Development and external validation of an international, multicenter machine learning algorithm for prediction of surgical outcome after microsurgery for unruptured intracranial aneurysms

The *Prediction of adverse events after microsurgery for intracranial unruptured aneurysms* (PRAEMIUM) study

Victor E. Staartjes[1], BMed; Luca Regli[1], MD; Giuseppe Esposito[1], MD;

ClinicalTrials.gov:

ID: PRAEMIUM Prediction of Outcomes After Surgery for Unruptured Intracranial Aneurysms NCT04819074

1: Department of Neurosurgery & Machine Intelligence in Clinical Neuroscience (MICN) Laboratory, Clinical Neuroscience Center, University Hospital Zurich, University of Zurich, Switzerland

Introduction

Unruptured intracranial aneurysms (UIAs) are incidentally detected at an increasing rate, mostly owing to the rise in availability of non-invasive cranial imaging.[1] Decision-making in UIAs is complex and requires consideration of many risk factors for aneurysm growth and rupture to balance the benefits and risks of treatment versus observation. This is due to: 1) the high morbidity and case fatality inherent to aneurysmal subarachnoid hemorrhage (SAH) 2) the relatively low rupture rate of unruptured aneurysms; 3) the potential morbidity and mortality rate associated with either microsurgical or endovascular treatment.[2–8]

Some consistent risk factors for rupture have been identified, including involvement of the posterior circulation, larger diameter, higher age, and some specific populations such as Japanese and Finnish patients.[9] Many other risk factors have been suggested based on varying levels of evidence.[3,5,8] However, it is difficult to integrate this considerable number of factors into a single risk assessment and to present a clear clinical decision making algorithm to patients. A range of scoring systems have been developed and validated to approximate the risk of rupture (PHASES)[5] and growth (ELAPSS)[2], or to balance the risks and benefits of microsurgical treatment versus follow-up imaging directly (UIATS)[3] by integrating some of these risk factors.[10] Still, these scores are focused on predicting rupture events instead of neurological outcome. In addition, they usually are focused on solely one outcome, instead of providing a wide range of objective predictive analytics that may then improve shared decision-making.[11]

Machine learning (ML) methods have been extraordinarily effective at integrating many clinical patient variables into one holistic risk prediction tailored to each patient. We have conducted a pilot study to assess the feasibility of predicting surgical outcomes after surgery for UIAs [12] in a small single-center sample. We found that prediction was feasible with good performance metrics, and also identified the most important factors to be included in such models. To the authors' best knowledge, a robust, multicenter, externally validated prediction model or predictive score for surgical outcome after microsurgery for UIAs does not yet exist.

Accurate preoperative identification of patients at high risk for adverse outcomes would be clinically advantageous, as it would allow enhanced resource preparation, better surgical decision-making, enhanced patient education and informed consent, and potentially even modification of certain modifiable risk factors. The aim of the *Prediction of adverse events after microsurgery for intracranial unruptured aneurysms* (PRAEMIUM) study is therefore to develop and externally validate a clinically applicable, robust ML-based prediction tool based on multicenter data from a range of international centers.

Methods

Overview

Data will be collected by a range of international centers. Overall, the model will be built and publication will be compiled according to the transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD) guidelines.[13]

The University Hospital Zurich (USZ) center is the sponsor of the PRAEMIUM study.
(**Sponsor**: Prof. Dr. Luca Regli; **Principal Investigators**: PD Dr. med. G. Esposito, Victor Staartjes)

Ethical Considerations

Each center will be responsible for their own ethics board / institutional review board (IRB) approval. They must gain approval for retrospective or prospective data collection and sharing of the completely anonymized data with the sponsor. They must also organize a data use agreement to the sponsor institution. The sponsor can aid by providing said protocol and data use agreement, which will also be registered online on ClinicalTrials.gov.

Inclusion and Exclusion Criteria

We will include all adult patients (18 years or older) undergoing microsurgical treatment for UIAs. No specific exclusion criteria will be set. Patients with prior SAH may only be included when surgical treatment occurred at least 4 weeks after ictus. **Only patients treated from January 1st 2010 onwards can be included in this study**.

Authorship

Each center must contribute data from at least 100 patients to be included. Each eligible institution designates 1 primary author per center who will be included in the primary author list. Each eligible center may also designate 2 additional secondary authors, who will be included in the author list as part of the PRAEMIUM study group, and who

will be fully indexed in PubMed as contributing authors. The sponsor institution is entitled to 3 primary authors and to 4 secondary authors.

Data Collection

Each center will collect their data either retrospectively, or from a prospective registry, or from a prospective registry supplemented by retrospectively collected variables. Data from patients operated from January 1st 2010 and onwards will be eligible for inclusion. **Data collection should be completed, and anonymized data should be sent to the sponsor institution (**praemium@usz.ch**) by 1st of August 2021.**

A standardized Excel spreadsheet will be provided by the sponsor. The data will be entered in standardized and anonymized form. This spreadsheet will only contain a study-specific patient number. The data set is anonymized source data that includes clinical data extracted from electronic health records (retrospectively or from a prospective registry of already existing data). The data will be anonymized upon entering them into the PRAEMIUM Excel spreadsheet, after which the patients will be numbered consecutively and there will be no way to trace the data back to individual patients. No identifiable data such as date of birth will be included. Whenever the PRAEMIUM Excel spreadsheet is transferred, it will be encrypted using a password and sent through a secure institutional e-mail server to **PRAEMIUM@usz.ch**. The password will be sent in a separate e-mail. Some missing data is acceptable, but should be kept to a minimum (i.e. must be < 10%)

Endpoint Definitions

Models will be developed for the following three endpoints at discharge: **Poor neurological outcome** (1), as well as presence of (2) **new sensorimotor neurological deficits** and (3) any **complications** (surgical or non-surgical). Neurological outcome was assessed by the modified Rankin scale (mRS), and a favorable neurological outcome was defined as mRS 0, 1, or 2.[14] Complications will be assessed using the modified 2009 Clavien-Dindo grading (CDG), and occurrence of a complication was defined as any deviation from CDG 0.[15] The Clavien-Dindo grading system is a classification of surgical complications: Grad 0 signifying no complication, Grade I identifying complications with any deviation from the normal intra- or postoperative course requiring medical treatment, and so forth. Detailed definitions are provided in the Excel spreadsheet. Surgery-related as well as none-surgery-related complications are counted. In case of multiple complications, only the complication with the highest CDG was counted per patient.

Input Feature Definitions

All features are measured preoperatively. Recorded baseline variables will include age, gender, maximum aneurysm diameter, anatomical location (artery), total number of aneurysms per patient, if multiple aneurysms were treated

during the index session, calcification of the aneurysm wall or neck, aneurysm morphology (saccular, dissecting, fusiform, or other), involvement of critical perforating or branch vessels, and intraluminal thrombosis.
We will also capture prior SAH, mRS at admission, prior aneurysm treatment, presence of anticoagulation/antiplatelet therapy preoperatively, and hypertension, as well as American Society of Anesthesiologists (ASA) grading, the PHASES[5], ELAPSS[2], and UIATS[3] scores including the UIATS "pro-repair" and "pro-conservative treatment" subscores. The unruptured intracranial aneurysm treatment score (UIATS) consists of two subscores: One that represents the strength of recommendation for invasive repair of an unruptured aneurysm, and one that represents the strength of recommendation for conservative management of an unruptured aneurysm. The final overall UIATS score is subsequently calculated as the difference between the two subscores. Also included was the surgical approach: minimally invasive or standard approach, and whether a bypass was performed.[16,19,23]

Sample Size

While even the largest cohort with millions of patients is not guaranteed to result in a robust clinical prediction model if no relevant input variables are included ("garbage in, garbage out" – do not expect to predict the future from age, gender, and body mass index), the relationship among predictive performance and sample size is certainly directly proportional, especially for some data-hungry ML algorithms. To ensure generalizability of the clinical prediction model, the sample size should be both representative enough of the patient population, and should take the complexity of the algorithm into account. For instance, a deep neural network – as an example of a highly complex model – will often require thousands of patients to converge, while a logistic regression model may achieve stable results with only a few hundreds of patients. In addition, the number of input variables plays a role. Roughly, it can be said that a bare minimum of 10 positive cases are required per included input variable to model the relationships. Often, erratic behavior of the models and high variance in performance among splits is observed when sample sizes are smaller than calculated with this rule of thumb. Of central importance is also the proportion of patients who experience the outcome. For very rare events, a much larger total sample size is consequentially needed. For instance, a prediction based on 10 input features for an outcome occurring in only 10% of cases would require at least 1000 patients including at least 100 who experienced the outcome, according to the above rule of thumb. In general and from personal experience, we do not recommend developing ML models on cohorts with less than 100 positive cases and reasonably more cases in total, regardless of the rarity of the outcome. Also, one might consider the available literature on risk factors for the outcome of interest: If epidemiological studies find only weak associations with the outcome, it is likely that one will require more patients to arrive at a model with good predictive performance, as opposed to an outcome which has several highly associated risk factors, which may be easier to predict. Larger sample sizes also allow for more generous evaluation through a larger amount of patient data dedicated to training or validation, and usually results in better calibration measures.

For this study, based on our expertise and on the rules of thumb mentioned above, we estimate that a minimum of 250 patients with positive outcome are required to extract generalizable features. With an incidence of these outcomes of

around 10%[12] that means that a minimum of 2500 patients are required for training. For adequate evaluation of calibration at external validation, we estimate that – as a bare minimum – another 1500 patients (thus, approximately 150 with positive outcomes) will be required. Thus, in total, we estimate that **4000 patients** are necessary to arrive at a robust model. More data will likely lead to greater performance and better calibration.

Predictive Modeling

A KNN imputer will be co-trained to impute any missing data that may occur in future application of the model. If there is missing data in the training set, it will be imputed using said KNN imputer.[24] Features or patients with a missingness greater than 25% will be excluded. Data will be standardized and one-hot-encoded. In case of major class imbalance – which is expected for the abovementioned endpoint – random upsampling or synthetic minority oversampling (SMOTE) will be applied to the training set.[25,26] All features will initially be provided to the model for training. If necessary, we will apply recursive feature elimination (RFE) to select input features on the training data.[27]

We will trial the following algorithms for binary classification: Generalized linear model (GLM), generalized additive model (GAM), stochastic gradient boosting machine (GBM), naïve Bayes classifier, artificial neural network, support vector machine (SVM), and random forest. Each model will be fully trained and hyperparameter tuned where applicable. The final model will be selected based upon AUC, sensitivity, and specificity, as well as calibration metrics on the resampled training performance. Training will occur in repeated 5-fold cross-validation with 10 repeats.

The one final model will then be assessed on the external validation data only once. 95% confidence intervals for external validation metrics will be derived using the bootstrap.

The threshold for binary classification will be identified on the training data alone using the AUC-based "closest-to-(0,1)-criterion" or Youden's index to optimize both sensitivity and specificity.[28] All analyses will be carried out in R Version 3.6.2.[29]

Evaluation

The performance of classification models can roughly be judged along two dimensions: Model discrimination and calibration.[30] The term *discrimination* denotes the ability of a prediction model to correctly classify whether a certain patient is going to or is not going to experience a certain outcome. Thus, discrimination described the accuracy of a binary prediction – yes or no. *Calibration*, however, describes the degree to which a model's predicted probabilities (ranging from 0% to 100%) correspond to the actually observed incidence of the binary endpoint (true posterior). Many publications do not report calibration metrics, although these are of central importance, as a well-calibrated predicted probability (e.g. your predicted probability of experiencing a complication is 18%) is often much more valuable to clinicians – and patients! – than a binary prediction (e.g. you are likely not going to experience a complication).[30]

Resampled training performance as well as performance on the external validation set will be assessed for discrimination and calibration. In terms of discrimination, we will evaluate AUC, accuracy, sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), and F1 Score. In terms of calibration, we will assess the Brier score, expected-observed (E/O)-ratio, calibration slope and intercept, the Hosmer-Lemeshow goodness-of-fit test, as well as visual inspection of calibration plots for both datasets, which will also be included in the publication.

Interpretability

The degree and choice of methods for interpretability will depend on the finally chosen algorithm. Some algorithms can natively provide explanations as to which factors influence the outcome in what way. Thus, in case e.g. a GLM, GAM, or naïve Bayes classifier is chosen, the parameters / partial dependence values will be provided. For simple decision trees, diagrams of the decision-making process can be provided. Other models with higher degrees of complexity, such as neural networks or stochastic gradient boosting machines cannot natively provide such explanations. In that case, we will provide both AUC-based variable importance as well as model-agnostic local interpretations of variable importance using the LIME principle.[31]

Expected Results

We expect to arrive at a generalizable model based on multicenter international data that is likely to predict the abovementioned binary outcomes consistently with an AUC of at least 0.70 (more realistically 0.8 or greater) and that is well-calibrated.[30] A web-based prediction tool will also be created for each of the two models using the *shiny* [32] environment, much akin to e.g. https://neurosurgery.shinyapps.io/impairment. This web-based app will be available for free on any internet-capable device (mobile or desktop), and should be stable on most devices due to the server-based computing. The costs for maintaining the server will be carried by the sponsor.

The collected data will be stored by the sponsor for 10 years. The large dataset will be open to further analysis and will be provided to any of the contributing centers at reasonable request and after approval by all other centers. The goal is to enable other analyses using the collected dataset. If any additional analyses lead to publication, all contributors will be included as co-authors and all co-authors will have the opportunity to review said manuscript beforehand. Any contributing study center has the right to veto publication of any subsequent analyses that includes their data.

References

1. Vlak MH, Algra A, Brandenburg R, Rinkel GJ. Prevalence of unruptured intracranial aneurysms, with emphasis on sex, age, comorbidity, country, and time period: a systematic review and meta-analysis. *Lancet Neurol*. 2011;10(7):626-636. doi:10.1016/S1474-4422(11)70109-0

2. Backes D, Rinkel GJE, Greving JP, et al. ELAPSS score for prediction of risk of growth of unruptured intracranial aneurysms. *Neurology*. 2017;88(17):1600-1606. doi:10.1212/WNL.0000000000003865

3. Etminan N, Brown RD, Beseoglu K, et al. The unruptured intracranial aneurysm treatment score: A multidisciplinary consensus. *Neurology*. 2015;85(10):881-889. doi:10.1212/WNL.0000000000001891

4. Greving JP, Rinkel GJE, Buskens E, Algra A. Cost-effectiveness of preventive treatment of intracranial aneurysms: new data and uncertainties. *Neurology*. 2009;73(4):258-265. doi:10.1212/01.wnl.0b013e3181a2a4ea

5. Greving JP, Wermer MJH, Brown RD, et al. Development of the PHASES score for prediction of risk of rupture of intracranial aneurysms: a pooled analysis of six prospective cohort studies. *The Lancet Neurology*. 2014;13(1):59-66. doi:10.1016/S1474-4422(13)70263-1

6. Juvela S, Porras M, Heiskanen O. Natural history of unruptured intracranial aneurysms: a long-term follow-up study. *Journal of Neurosurgery*. 1993;79(2):174-182. doi:10.3171/jns.1993.79.2.0174

7. Nieuwkamp DJ, Setz LE, Algra A, Linn FHH, de Rooij NK, Rinkel GJE. Changes in case fatality of aneurysmal subarachnoid haemorrhage over time, according to age, sex, and region: a meta-analysis. *Lancet Neurol*. 2009;8(7):635-642. doi:10.1016/S1474-4422(09)70126-7

8. Wermer MJH, van der Schaaf IC, Algra A, Rinkel GJE. Risk of rupture of unruptured intracranial aneurysms in relation to patient and aneurysm characteristics: an updated meta-analysis. *Stroke*. 2007;38(4):1404-1410. doi:10.1161/01.STR.0000260955.51401.cd

9. Backes D, Rinkel GJE, Laban KG, Algra A, Vergouwen MDI. Patient- and Aneurysm-Specific Risk Factors for Intracranial Aneurysm Growth. *Stroke*. 2016;47(4):951-957. doi:10.1161/STROKEAHA.115.012162

10. Esposito G, Regli L. Surgical decision-making for managing complex intracranial aneurysms. *Acta Neurochir Suppl*. 2014;119:3-11. doi:10.1007/978-3-319-02411-0_1

11. Obermeyer Z, Emanuel EJ. Predicting the Future - Big Data, Machine Learning, and Clinical Medicine. *N Engl J Med*. 2016;375(13):1216-1219. doi:10.1056/NEJMp1606181

12. Staartjes VE, Sebök M, Blum PG, et al. Development of machine learning-based preoperative predictive analytics for unruptured intracranial aneurysm surgery: a pilot study. *Acta Neurochir*. Published online May 1, 2020. doi:10.1007/s00701-020-04355-0

13. Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *BMJ*. 2015;350:g7594.

14. Broderick JP, Adeoye O, Elm J. Evolution of the Modified Rankin Scale and Its Use in Future Stroke Trials. *Stroke*. 2017;48(7):2007-2012. doi:10.1161/STROKEAHA.117.017866

15. Clavien PA, Barkun J, de Oliveira ML, et al. The Clavien-Dindo classification of surgical complications: five-year experience. *Ann Surg*. 2009;250(2):187-196. doi:10.1097/SLA.0b013e3181b13ca2

16. Esposito G, Amin-Hanjani S, Regli L. Role of and Indications for Bypass Surgery After Carotid Occlusion Surgery Study (COSS)? *Stroke*. 2016;47(1):282-290. doi:10.1161/STROKEAHA.115.008220

17. Esposito G, Fierstra J, Regli L. Distal outflow occlusion with bypass revascularization: last resort measure in managing complex MCA and PICA aneurysms. *Acta Neurochir (Wien)*. 2016;158(8):1523-1531. doi:10.1007/s00701-016-2868-3

18. Esposito G, Dias SF, Burkhardt J-K, et al. Selection Strategy for Optimal Keyhole Approaches for Middle Cerebral Artery Aneurysms: Lateral Supraorbital Versus Minipterional Craniotomy. *World Neurosurg*. 2019;122:e349-e357. doi:10.1016/j.wneu.2018.09.238

19. Esposito G, Durand A, Van Doormaal T, Regli L. Selective-targeted extra-intracranial bypass surgery in complex middle cerebral artery aneurysms: correctly identifying the recipient artery using indocyanine green videoangiography. *Neurosurgery*. 2012;71(2 Suppl Operative):ons274-284; discussion ons284-285. doi:10.1227/NEU.0b013e3182684c45

20. Esposito G, Fierstra J, Regli L. Partial Trapping Strategies for Managing Complex Intracranial Aneurysms. *Acta Neurochir Suppl*. 2016;123:73-75. doi:10.1007/978-3-319-29887-0_10

21. Jafar JJ, Russell SM, Woo HH. Treatment of giant intracranial aneurysms with saphenous vein extracranial-to-intracranial bypass grafting: indications, operative technique, and results in 29 patients. *Neurosurgery*. 2002;51(1):138-144; discussion 144-146. doi:10.1097/00006123-200207000-00021

22. Lawton MT, Hamilton MG, Morcos JJ, Spetzler RF. Revascularization and aneurysm surgery: current techniques, indications, and outcome. *Neurosurgery*. 1996;38(1):83-92; discussion 92-94. doi:10.1097/00006123-199601000-00020

23. Esposito G, Dias S, Burkhardt J-K, Bozinov O, Regli L. Role of Indocyanine Green Videoangiography in Identification of Donor and Recipient Arteries in Cerebral Bypass Surgery. *Acta Neurochir Suppl*. 2018;129:85-89. doi:10.1007/978-3-319-73739-3_12

24. Templ M, Kowarik A, Alfons A, Prantner B. *VIM: Visualization and Imputation of Missing Values*.; 2019. Accessed January 5, 2020. https://CRAN.R-project.org/package=VIM

25. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: Synthetic Minority Over-sampling Technique. *JAIR*. 2002;16:321-357. doi:10.1613/jair.953

26. Staartjes VE, Schröder ML. Letter to the Editor. Class imbalance in machine learning for neurosurgical outcome prediction: are our models valid? *J Neurosurg Spine*. 2018;29(5):611-612. doi:10.3171/2018.5.SPINE18543

27. Granitto PM, Furlanello C, Biasioli F, Gasperi F. Recursive feature elimination with random forest for PTR-MS analysis of agroindustrial products. *Chemometrics and Intelligent Laboratory Systems*. 2006;83(2):83-90. doi:10.1016/j.chemolab.2006.01.007

28. Perkins NJ, Schisterman EF. The inconsistency of "optimal" cutpoints obtained using two criteria based on the receiver operating characteristic curve. *Am J Epidemiol*. 2006;163(7):670-675. doi:10.1093/aje/kwj063

29. R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing; 2020. https://www.R-project.org/

30. Staartjes VE, Kernbach JM. Letter to the Editor. Importance of calibration assessment in machine learning-based predictive analytics. *J Neurosurg Spine*. Published online February 21, 2020:1-2. doi:10.3171/2019.12.SPINE191503

31. Tulio Ribeiro M, Singh S, Guestrin C. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. *arXiv e-prints*. 2016;1602:arXiv:1602.04938.

32. Chang W, Cheng J, Allaire JJ, et al. *Shiny: Web Application Framework for R*.; 2020. Accessed May 28, 2020. https://CRAN.R-project.org/package=shiny