Layman’s summary

There is a small chance of rupture of an cerebral aneurysm, but when it does it causes a severe stroke (aneurysmal subarachnoid hemorrhage) that generally high mortality and morbidity. Treatment of cerebral aneurysms is relative succesfull however also not without risks so when they are detected by doctors it is important to estimate the risk of rupture of these aneurysms to determine so that they can determine if it is necessary to surgically treat them. An increase in models that predict the risk of rupture has taken place recently but the clinical usability of these models is still low however prediction model for outcome after aSAH are more common. In this review studies in which these models are developed were selected and assessed for potential methodological biases using the PROBAST tool. Results from this analyses show that biases are mostly introduced during data selection by combining data sets and during the analysis part. Most studies also did not publish the data or source code that was used which could lead to replication problems.

Using the PROBAST tool to identify potential methodological biases in studies developing prediction models for the outcome of aneurysmal subarachnoid hemorrhage.

L. Edwards, supervisor: dr. Y.M. Ruigrok, daily supervisor: MSc. J. Kanning

Abstract

Predicting the outcome of aneurysmal subarachnoid hemorrhage could play an important role in the management of aneurysms. As developing prediction models becomes easier an uptake in model development is taking place in science. This also increases the amount of methodological errors made during development leading to difficulty in reproduction or clinical usage of these models. In this review papers that developed a model predicting the outcome of aneurysmal subarachnoid hemorrhage were selected from PubMed. These papers were then assessed using the PROBAST tool to determine potential methodological biases that were introduce during their development.

Introduction

Aneurysmal subarachnoid hemorrhage (aSAH) is a complication of a bleeding inside the brain due to rupture of one or multiple aneurysm(s) where blood is leaking into the subarachnoid space. This type of stroke represents 80% of the non traumatic subarachnoid hemorrhage cases (1,2). Although the incident rate declined over time from 10 to 6 cases per 100.000 person from the 1980 to 2010, aSAH still accounts for many lost life years of relatively young patients as it is generally characterized by high case mortality as approximately one third of patients die within weeks after occurrence and most surviving patients experience long lasting disability such as cognitive failure, anxiety, depression and sleep problems (3,4). Due to better preventive management strategies such as coiling and clipping of aneurysms in recent years case fatality decreased by 17% (4). As these management strategies still induce risk for the patients, assessing the risk of rupture for aneurysms could help improving aneurysm management. Risk factors often associated with rupture of intracranial aneurysms include smoking, high blood pressure, a family history of aSAH, ethnicity, excessive alcohol usage, diabetes and high age (5). These properties associated with rupture can then be used to develop models that assess risk of rupture for patients. In a similar way relationships can be inferred between properties and the outcome of rupture to make predictions about the outcome of aSAH. Several models that predict outcome after aSAH have been developed (6).

Predicting the outcome of an event from a range of predictors can be done by prediction models which can be trained on data sets and have been used in highly diverse fields (7). The use of machine learning in health science to infer relationships in data sets is increasing rapidly (8). These prediction models are derived from indicators that have a statistical relationship with the clinical outcome and can be assessed during intake of a patient (9). For different type of diseases these models are often not used clinically due to lack of external validation or poor performance of these models on new cohorts as they were developed using methodological errors by the researchers (10). Outside of health science question arise about the application of predictive models where researchers apply these techniques without properly understanding the limitations of the developed model. This problem is also compared to the replication crisis that affected social and medical sciences (11).

Besides following the correct methodological practices when applying machine learning techniques, reproducibility of the developed model is important to mitigate the ongoing replication crisis in science that arose due to difficulty in reproducing many studies (12). This includes availability of data and source code so that the exact methodology can be followed or external data sets can be tested (12). Models developed in the machine learning field of health science scored worse in this regard compared to other fields in machine learning such as language processing, computer vision and general machine learning (13). More generally machine learning platforms do not provide sufficient

tools to easily create reproducibility for research as developed models using the same data produce highly different results between different machine learning packages (14).

Individual studies or developed models are important to understand new relationships between predictors and the outcome of an disease. However these are then only applicable to small cohorts of patients for example one or a small number of hospitals. Systematic reviews can hereby play an important role in health science and the development of clinical procedures (15). They can provide reliable evidence for the effects of a procedure by combining multiple studies (16). Systematic reviews of prediction models is a relatively new and developing area (17). These type of reviews is needed as the number of performed studies that either develop a prediction models or externally validate a developed model keep increasing. It is not yet known if this increase could also lead to an increase of studies that do not follow methodological standards or where procedures are not fully reported. This signals the need for systematic reviews testing for these errors due to researchers not being familiar with machine learning procedures (18). To asses these type of methodological errors the PROBAST tool was developed using the Delphi method (19). The PROBAST tool aims to help find the risk of biases in four domains (participants, predictors, outcome and analysis) that are introduced in the development of prediction models (20).

In this systematic review studies that developed a prediction model for the outcome of aSAH were selected from PubMed. The selected articles were assessed using the PROBAST tool to find possible risk of biases introduced during the development of the model. An extra domain was added to asses reproducibility of each study.

Methods

Study design

To analyze the quality of methodology followed in models developed to predict the outcome aSAH the PROBAST tool was used. This tool is intended to be used to assess the risk of bias in multivariate prediction models. The PubMed database was used to search for papers published before September 13, 2022. The PROBAST tool was extended by three questions to assess reproducibility of these studies (Table S2).

Data sources and strategy

A search on the PubMed database was performed using the following keyword selection: predictive AND models AND subarachnoid AND hemorrhage OR haemorrhage (Figure 1). For each search result the title, abstract and keywords were used to assess the relevance of said paper. Cited sources in each selected paper were scanned to select papers that were not included in the search query.

Inclusion and exclusion criteria

Included in this review are papers for which the full text was available and when necessary supplemental methods. Only papers were selected that developed a model to predict outcome of aSAH using machine learning algorithms (Figure 1). In this study machine learning models were defined as models that have the ability to either detect or abstract relationships in the data (21). Examples of models that meet this definition are: logistic regression, tree based algorithms and neural networks. The journal in which these papers were published were assessed and only papers published in a journal with an impact factor that was equal or higher than four were accepted to keep the focus of this review on published studies that have a relatively high impact and thus have a bigger chance to share potential methodological biases which could then be passed on to other studies. Another reason to filter papers based on journal impact factor and thus limiting the number of total selected papers were time constraints. Papers that only validated an earlier developed model were omitted from this review due to time constraints.

Model development assessment

To test the produced models in the papers for potential biases the PROBAST tool was used (20). This tool consists of four steps. First a research question was proposed by the reviewers. Second the type of prediction model used in each paper was noted. Third for each domain the risk of bias was assessed and fourth the overall risk of bias was assessed. The PROBAST tool consists tests biases in four domains: participants, predictors, outcome and analysis (Table S2). A fifth domain was added to test the reproducibility of each research paper (Table S2). This domain consisted of the following three questions: [1] is the research data available publicly, [2] is the source code available publicly and [3] is the developed model validated on external data sets in the same study. All questions used to test papers according to the PROBAST tool can be found in supplementary table 1.

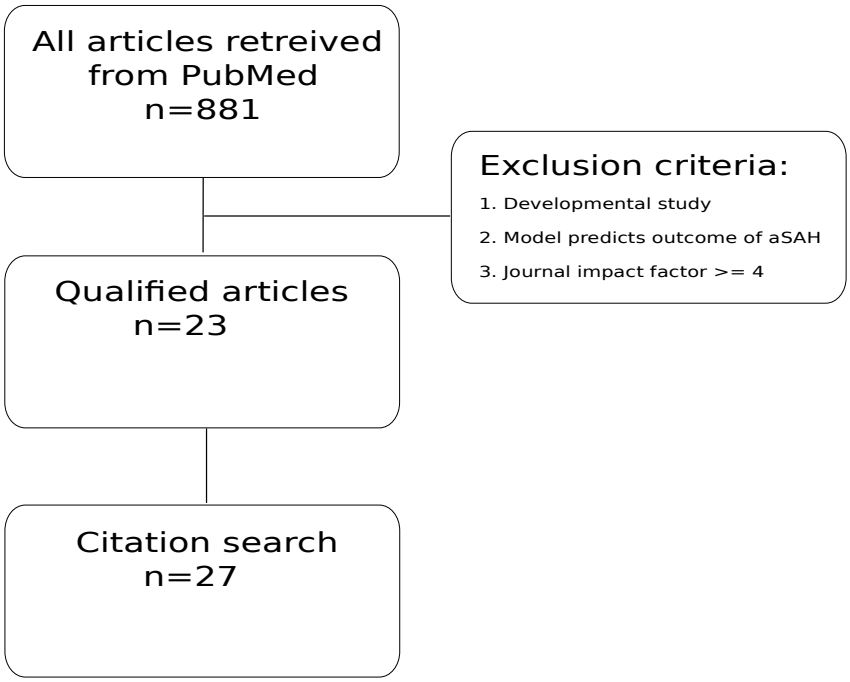


Figure 1: Search method and selection criteria of the included studies.

Results

In total 27 papers were included in this review. Out of these, 18 used retrospective cohorts (22–39) and 9 used prospective cohorts (40–48) (Table S1). Sources of data were mainly from European or North American origin including the United States (28–30,37,41,47,48), Canada (36), United Kingdom (23), Germany (31), Spain (22,34), Switzerland (24,26,27), The Netherlands (44–46) and two studies from China (35,40) (Table S1). Several studies used databases such as the SAHIT database (49) which gather data from patients located in multiple countries (25,32,33,38,39). These kind of databases are generally also focused on European and North American participants. Sampling date range from two years (37,40) to 18 years (41,48) with a median of 7 years (Table S1). The number of enrolled participants varied highly between studies ranging from 147 (35) to 10,936 (25) with an median of 548 enrolled participants. 14 papers created models using multivariate logistic regression (22,25,28,32,35,37,39,40,43–48) and 13 papers used other machine learning methods to create develop models such as decision trees (27,29,33,34), support vector machines (27,29,41) or neural nets (29–31,38) (Table S1). Most studies used a binary outcome using a modified Rankin scale (positive: mrs 0-3, negative mrs 4-6).

When considering multiple data sources that can be used to develop models it is important that these sources are compatible with each other in order to maintain the integrity of the model meaning that data is collected and assessed in a similar way. This potential bias is covered in the first domain of the PROBAST tool. Several studies used different data sets for development and validation, leading to usage of different outcome metrics not being compatible with each other. This required the need to correct for these differences leading to artificial outcome measures (23,32,46) (Table S3, Domain 1 & 3) or different methodology for assessing predictors (22,32,46) (Table S3, Domain 1 & 2). In another case the cohort was divided in a development and an internal validation cohort using date of admission as a separator between cohorts (22) (Table S3, Domain 1 & 2). Due to possible changes in procedures over time this could lead to potential biases. Inclusion and exclusion of participants was generally well documented (Table S3, Domain 1). Whereas domain one to three regard biases introduced via the data selection, domain four of the PROBAST tool focuses on potential biases that could be introduced via the statistic analyses that was performed. Overall there were enough participants that had the predetermined outcome during each study meaning that there was no need to correct for class imbalances in the data. Transformations in categorical, continuous variables or the dichotomization of variables could help the performance of machine learning algorithms. Generally the way in which variables were handled was well documented except for one paper where this was not mentioned (26). Missing data for enrolled participants can be handled in several ways. Generally some form of imputation was involved however certain studies performed complete case analyses by assuming that data is missing randomly (42,43). Data is often not missing at random and therefore complete case analysis could introduce bias by excluding participants with missing data. Six studies did not mention how missing data was handled (22,25,35–37,40,47) (Table S3, Domain 4). When benchmarking developed models it is preferred to use a validation cohort that is independent of the development cohort. In eight papers no mention was made of a separate validation cohort (22,28,32,37,40,43,45,48) (Table S3, Domain 4). Using predictors that are multicolinear could introduce bias in the coefficients of the used predictors of developed models. Using such predictors should be avoided when developing models that are intended to be used to interpret the resulting predictions. When only outcome prediction is important checking for multicolinearity is not required. Out of the 27 papers, 18 did not fully check or did not mention multicolinearity between predictors (22,24–26,28–30,32,35–37,40,42,44–48) (Table S3, Domain 4). During the benchmarking of the model over fitting of the model meaning that the model corresponds too well to the development data leading to decreased predicting performance on different data sets, should be taken into account. Usually this is done via comparison of AUROC, precision recall curves or confusion matrixes between the development and validation data sets. In total eight papers did not mention such analysis in their study

(22,23,28,32,37,42,44,45) (Table S3, Domain 4). Optimism of the developed model could be determined by applying cross validation or bootstrapping to the data set during development. Twelve out of the 27 studies did not apply or mention one of these techniques (24,26,28,29,35–40,45,47) (Table S3, Domain 4). Overall this domain contained the most potential to bias. By having 23 out of the 27 studies containing at least one way in which they could introduce a bias in to the developed model via the performed analyses for this domain.

The newly added fifth domain, containing questions about reproducibility of the work performed in the papers, showed that reproducibility is not reported on frequently by researchers while publishing their work. Only one study published the research data (25). and another would give the data upon reasonable request (23) (Table S3, Domain 4). External validation to determine the applicability of a developed model in circumstances wherein the model was not developed using a data set that is independently generated, was only performed in six of the reviewed papers (25,33–35,44,46) (Table S3, Domain 4).

Looking at the total picture across all five domains there would be no paper that scores good in all domains. However after excluding domain five only four papers do not introduce a potential bias during model development according to PROBAST criteria (27,31,33,41) showing that following correct procedures, when developing models using machine learning, is still not widely applied (Table S3, Domain 4). Overall most potential biases were introduced during the analyses (domain 4) (Figure 2).

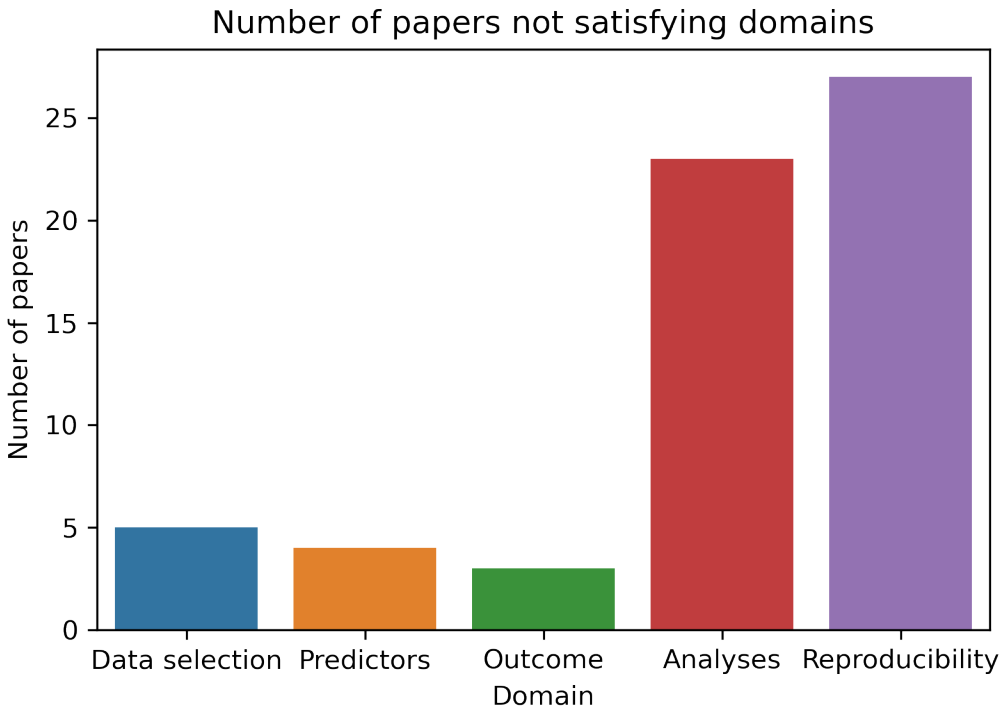


Figure 2: Overview of the number of papers not satisfying the PROBAST domain 1-4 and the newly added 5th domain.

Discussion

This study aimed to review the methodological process of developed models for the outcome of aSAH using machine learning algorithms. Only four out of the 27 reviewed papers scored well on the PROBAST tool that was used to asses potential biases introduced during model development. Most biases were introduced during the analyses phase (domain 4) mainly omitting to check for multicolinearity between predictors in the data set or check for over fitting of the model. The newly added fifth domain to the PROBAST tool covering study reproducibility showed that very few studies publish data or their code. External validation of the developed model also occurred only in 6 out of the 27 papers. This number could be higher as some of the models were validated externally in a different paper while this review only focused on papers were a model was developed.

A strength of this study was that the scope of the review was based on a validated assessment tool (PROBAST). Predefined inclusion and exclusion criteria were used and the study design was determined beforehand. This systematic review aimed to give a complete overview of the published literature of prediction models for the outcome of aSAH. An earlier published systematic review published in 2013 aimed to do the same and included 11 studies in their review which found similar results as this review found such as lack of outcome metric evaluation and lack of external validation (6). As the use of prediction models is becoming more widespread a new overview of prediction models for the outcome of subarachnoid hemorrhage was needed. In this review 27 papers matched the inclusion criteria leading to a good overview of the current developments in this field. Using the combination of PubMed and a complement citation search in each paper resulted in a more complete paper selection. Another strength of this review is the use of PROBAST to assess the methodology of developing prediction models making it possible to compare the results of this systematic review to other fields by comparing the specific PROBAST domains.

12 out of 27 studies did not make any mention of cross validating or bootstrapping the developed model. Both are re-sampling methods used to determine model performance. Generally cross validation is used to determine model optimism by dividing the development group into smaller groups (folds) so that the model can be validated on the hold-out data (50). This is in contrast with a bootstrapping method where new groups are made using re-sampling of the development group. Bootstrapping is therefore mainly used to determine model stability (51). In health science data sets are generally limited by the amount of participants that have a certain condition such as aSAH. This was also mentioned in most studies reviewed for this paper as a reason for the relative small data sets. Therefore it could be a logical conclusion that applying cross validation and thus making the validation groups smaller is not feasible. For small data sets leave one out cross validation (LOOCV) could be considered. When applying LOOCV the model is trained on n-1 samples leaving one sample for the validation set. One drawback of this method is that a high variance could be introduced in the resulting performance by validating on one sample (52). Lack of cross validation and bootstrapping could lead to a failure to detect over fitting making the developed models less suitable for outcome prediction (53).

18 out of the 27 studies did not check for or reported multicolinearity. Multicolinearity in data sets occurs due to correlations between predictors in the data. This can be avoided by checking for correlations during predictor selection and only select non correlated predictors or combine predictors when necessary (54). A result of multicolinearity in a data set can be that the resulting predictor coefficients of the model are skewed. When the aim of the model is to explain outcome predictions this could form a problem (55). Since the main focus of the models developed by the reviewed studies is the prediction of clinical outcome and not risk factor detection this could be an valid argument for not checking for multicolinearity.

The popularity of area under the curve as an outcome measurement has increased over time (56) which also shows in this review as most studies reported area under the curve as the outcome measurement. There is however also criticism on the use of this outcome measurement. This is because a threshold for area under the curve is calibrated for every developed model, comparison between

models using AUROC as a measure would be the same as comparison between multiple metrics (56,57). It could be considered to use the H measurement, which was proposed to to overcome this effect (58). Another issue with area under the curve as a outcome measure is that when the data set is unbalanced, area under the curve could give a skewed result. A high area under the curve can be obtained while due to the low number of outcomes, false positives are masked leading to an high area under the curve. For imbalanced data sets it is recommended to use the precision recall curve or scorers such as accuracy or F-beta as an outcome measure (59). As there were no imbalanced data sets in the reviewed studies AUROC seems to be an acceptable outcome measure as most other studies report on this outcome measure.

In the field of machine learning models are often only validated internally (60) which is generally done by splitting the data set in an developmental data set and an validation data set and then apply methods such as bootstrapping or cross validation (61). Only six studies in this review performed external validation. External validation is critical when the model is intended to be used clinically as the developed model has to be widely applicable when used for this purpose (27). Most studies reported lack of external data sets that had the same outcome measures or predictors as reason to not externally validate their model. Also some of the other studies might have published the external validation of their model in an different paper which were not included in this review.

Only 2 out of the 27 reviewed studies published their data sets in a online repository and only one study published their source code. Good data stewardship is important when it comes to the reproducibility of performed research (62). In order to increase and spread knowledge about data stewardship the FAIR principles were proposed. The goal of these principles was to make data more easier to find, access, compatible and reusable. In principle this has to be done in such a way that repositories are more accessible for machines by implementing for example well documented API's and in this way making data also more accessible for humans (12). A difficulty in health science regarding open data is that due to patient privacy data can not be published publicly which could explain the low amount of reviewed studies publishing their data sets (63). A possible solution to this difficulty could be the synthesis of artificial data sets that contain the exact same properties as the original data set (64–66). This data set can then be published using the FAIR principles.

The median of enrolled participants of all reviewed papers was 548 showing that overall the size of the data sets were small for machine learning applications. Although sample size of the data set in not a PROBAST criteria many of the studies reported small sample size as an factor in how that study was performed. Small sample size is an issue that is also present in other fields of health science (67). In health science the size of the data sets are mainly limited by the costs of large scale experiments necessary for data collection or the occurrence of a specific outcome (68). The reviewed papers also mention this as the main constraint for generating data sets that include more participants. Data collection over longer time scale often was hard as procedures changed over time making the use of smaller data sets more acceptable.

Although the results published in this study could provide valuable insight in the application of machine learning algorithms within health science this study also contains several limitations. This study assessed the methodological approach of studies used in the model development but did not assess the quality of the models itself and thus their ability to make clinical predictions. Furthermore the selection and assessment of studies was performed by one researcher. Assessment by multiple researchers could help reduce potentiality systematic errors that could have been introduced during the process. Only studies that have been published in English have been included in this study which could potentially cause a bias in the selected literature and thus participants of the selected studies (69). For this review the online repository PubMed was used. As PubMed mainly focuses on medical journals most relevant studies should have been found however it can not be excluded that some journals publishing an relevant study was not available in this repository (70).

Conclusion

Potential methodological biases of model development studies were identified using the PROBAST tool. The reviewed studies show that mainly biases are introduced during the analyses for example accounting for missing data and considering different outcome metrics. Also during data collection biases were introduced for example when multiple data sets are used using different methodology of data collection. Following the FAIR principles data should be easily accessible which was not applicable to most studies as they did not publish their data sets to an repository.

Supplements

Table S1: Summary data of selected papers.

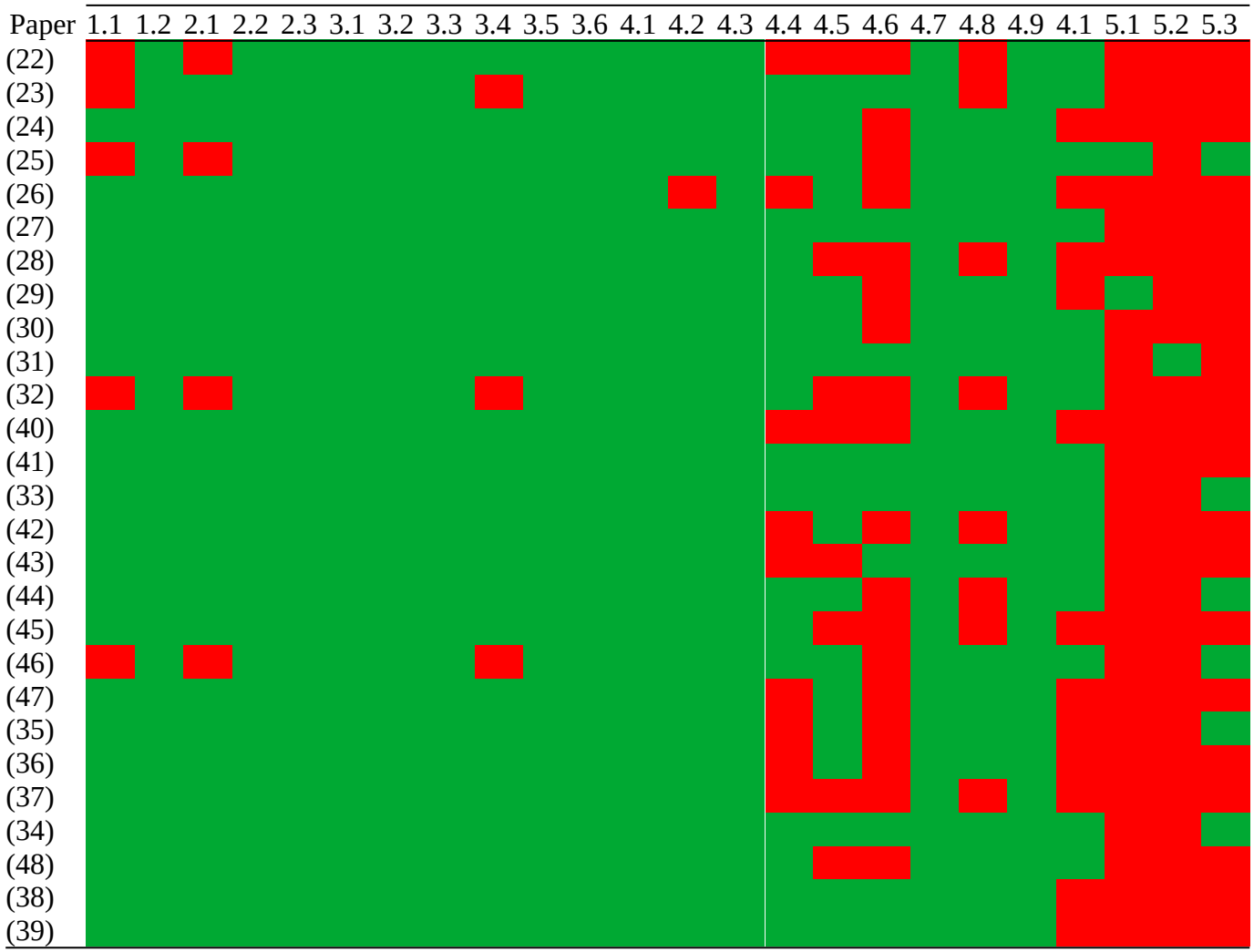
Paper	Model type	Country	Cohort	Sampling time (years)	Cohort size
(22)	Logistic regression	Spain	Retrospective	10	536
(23)	RF, SVM	UK	Retrospective	8	1017
(24)	CHAID	CH	Retrospective	6	548
(25)	Logistic regression	Multiple	Retrospective		10936
(26)	ML	CH	Retrospective	6	1866
(27)	SVM, FAM, RF, GLM, GBM	CH	Retrospective	6	156
(28)	Logistic regression	US	Retrospective	3	430
(29)	SVM, FAM, RF, GLM, GBM, MLP	US	Retrospective	14	2467
(30)	MLP	US	Retrospective	7	451
(31)	GLM, MLP	DE	Retrospective	6	388
(32)	Logistic regression	Multiple	Retrospective	10	357
(40)	Logistic regression	CN	Prospective	2	366
(41)	SVM	US	Prospective	18	1595
(33)	RF	Multiple	Retrospective		266
(42)	ffANN	NL	Prospective	7	585
(43)	Logistic regression	EU	Prospective		2143
(44)	Logistic regression	NL	Prospective	16	1620
(45)	Logistic regression	NL	Prospective	16	1620
(46)	Logistic regression	NL	Prospective	5	409
(47)	Logistic regression	US	Prospective		527
(35)	Logistic regression	CN	Retrospective	6	147
(36)	ML	CA	Retrospective	14	10322
(37)	Logistic regression	US	Retrospective	2	161
(34)	RF	SP	Retrospective	11	441
(48)	Logistic regression	US	Prospective	18	1619
(38)	MLP	Multiple	Retrospective	7	3550
(39)	Logistic regression	Multiple	Retrospective	6	3551

Table S2: Questions used from the PROBAST tool (domain 1-4) (20) and questions added (domain 5) to review the selected papers.

Domain 1. Data selection.	1.1	Were appropriate data sources used (e.g. compatible with each other)
	1.2	Were all inclusions and exclusions of participants appropriate?
Domain 2. Predictors.	2.1	Were predictors defined and assessed in a similar way for all participants?
	2.2	Were predictor assessments made without knowledge of outcome data?
	2.3	Are all predictors available at the time the model is intended to be used?
Domain 3. Outcome.	3.1	Was the outcome determined appropriately?
	3.2	Was a pre-specified or standard outcome definition used -> were multiple outcome measurements considered
	3.3	Were predictors excluded from the outcome definition?
	3.4	Was the outcome defined and determined in a similar way for all participants?
	3.5	Was the outcome determined without knowledge of predictor information?
	3.6	Was the time interval between predictor assessment and outcome determination appropriate?
Domain 4. Analyses.	4.1	Were there a reasonable number of participants with the outcome? / Was there accounted for imbalanced data sets?
	4.2	Were continuous and categorical predictors handled appropriately?
	4.3	Were all enrolled participants included in the analysis?
	4.4	Were participants with missing data handled appropriately?
	4.5	Were test and training groups well defined?
	4.6	Was selection of predictors based on univariable analysis avoided?
	4.7	Were relevant model performance measures considered?
	4.8	Were model over fitting and optimism in model performance accounted for?
	4.9	Do predictors and their assigned weights in the final model correspond to the results from multivariable analysis?
	4.1 0	Was cross validation/bootstrapping used?
Domain 5. Reproducibility.	5.1	Is the research data available publicly?

5.2	Is the source code available publicly?
5.3	Was the result validated on different external data sets in the same study?

Table S3: Results of the questions from the PROBAST tool and the added domain 5 for each selected paper with green indicating that a question was satisfied and red indicating that a question was unsatisfied.



1. Suarez JI, Tarr RW, Selman WR. Aneurysmal subarachnoid hemorrhage. N Engl J Med. 2006;354(4):387–96.

2. van Gijn J, Rinkel G. Subarachnoid haemorrhage: diagnosis, causes and management. Brain. 2001;124(2):249–78.

3. Al-Khindi T, Macdonald RL, Schweizer TA. Cognitive and functional outcome after aneurysmal subarachnoid hemorrhage. Stroke. 2010;41(8):e519–36.

4. Nieuwkamp DJ, Setz LE, Algra A, Linn FH, de Rooij NK, Rinkel GJ. Changes in case fatality of aneurysmal subarachnoid haemorrhage over time, according to age, sex, and region: a meta-analysis. Lancet Neurol. 2009;8(7):635–42.

5. Rehman S, Sahle BW, Chandra RV, Dwyer M, Thrift AG, Callisaya M, et al. Sex differences in risk factors for aneurysmal subarachnoid haemorrhage: Systematic review and meta-analysis. J Neurol Sci. 2019;406:116446.

6. Jaja BN, Cusimano MD, Etminan N, Hanggi D, Hasan D, Ilodigwe D, et al. Clinical prediction models for aneurysmal subarachnoid hemorrhage: a systematic review. Neurocrit Care. 2013;18(1):143–53.

7. Bhavya S, Pillai AS. Prediction models in healthcare using deep learning. In: International Conference on Soft Computing and Pattern Recognition. Springer; 2019. p. 195–204.

8. Campbell C. Machine learning methodology in bioinformatics. In: Springer handbook of bio-/neuroinformatics. Springer; 2014. p. 185–206.

9. Steyerberg EW. Applications of prediction models. In: Clinical prediction models. Springer; 2009. p. 11–31.

10. den Boer S, de Keizer NF, de Jonge E. Performance of prognostic models in critically ill cancer patients–a review. Crit Care. 2005;9(4):1–6.

11. Hullman J, Kapoor S, Nanayakkara P, Gelman A, Narayanan A. The worst of both worlds: A comparative analysis of errors in learning from data in psychology and machine learning. ArXiv Prepr ArXiv220306498. 2022;

12. Wilkinson MD, Dumontier M, Aalbersberg IjJ, Appleton G, Axton M, Baak A, et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data. 2016;3(1):1–9.

13. McDermott MB, Wang S, Marinsek N, Ranganath R, Foschini L, Ghassemi M. Reproducibility in machine learning for health research: Still a ways to go. Sci Transl Med. 2021;13(586):eabb1655.

14. Gundersen OE, Shamsaliei S, Isdahl RJ. Do machine learning platforms provide out-of-the-box reproducibility? Future Gener Comput Syst. 2022;126:34–47.

15. Steinberg E, Greenfield S, Wolman DM, Mancher M, Graham R, others. Clinical practice guidelines we can trust. National Academies Press; 2011.

16. Egger M, Higgins JP, Smith GD. Systematic Reviews in Health Research: Meta-Analysis in Context. John Wiley & Sons; 2022.

17. Debray TP, Damen JA, Snell KI, Ensor J, Hooft L, Reitsma JB, et al. A guide to systematic review and meta-analysis of prediction model performance. bmj. 2017;356.

18. Collins GS, Moons KG, Debray TP, Altman DG, Riley RD. Systematic Reviews of Prediction Models. Syst Rev Health Res Meta-Anal Context. 2022;347–76.

19. Brown BB. Delphi process: a methodology used for the elicitation of opinions of experts. Rand Corp Santa Monica CA; 1968.

20. Wolff RF, Moons KG, Riley RD, Whiting PF, Westwood M, Collins GS, et al. PROBAST: a tool to assess the risk of bias and applicability of prediction model studies. Ann Intern Med. 2019;170(1):51–8.

21. Murdoch WJ, Singh C, Kumbier K, Abbasi-Asl R, Yu B. Definitions, methods, and applications in interpretable machine learning. Proc Natl Acad Sci. 2019;116(44):22071–80.

22. Mourelo-Fariña M, Pértega S, Galeiras R. A model for prediction of in-hospital mortality in patients with subarachnoid hemorrhage. Neurocrit Care. 2021;34(2):508–18.

23. Gaastra B, Barron P, Newitt L, Chhugani S, Turner C, Kirkpatrick P, et al. CRP (C-reactive protein) in outcome prediction after subarachnoid hemorrhage and the role of machine learning. Stroke. 2021;52(10):3276–85.

24. Hostettler IC, Muroi C, Richter JK, Schmid J, Neidert MC, Seule M, et al. Decision tree analysis in subarachnoid hemorrhage: prediction of outcome parameters during the course of aneurysmal subarachnoid hemorrhage using decision tree analysis. J Neurosurg. 2018;129(6):1499–510.

25. Jaja BN, Saposnik G, Lingsma HF, Macdonald E, Thorpe KE, Mamdani M, et al. Development and validation of outcome prediction models for aneurysmal subarachnoid haemorrhage: the SAHIT multinational cohort study. Bmj. 2018;360.

26. Maldaner N, Zeitlberger AM, Sosnova M, Goldberg J, Fung C, Bervini D, et al. Development of a complication-and treatment-aware prediction model for favorable functional outcome in aneurysmal subarachnoid hemorrhage based on machine learning. Neurosurgery. 2021;88(2):E150–7.

27. Staartjes VE, Sebök M, Blum PG, Serra C, Germans MR, Krayenbühl N, et al. Development of machine learning-based preoperative predictive analytics for unruptured intracranial aneurysm surgery: a pilot study. Acta Neurochir (Wien). 2020;162(11):2759–65.

28. Dasenbrock H, Gormley WB, Lee Y, Mor V, Mitchell SL, Fehnel CR. Long-term outcomes among octogenarians with aneurysmal subarachnoid hemorrhage. J Neurosurg. 2018;131(2):426–34.

29. Yu D, Williams GW, Aguilar D, Yamal JM, Maroufy V, Wang X, et al. Machine learning prediction of the adverse outcome for nontraumatic subarachnoid hemorrhage patients. Ann Clin Transl Neurol. 2020;7(11):2178–85.

30. Savarraj JP, Hergenroeder GW, Zhu L, Chang T, Park S, Megjhani M, et al. Machine learning to predict delayed cerebral ischemia and outcomes in subarachnoid hemorrhage. Neurology. 2021;96(4):e553–62.

31. Dengler NF, Madai VI, Unteroberdörster M, Zihni E, Brune SC, Hilbert A, et al. Outcome prediction in aneurysmal subarachnoid hemorrhage: a comparison of machine learning methods and established clinico-radiological scores. Neurosurg Rev. 2021;44(5):2837–46.

32. Ten Brinck MF, Shimanskaya VE, Aquarius R, Bartels RH, Meijer FJ, Koopmans PC, et al. Outcomes after Flow Diverter Treatment in Subarachnoid Hemorrhage: A Meta-Analysis and Development of a Clinical Prediction Model (OUTFLOW). Brain Sci. 2022;12(3):394.

33. Liu J, Xiong Y, Zhong M, Yang Y, Guo X, Tan X, et al. Predicting long-term outcomes after poor-grade aneurysmal subarachnoid hemorrhage using decision tree modeling. Neurosurgery. 2020;87(3):523–9.

34. de Toledo P, Rios PM, Ledezma A, Sanchis A, Alen JF, Lagares A. Predicting the outcome of patients with subarachnoid hemorrhage using machine learning techniques. IEEE Trans Inf Technol Biomed. 2009;13(5):794–801.

35. Shen J, Yu J, Huang S, Mungur R, Huang K, Pan X, et al. Scoring model to predict functional outcome in poor-grade aneurysmal subarachnoid hemorrhage. Front Neurol. 2021;12:601996.

36. English SW, McIntyre L, Fergusson D, Turgeon A, Dos Santos MP, Lum C, et al. Subarachnoid hemorrhage admissions retrospectively identified using a prediction model. Neurology. 2016;87(15):1557–64.

37. Hemphill III JC, Bonovich DC, Besmertis L, Manley GT, Johnston SC. The ICH score: a simple, reliable grading scale for intracerebral hemorrhage. Stroke. 2001;32(4):891–7.

38. Lo BW, Macdonald RL, Baker A, Levine MA. Clinical outcome prediction in aneurysmal subarachnoid hemorrhage using Bayesian neural networks with fuzzy logic inferences. Comput Math Methods Med. 2013;2013.

39. Lo BW, Fukuda H, Angle M, Teitelbaum J, Macdonald RL, Farrokhyar F, et al. Clinical outcome prediction in aneurysmal subarachnoid hemorrhage–Alterations in brain–body interface. Surg Neurol Int. 2016;7(Suppl 18):S527.

40. Zheng K, Zhong M, Zhao B, Chen SY, Tan XX, Li ZQ, et al. Poor-grade aneurysmal subarachnoid hemorrhage: risk factors affecting clinical outcomes in intracranial aneurysm patients in a multi-center study. Front Neurol. 2019;10:123.

41. Park S, Megjhani M, Frey HP, Grave E, Wiggins C, Terilli KL, et al. Predicting delayed cerebral ischemia after subarachnoid hemorrhage using physiological time series data. J Clin Monit Comput. 2019;33(1):95–105.

42. de Jong G, Aquarius R, Sanaan B, Bartels RH, Grotenhuis JA, Henssen DJ, et al. Prediction models in aneurysmal subarachnoid hemorrhage: forecasting clinical outcome with artificial intelligence. Neurosurgery. 2021;88(5):E427–34.

43. Risselada R, Lingsma H, Bauer-Mehren A, Friedrich C, Molyneux A, Kerr R, et al. Prediction of 60 day case-fatality after aneurysmal subarachnoid haemorrhage: results from the International Subarachnoid Aneurysm Trial (ISAT). Eur J Epidemiol. 2010;25(4):261–6.

44. van Donkelaar CE, Bakker NA, Veeger NJ, Uyttenboogaart M, Metzemaekers JD, Eshghi O, et al. Prediction of outcome after subarachnoid hemorrhage: timing of clinical assessment. J Neurosurg. 2017;126(1):52–9.

45. van Donkelaar CE, Bakker NA, Birks J, Veeger NJ, Metzemaekers JD, Molyneux AJ, et al. Prediction of outcome after aneurysmal subarachnoid hemorrhage: development and validation of the SAFIRE grading scale. Stroke. 2019;50(4):837–44.

46. van der Steen W, Marquering H, Ramos L, van den Berg R, Coert B, Boers A, et al. Prediction of outcome using quantified blood volume in aneurysmal SAH. Am J Neuroradiol. 2020;41(6):1015–21.

47. Ban VS, El Ahmadieh TY, Aoun SG, Plitt AR, Lyon KA, Eddleman C, et al. Prediction of outcomes for ruptured aneurysm surgery: the Southwestern aneurysm severity index. Stroke. 2019;50(3):595–601.

48. Witsch J, Frey HP, Patel S, Park S, Lahiri S, Schmidt JM, et al. Prognostication of long-term outcomes after subarachnoid hemorrhage: The FRESH score. Ann Neurol. 2016;80(1):46–58.

49. Macdonald RL, Cusimano MD, Etminan N, Hanggi D, Hasan D, Ilodigwe D, et al. Subarachnoid hemorrhage international trialists data repository (SAHIT). World Neurosurg. 2013;79(3–4):418–22.

50. Hawkins DM. The problem of overfitting. J Chem Inf Comput Sci. 2004;44(1):1–12.

51. Ng V, Cardie C. Bootstrapping coreference classifiers with multiple machine learning algorithms. In: Proceedings of the 2003 conference on Empirical methods in natural language processing. 2003. p. 113–20.

52. Wong TT. Performance evaluation of classification algorithms by k-fold and leave-one-out cross validation. Pattern Recognit. 2015;48(9):2839–46.

53. Berrar D. Cross-Validation. 2019.

54. Garg A, Tai K. Comparison of statistical and machine learning methods in modelling of data with multicollinearity. Int J Model Identif Control. 2013;18(4):295–312.

55. Chan JYL, Leow SMH, Bea KT, Cheng WK, Phoong SW, Hong ZW, et al. Mitigating the Multicollinearity Problem and Its Machine Learning Approach: A Review. Mathematics. 2022;10(8):1283.

56. Hand DJ, Anagnostopoulos C. When is the area under the receiver operating characteristic curve an appropriate measure of classifier performance? Pattern Recognit Lett. 2013;34(5):492–5.

57. Hilden J. The area under the ROC curve and its competitors. Med Decis Making. 1991;11(2):95–101.

58. Hand DJ. Measuring classifier performance: a coherent alternative to the area under the ROC curve. Mach Learn. 2009;77(1):103–23.

59. Jeni LA, Cohn JF, De La Torre F. Facing imbalanced data–recommendations for the use of performance metrics. In: 2013 Humaine association conference on affective computing and intelligent interaction. IEEE; 2013. p. 245–51.

60. Zhang JM, Harman M, Ma L, Liu Y. Machine learning testing: Survey, landscapes and horizons. IEEE Trans Softw Eng. 2020;

61. Cabitza F, Campagner A, Soares F, de Guadiana-Romualdo LG, Challa F, Sulejmani A, et al. The importance of being external. methodological insights for the external validation of machine learning models in medicine. Comput Methods Programs Biomed. 2021;208:106288.

62. Jacobsen A, de Miranda Azevedo R, Juty N, Batista D, Coles S, Cornet R, et al. FAIR principles: interpretations and implementation considerations. Vol. 2, Data intelligence. MIT Press One Rogers Street, Cambridge, MA 02142-1209, USA journals-info …; 2020. p. 10–29.

63. Queralt-Rosinach N, Kaliyaperumal R, Bernabé CH, Long Q, Joosten SA, van der Wijk HJ, et al. Applying the FAIR principles to data in a hospital: challenges and opportunities in a pandemic. J Biomed Semant. 2022;13(1):1–19.

64. Montjoye YA de, Farzanehfar A, Hendrickx J, Rocher L. Solving artificial intelligence's privacy problem. Field Actions Sci Rep J Field Actions. 2017;(Special Issue 17):80–3.

65. Zhu T, Ye D, Wang W, Zhou W, Yu P. More than privacy: Applying differential privacy in key areas of artificial intelligence. IEEE Trans Knowl Data Eng. 2020;

66. Triastcyn A, Faltings B. Generating artificial data for private deep learning. ArXiv Prepr ArXiv180303148. 2018;

67. Shaikhina T, Lowe D, Daga S, Briggs D, Higgins R, Khovanova N. Machine learning for predictive modelling based on small data in biomedical engineering. IFAC-Pap. 2015;48(20):469–74.

68. Jiang J, Wang R, Wang M, Gao K, Nguyen DD, Wei GW. Boosting tree-assisted multitask deep learning for small scientific datasets. J Chem Inf Model. 2020;60(3):1235–44.

69. Burdett S, Stewart LA, Tierney JF. Publication bias and meta-analyses: a practical example. Int J Technol Assess Health Care. 2003;19(1):129–34.

70. Liljekvist MS, Andresen K, Pommergaard HC, Rosenberg J. For 481 biomedical open access journals, articles are not searchable in the Directory of Open Access Journals nor in conventional biomedical databases. PeerJ. 2015;3:e972.