Development and external validation of a clinical prediction model for functional impairment after intracranial tumor surgery

Victor E. Staartjes, BMed,[1,2] Morgan Broggi, MD, PhD,[3] Costanza Maria Zattra, MD,[3] Flavio Vasella, MD,[1] Julia Velz, MD,[1] Silvia Schiavolin, PsyD,[4] Carlo Serra, MD,[1] Jiri Bartek Jr., MD, PhD,[5–7] Alexander Fletcher-Sandersjöö, MD,[5,6] Petter Förander, MD, PhD,[5,6] Darius Kalasauskas, MD,[8] Mirjam Renovanz, MD,[8] Florian Ringel, MD,[8] Konstantin R. Brawanski, MD,[9] Johannes Kerschbaumer, MD,[9] Christian F. Freyschlag, MD,[9] Asgeir S. Jakola, MD, PhD,[10,11] Kristin Sjåvik, MD, PhD,[12] Ole Solheim, MD, PhD,[13] Bawarjan Schatlo, MD,[14] Alexandra Sachkova, MD,[14] Hans Christoph Bock, MD,[14] Abdelhalim Hussein, MD,[14] Veit Rohde, MD,[14] Marike L. D. Broekman, MD, PhD,[15,16] Claudine O. Nogarede, MSc,[15,16] Cynthia M. C. Lemmens, MD,[17] Julius M. Kernbach, MD,[18] Georg Neuloh, MD,[18] Oliver Bozinov, MD,[1] Niklaus Krayenbühl, MD,[1] Johannes Sarnthein, PhD,[1] Paolo Ferroli, MD,[3] Luca Regli, MD,[1] and Martin N. Stienen, MD, FEBNS[1]

[1]Department of Neurosurgery and Machine Intelligence in Clinical Neuroscience (MICN) Laboratory, Clinical Neuroscience Center, University Hospital Zurich, University of Zurich, Switzerland; [2]Amsterdam UMC, Vrije Universiteit Amsterdam, Neurosurgery, Amsterdam Movement Sciences, Amsterdam, The Netherlands; [3]Department of Neurosurgery, Fondazione IRCCS Istituto Neurologico Carlo Besta, Milan; [4]Neurology, Public Health and Disability Unit, Fondazione IRCCS Istituto Neurologico Carlo Besta, Milan, Italy; [5]Department of Neurosurgery, Karolinska University Hospital, Stockholm; [6]Department of Clinical Neuroscience and Medicine, Karolinska Institutet, Stockholm, Sweden; [7]Department of Neurosurgery, Rigshospitalet, Copenhagen, Denmark; [8]Department of Neurosurgery, University Medical Center, Johannes Gutenberg University Mainz, Germany; [9]Department of Neurosurgery, Medical University of Innsbruck, Austria; [10]Department of Neurosurgery, Sahlgrenska University Hospital, Gothenburg; [11]Institute of Neuroscience and Physiology, Sahlgrenska Academy, Gothenburg, Sweden; [12]Department of Neurosurgery, University Hospital of North Norway, Tromsö; [13]Department of Neurosurgery, St. Olav's University Hospital, Trondheim, Norway; [14]Department of Neurosurgery, Georg August University, University Medical Center, Göttingen, Germany; [15]Department of Neurosurgery, Haaglanden Medical Center, The Hague; [16]Department of Neurosurgery, Leiden University Medical Center, Leiden; [17]Department of Neurology, Haaglanden Medical Center, The Hague, The Netherlands; and [18]Department of Neurosurgery, Faculty of Medicine, RWTH Aachen University, Aachen, Germany

OBJECTIVE Decision-making for intracranial tumor surgery requires balancing the oncological benefit against the risk for resection-related impairment. Risk estimates are commonly based on subjective experience and generalized numbers from the literature, but even experienced surgeons overestimate functional outcome after surgery. Today, there is no reliable and objective way to preoperatively predict an individual patient's risk of experiencing any functional impairment.

METHODS The authors developed a prediction model for functional impairment at 3 to 6 months after microsurgical resection, defined as a decrease in Karnofsky Performance Status of ≥ 10 points. Two prospective registries in Switzerland and Italy were used for development. External validation was performed in 7 cohorts from Sweden, Norway, Germany, Austria, and the Netherlands. Age, sex, prior surgery, tumor histology and maximum diameter, expected major brain vessel or cranial nerve manipulation, resection in eloquent areas and the posterior fossa, and surgical approach were recorded. Discrimination and calibration metrics were evaluated.

RESULTS In the development (2437 patients, 48.2% male; mean age ± SD: 55 ± 15 years) and external validation (2427 patients, 42.4% male; mean age ± SD: 58 ± 13 years) cohorts, functional impairment rates were 21.5% and

ABBREVIATIONS AUC = area under the curve; EOR = extent of resection; KPS = Karnofsky Performance Status; ML = machine learning; PROM = patient-reported outcome measure.

SUBMITTED February 29, 2020. **ACCEPTED** April 6, 2020.

INCLUDE WHEN CITING Published online June 12, 2020; DOI: 10.3171/2020.4.JNS20643.

28.5%, respectively. In the development cohort, area under the curve (AUC) values of 0.72 (95% CI 0.69–0.74) were observed. In the pooled external validation cohort, the AUC was 0.72 (95% CI 0.69–0.74), confirming generalizability. Calibration plots indicated fair calibration in both cohorts. The tool has been incorporated into a web-based application available at https://neurosurgery.shinyapps.io/impairment/.

CONCLUSIONS Functional impairment after intracranial tumor surgery remains extraordinarily difficult to predict, although machine learning can help quantify risk. This externally validated prediction tool can serve as the basis for case-by-case discussions and risk-to-benefit estimation of surgical treatment in the individual patient.

https://thejns.org/doi/abs/10.3171/2020.4.JNS20643

KEYWORDS predictive analytics; outcome prediction; machine learning; functional impairment; neurosurgery; oncology

Patients frequently ask whether they will “stay the same” after the resection of an intracranial tumor—an intricate question often challenging to answer satisfactorily. Clinicians cautiously estimate the likelihood of functional impairment after microsurgical resection by integrating radiological information, anatomo-topographical features, the expected histopathological tumor type, and the complexity of the required surgical approach in view of patient-intrinsic characteristics, generalized numbers from the literature, and the surgeon’s own expertise and experience. The answer to this question plays a critical role in the shared decision-making process.

Among multiple centers and surgeons, considerable diversity exists in treatment protocols, surgical techniques, experience, and equipment, which relate to the achieved extent of resection (EOR), survival, and functional and patient-reported outcome measures (PROMs).[1–7] Today, evidence is accumulating regarding the lower oncological benefit of complete resection in cases of postoperative neurological and/or functional worsening,[8,9] emphasizing the importance of periprocedural safety and the regimen of maximum safe resection, which means aiming for the greatest EOR that allows for preservation of neurological function.[5]

Functional impairment after intracranial tumor surgery is an extraordinarily difficult outcome to predict, and neurooncological surgeons often overestimate postoperative functional outcome.[2,10] Currently, risk estimation is based on prior experiences and generalizable rates from the literature, but outcome prediction tailored to a patient’s specific features is increasingly becoming a part of modern-precision “personalized medicine.”[11–13] Recently, machine learning (ML) methods have been applied to generate patient-specific predictive analytics for outcomes in neurosurgery, and these often outperform classification schemes and conventional modeling techniques such as logistic regression.[11–16] The present study aimed to develop and externally validate a novel prediction model that forecasts individualized postoperative functional impairment from a set of variables usually available at the time of preoperative informed patient consent.

Methods

Overview

From a large bicentric sample of patients who underwent microsurgical resection of intracranial tumors, we developed an ML-based prediction tool for new postoperative functional impairment. The prediction tool was externally validated with data from 7 European centers. This study was compiled according to the TRIPOD (transparent reporting of a multivariable prediction model for individual prognosis or diagnosis) statement.

Ethical Considerations

The scientific workup of registry data was approved by the IRBs of all informed institutions. The study was registered at the University Hospital Zurich (clinicaltrials.gov identifier NCT01628406). Patients provided informed consent or informed consent was waived, depending on the demands of the local IRB.

Data Sources

Prospective institutional databases from 2 centers were retrospectively analyzed. Consecutive patients undergoing microsurgical resection of intracranial tumors via microscopic craniotomy or transsphenoidal surgery were included. Diagnostic biopsies were excluded. We pooled data from patients undergoing surgery between January 2013 and December 2017 at the Department of Neurosurgery, University Hospital Zurich, Switzerland, and between January 2014 and December 2017 at the Department of Neurosurgery, Fondazione IRCCS Istituto Neurologico Carlo Besta in Milan, Italy. The methodological details of these 2 patient registries were described previously.[2,6,17] Physicians who collected the registry and outcome data in these registries were specifically trained; internal standard operating procedures additionally helped with harmonizing the data collection. Data quality in the registries was regularly reviewed and improved as required. All patients in the derivation cohort had the required variables recorded; there was no need to delete cases or impute missing data.

The use of intraoperative technology to increase EOR while monitoring neurological function, e.g., intraoperative imaging (ultrasound, magnetic resonance imaging, neuronavigation, fluorescence guidance, etc.), electrophysiological monitoring, or awake surgery, is routinely applied in addition to the use of surgical tools (e.g., intraoperative microscope, ultrasonic aspirators).[3,5,18–21]

The model was evaluated in 7 centers from 5 countries. Göttingen (2014–2017), Innsbruck (2015–2018), and Leiden and the Hague (2015–2018) data were derived from prospective registries. Trondheim data (2007–2015) were based on a prospective registry supplemented with retrospectively collected data. Stockholm (2007–2015), Mainz (2007–2018), and Aachen (2018) data were retrospectively collected. To improve the realistic representation of external validation model performance, neurosurgeons who collected data for the external validation cohort were not specifically trained, apart from receiving the same detailed variable definitions as described in this *Methods* section and as listed in the web-based application. All participating centers pursue a "maximum safe resection" philosophy.[5]

Outcome Measures

The primary outcome measure was "new postoperative functional impairment," defined as a 10-point or greater decrease in Karnofsky Performance Status (KPS) at 3 to 6 months postoperatively, compared with preoperative functional status.[2] There is no established minimal clinically important difference for KPS after intracranial tumor surgery. We deliberately chose the 10-point cutoff,[2] as opposed to a dynamic cutoff with different steps depending on baseline status,[22] in order not to overlook subtle differences in performance, since even minor decreases in performance as judged by clinical scales can be perceived as devastating by patients.[7]

Recorded variables included KPS at admission and at 3 to 6 months, age, sex, prior surgery, tumor type and maximum diameter, expected major vessel or cranial nerve manipulation, surgery in the posterior fossa, resection in an eloquent area, and whether a transsphenoidal or transcranial resection was performed. We defined major brain vessel manipulation as the expected manipulation of major vessels encased by or in proximity to the tumor. Major vessels included the internal carotid artery; the anterior, middle, and posterior cerebral arteries; the basilar and vertebral arteries; and the large venous sinuses and internal, Trolard, and Labbé veins. Eloquent areas were defined as motor, sensory, language, or visual areas, as well as the hypothalamus, thalamus, internal capsule, brainstem, and pineal region.[2] These variables were chosen as inputs for the model due to their demonstrated relationships to functional impairment, and their number was limited to ensure the practical applicability of the prediction model.[2]

Model Development and Validation

Continuous data are reported as mean ± SD or median (IQR) and categorical data as numbers (percentages). Nondichotomous categorical input variables were one-hot encoded. Numerical input variables were standardized using centering and scaling.

A logistic generalized additive model based on locally estimated scatterplot smoothing was developed on the derivation cohort to predict any functional impairment, using the "caret" and "gam" packages.[23–26] The model parameters were fitted in 50 bootstrap resamples with replacement, hyperparameters were tuned, and the final model was selected based on the area under the curve (AUC). The final model had a span of 0.5. A *k*-nearest neighbors algorithm was trained on the derivation set to impute any potential missing data during prediction on new data.[27] The threshold for binary classification was selected on the derivation cohort based on the "closest to(0,1)" criterion.[28]

The prediction model was subsequently externally validated. No recalibration was carried out.[29] When predicting on the external validation cohort, the cotrained *k*-nearest neighbors algorithm was applied to impute missing data.[27] Calibration was visually assessed using calibration plots. Quantile-based 95% confidence intervals of the discrimination and calibration metrics were obtained in 1000 bootstrap resamples.

All analyses were carried out in R version 3.5.2 (The R Foundation for Statistical Computing). The *Supplementary Methods* contains the statistical code.

Results

Derivation Cohort

A total of 2437 patients were available in the 2 prospective registries. There were no missing data. The mean patient age was 55 ± 15 years, and 1175 patients (48.2%) were male. The median KPS at admission was 90 (IQR 80–90), and 440 patients (18.1%) had undergone prior surgery. The majority of patients (2148, 88.1%) underwent open craniotomy, while 289 patients (11.9%) underwent transsphenoidal surgery. New functional impairment was observed in 525 patients (21.5%). Early mortality occurred in 85 patients (3.5%). Detailed patient characteristics are provided in Table 1.

External Validation Cohort

Seven centers in 5 countries provided data for external validation. The external validation cohort comprised 2427 patients. Patient characteristics per center are provided in Supplementary Table S1. Overall, 392 of 26,697 baseline data fields (1.5%) were incomplete, and the primary outcome was available for all patients. The mean patient age was 58 ± 13 years, and 1023 patients (42.4%) were male. The median admission KPS was 80 (IQR 70–90). Three hundred six patients (12.6%) had undergone prior surgery. Open craniotomy was carried out in 2326 patients (95.8%), while 101 patients (4.2%) underwent transsphenoidal surgery. In the external validation cohort, the rate of functional impairment was 28.5% (n = 692). Early mortality occurred in 74 cases (3.1%).

Model Performance

The prediction model resulted in an AUC of 0.72 (95% CI 0.69–0.74) on the derivation cohort (Fig. 1). A threshold of 0.205 for binary classification of functional impairment was determined based on the AUC. A sensitivity and specificity of 0.73 (95% CI 0.69–0.77) and 0.59 (95% CI 0.57–0.62), respectively, were observed (Table 2). The prediction model was well calibrated on the development cohort, with a calibration slope of 1.01 (95% CI 0.87–1.15) and intercept of −0.00 (95% CI −0.10 to 0.10) (Fig. 2).

In the external validation cohort, a pooled AUC of 0.72 (95% CI 0.69–0.74) was observed. The sensitivity and specificity amounted to 0.62 (95% CI 0.59–0.66) and 0.70 (95% CI 0.67–0.72), respectively. Among the external validation centers, AUC values ranged from 0.54 (95% CI 0.47–0.61) to 0.78 (95% CI 0.73–0.82). In terms of calibration, a slope of 0.88 (95% CI 0.77–0.99) and intercept of 0.58 (95% CI 0.48–0.67) were observed. Location in

TABLE 1. Patient characteristics and incidence of functional impairment

	Cohort	
Variable	Development (n = 2437)	External Validation (n = 2427)
Male sex	1175 (48.2)	1023 (42.4)
No. missing	0 (0.0)	12 (0.5)
Age, yrs		
Mean ± SD	54.6 ± 15.3	58.2 ± 13.3
Median (IQR)	55 (44–67)	59 (49–68)
Range	18–92	18–91
No. missing	0 (0.0)	2 (0.1)
Maximum tumor diameter, cm		
Mean ± SD	3.5 ± 1.6	3.7 ± 1.7
Median (IQR)	3.2 (2.3–4.5)	3.5 (2.5–4.9)
Range	0.1–10.0	0.3–10.2
No. missing	0 (0.0)	3 (0.1)
Histology		
Meningioma	636 (26.1)	1348 (55.5)
Glioblastoma	514 (21.1)	554 (22.8)
Metastasis	324 (13.3)	259 (10.7)
Adenoma	243 (10.0)	103 (4.2)
Low-grade glioma	121 (5.0)	44 (1.8)
Schwannoma	120 (4.9)	35 (1.4)
Anaplastic astrocytoma	112 (4.6)	48 (2.0)
Craniopharyngioma	39 (1.6)	2 (0.1)
(Epi-)dermoid cyst	30 (1.2)	6 (0.2)
Chordoma	25 (1.0)	0 (0.0)
Other	273 (11.2)	28 (1.2)
No. missing	0 (0.0)	0 (0.0)
Prior surgery	440 (18.1)	306 (12.6)
No. missing	0 (0.0)	2 (0.1)
Open craniotomy	2148 (88.1)	2326 (95.8)
No. missing	0 (0.0)	0 (0.0)
Surgery in eloquent area	1197 (49.1)	879 (36.2)
No. missing	0 (0.0)	1 (0.0)
Brain vessel manipulation	898 (36.8)	995 (41.0)
No. missing	0 (0.0)	185 (7.6)
Cranial nerve manipulation	715 (29.3)	487 (20.1)
No. missing	0 (0.0)	185 (7.6)
Surgery in posterior fossa	413 (16.9)	361 (14.9)
No. missing	0 (0.0)	1 (0.0)
KPS at admission		
Mean ± SD	84.3 ± 13.9	82.0 ± 13.9
Median (IQR)	90 (80–90)	80 (70–90)
Range	20–100	10–100
No. missing	0 (0.0)	1 (0.0)
New functional impairment*	525 (21.5)	692 (28.5)

Values represent the number of patients (%) unless stated otherwise.

* New functional impairment was defined as a ≥ 10-point decrease in KPS from baseline to the 3-month follow-up.

an eloquent area, surgical approach, tumor histology, KPS at admission, and sex demonstrated the highest variable importance in the prediction model (Supplementary Table S2). Partial dependence plots for each variable are provided in Supplementary Figure S1.

Model Deployment

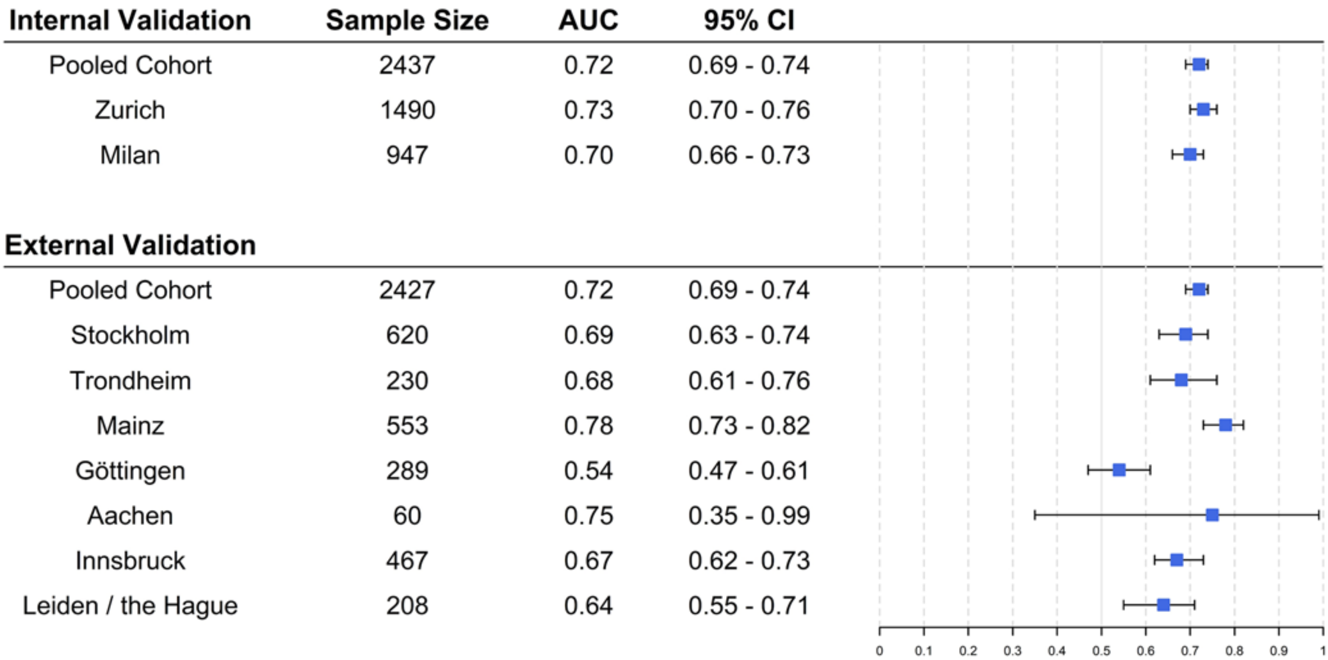
The prediction model was integrated into a free, user-friendly, web-based application accessible at https://neurosurgery.shinyapps.io/impairment/.

Discussion

Prediction tools can assist in the shared surgical decision-making process.[11–14] Compared with other pathologies, where scoring systems are broadly applied to estimate postoperative outcome (e.g., for arteriovenous malformations[30] or intracranial aneurysms[15]), there is little research on classification or prediction tools for postoperative functional impairment after resection of intracranial tumors. In addition, what is known about postoperative functional impairment usually focuses on a particular histopathological entity instead of principles that apply to various kinds of intracranial neoplastic lesions. The Milan Complexity Scale is a classification system based on objective surgical complexity, which correlates with the risk of functional impairment.[2] The scale can help judge case complexity and thus provides benchmarks for complication risk, resident training, and health system management.[31] We expanded on this concept by applying ML techniques to multicentric data and incorporating additional variables in a nonlinear fashion. Learning of nonlinear structures in the data may reveal patterns that linear models are blind to, potentially leading to better predictions.[14]

No tools exist to enable the prediction of an individual patient's risk of functional impairment after intracranial tumor surgery. Experienced clinicians are proficient at judging this risk by integrating clinical and imaging findings and the proposed procedure into their personal pool of experience. However, studies assessing the accuracy of these subjective predictions have raised concern about the accuracy of the information available to patients at preoperative informed consent. It appears that neurosurgeons tend to overestimate patients' postoperative functional status.[10] Our study provides a first objective benchmark of this accuracy and the functional result that can be expected by patients. The free web-based application can be used by physicians and patients alike as a basis for individual case-by-case discussions of the risk-to-benefit estimation of surgical treatment.

From specific pathologies such as pituitary adenomas, we know that classification systems and experienced clinicians are usually adept at identifying patients who are at either very high or low risk of a certain endpoint.[2,16] Thus, they excel at identifying extreme cases, such as large glioblastomas in eloquent areas, but are less successful in differentiating between good and bad outcomes in cases with moderate risk, such as diffuse low-grade gliomas in non-eloquent areas but adjacent to critical structures. The hope is that ML enables better differentiation in these moderate cases, leading to more accurate predictions.[16] This notion



Internal Validation	Sample Size	AUC	95% CI
Pooled Cohort	2437	0.72	0.69 - 0.74
Zurich	1490	0.73	0.70 - 0.76
Milan	947	0.70	0.66 - 0.73
External Validation			
Pooled Cohort	2427	0.72	0.69 - 0.74
Stockholm	620	0.69	0.63 - 0.74
Trondheim	230	0.68	0.61 - 0.76
Mainz	553	0.78	0.73 - 0.82
Göttingen	289	0.54	0.47 - 0.61
Aachen	60	0.75	0.35 - 0.99
Innsbruck	467	0.67	0.62 - 0.73
Leiden / the Hague	208	0.64	0.55 - 0.71

FIG. 1. AUC values of the prediction model among the different centers. AUC values are provided with bootstrapped 95% confidence intervals. Figure is available in color online only.

is corroborated by a systematic review demonstrating that artificial intelligence, including ML, is often superior to experienced raters (coined "natural intelligence") in terms of neurosurgical decision-making.[32] Notably, in studies in which clinical experts assisted by ML models were compared with clinical experts alone, the ML-assisted group consistently performed better.[32]

This underlines that prediction models such as ours are not meant to be used as absolute red or green lights, but rather as a supplement to neurosurgeons' clinical expertise. The current model mainly provides the ability to rule out functional impairment at 3 to 6 months postoperatively, due to its relatively high negative predictive value. However, the objective risk estimates produced by the model are more informative than the derived binary classifications. For example, a predicted risk of functional impairment of 55% may not accurately classify patients in a binary fashion but may be useful to communicate a relatively high risk of impairment to a patient. The risk estimates our model calculates appear well calibrated. In the external validation cohort, major heterogeneities were observed, including a higher rate of new functional impairment, which explains the calibration intercept of 0.58 observed at external validation. This would mean that—because the incidence of functional impairment was 33% higher in the external validation cohort—the model slightly underestimates functional impairment in this new cohort. For example, in a different cohort with a massively increased incidence of functional impairment of 42%, the model would predict an impairment risk of 10%, while the actual risk would be around 20%. This phenomenon is frequently observed and in fact is unavoidable unless the variables that explain the increased rate of functional impairment, such as potentially center caseload or surgeon experience, are included in the model.[29,33] The calibration intercept at external validation can be artificially improved by recalibrating onto the new population by changing mode intercepts. We chose not to recalibrate our model to the external validation data in order to evaluate its external validity in a more realistic setup. Still, the calibration of our model appears to generalize well in terms of slope, and when applying the prediction model to different demographics with different rates of new functional

TABLE 2. Discrimination and calibration metrics of the ML-based prediction model

	Cohort	
Metric	Development (n = 2437)	External Validation (n = 2427)
Discrimination		
AUC	0.72 (0.69 to 0.74)	0.72 (0.69 to 0.74)
Accuracy	0.62 (0.60 to 0.64)	0.68 (0.66 to 0.69)
Sensitivity	0.73 (0.69 to 0.77)	0.62 (0.59 to 0.66)
Specificity	0.59 (0.57 to 0.62)	0.70 (0.67 to 0.72)
PPV	0.33 (0.30 to 0.36)	0.45 (0.42 to 0.48)
NPV	0.89 (0.87 to 0.90)	0.82 (0.80 to 0.84)
Calibration		
Intercept	−0.00 (−0.10 to 0.10)	0.58 (0.48 to 0.67)
Slope	1.01 (0.87 to 1.15)	0.88 (0.77 to 0.99)

NPV = negative predictive value; PPV = positive predictive value.
Metrics are provided with bootstrapped 95% confidence intervals.

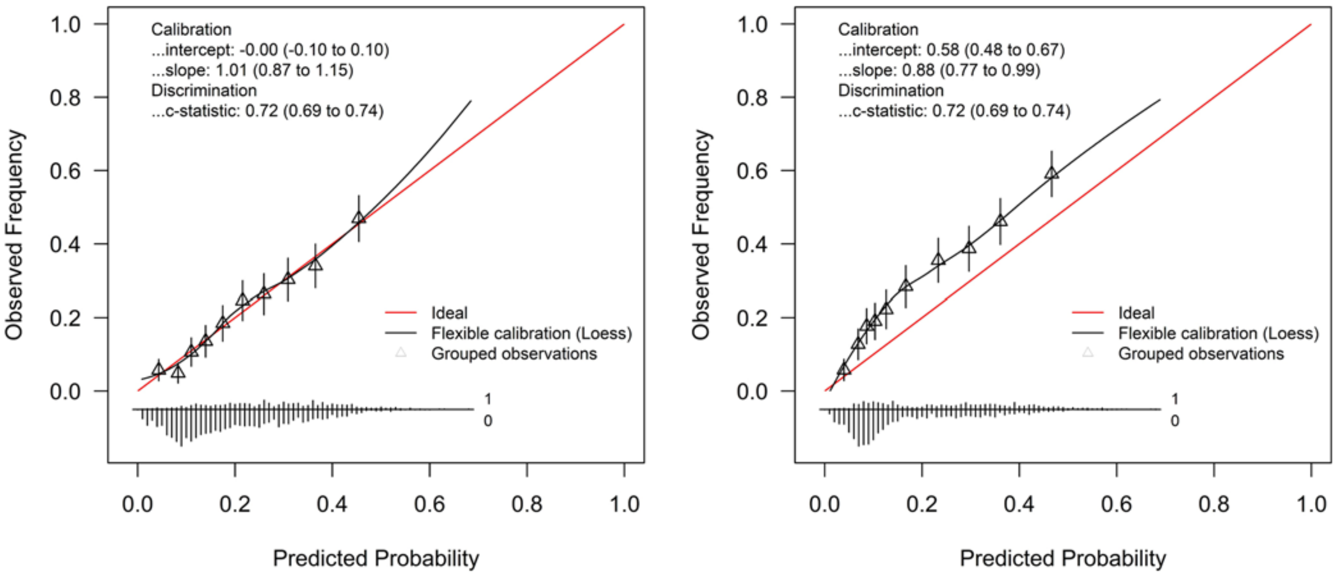


FIG. 2. Calibration curves of the prediction model on the internal (**left**) and external (**right**) validation cohorts. The predicted probabilities for functional impairment are distributed into 10 equally sized groups and contrasted with the actual observed frequencies of functional impairment. Calibration intercept and slope are calculated. A perfectly calibrated model has a calibration intercept of 0 and slope of 1. The calibration intercept is influenced by the frequency of the outcome of interest in a certain population. Metrics are provided with bootstrapped 95% confidence intervals. Figure is available in color online only.

impairment, the model can be recalibrated by updating its intercept accordingly or by other rescaling techniques.[29,33]

Even with a large amount of development data and the application of ML techniques, functional impairment after intracranial tumor surgery remains difficult to predict with high reliability. One likely cause is the lack of functional anatomo-topographical data as inputs for our model, which was designed to include only a few simple, preoperatively and easily available variables. This was intended to keep it applicable to primary care and other nonneurosurgery physicians, who are typically the first and most important contact for patients facing the new diagnosis of an intracranial tumor. The introduction of anatomical features and the ability to account for intraoperative parameters and complications in a second postoperative model would surely improve performance to some extent.

In the case of intracranial tumor surgery, a key factor for variability is the use of different treatment protocols. Different surgical approaches, availability of intraoperative imaging, functional mapping, and use of fluorescents, as well as varying “aggressiveness” in terms of resection but also handling of critical structures, introduce biases that are difficult to statistically account or adjust for.[3–5,18–21] Depending on case complexity, surgical experience may also influence outcome.[31] Even an externally validated prediction model lacks generalizability to cohorts with radically different treatment protocols.

An often-cited drawback of ML models is the inability to understand why a certain prediction has been generated. Whereas logistic regression models provide interpretable odds ratios, ML models are often considered “black boxes”—that is, inputs and outputs are known, but the internal decision-making process is not necessarily interpretable. Some insight can be gained by assessing overall variable importance (Supplementary Table S2). Additionally, generalized additive models are somewhat of an exception, since one can exploit their inherent additivity to examine each variable for the purpose of inference (see Supplementary Figure S1).[23,24] Surgery in eloquent areas may double the rate of postoperative functional impairment as high-grade tumors do,[2,7,34] and preoperative status has been demonstrated to relate to complications and outcome.[2,6,35] It is not always feasible for clinicians to integrate these many independent risk factors into a single communicable risk for outcomes such as impairment. Prediction tools represent an interface between these patient factors with complex interactions and output a risk that is interpretable and clinically useful to clinicians and patients alike.[12,13]

Decision-making for intracranial tumor surgery requires balancing oncological benefit against the risk of resection-related impairment. Our study demonstrates that ML-based prediction of functional impairment is feasible and externally valid with simple inputs. Integrating artificial intelligence as supportive means into the clinical routine is likely to provide valuable improvements in patient information, objective risk assessment, and shared surgical decision-making.

Strengths and Limitations

Our study used data sets from 9 large institutional registries of national referral centers, encompassing several different cultural and linguistic regions. Variable definitions were unified in all centers, allowing us to generate results with fair external validity and generalizability. The primary outcome of our study was based on a clearly defined and well-established outcome measure that correlates with PROMs.[6,7,36] The final model is accessible as a free web-based tool, allowing clinicians and patients to access the objective risk estimates.

A range of tumor types was analyzed, which may bias our prediction model toward more common tumor types, whereas performance may be limited for the less frequently included tumor types. However, the resulting model enables outcome prediction for most major classes of intracranial tumors. In addition, one might expect especially pituitary adenomas and recurrent craniotomies to exhibit an inherently different risk profile, potentially limiting performance of the model. However, we found that their inclusion did not alter overall model performance. In addition, the local regression algorithm on which our model relies is limited in terms of extrapolation to unseen, extreme input variable values.[23,24] For this reason, predictions made from inputs not available in the derivation data, such as ages older than 92 years and tumor sizes greater than 10 cm, should be cautiously interpreted.

Although external validation was successful, no conclusions can be drawn regarding performance in centers with radically different resection protocols and vastly different rates of new functional impairment. The high negative predictive value can be seen as one of the model's strengths. However, predictive values are inherently dependent on the prevalence of the outcome and, as such, the setting in which the prognostic model is used.[29] The predictive values should therefore be interpreted with caution, especially when generalizing to other centers.

Although all participating centers followed a "maximum safe resection" philosophy, potential nuances in EOR may persist, which were not accounted for.[5] We only assessed outcomes at 3 to 6 months postoperatively, and the outcome definition did not include further, relevant aspects such as quality of life, cognitive or work status, and PROMs. Additionally, as with most outcome measures, the interrater agreement of the KPS has been debated, with generally better interrater agreement compared with ECOG (Eastern Cooperative Oncology Group) and palliative performance status.[37] Lastly, the study protocol of this analysis was not prospectively registered.

Conclusions

Functional impairment after intracranial tumor surgery is extraordinarily difficult to predict preoperatively. An ML-based approach resulted in a prediction model capable of forecasting individualized risk for any functional impairment at 3 to 6 months postoperatively with fair performance. Extensive external validation demonstrated the high generalizability of the prediction model. To our knowledge, this study is the first externally validated attempt at preoperatively quantifying the "patient-specific" surgical risk for any functional impairment after intracranial tumor surgery. The web-based application can be used by physicians and patients alike, serving as a basis for case-by-case discussions on the risk-to-benefit estimation of surgical treatment.

Acknowledgments

We thank the patients whose anonymized data were used for this research.

The following persons contributed in establishing the data collection. David Y. Zhang and Dominik Seggewiss: Department of Neurosurgery, Clinical Neuroscience Center, University Hospital Zurich, University of Zurich, Switzerland; Stefano Villa: Department of Neurosurgery, Fondazione IRCCS Istituto Neurologico Carlo Besta, Milan, Italy; Rishi Nandoe Tewarie and Fred Kloet: Department of Neurosurgery, Haaglanden Medical Center, The Hague, The Netherlands, and Department of Neurosurgery, Leiden University Medical Center, Leiden, The Netherlands.

References

1. Barker FG II, Curry WT Jr, Carter BS. Surgery for primary supratentorial brain tumors in the United States, 1988 to 2000: the effect of provider caseload and centralization of care. *Neuro Oncol*. 2005;7(1):49–63.
2. Ferroli P, Broggi M, Schiavolin S, et al. Predicting functional impairment in brain tumor surgery: the Big Five and the Milan Complexity Scale. *Neurosurg Focus*. 2015;39(6):E14.
3. Yordanova YN, Moritz-Gasser S, Duffau H. Awake surgery for WHO Grade II gliomas within "noneloquent" areas in the left dominant hemisphere: toward a "supratotal" resection. Clinical article. *J Neurosurg*. 2011;115(2):232–239.
4. Sanai N, Berger MS. Glioma extent of resection and its impact on patient outcome. *Neurosurgery*. 2008;62(4):753–764, 264–266.
5. Marko NF, Weil RJ, Schroeder JL, et al. Extent of resection of glioblastoma revisited: personalized survival modeling facilitates more accurate survival prediction and supports a maximum-safe-resection approach to surgery. *J Clin Oncol*. 2014;32(8):774–782.
6. Stienen MN, Zhang DY, Broggi M, et al. The influence of preoperative dependency on mortality, functional recovery and complications after microsurgical resection of intracranial tumors. *J Neurooncol*. 2018;139(2):441–448.
7. Schiavolin S, Raggi A, Scaratti C, et al. Patients' reported outcome measures and clinical scales in brain tumor surgery: results from a prospective cohort study. *Acta Neurochir (Wien)*. 2018;160(5):1053–1061.
8. Rahman M, Abbatematteo J, De Leo EK, et al. The effects of new or worsened postoperative neurological deficits on survival of patients with glioblastoma. *J Neurosurg*. 2017;127(1):123–131.
9. Jakola AS, Gulati S, Weber C, et al. Postoperative deterioration in health related quality of life as predictor for survival in patients with glioblastoma: a prospective study. *PLoS One*. 2011;6(12):e28592.
10. Sagberg LM, Drewes C, Jakola AS, Solheim O. Accuracy of operating neurosurgeons' prediction of functional levels after intracranial tumor surgery. *J Neurosurg*. 2017;126(4):1173–1180.
11. Obermeyer Z, Emanuel EJ. Predicting the future—big data, machine learning, and clinical medicine. *N Engl J Med*. 2016;375(13):1216–1219.
12. Galovic M, Stauber AJ, Leisi N, et al. Development and validation of a prognostic model of swallowing recovery and enteral tube feeding after ischemic stroke. *JAMA Neurol*. 2019;76(5):561–570.
13. Khor S, Lavallee D, Cizik AM, et al. Development and validation of a prediction model for pain and functional outcomes after lumbar spine surgery. *JAMA Surg*. 2018;153(7):634–642.
14. Senders JT, Staples PC, Karhade AV, et al. Machine learning and neurosurgical outcome prediction: a systematic review. *World Neurosurg*. 2018;109:476–486.e1.
15. Jaja BNR, Saposnik G, Lingsma HF, et al. Development and validation of outcome prediction models for aneurysmal subarachnoid haemorrhage: the SAHIT multinational cohort study. *BMJ*. 2018;360:j5745.

16. Staartjes VE, Serra C, Muscas G, et al. Utility of deep neural networks in predicting gross-total resection after transsphenoidal surgery for pituitary adenoma: a pilot study. *Neurosurg Focus*. 2018;45(5):E12.
17. Sarnthein J, Stieglitz L, Clavien P-A, Regli L. A patient registry to improve patient safety: recording general neurosurgery complications. *PLoS One*. 2016;11(9):e0163154.
18. Stummer W, Stepp H, Wiestler OD, Pichlmeier U. Randomized, prospective double-blinded study comparing 3 different doses of 5-aminolevulinic acid for fluorescence-guided resections of malignant gliomas. *Neurosurgery*. 2017;81(2):230–239.
19. Kubben PL, ter Meulen KJ, Schijns OE, et al. Intraoperative MRI-guided resection of glioblastoma multiforme: a systematic review. *Lancet Oncol*. 2011;12(11):1062–1070.
20. Gronningsaeter A, Kleven A, Ommedal S, et al. SonoWand, an ultrasound-based neuronavigation system. *Neurosurgery*. 2000;47(6):1373–1380.
21. Sanai N, Mirzadeh Z, Berger MS. Functional outcome after language mapping for glioma resection. *N Engl J Med*. 2008;358(1):18–27.
22. Nghiemphu PL, Liu W, Lee Y, et al. Bevacizumab and chemotherapy for recurrent glioblastoma: a single-institution experience. *Neurology*. 2009;72(14):1217–1222.
23. Hastie T, Tibshirani R. *Generalized Additive Models*. 1st ed. Chapman & Hall; 1990.
24. Hastie T, Tibshirani R, Friedman J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Science & Business Media; 2013.
25. Hastie T. gam: generalized additive models. 2019. Accessed April 22, 2020. https://CRAN.R-project.org/package=gam
26. Kuhn M. Building predictive models in R using the caret package. *J Stat Softw*. 2008;28(5).
27. Batista GEAPA, Monard MC. An analysis of four missing data treatment methods for supervised learning. *Appl Artif Intell*. 2003;17(5–6):519–533.
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.
29. Janssen KJM, Moons KGM, Kalkman CJ, et al. Updating methods improved the performance of a clinical prediction model in new patients. *J Clin Epidemiol*. 2008;61(1):76–86.
30. Spetzler RF, Martin NA. A proposed grading system for arteriovenous malformations. *J Neurosurg*. 1986;65(4):476–483.
31. Vasella F, Velz J, Neidert MC, et al. Safety of resident training in the microsurgical resection of intracranial tumors: data from a prospective registry of complications and outcome. *Sci Rep*. 2019;9(1):954.
32. Senders JT, Arnaout O, Karhade AV, et al. Natural and artificial intelligence in neurosurgery: a systematic review. *Neurosurgery*. 2018;83(2):181–192.
33. van Rein EAJ, van der Sluijs R, Voskens FJ, et al. Development and validation of a prediction model for prehospital triage of trauma patients. *JAMA Surg*. 2019;154(5):421–429.
34. Duffau H, Capelle L, Denvil D, et al. Functional recovery after surgical resection of low grade gliomas in eloquent brain: hypothesis of brain compensation. *J Neurol Neurosurg Psychiatry*. 2003;74(7):901–907.
35. Chang SM, Parney IF, McDermott M, et al. Perioperative complications and neurological outcomes of first and second craniotomies among patients enrolled in the Glioma Outcome Project. *J Neurosurg*. 2003;98(6):1175–1181.
36. Reponen E, Tuominen H, Korja M. Evidence for the use of preoperative risk assessment scores in elective cranial neurosurgery: a systematic review of the literature. *Anesth Analg*. 2014;119(2):420–432.
37. Chow R, Chiu N, Bruera E, et al. Inter-rater reliability in performance status assessment among health care professionals: a systematic review. *Ann Palliat Med*. 2016;5(2):83–92.

Disclosures

The authors report no conflict of interest concerning the materials or methods used in this study or the findings specified in this paper.

Author Contributions

Conception and design: Staartjes, Broggi, Ferroli, Regli, Stienen. Acquisition of data: all authors. Analysis and interpretation of data: Staartjes, Regli, Stienen. Drafting the article: Staartjes. Critically revising the article: all authors. Reviewed submitted version of manuscript: all authors. Approved the final version of the manuscript on behalf of all authors: Staartjes. Statistical analysis: Staartjes, Kernbach. Administrative/technical/material support: Staartjes, Regli, Stienen. Study supervision: Regli, Stienen.

Supplemental Information

Online-Only Content

Supplemental material is available with the online version of the article.

Supplementary Data. https://thejns.org/doi/suppl/10.3171/2020.4.JNS20643.

Correspondence

Victor E. Staartjes: Machine Intelligence in Clinical Neuroscience (MICN) Laboratory, University Hospital Zurich, Switzerland. victoregon.staartjes@usz.ch.