were obtained at a single hospital or medical centre. The data used in one of the studies (33.3%) was collected from multiple hospitals or medical centres.
3. Were there sufficient data to train the algorithm?	All three studies (100%) reported the number of patients in the training dataset. None of the studies prespecified a sample size.
4. How well does the algorithm perform?	Two of the studies (66.7%) used resampling methods. One of the studies used 10-fold cross-validation. One of the studies performed 10 iterations of 5-fold cross-validation. Across all three of the studies (100%), the reported AUC-ROC ranged from 0.83 to 0.95. One of the studies (33.3%) performed external validation.
5. Is the algorithm transferable to new clinical settings?	None of the studies (0%) assessed algorithm performance in a real-world context.
6. Are the outputs of the algorithm clinically intelligible?	Three of the studies (100%) used machine learning models. Only one of the studies (33.3%) provided a heat map of feature importance.
7. How will this algorithm fit into and complement current workflow?	None of the studies reported how their algorithms impacted real-world clinical workflows, although all of the studies described how their algorithms could be used to positive effect within a clinical setting.
8. Has use of the algorithm been shown to improve patient care and outcomes?	None of the algorithms in these studies were subjected to clinical trials aimed at demonstrating improved care or patient outcomes.
9. Could the algorithm cause patient harm?	No comments were made about potential harm.
10. Does use of the algorithm raise ethical, legal or social concerns?	No comments were made about any such concerns.

There could be a risk of measurement bias within the study by Myers et al. as the patient data were collected over a period of 24 years, between 1989 and 2000 and between 2006 and 2013. It seems reasonable to suggest that over such a period there may have been changes in patient care practices that influenced the integrity of the model. However, Myers et al. attempted to reconcile this by training their model using the oldest of the recorded data (collected between 1989 and 1996) and testing their model on the most recent data (collected between 2006 and 2013). It appears that the authors assumed that if their model trained on the oldest data performed well when tested on the newest data, this was evidence of the model's robustness against changes in care management.

4. Discussion

4.1. Summary of Evidence

The review identifies logistic regression, Gaussian process (GP) and long short-term memory (LSTM) models as the current leading machine learning techniques for predicting ICP crises in TBI patients. Notably, each of the three studies presented a different solution, with the most recent paper by Schweingruber et al. demonstrating superior performance using an LSTM model architecture, achieving an AUC-ROC of 0.95. This result outperforms the GP approach utilized by Güiza et al. by 8.56% and the logistic regression model described by Myers et al. by 9.95%. These findings emphasize the potential of advanced neural network models, such as LSTM, for enhancing the accuracy of ICP crisis prediction.

This systematic review suggests that the logistic regression, GP and LSTM models are the current leading machine learning techniques for predicting ICP crises in TBI patients. Each of the three studies presents a different solution. Of the different approaches, the most recent paper by Schweingruber et al. performed best, using an LSTM model architecture that returned an AUC-ROC of 0.95, which was 8.56% greater than the GP approach used by Güiza et al. and 9.95% greater than the logistic regression model described by Myers et al.

However, the review also draws attention to the lack of consistency in defining an ICP crisis across the three studies. Each study employed different thresholds of ICP, and varying durations above a defined threshold, to classify a crisis. This discrepancy raises concerns regarding the comparability of results and the establishment of standardized criteria for ICP crisis prediction. To address this issue, future research should prioritize consensus building efforts to define an ICP crisis uniformly, ensuring robust and reproducible findings.

Furthermore, the review highlights the existence of some agreement regarding the clinically relevant prediction horizon. Two of the studies proposed a prediction horizon of 30 min, while the third suggested a two-hour horizon. Our review not only identifies partial alignment among the studies, regarding the clinically relevant prediction horizon for ICP prediction in TBI patients, but also emphasizes the need for further clarity in determining the optimal prediction horizon with clinical relevance. To establish consensus guidelines for the appropriate prediction period, various factors should be considered, such as the nature of the injury, patient characteristics, available healthcare professionals and treatment interventions.

The nature of the injury and patient characteristics play a role in defining a suitable prediction horizon for predicting an ICP crisis. TBI encompasses a wide spectrum of severity and complexity, ranging from mild to severe cases. The extent of brain damage, the presence of associated injuries and individual patient factors can influence the progression and severity of the injury. In cases of severe TBI, where the risk of an ICP crisis is higher, a shorter prediction horizon, such as 30 min, may be more appropriate, to enable timely interventions and prevent adverse outcomes.

The available healthcare resources and the context of a busy clinical setting are important considerations. In such settings, healthcare professionals often face resource constraints and time limitations. Allocating resources effectively, to address critical patient needs, is crucial for optimal patient care. A prediction horizon of two hours, as suggested by one of the studies, may potentially strain already stretched healthcare resources, as continuous monitoring and interventions over an extended period may not be feasible.

Thus, it seems reasonable to suggest that a shorter prediction horizon, such as 30 min, could be more suitable in busy clinical settings with limited resources. This duration allows healthcare professionals to promptly respond to impending ICP crises while optimising resource utilization. However, the determination of the most appropriate prediction horizon should be a collaborative effort involving clinicians, researchers and relevant stakeholders, to ensure that it considers both the clinical needs and the practical realities of healthcare delivery.

In addition to assessing the methods and outcomes of the reviewed studies, an additional analysis was conducted to evaluate the clinical relevance of the different models. A significant concern arises from the lack of emphasis on clinical intelligibility in research efforts. Considering the potential value it holds, it is reasonable to suggest machine learning will play a larger role in healthcare. As a result, establishing trust through intelligibility in the initial phase (we suggest the idea that intelligibility becomes of less importance as confidence increases, similarly to how we rarely question the directions generated by maps on our phone) between machine learning products and healthcare professionals becomes crucial, in order to encourage adoption and leverage possible potential benefits. Within this context, it is crucial to consider the trade-off between the slightly lower performance yet higher interpretability offered by white box models versus the potential for higher performance but reduced interpretability associated with black box models. This trade-off is evident in this review, where Shweingruber et al. employed an LSTM model, considered a black box model, while Myers et al. utilized a white box approach using Logistic Regression.

White box models refer to models that provide a clear understanding of how they arrive at their predictions. These models are often based on well-defined rules that can be easily interpreted by humans. In the case of predicting ICP crises, white box models may offer valuable insights into the decision making process, allowing clinicians to understand the variables that contribute most significantly to the predictions. The greater interpretability can cultivate confidence, aiding in the collaboration between clinicians and machine learning tools.

Conversely, black box models may provide greater prediction accuracy; however, their internal workings are often opaque, making it challenging to understand why a particular prediction was made. Given this lack of interpretability, in a clinical setting —where decisions hold vital implications for patients—clinicians may hesitate to adopt tools leveraging black box models if the reasoning behind the predictions is unclear.

It is therefore essential to strike a balance between model performance and interpretability in a clinical setting. White box models can offer a clearer understanding of the decision making process, but they may sacrifice some predictive performance compared to black box models. While the reviewed studies showcased the effectiveness of machine learning algorithms in predicting ICP crises, there is a paucity of investigation into the interpretability of these algorithms. Ensuring clinical intelligibility should be a priority in research endeavours, to maximise trust and increase the possibility of effective use in real-world scenarios.

4.2. Limitations

No attempt was made to identify or translate non-English-language publications, which may have limited the inclusion of some relevant studies. There is a possibility of publication bias, as the review included only studies published since 2012 and those which focused on TBI patients, which could act as a possible restriction on the number of studies reviewed. The variability in the sizes of the datasets and approaches used to train and test the models, together with the fact that one of the studies contained other diagnoses alongside TBI, may have affected the outcomes of the review. The literature has shown the prevalence and effect of TBI on patients and on society. As the studies in this review describe, it seems reasonable to suggest that the ability to successfully predict ICP crises in TBI patients and subsequent timely interventions from healthcare professionals may enhance patient outcomes. Although this review identified GP, logistic regression and

LSTM machine learning models as the most effective techniques for predicting ICP in TBI patients, further research is required, to collect more data for the use of similar predictive models in a clinical setting. The heterogeneity of the methodologies of the different studies and predictive models did not allow for the undertaking of a meta-analysis and, thereby, for a definitive conclusion.

5. Conclusions

This review provides insights into the current state of machine learning approaches to ICP prediction in patients with TBI. While the reviewed studies demonstrated accuracy in predicting ICP crises, we identified several areas for future research, to advance the field and maximise the clinical applicability of prediction models. Further research efforts should focus on addressing three key aspects:

1. Definition of ICP crises: The absence of a universally accepted definition for ICP crises poses a challenge to achieving consistency and comparability across studies. Establishing a standardized definition that encompasses specific thresholds and durations of elevated ICP would facilitate meaningful comparisons of prediction models. Future research should prioritize the establishment of a consensus definition, through expert panels and collaborative efforts involving multiple research institutions;
2. Optimal Prediction Horizon: Determining the optimal prediction horizon is essential for timely interventions and improved patient outcomes. Our review shows divergent views on the ideal time window for predicting ICP crises, with variations ranging from 30 min to 2 h. Resolving this discrepancy and establishing a widely accepted prediction horizon would guide researchers and clinicians in the development and application of prediction models. We suggest that 2 h is too long a time window, as it would potentially consume too many scarce healthcare resources. Studies that involve healthcare professionals with experience caring for TBI patients could provide crucial insights into the most clinically relevant prediction horizon;
3. Model Intelligibility: While the reviewed studies focused primarily on the performance aspects of machine learning algorithms, there is a need for ethical considerations and the impact on patients to be given increased attention. Research should prioritize the development of clinically interpretable models that provide transparency and intelligibility to patients and healthcare professionals. An evaluation of potential challenges—including ethical, legal and social implications—is needed, to ensure responsible and accountable deployment of machine learning approaches in clinical practice.

While machine learning methods show promise in predicting ICP crises in TBI patients, this systematic review highlights the need for further research, to address critical gaps in the field. Establishing a consensus definition of ICP crises, determining the optimal prediction horizon and incorporating ethical considerations will strengthen the clinical relevance and applicability of prediction models, thus enhancing patient care and contributing to improved outcomes in the management of TBI.

Author Contributions: Conceptualization, G.R.E.B. and P.A.K.; methodology, G.R.E.B. and M.R.; formal analysis, G.R.E.B. and M.R.; writing—original draft preparation, G.R.E.B.; writing—review and editing, G.R.E.B. and P.A.K.; visualisation, G.R.E.B. All authors have read and agreed to the published version of the manuscript.

Funding: The study was financially supported by the George Daniel Doctoral Studentship at City University of London.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

Table A1. Table of keywords used in the systematic search of articles.

Systematic Search Keywords			
Traumatic Brain Injury	**Intracranial Pressure**	**Machine Learning**	**Prediction**
traumatic brain injuries	Intracranial Pressure	Artificial Intelligence	Prediction
brain injury	intra-cranial pressure	machine learning	predicting
brain injuries	ICP	ML	predict
brain trauma	intracranial hypertension	algorithm	forecast
traumatic encephalopathy	intra-cranial hypertension	algorithms	forecasting
traumatic brain injury	intracranial hypotension	AI	early warning
encephalopathy	intra-cranial hypotension	deep learning	
TBI		supervised machine learning	
TBIs		unsupervised machine learning	
		supervised learning	
		unsupervised learning	
		multi-layer perceptron	
		MLP	
		artificial neural network	
		ANN	
		neural network	
		NNs	
		NN	
		computational	
		computational approaches	
		mathematical concepts	
		statistical modelling	

References

1. Brazinova, A.; Rehorcikova, V.; Taylor, M.S.; Buckova, V.; Majdan, M.; Psota, M.; Peeters, W.; Feigin, V.; Theadom, A.; Holkovic, L.; et al. Epidemiology of Traumatic Brain Injury in Europe: A Living Systematic Review. *J. Neurotrauma* **2018**, *38*, 1411–1440. [CrossRef]
2. Dewan, M.C.; Rattani, A.; Gupta, S.; Baticulon, R.E.; Hung, Y.C.; Punchak, M.; Agrawal, A.; Adeleye, A.O.; Shrime, M.G.; Rubiano, A.M.; et al. Estimating the global incidence of traumatic brain injury. *J. Neurosurg.* **2018**, *130*, 1080–1097. [CrossRef]
3. James, S.L.; Theadom, A.; Ellenbogen, R.G.; Bannick, M.S.; Mountjoy-Venning, W.; Lucchesi, W.C.; Lydia, R.D.; Karch, A. Global, regional, and national burden of traumatic brain injury and spinal cord injury, 1990–2016: A systematic analysis for the Global Burden of Disease Study 2016. *Lancet Neurol.* **2018**, *18*, 56–87. [CrossRef] [PubMed]
4. van Dijck, J.T.J.M.; Dijkman, M.D.; Ophuis, R.H.; de Ruiter, G.C.W.; Peul, W.C.; Polinder, S. In-hospital costs after severe traumatic brain injury: A systematic review and quality assessment. *PLoS ONE* **2019**, *14*, e0216743.
5. Lecky, F.E.; Otesile, O.; Marincowitz, C.; Majdan, M.; Nieboer, D.; Lingsma, H.F.; Maegele, M.; Citerio, G.; Stocchetti, N.; Steyerberg, E.W.; et al. The burden of traumatic brain injury from low-energy falls among patients from 18 countries in the CENTER-TBI Registry: A comparative cohort study. *PLoS Med.* **2021**, *18*, e1003761. [CrossRef] [PubMed]
6. Kureshi, N.; Erdogan, M.; Thibault-Halman, G.; Fenerty, L.; Green, R.S.; Clarke, D.B. Long-Term Trends in the Epidemiology of Major Traumatic Brain Injury. *J. Community Health* **2021**, *46*, 1197–1203. [CrossRef]
7. Hsia, R.Y.; Markowitz, A.J.; Lin, F.; Guo, J.; Madhok, D.Y.; Manley, G.T. Ten-year trends in traumatic brain injury: A retrospective cohort study of California emergency department and hospital revisits and readmissions. *BMJ Open* **2018**, *8*, e022297. [CrossRef] [PubMed]
8. Mokri, B. The Monro-Kellie hypothesis: Applications in CSF volume depletion. *Neurology* **2001**, *56*, 1746–1748. [CrossRef]
9. Ghajar, J. Traumatic brain injury. *Lancet* **2000**, *356*, 923–929. [CrossRef]

10. Rangel-Castilla, L.; Gopinath, S.; Robertson, C.S. Management of intracranial hypertension. *Neurol. Clin.* **2008**, *26*, 521–541. [CrossRef]
11. Stevens, R.D.; Shoykhet, M.; Cadena, R. Emergency Neurological Life Support: Intracranial Hypertension and Herniation. *Neurocrit. Care* **2015**, *23* (Suppl. S2), S76–S82. [CrossRef] [PubMed]
12. Adams, C.A.; Stein, D.M.; Morrison, J.J.; Scalea, T.M. Does intracranial pressure management hurt more than it helps in traumatic brain injury? *Trauma Surg. Acute Care Open* **2018**, *3*, e000142. [CrossRef] [PubMed]
13. Engel, D.C.; Mikocka-Walus, A.; Cameron, P.A.; Maegele, M. Pre-hospital and in-hospital parameters and outcomes in patients with traumatic brain injury: A comparison between German and Australian trauma registries. *Injury* **2010**, *41*, 901–906. [CrossRef] [PubMed]
14. Carney, N.; Totten, A.M.; O'Reilly, C.; Ullman, J.S.; Hawryluk, G.W.J.; Bell, M.J.; Bratton, S.L.; Chesnut, R.; Harris, O.A.; Kissoon, N.; et al. Guidelines for the Management of Severe Traumatic Brain Injury, Fourth Edition. *Neurosurgery* **2017**, *80*, 6–15. [CrossRef] [PubMed]
15. Guillaume, J.; Janny, P. Continuous intracranial manometry; importance of the method and first results. *Rev. Neurol.* **1951**, *84*, 131–142. [PubMed]
16. Lundberg, N. Continuous recording and control of ventricular fluid pressure in neurosurgical practice. *Acta Psychiatr. Scand. Suppl.* **1960**, *36*, 1–193. [CrossRef] [PubMed]
17. Armstead, W.M. Cerebral Blood Flow Autoregulation and Dysautoregulation. *Anesthesiol. Clin.* **2016**, *34*, 465–477. [CrossRef] [PubMed]
18. Moher, D.; Liberati, A.; Tetzlaff, J.; Altman, D.G.; PRISMA Group. Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *Int. J. Surg.* **2010**, *8*, 336–341. [CrossRef]
19. Higgins, J.P.; Thomas, J.; Chandler, J.; Cumpston, M.; Li, T.; Page, M.J.; Welch, V.A. *Cochrane Handbook for Systematic Reviews of Interventions*; John Wiley & Sons.: New York, NY, USA, 2019.
20. Bajaj, S.S.; Martin, A.F.; Stanford, F.C. Health-based civic engagement is a professional responsibility. *Nat. Med.* **2021**, *27*, 1661–1663. [CrossRef]
21. Collins, G.S.; Dhiman, P.; Andaur Navarro, C.L.; Ma, J.; Hooft, L.; Reitsma, J.B.; Logullo, P.; Beam, A.L.; Peng, L.; Van Calster, B.; et al. Protocol for development of a reporting guideline (TRIPOD-AI) and risk of bias tool (PROBAST-AI) for diagnostic and prognostic prediction model studies based on artificial intelligence. *BMJ Open* **2021**, *11*, e048008. [CrossRef]
22. Whiting, P.F.; Rutjes, A.W.S.; Westwood, M.E.; Mallett, S.; Deeks, J.J.; Reitsma, J.B.; Leeflang, M.M.G.; Sterne, J.A.C.; Bossuyt, P.M.M.; QUADAS-2 Group. QUADAS-2: A revised tool for the quality assessment of diagnostic accuracy studies. *Ann. Intern. Med.* **2011**, *155*, 529–536. [CrossRef] [PubMed]
23. Schweingruber, N.; Mader, M.M.D.; Wiehe, A.; Röder, F.; Göttsche, J.; Kluge, S.; Westphal, M.; Czorlich, P.; Gerloff, C. A recurrent machine learning model predicts intracranial hypertension in neurointensive care patients. *Brain* **2022**, *145*, 2910–2919. [CrossRef]
24. Güiza, F.; Depreitere, B.; Piper, I.; Van den Berghe, G.; Meyfroidt, G. Novel methods to predict increased intracranial pressure during intensive care and long-term neurologic outcome after traumatic brain injury: Development and validation in a multicenter dataset. *Crit. Care Med.* **2013**, *41*, 554–564. [CrossRef] [PubMed]
25. Myers, R.B.; Lazaridis, C.; Jermaine, C.M.; Robertson, C.S.; Rusin, C.G. Predicting Intracranial Pressure and Brain Tissue Oxygen Crises in Patients With Severe Traumatic Brain Injury. *Crit. Care Med.* **2016**, *44*, 1754–1761. [CrossRef] [PubMed]
26. Hochreiter, S.; Schmidhuber, J. Long Short-Term Memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef]
27. Pollard, T.J.; Johnson, A.E.W.; Raffa, J.D.; Celi, L.A.; Mark, R.G.; Badawi, O. The eICU Collaborative Research Database, a freely available multi-center database for critical care research. *Sci. Data* **2018**, *5*, 180178. [CrossRef] [PubMed]
28. Goldberger, A.L.; Amaral, L.A.; Glass, L.; Hausdorff, J.M.; Ivanov, P.C.; Mark, R.G.; Mietus, J.E.; Moody, G.B.; Peng, C.K.; Stanley, H.E. PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. *Circulation* **2000**, *101*, E215–E220. [CrossRef]
29. Piper, I.; Citerio, G.; Chambers, I.; Contant, C.; Enblad, P.; Fiddes, H.; Howells, T.; Kiening, K.; Nilsson, P.; Yau, Y.H.; et al. The BrainIT group: Concept and core dataset definition. *Acta Neurochir.* **2003**, *145*, 615–628; discussion 628–629. [CrossRef]
30. Scott, I.; Carter, S.; Coiera, E. Clinician checklist for assessing suitability of machine learning applications in healthcare. *BMJ Health Care Inform.* **2021**, *28*, e100251. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.