Title: Predicting cognitive function three months after surgery in patients with a glioma

Running title: Predicting cognitive function after treatment

Authors: Sander Martijn Boelders[1,2], Bruno Nicenboim[2], Elke Butterbrod[1]. Wouter de Baene[3], Eric Postma[2], Geert-Jan Rutten[1], Lee-Ling Ong[2], Karin Gehring[1,3]

Affiliations:

[1]Department of Neurosurgery, Elisabeth-TweeSteden Hospital, Tilburg, The Netherlands
[2]Department of Cognitive Sciences and AI, Tilburg University, Tilburg, The Netherlands
[3]Department of Cognitive Neuropsychology, Tilburg University, Tilburg, The Netherlands

Corresponding author: Karin Gehring, PhD, Department of Neurosurgery, Elisabeth-Tweesteden Hospital/Cognitive Neuropsychology, Tilburg University, P.O. Box 90153, Warandelaan 2, Tilburg, The Netherlands, 5000 LE, (k.gehring@tilburguniversity.edu).

Introduction: Patients with a glioma often suffer from cognitive impairments both before and after anti-tumor treatment. Ideally, clinicians can rely on predictions of post-operative cognitive functioning for individual patients based on information obtainable before surgery. Such predictions would facilitate selecting the optimal treatment considering patients' onco-functional balance.

Method: Cognitive functioning three months after surgery was predicted for 317 patients with a glioma across eight cognitive tests. Nine multivariate Bayesian regression models were used following a machine-learning approach while employing pre-operative neuropsychological test scores and a comprehensive set of clinical predictors obtainable before surgery. Model performances were compared using the Expected Log Pointwise Predictive Density (ELPD), and pointwise predictions were assessed using the Coefficient of Determination (R^2) and Mean Absolute Error. Models were compared against models employing only pre-operative cognitive functioning and the best-performing model was interpreted. Moreover, an example prediction including uncertainty for clinical use was provided.

Results: The best-performing model obtained a median R^2 of 34.20%. Individual predictions, however, were uncertain. Pre-operative cognitive functioning was the most influential predictor. Models including clinical predictors performed similarly to those using only pre-operative functioning (ΔELPD 14.4±10.0, ΔR^2 -0.53%.).

Conclusion: Post-operative cognitive functioning cannot yet reliably be predicted from pre-operative cognitive functioning and the included clinical predictors. Moreover, predictions relied strongly on pre-operative cognitive functioning. Consequently, clinicians should not rely on the included predictors to infer patients' cognitive functioning after treatment. Moreover, it stresses the need to collect larger cross-center multimodal datasets to obtain more certain predictions for individual patients.

Keywords: Cognitive function after treatment, glioma, individual predictions, machine learning, Bayesian regression

Importance of the study: Patients with a glioma often suffer from cognitive impairments both before and after anti-tumor treatment. Ideally, clinicians would be able to rely on predictions of cognitive functioning after treatment for individual patients based on information that is obtainable before surgery. Such predictions would facilitate selecting the optimal treatment considering patients' onco-functional balance and could improve patient counseling. First, our study shows that cognitive functioning three months after surgery cannot be reliably predicted from pre-operative cognitive functioning and the included clinical predictors, with pre-operative cognitive functioning being the most important predictor. Consequently, clinicians should not rely on the included predictors to infer individual patients' cognitive functioning after surgery. Second, results demonstrate how individual predictions resulting from Bayesian models, including their uncertainty estimates, may ultimately be used in clinical practice. Third, our results show the importance of collecting additional predictors and stress the need to collect larger cross-center multimodal datasets.

Key points:

- Cognitive functioning after treatment cannot yet reliably be predicted
- Pre-operative cognitive functioning was the most important predictor
- Additional predictors and larger cross-center datasets are needed

Introduction

Patients with a glioma often suffer from cognitive impairments, both before and after anti-tumor treatment[1,2], which may contribute to a decreased quality of life[3–5]. Cognitive impairments after anti-tumor treatment are likely caused by the damage inflicted by the tumor before surgery[6,7], the surgical resection[8], and adjuvant therapies[9,10]. Moreover, cognitive functioning after anti-tumor treatment has been related to numerous patient characteristics, such as age, education, and medicine use[11], and cognitive functioning before surgery appears to be one of the strongest indicators of post-operative functioning[12,13]. Unfortunately, the exact mechanisms by which glioma affect cognitive functioning after treatment remain poorly understood.

The consideration of cognitive functioning is becoming increasingly important in determining the optimal treatment in view of patients' onco-functional balance. This onco-functional balance refers to weighing the oncological benefit of treatment against its adverse side effects on the functional status and quality of life of the patient[14]. Ideally, clinicians would be able to use predictions of cognitive functioning after treatment to facilitate selecting the optimal treatment[13,15–17].

Unfortunately, achieving accurate predictions of cognitive functioning at the individual level is challenging due to two sources of uncertainty: aleatoric and epistemic uncertainty[18]. Here, aleatoric uncertainty refers to the inherent randomness present in most real-world settings, such as the variability in measurements of cognitive functioning[19]. Epistemic uncertainty stems from an incomplete understanding of the causal mechanisms behind observed data, such as how surgery impacts cognitive functioning. Given that predictions may be unreliable due to aleatoric and epistemic uncertainty, it is essential to use methods to quantify uncertainty in individual predictions such that clinicians know when predictions can be relied upon[20].

Bayesian models offer two main advantages. First, they can be used to model the uncertainty in individual predictions[21]. Bayesian models do this by learning distributions of possible values for each parameter, rather than point estimates. By combining these parameter distributions with the predictors for a new data point, Bayesian models produce a probability distribution of potential outcomes. This distribution can be used to obtain a point estimate and reflects the uncertainty in the prediction.

Second, Bayesian models allow for incorporating prior knowledge into parameter estimates using priors. These priors represent our beliefs about the parameters before having seen any data, potentially improving model performance[22]. Even though the mechanisms by which glioma affect cognitive functioning are poorly understood, weakly informative priors can still be used to provide some guidance to the models. Bayesian models already have significant traction for making predictions in clinical applications[22] and for describing neurological functions[19,23], and are becoming increasingly accessible[24].

Predictions of cognitive functioning after treatment could aid in selecting the optimal treatment. Unfortunately, previous studies could only partially explain cognitive functioning after treatment at the individual level and are limited by a small sample size, a small number of included predictors, and do not model uncertainty in individual predictions[13]. To address these limitations, the current study aims to predict cognitive functioning after treatment in a large sample of patients with a glioma (n=317) using

Bayesian models employing a comprehensive set of predictors available before surgery. The current study is an extension to our previous study where we employed machine-learning models to predict pre-operative cognitive functioning using the same set of predictors[25].

Method

Participants

Patients were included when they had an oligodendroglioma or astrocytoma (WHO grade 2, 3, and 4) and underwent elective surgery between 2010 and 2019 at the Elisabeth-TweeSteden Hospital, Tilburg, The Netherlands, and had a valid pre-operative cognitive screening as performed during clinical care. Patients were not included when they had reduced testability (e.g. no serious visual or motor deficits) for the neuropsychological screening, were under 18, had a progressive neurological disease, or had a psychiatric or acute neurological disorder within the previous two years. This study was part of a protocol registered with the Medical Ethics Committee Brabant (file number NW2020-32). This is the same sample as (in part) included in[12,25–29].

Interview and cognitive testing

Informed consent was obtained prior to performing a standardized interview. This interview was performed to collect age, sex, and education (the Dutch Verhage scale), and to measure symptoms of anxiety and depression using the Dutch translation of the Hospital Anxiety and Depression Scale (HADS)[30] for use as predictors.

Cognitive functioning was assessed immediately before and three months after surgical resection of the tumor using the CNS Vital Signs (CNS VS)[31] computerized neuropsychological test battery. The psychometric properties of this battery were shown to be comparable to the pen-and-paper tests that it is based on in patients with various neuropsychiatric disorders and healthy individuals[32–35]. A well-trained technician (neuropsychologist or neuropsychologist in training) provided test instructions and reported on the validity of each test. Requirements for the validity included the patient understanding the test, showing sufficient effort, having no vision or motor impairments that significantly affected task performance, and the absence of any (external) distractions. Invalid tests were excluded on a test-by-test basis. Test scores were calculated from the CNS VS results according to the formulas presented in Appendix 1 and were defined such that a higher score represents a better performance.

Clinical characteristics

Five of the variables used for prediction were collected from patients' electronic medical records. This set consisted of the involved hemisphere, the use of antiepileptic drugs, comorbidities, the ASA score (assessment of the patient's physical status before surgery[36]), and the symptoms the patient presented with. These presenting symptoms were categorized into five binary categories indicating whether the symptom was present or not: behavioral/self-reported cognitive problems, language problems, epilepsy/loss of consciousness, motor deficits (including paresis), and headache.

Three tumor characteristics were also included as predictors. These were the tumor grade which describes the malignancy of the tumor (classified according to the WHO guidelines as used at the time of treatment[37,38]), histopathological diagnosis (oligodendroglioma, astrocytoma as based on cell

origin/molecular markers), and IDH1 mutation status. Note that we used the measured values for these tumor characteristics whereas they can only be estimated preoperatively[39].

In our clinical practice, the IDH mutation status of patients with a grade 4 glioblastoma aged over 55 is not always tested due to the very low incidence rate of IDH mutant gliomas for these patients[40–42]. Therefore, missing IDH mutation statuses for this subset of patients were set to wild-type. A detailed explanation is provided in Appendix 2.

Tumor volume and location

Tumor volume and location were also used as predictors. Tumors were segmented automatically from routine MRI scans. All segmentations were manually validated and redone semi-automatically when deemed incorrect. For low-grade gliomas, the tumor region was defined as the hyperintense area on the FLAIR scan, while for high-grade gliomas, it was defined as the hyperintense area on the T1 contrast scan. Additional details regarding segmentation are provided in Appendix 3. Tumor volume was quantified by the number of voxels (mm^3) in the segmentation. Tumor location was determined by calculating the percentage of overlap between segmentations and the four lobes individually for each hemisphere[26].

Analysis

Follow-up Participation

Not all included patients with a valid pre-operative screening returned for or were able to complete the follow-up neuropsychological assessment three months after surgery. Therefore, these patients were not included in the prediction models. The number of patients without a valid follow-up measurement, along with their reasons was reported. Additionally, patient characteristics and cognitive test scores were statistically compared between those with and without a valid follow-up measurement. This was done using either a t-test, Mann–Whitney U test, or Chi-Square test. No corrections for multiple testing were applied given the descriptive nature of these analyses.

Modeling

Variable reduction

The complete set of predictors is listed in Table 1. Although the models used in this study can handle large numbers of predictors due to the use of shrinkage priors (see model specification below), an excessive number of predictors can hinder model convergence and increase uncertainty in individual predictions. Therefore, the number of predictors was reduced by considering the number of patients in different categories (at least 10% per category), variance inflation factors (< 0.5), pairwise correlations (> 0.6 for to-be-combined variables), and the interpretability of combined predictors[25].

Preprocessing

All predictors were normalized to have a mean of zero and a standard deviation of one to ensure all predictors contribute equally during model fitting and to aid the interpretation of model parameters. Test scores three months after surgery were normalized relative to pre-operative scores to ensure they were on the same scale, facilitating interpretation. Moreover, pre-operative scores of patients without valid follow-up screenings, who were not included in the models, were normalized relative to those with valid follow-ups for descriptive purposes. As no statistical analyses are performed in the current study,

cognitive test scores were not normalized relative to healthy participants, corrected for effects of age, sex, and education as found in healthy participants, nor corrected for test-retest effects.

Model specification

Three Bayesian models (models 1, 2, and 3) were evaluated for predicting cognitive functioning three months after surgery. These models were the following:

Model 1 was a multiple multivariate linear regression model (i.e., a model with multiple predictors and multiple outcomes). A multivariate approach was used to allow for the joint estimation of model parameters for the eight different test scores.

Model 2 was similar to the first but included interaction effects between predictors and the histopathological diagnoses (oligodendroglioma, astrocytoma, or glioblastoma). These interactions were included as predictors of cognitive function, while related, may vary across different diagnoses[43].

Model 3 was also a multiple multivariate linear model but allowed coefficients to differ between histopathological diagnoses using partial pooling. This method allows coefficients to vary across groups while pulling them toward the population average.

All three models were evaluated while modeling residual correlations between the test scores as cognitive test scores are known to be correlated[44]. Moreover, all models were fitted with **(a)** no (additional) interaction effects, **(b)** an (additional) interaction effect between age and tumor volume, or **(c)** an (additional) interaction effect between education level and tumor volume. These interaction effects were added since evidence for the role of tumor volume by itself on postoperative cognitive function is mixed[11] and may be moderated by proxies of neuroplasticity and cognitive reserve such as age and education level[45].

Models were defined using the Bayesian Regression Models using Stan (BRMS) package (v2.20.1)[24,46,47]. A formal description of the models including the BRMS syntax is presented in Appendix 4. Models were fitted using the Hamiltonian Monte Carlo algorithm in STAN (v2.21)[48]. Missing test scores were estimated within the Bayesian models themselves. Missing predictors were imputed before fitting the models with multiple imputation using MICE[49] (v3.16.0). Thirty different imputed datasets were created and models were fitted individually on each of the imputed datasets. Afterward, the model parameters were pooled to account for the uncertainty in imputation.

For all models, weakly informative priors were used for the coefficients, intercept, residuals, and random effects instead of informative priors for two reasons. First, previous studies employed various neuropsychological tests, which differ in both their sensitivity and the cognitive domains they measure. Consequently, model parameters may not translate to the tests used in this study. Second, we do not expect our model parameters to be independent, complicating the determination of informative priors. Alongside the weakly informative priors, expectations regarding the number of non-zero coefficients and the magnitude of coefficients were set using horseshoe priors[50,51] because of the small sample-to-variable ratio. For a detailed rationale behind each individual prior used we refer to Appendix 5. To verify that the priors correctly modeled our expectations, prior predictive checks were performed for each model and described for the best-performing model (see below).

The role of pre-operative cognitive functioning

To assess the added value of clinical predictors beyond pre-operative cognitive functioning, three additional models were fitted using only pre-operative cognitive functioning as predictors. These models were labeled as 1d, 2d, and 3d and mirror the structure of models 1a, 2a, and 3a respectively while not including the clinical predictors. Note that models 2b and 3b still include the interaction effects with or structure across different histopathological diagnoses.

Model convergence and evaluation

To explore how well the sampling process explored the parameter space, the effect size (ESS) was evaluated. To determine if the model converged, the Rhat values were inspected. An ESS of above 1000 and a Rhat below 1.05 were interpreted as sufficient.

Models were compared using the expected log pointwise predictive density as determined using the expected leave-one-out cross-validation (ELPD-LOO)[52] with Pareto smoothed importance sampling (PSIS)[53]. The ELPD-LOO is a Bayesian measure for model comparison that approximates the out-of-sample generalizability of model predictions based on the full posterior distributions. The best-performing model was defined as having the highest ELPD-LOO.

To facilitate comparison with studies using frequentist machine-learning models, point-wise predictions resulting from the best-performing model were evaluated using 10-fold cross-validation. Here, point-wise predictions were defined as the mean of the posterior predictive distribution, and were evaluated using the frequentist versions of the mean absolute error (MAE) and coefficient of determination (R^2) score. Normalization and imputation of the predictors were performed within the cross-validation loop to prevent information leakage[54]. To assess the added value of clinical predictors, point-wise predictions were additionally evaluated for the model that obtained the highest ELPD-LOO while only employing pre-operative cognitive functioning (and potentially tumor histopathology), and the plain multivariate model that only employed pre-operative cognitive functioning (model 1d). To evaluate whether the best-performing model is a good fit for the observed data, the posterior predictive distributions resulting from this model were visualized.

Sensitivity to selected priors

To ensure the priors were only weakly informative, fitted model parameters including their credibility intervals (i.e. the posterior distributions) were inspected for the best-performing model. Moreover, to test the sensitivity to the selected priors and to distinguish between the effect of the horseshoe prior and all other priors, model performance was compared against three additional versions of the best-performing model. These were the same model with only the default priors in BRMS, and versions with weakly informative priors where either the horseshoe prior or all weakly informative priors were replaced with their default.

Model interpretation and application

To interpret the relationships captured by the best-performing model, its fitted model parameters including their credibility intervals were inspected. To interpret the certainty of the individual out-of-sample predictions, the amount of uncertainty in predictions as obtained using the 10-fold cross-validation (i.e. posterior predictive distributions) was described. Last, to inspect whether the model

made any systematic errors, the point-wise out-of-sample predictions were plotted against the measured values.

To illustrate the application of Bayesian models and their uncertainty estimates can be applied in clinical practice, an out-of-sample prediction resulting from the best-performing model was visualized. This was achieved by showing the point estimate, the posterior predictive distribution describing the uncertainty, and the true measured value. The prediction was selected to have a standard deviation in the posterior predictive distribution closest to the population median, thus having a median amount of uncertainty. Multiple example predictions for all outcome measures selected to differ in their amount of uncertainty were provided as an appendix.

The Bayesian Analysis Reporting Checklist by Kruschke[55] was followed and is provided as an online supplement. Moreover, documented R (v4.0.4)[56] code and dummy data are provided as an online supplement.

Results

Descriptive statistics and follow-up participation

A total of 317 patients were included in the study. Eighty of these patients did not participate in the three-month follow-up. The reasons for not participating in the follow-up were: not responding to, not showing up for, or canceling the appointment without reporting a reason (n=24); being clinically unable to show up for or perform the assessment (23); having passed away (12); being treated in a different hospital (7); logistical reasons (2); or undergoing a re-resection (1). For eleven patients, the reason could not be determined retrospectively. Finally, for seven of the remaining 237 follow-up measurements, all tests in the battery were deemed invalid by the test technician. Descriptive statistics for the remaining sample (n=230) are provided in Table 1. Moreover, descriptive statistics for the sample that did not participate in the follow-up or had a follow-up measurement that was not deemed valid are presented in Appendix 6.

Table 1: Sample characteristics

Variable name	count	Mean / %	std	min	25%	50%	75%	max	Missi (%)
Age	230	50.92	14.69	18.00	39.25	53.00	61.75	80.00	0.0
Education	230	5.13	1.12	1.00	4.00	5.00	6.00	7.00	0.0
Sex (m)	230	0.64%	0.48						0.0
Astrocytoma	230	0.32%	0.47						0.0
Glioblastoma	230	0.50%	0.50						0.0
Oligodendroglioma	230	0.16%	0.37						0.0
WHO grade 2	230	0.35%	0.48						0.0
WHO grade 3	230	0.11%	0.32						0.0
WHO grade 4	230	0.53%	0.50						0.0
IDH1 mutation status (mutant)	213	0.52	0.50	0.00	0.00	1.00	1.00	1.00	7.3
Lateralization left	230	0.41%	0.49						0.0
Lateralization right	230	0.61%	0.49						0.0
Frontal Lobe left (mm^3)	224	9179.84	21528.38	0.00	0.00	0.00	7335.25	164182.00	2.6
Occipital lobe left (mm^3)	224	747.63	3507.75	0.00	0.00	0.00	0.00	22438.00	2.6
Parietal lobe left (mm^3)	224	1639.32	5421.57	0.00	0.00	0.00	0.25	39587.00	2.6
Temporal lobe left (mm^3)	224	3588.32	11180.60	0.00	0.00	0.00	0.00	74049.00	2.6

Frontal lobe right (mm^3)	224	10722.05	19930.29	0.00	0.00	452.50	10880.50	99580.00	2.6
Occipital lobe right (mm^3)	224	954.83	4576.02	0.00	0.00	0.00	0.00	40452.00	2.6
Parietal lobe right (mm^3)	224	3850.77	10918.67	0.00	0.00	0.00	538.50	77898.00	2.6
Temporal lobe right (mm^3)	224	7250.19	16617.50	0.00	0.00	0.00	2255.75	91482.00	2.6
Tumor volume (mm^3)	224	51589.21	45275.83	305.00	20654.50	38027.00	71466.75	264510.00	2.6
ASA I	229	0.49%	0.50						0.4
ASA II	229	0.46%	0.50						0.4
ASA III	229	0.05%	0.21						0.4
Comorbidity	230	0.42%	0.49						0.0
Corticosteroid use	230	0.56%	0.50						0.0
Antiepileptic drug use	230	0.50%	0.50						0.0
HADS anxiety	209	7.09	4.34	0.00	4.00	6.00	10.00	19.00	9.1
HADS depression	209	4.96	3.58	0.00	2.00	4.00	7.00	17.00	9.1
Presents with attention, executive function, memory, and/or behavioral problems	230	0.19%	0.39						0.0
Presents with language problems	230	0.13%	0.34						0.0
Presents with epilepsy or loss of consciousness	230	0.47%	0.50						0.0
Presents with motor deficits	230	0.21%	0.41						0.0
Presents with headache	230	0.24%	0.43						0.0
Pre-operative cognitive test scores									
Verbal memory recognition	216	0.00	1.00	-2.93	-0.68	0.07	0.82	1.76	6.0
Visual memory recognition	227	0.00	1.00	-3.41	-0.61	0.19	0.79	1.79	1.3
Symbol digit coding	226	0.00	1.00	-2.67	-0.60	0.11	0.57	2.51	1.7
Simple reaction time	219	0.00	1.00	-4.60	-0.29	0.35	0.64	1.05	4.7
Stroop interference	211	0.00	1.00	-3.39	-0.56	0.00	0.74	2.63	8.2
Continuous performance test	228	0.00	1.00	-5.83	-0.50	0.12	0.70	2.25	0.8
Shifting attention task	212	0.00	1.00	-2.01	-0.74	-0.11	0.70	2.50	7.8
Finger tapping test	219	0.00	1.00	-4.44	-0.47	0.14	0.61	3.89	4.7
Cognitive test scores after treatment									
Verbal memory recognition	218	-0.01	1.00	-2.93	-0.68	0.07	0.63	1.76	5.2
Visual memory recognition	227	-0.11	0.96	-3.41	-0.81	-0.01	0.59	1.59	1.3
Symbol digit coding	226	0.11	1.05	-3.13	-0.53	0.11	0.89	2.51	1.7
Simple reaction time	220	-0.07	1.05	-6.24	-0.34	0.28	0.58	1.09	4.3
Stroop interference	220	-0.02	0.97	-3.26	-0.69	0.07	0.68	2.73	4.3
Continuous performance test	225	-0.14	1.03	-4.19	-0.66	0.00	0.59	1.93	2.1
Shifting attention task	215	0.13	1.05	-1.87	-0.60	0.10	0.81	3.34	6.5
Finger tapping test	218	0.03	0.96	-4.32	-0.45	0.07	0.58	4.54	5.2

Table 1: Sample characteristics of the sample used for model fitting. Pre-operative test scores were normalized to have zero mean and unit variance. Cognitive test scores three months after surgery were scaled relative to the pre-operative test scores.

Patients who did not participate in the follow-up or had an invalid follow-up measurement more often had a comorbidity (chi=6.89, p=0.009), and had lower ASA scores (U=1166, p=0.021). Note that no differences in age (U=11191, p=0.103), sex (chi=0.018, p= 0.893), and education (U=10267, p=0.708) were found. Regarding cognitive test scores, patients who did not participate in the follow-up or had an invalid follow-up measurement had a significantly lower score pre-operatively on the measure of verbal memory recognition (U=7315, p=0.020), the symbol digit coding task (T=-4.50, p=0.000), the measure of Simple reaction time (U=6906, p=0.000), and the shifting attention task (U=6693, p=0.013), but not the other four measures (all p's>0.127).

Variable reduction

Based on the variance inflation factor, pairwise correlations, and number of patients per category, tumor lateralization was grouped into right-lateralized and left-lateralized + bilateral; tumor grades were grouped into low-grade (grade 2) and high-grade (grade 3 + 4); ASA scores were grouped into ASA I and ASA II + III; use of antiepileptic drugs was merged with 'presenting with epilepsy or loss of consciousness'; and HADS anxiety and depression were combined. The resulting set of predictors used as predictors is the same as used in our previous study[25].

Model convergence and evaluation

The prior predictive check for the best-performing model (model 2c, see below) is reported in Appendix 7 and was highly similar for all other models. The prior predictive check showed that the simulated data covered a wide but reasonable range of outcomes. This indicates that the selected priors were weakly informative.

BRMS did not report problems with model fitting. The Rhat values generally were below 1.05 with a small number of exceptions ranging up to 1.13, indicating good convergence. The ESS for the bulk and tail of the distributions for the different models generally were above 1000 with a small number of exceptions as low as 575 and 836 for the bulk and tail ESS respectively, indicating effective sampling.

Table 2 (part 1) presents the ELPD-LOO for each model including the difference relative to the best-performing model. Model 2c achieved the best performance (ELPD-LOO = -1624) and includes an interaction effect between education and tumor volume and interactions with the histopathological diagnosis. The five runner-ups (model 2a, 2b, 3a, 3b, and 3c) performed similarly with decreases in ELPD-LOO ranging from -2.2 to -11.0. These differences were smaller than the standard deviations in this difference which were between 7.9 and 14.7. Therefore, we cannot distinguish between these models according to their ELPD-LOO. Models that performed partial pooling (Models 3) performed the worst, with a decrease in ELPD-LOO of at least -127.6 (SE=17.7). Finally, there was no clear effect of including the interaction effect between tumor volume and age or education level. This can be seen from the small differences between variants a, b, and c of the different models.

Table 2 (part 2) presents the performance of models 1d, 2d, and 3d which only employed pre-operative cognitive functioning (and histopathological diagnosis) and were evaluated to test the added value of the clinical predictors. Of these models, model 2d performed best with an ELPD-LOO of -1638.7. Comparing its performance to the best-performing model overall (2c), which has the same structure, only a slight decrease in performance is observed with a difference of -14.4 (SE=10.0).

Table 2	Interaction effect		ELPD LOO estimate		ELPD LOO comparison	
Model name	Age times volume	Education times volume	Estimate	SE	Difference	SE Difference
Part 1: Using pre-operative cognitive functioning and all clinical predictors						
Model 2c (interactions hist.diag)		☒	-1624.3	46.6	*Baseline*	
Model 1b (plain)	☒		-1626.4	47.5	-2.2	9.2

Model 1c (plain)	?	-1627.3	47.5	-3.0	10.2
Model 2a (interactions hist.diag)		-1630.1	46.5	-5.8	7.8
Model 2b (interactions hist.diag)	?	-1632.0	45.9	-8.7	11.0
Model 1a (plain)		-1635.3	48.7	-11.0	14.7
Model 3c (partial pooling)	?	-1752.0	47.6	-127.6	17.7
Model 3a (partial pooling)		-1772.4	51.8	-148.0	31.5
Model 3b (partial pooling)	?	-1773.8	54.6	-149.5	38.1
Part 2: Only using only pre-operative cognitive functioning (and histopathological diagnosis)					
Model 2d (interactions hist.diag)		-1638.7	46.7	-14.4	10.0
Model 1d (plain)		-1643.7	47.5	-19.4	16.9
Model 3d (partial hist.diag)		-1670.1	49.5	-45.7	113.0

Table 2: Model performance sorted from best to worst individually for models including the clinical predictors (part 1), and models not including clinical predictors (part 2). Performance is described as the Expected Log Predictive Density - Leave-One-Out (ELPD-LOO) and the standard error of this estimate is reported. Moreover, the difference of all models relative to the best-performing model is reported including the expected standard error of this difference. hist.diag: Histopathological diagnosis

Table 3 describes the out-of-sample prediction performance of the pointwise predictions using the mean absolute error (MAE) and the coefficient of determination (R^2). These results show that the best-performing model as determined using the ELPD-LOO (model 2c) obtained a median R^2 of 34.20% of variance and a median MAE of 0.599. Performance for the individual tests ranged between an R^2 of 13.77% and an MAE of 0.693 for the Stroop interference ratio and an R^2 of 73.22% and an MAE of 0.420 for the symbol digit coding task.

	Model 2c (Best model): interactions hist.diag and education times size		**Model 2d**: Best model when using only pre-operative functioning and histopathology:		**Model 1d**: Model using only pre-operative functioning:	
Table 3	R^2 score	MAE	R^2 score	MAE	R^2 score	MAE
Verbal memory recognition	26.78%	0.661	28.26%	0.660	**28.73%**	**0.659**
Visual memory recognition	**27.93%**	**0.666**	27.71%	0.674	28.78%	0.670
Symbol digit coding	**73.22%**	**0.420**	69.47%	0.441	69.47%	0.441
Simple reaction time	25.48%	0.602	**28.27%**	**0.590**	27.32%	0.592
Stroop interference ratio	13.77%	**0.693**	12.59%	0.706	**13.96%**	0.697
Continuous performance test	**51.06%**	0.539	49.91%	**0.537**	49.49%	0.543
Shifting attention task	48.64%	0.597	**48.97%**	**0.587**	48.89%	0.589
Finger tapping test	40.47%	0.508	**41.19%**	**0.499**	41.01%	0.503
Median	34.20%	0.599	34.73%	**0.589**	**34.89%**	0.590

Table 3: The mean absolute error (MAE) and the coefficient of determination (R2) individually for each test score and the median across the different test scores. Hist.diag: Histopathological diagnosis

The median R^2 score of the best-performing model (2c) and its MAE were lower when compared to both models 1d and 2d which only relied on pre-operative cognitive functioning (and histopathological diagnosis). This difference, however, is very small with a change of 0.54 percentage points in R2 and a change of 0.01 in MAE.

When considering individual test measures, the best-performing model (Model 2c) obtained the highest R^2 score and MAE for the measure of verbal memory recognition and the symbol digit coding task.

Moreover, this model obtained the highest R^2 score for the continuous performance test and the highest MAE for the Stroop interference ratio.

To evaluate whether the best-performing model (2c) accurately describes the distribution of the observed data, the posterior predictive checks for this model are visualized in Appendix 8. This figure shows that the simulated data matches the observed data for most draws from the model parameters and training data. This indicates that the model was able to adequately describe the observed data. For most tests, however, and especially the measure of simple reaction time, the model was not able to completely capture the skewness.

Sensitivity to selected priors

The parameter estimates after model fitting (i.e. the posterior distributions) of the best-performing model (model 2c) are visualized in Appendix 9. This visualization shows that they were within the specified priors, indicating that the priors were suitable.

Two of the three sensitivity checks as performed for model 2c did not converge. These were the variant with all default priors, and the variant with weakly-informative priors but no horseshoe prior. The variant with all default priors except for the horseshoe prior converged and performed slightly worse when compared to model 2c with a difference in ELDP-LOO of -12.53 (SE=9.85). This shows that the horseshoe prior was crucial for model convergence while the weakly-informative priors only had a small positive impact on model fit.

Model interpretation and application

For the estimated model parameters for the best-performing model (2c), we refer back to Appendix 9. Note that the relationships captured by the model are solely descriptive of how the model obtains its predictions.

Results showed that the most important predictor of a given measure of cognitive functioning after treatment was this same measure before treatment with coefficients ranging between 0.35 [95% CI: 0.20, 0.48] and 0.74 [95% CI: 0.63, 0.84] (Appendix 9A). One notable exception from this was the Stroop interference ratio, whose predictions relied mostly on the pre-operative measure of the shifting attention task with a coefficient of 0.23 [95% CI: 0.20, 0.40], followed by the pre-operative measure of the Stroop interference ratio with a coefficient of 0.11 [95% CI: 0.00, 0.26].

The contribution of most clinical predictors was negligible with coefficients generally being around zero with only 3.34% of the coefficients being above $|0.05|$. Moreover, the credibility intervals associated with these coefficients were large relative to the magnitudes of the coefficients. Additionally, the included interaction effect between education level and size contributed little to the predictions with coefficients of at most $|0.02|$.

The expected variability in the measures of cognitive functioning after treatment (i.e. standard deviation of the likelihood) ranged between 0.53 [95% CI: 0.0.48, 0.59] for symbol digit coding and 0.84 [95% CI: 0.76, 0.93] for verbal memory recognition (Appendix 9C). The median amount of uncertainty in the resulting predictions as obtained using 10-fold cross-validation (i.e. standard deviations of the posterior predictive distributions) ranged between 0.56 for the symbol digit coding task and 0.94 for the Stroop interference ratio (Appendix 10).

To inspect whether the model made any systematic errors, out-of-sample predictions were plotted against the measured values in Figure 1. This figure shows that there are no systematic deviations in model performance as can be seen from most points being clustered around the line representing perfect predictions. For Simple reaction time, however, there is some heteroscedasticity. This can be seen from predictions for patients who scored poorly on this measure showing more variance in prediction error.

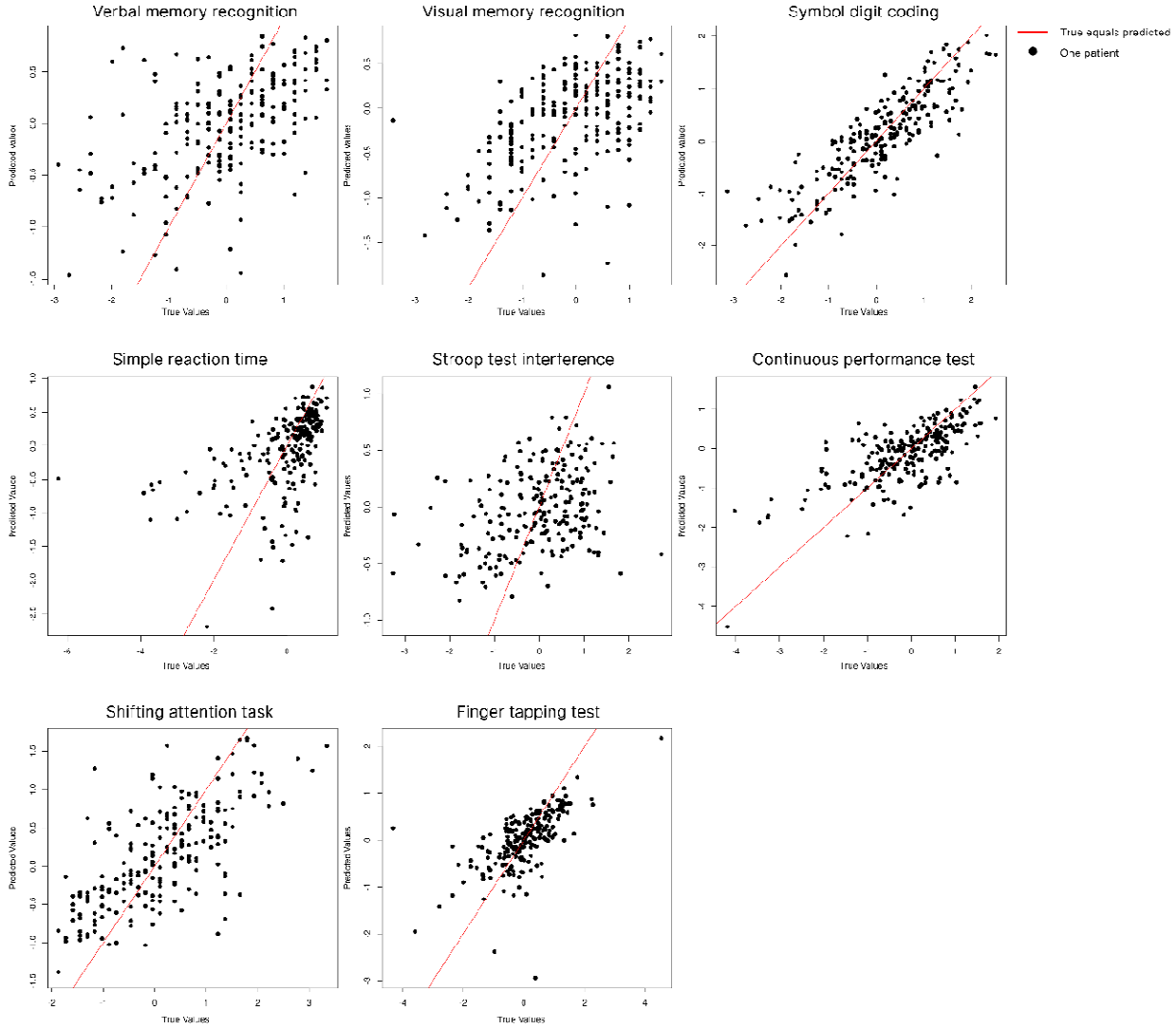

Figure 1: Scatter plots of predicted values obtained using 10-fold cross-validation from the best-performing model (2c) versus the measured values, individually for each outcome measure. Each dot represents a patient, its position along the x-axis represents their measured value, and the position along the y-axis the predicted test score. The red line (x=y) represents perfect predictions, and the distance along the x-axis represents the error of the prediction.

A demonstration of how uncertainty estimates can be used in clinical practice is provided in Figure 2. The large range of outcomes covered by the uncertainty estimate shown in this figure (in blue) indicates that clinicians should not rely on the point estimate (in red) as there is a large chance that it will not be close to the true value (in green). Therefore, clinicians should not rely on this prediction for decision-making

and can inform patients we don't know how the treatment will affect their cognitive functioning. Additional examples for all outcome measures are presented in Appendix 11, showing similar uncertainty in the predictions.

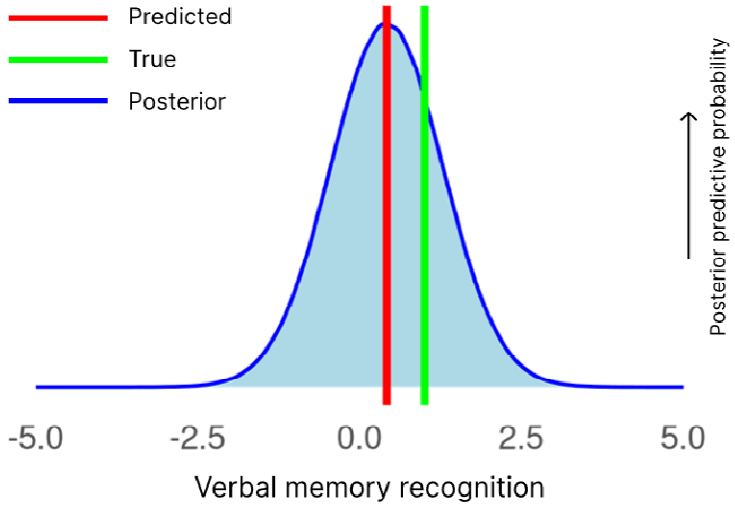

Figure 2: Example predictions of verbal memory recognition obtained using 10-fold cross-validation for the best-performing model (2c). The example prediction was selected to be at the median in terms of the amount of uncertainty in the prediction. The blue distribution represents the posterior predictive distribution resulting from the model, which represents the probability of each outcome, the red line represents the point estimate obtained from this distribution, and the green line represents the measured value. Three example predictions for all cognitive tests can be found in Appendix 11.

Discussion

Results show that predicting cognitive functioning three months after surgery using pre-operative cognitive functioning and the included clinical predictors is not yet possible. The amount of variance explained in cognitive functioning three months after surgery ranged between 13.77% (Stroop interference ratio) and 73.22% (Symbol digit coding) with a median of 34.20%. Moreover, the uncertainty in individual predictions ranged between a median standard deviation of 0.72 and 0.94 (relative to the population standard deviations). This performance likely is insufficient for clinical application, though further research is needed to establish thresholds for clinical utility. These findings align with the study by Zangrossi and colleagues which explained between 0.81% and 62.41% (median 39.09%) of variance in cognitive performance one week after awake surgery[13] while employing age, education, and pre-operative neuropsychological test scores.

The best-performing model relied strongly on pre-operative cognitive functioning to predict cognitive functioning three months after surgery, in line with previous studies[12,13]. Moreover, the model using only pre-operative cognitive functioning and interactions with the histopathological diagnosis as predictors (model 2d) performed similarly to the best-performing model overall (2c). Additionally, the included interaction effects between tumor volume and age or education level had a negligible influence on

predictions. These findings show that the added value of the clinical predictors and included interaction effects as used in the current study are limited when predicting cognitive functioning after treatment.

The maximum amount of variance that a perfect model can explain is unknown and limited by the aleatoric uncertainty. One part of this aleatoric uncertainty stems from the test-retest reliability. For CNS VS, the test-retest reliability has only been established for healthy participants[35] and likely is lower for patients with a brain tumor due to cognitive impairments and medication effects. The epistemic uncertainty likely results from relevant predictors that were not included. This set of predictors likely comprises predictors that only are available after treatment and therefore could not be used, such as surgical complications and the adjuvant treatments received[8,10,57–61], and promising predictors that can be obtained before surgery but are not (yet) routinely collected in our practice including measures of structural and functional connectivity[6,62] and information regarding edema[57]. Additionally, including different representations of predictors or interactions thereof may improve model performance. Finally, using models that can capture more complex relationships may reduce uncertainty, although this likely requires larger datasets[25].

The amount of variance explained in cognitive functioning after treatment differed up to 59.4 percentage points between tests. These substantial differences can likely be attributed to three factors. First, some cognitive domains may be more prone to change after surgery, causing the pre-operative test scores to be less informative. Second, the predictive power of the clinical predictors and pre-operative test scores to describe the change in functioning may differ across test scores. Third, the cognitive tasks used differ in their ability to reliably and repeatably measure the cognitive functions they are intended to measure, in line with differences in their test-retest reliability[35].

The current models assume that the decision for surgery is already made. Ideally, predictive models would avoid such assumptions, thereby enabling clinicians to compare predictions across various treatment decisions. This, however, requires either data from randomized control trials (RCTs)[63], simulating an RCT from retrospective data[64,65], or developing casual models[66,67]. Unfortunately, neither was possible as RCTs are undesirable, simulating an RCT requires all confounders influencing the treatment decision to be available, and causal models require the causal mechanism to be known.

Several limitations of the current study should be noted. First, models are solely based on patients with a valid three-month follow-up. Therefore, predictions are only valid if the patient will be able to undergo the follow-up, which is unknown before surgery. Consequently, the current model needs to be paired with a model that predicts whether a patient will complete the follow-up. Second, we used the histopathological diagnosis, WHO grade, and IDH1 status as determined post-operatively while they can merely be estimated pre-operatively[39], potentially further limiting the accuracy when applied in clinical practice. Third, the sample was gathered during clinical care and therefore did not include patients with severe impairments or in need of immediate surgical intervention. Finally, cognitive assessment was done using a brief computerized test battery which may be somewhat dependent on processing speed and does not measure language function, memory free recall, or visuoconstructive abilities. However, more comprehensive evaluations are not typically conducted during clinical care.

We believe our results to be highly important as they show that clinicians should not rely on the included clinical predictors to infer cognitive functioning three months after surgery. Additionally, our results demonstrate how estimates of uncertainty ultimately can be used in clinical practice to facilitate trust in predictions. Finally, our results show the importance of collecting larger datasets including additional predictors.

This need for larger datasets is especially important when including high-dimensional and noisy data such as structural and functional connectivity[68]. Moreover, the relatively low signal-to-noise ratio in scores resulting from brief neuropsychological screening[19,35], and the large individual differences between patients add to this need. Therefore, we hope future work will focus on standardizing data collection to obtain larger cross-center multimodal datasets. Such datasets have the potential to significantly improve the ability to predict cognitive functioning at the individual level while allowing models to generalize across centers. This need is being increasingly emphasized by numerous authors (e.g.[1,6,25,69,70])

Future studies can use information regarding the planned treatment (both primary and adjuvant) to improve predictions and could utilize virtual models of the brain to model the hypothesized effect of the planned surgery[6,71]. Moreover, future work could predict outcomes that are closer to patients daily functioning.

Conclusion

Predictions of cognitive functioning after treatment could aid in selecting the optimal treatment. The current study aimed to predict cognitive functioning three months after surgery (and adjuvant treatments) on the individual level using a comprehensive set of clinical predictors available before surgery and pre-operative cognitive functioning while employing Bayesian models. While predictions accounted for substantial variance in cognitive functioning three months after surgery, individual predictions were uncertain and likely of insufficient quality for use in clinical practice. Consequently, clinicians should not rely on the included predictors to infer patients' cognitive functioning after treatment. Pre-operative cognitive functioning was the most influential predictor and models including clinical predictors and pre-operative functioning performed roughly similarly to models using only pre-operative functioning, showing the limited added value of the clinical predictors and interaction effects as used in the current study. The current study further demonstrated how individual predictions including their uncertainty estimates may ultimately be used in clinical practice, allowing models to say ‘I don’t know’ instead of being confidently wrong. Finally, it stresses the need to collect larger cross-center multimodal datasets including additional predictors. Such datasets have the potential to significantly improve the ability to predict cognitive functioning at the individual level while allowing models to generalize across centers.

Funding

ZonMw (10070012010006, 824003007).

Conflict of interest

None to declare

Authorship

Experimental design: (SB, KG, EP, GR, LLO, BN), acquisition: (KG, GR, EB), analysis: (SB, LLO, KG, BN), interpretation: (SB, LLO, KG, BN, EP, EB, WDB, GR). All authors have been involved in the writing of the manuscript and approved the final version.

Acknowledgments

We would like to express our gratitude to Sacha van der Donk for her role as data manager on this project

Data availability

Data described in this work is not publicly available to protect the privacy of patients. All code used in this study is available as supplementary material.

References

1. De Roeck L, Gillebert RC, van Aert RCM, et al. Cognitive outcomes after multimodal treatment in adult glioma patients: A meta-analysis. *Neuro-Oncology*. Published online February 21, 2023:noad045. doi:10.1093/neuonc/noad045

2. Tariq R, Hussain N, Baqai MWS. Factors affecting cognitive functions of patients with high-grade gliomas: a systematic review. *Neurol Sci*. 2023;44(6):1917-1929. doi:10.1007/s10072-023-06673-4

3. Heffernan AE, Wu Y, Benz LS, Verhaak RGW, Kwan BM, Claus EB. Quality of life after surgery for lower grade gliomas. *Cancer*. 2023;129(23):3761-3771. doi:10.1002/cncr.34980

4. Svedung Wettervik T, Munkhammar ÅA, Jemstedt M, et al. Dynamics in cognition and health-related quality of life in grade 2 and 3 gliomas after surgery. *Acta Neurochir*. 2022;164(12):3275-3284. doi:10.1007/s00701-022-05408-2

5. Van Dyk K, Wall L, Heimberg BF, et al. Daily functioning in glioma survivors: associations with cognitive function, psychological factors and quality of life. *CNS Oncol*. 2022;11(2):CNS84. doi:10.2217/cns-2022-0002

6. Herbet G, Duffau H, Mandonnet E. Predictors of cognition after glioma surgery: connectotomy, structure-function phenotype, plasticity. *Brain*. Published online 2024:awae093.

7. Maas DA, Douw L. Multiscale network neuroscience in neuro-oncology: How tumors, brain networks, and behavior connect across scales. *Neuro-Oncology Practice*. 2023;10(6):506-517. doi:10.1093/nop/npad044

8. Dadario NB, Brahimaj B, Yeung J, Sughrue ME. Reducing the Cognitive Footprint of Brain Tumor Surgery. *Front Neurol*. 2021;12:711646. doi:10.3389/fneur.2021.711646

9. Butler JM, Rapp SR, Shaw EG. Managing the cognitive effects of brain tumor radiation therapy. *Curr Treat Options in Oncol*. 2006;7(6):517-523. doi:10.1007/s11864-006-0026-5

10. Li M, Caeyenberghs K. Longitudinal assessment of chemotherapy-induced changes in brain and cognitive functioning: A systematic review. *Neuroscience & Biobehavioral Reviews*. 2018;92:304-317. doi:10.1016/j.neubiorev.2018.05.019

11. Kirkman MA, Hunn BHM, Thomas MSC, Tolmie AK. Influences on cognitive outcomes in adult patients with gliomas: A systematic review. *Front Oncol*. 2022;12:943600. doi:10.3389/fonc.2022.943600

12. Rijnen SJM, Butterbrod E, Rutten GJM, Sitskoorn MM, Gehring K. Presurgical Identification of Patients With Glioblastoma at Risk for Cognitive Impairment at 3-Month Follow-up. *Neurosurgery*. Published online May 29, 2020:nyaa190. doi:10.1093/neuros/nyaa190

13. Zangrossi A, Silvestri E, Bisio M, et al. Presurgical predictors of early cognitive outcome after brain tumor resection in glioma patients. *NeuroImage: Clinical*. 2022;36:103219. doi:10.1016/j.nicl.2022.103219

14. Mandonnet E, Duffau H. An attempt to conceptualize the individual onco-functional balance: Why a standardized treatment is an illusion for diffuse low-grade glioma patients. *Critical Reviews in Oncology/Hematology*. 2018;122:83-91. doi:10.1016/j.critrevonc.2017.12.008

15. Pace A, Koekkoek JAF, van den Bent MJ, et al. Determining medical decision-making capacity in brain tumor patients: why and how? *Neuro-Oncology Practice*. 2020;7(6):599-612. doi:10.1093/nop/npaa040

16. Hewins W, Zienius K, Rogers JL, Kerrigan S, Bernstein M, Grant R. The Effects of Brain Tumours upon Medical Decision-Making Capacity. *Curr Oncol Rep*. 2019;21(6):55. doi:10.1007/s11912-019-0793-3

17. Halkett GKB, Lobb EA, Oldham L, Nowak AK. The information and support needs of patients diagnosed with High Grade Glioma. *Patient Education and Counseling*. 2010;79(1):112-119. doi:10.1016/j.pec.2009.08.013

18. Hüllermeier E, Waegeman W. Aleatoric and epistemic uncertainty in machine learning: an introduction to concepts and methods. *Mach Learn*. 2021;110(3):457-506. doi:10.1007/s10994-021-05946-3

19. Haines N, Sullivan-Toole H, Olino T. From Classical Methods to Generative Models: Tackling the Unreliability of Neuroscientific Measures in Mental Health Research. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*. 2023;8(8):822-831. doi:10.1016/j.bpsc.2023.01.001

20. Kompa B, Snoek J, Beam AL. Second opinion needed: communicating uncertainty in medical machine learning. *npj Digit Med*. 2021;4(1):4. doi:10.1038/s41746-020-00367-3

21. Van De Schoot R, Depaoli S, King R, et al. Bayesian statistics and modelling. *Nat Rev Methods Primers*. 2021;1(1):1. doi:10.1038/s43586-020-00001-2

22. Baldwin SA, Larson MJ. An introduction to using Bayesian linear regression with clinical data. *Behaviour Research and Therapy*. 2017;98:58-75. doi:10.1016/j.brat.2016.12.016

23. Parr T, Rees G, Friston KJ. Computational Neuropsychology and Bayesian Inference. *Front Hum Neurosci*. 2018;12:61. doi:10.3389/fnhum.2018.00061

24. Bürkner PC. brms: An R Package for Bayesian Multilevel Models Using Stan. *Journal of Statistical Software*. 2017;80(1):1-28. doi:10.18637/jss.v080.i01

25. Boelders SM, Gehring K, Postma EO, Rutten GJM, Ong LL. Cognitive functioning in untreated glioma patients: the limited predictive value of clinical variables. *Neuro-Oncology*. Published online December 1, 2023:noad221. doi:10.1093/neuonc/noad221

26. Boelders SM, De Baene W, Postma E, Gehring K, Ong LLS. *Predicting Cognitive Functioning in High-Grade Glioma: Evaluating Different Representations of Tumor Location in a Common Space*. In Review; 2024. doi:10.21203/rs.3.rs-3937209/v1

27. Butterbrod E, Bruijn J, Braaksma MM, et al. Predicting disease progression in high-grade glioma with neuropsychological parameters: the value of personalized longitudinal assessment. *J Neurooncol*. 2019;144(3):511-518. doi:10.1007/s11060-019-03249-1

28. Butterbrod E, Sitskoorn M, Bakker M, et al. The APOE ε4 allele in relation to pre- and postsurgical cognitive functioning of patients with primary brain tumors. *Eur J Neurol*. 2021;28(5):1665-1676. doi:10.1111/ene.14693

29. Rijnen SJM, Kaya G, Gehring K, et al. Cognitive functioning in patients with low-grade glioma: effects of hemispheric tumor location and surgical procedure. *Journal of Neurosurgery*. 2020;133(6):1671-1682. doi:10.3171/2019.8.JNS191667

30. Spinhoven Ph, Ormel J, Sloekers PPA, Kempen GIJM, Speckens AEM, Hemert AMV. A validation study of the Hospital Anxiety and Depression Scale (HADS) in different groups of Dutch subjects. *Psychol Med*. 1997;27(2):363-370. doi:10.1017/S0033291796004382

31. CNS Vital Signs. CNS Vital Signs Interpretation Guide. Accessed January 6, 2021. https://www.cnsvs.com/WhitePapers/CNSVS-BriefInterpretationGuide.pdf

32. Gualtieri C, Johnson L. Reliability and validity of a computerized neurocognitive test battery, CNS Vital Signs. *Archives of Clinical Neuropsychology*. 2006;21(7):623-643. doi:10.1016/j.acn.2006.05.007

33. Gualtieri CT, Hervey AS. The Structure and Meaning of a Computerized Neurocognitive Test Battery. *Frontiers in Psychological and Behavioral Science*. 2015;4:11.

34. Plourde V, Hrabok M, Sherman EMS, Brooks BL. Validity of a Computerized Cognitive Battery in Children and Adolescents with Neurological Diagnoses. *Archives of Clinical Neuropsychology*. 2018;33(2):247-253. doi:10.1093/arclin/acx067

35. Rijnen SJM, van der Linden SD, Emons WHM, Sitskoorn MM, Gehring K. Test-retest reliability and practice effects of a computerized neuropsychological battery: A solution-oriented approach. *Psychological Assessment*. 2018;30(12):1652-1662. doi:10.1037/pas0000618

36. Mayhew D, Mendonca V, Murthy BVS. A review of ASA physical status – historical perspectives and modern developments. *Anaesthesia*. 2019;74(3):373-379. doi:10.1111/anae.14569

37. Louis DN, Perry A, Reifenberger G, et al. The 2016 World Health Organization Classification of Tumors of the Central Nervous System: a summary. *Acta Neuropathol*. 2016;131(6):803-820. doi:10.1007/s00401-016-1545-1

38. Louis DN, Perry A, Wesseling P, et al. The 2021 WHO Classification of Tumors of the Central Nervous System: a summary. *Neuro-Oncology*. 2021;23(8):1231-1251. doi:10.1093/neuonc/noab106

39. van der Voort SR, Incekara F, Wijnenga MMJ, et al. Combined molecular subtyping, grading, and segmentation of glioma using multi-task deep learning. *Neuro-Oncology*. 2023;25(2):279-289. doi:10.1093/neuonc/noac166

40. Barresi V, Eccher A, Simbolo M, et al. Diffuse gliomas in patients aged 55 years or over: A suggestion for *IDH* mutation testing. *Neuropathology*. 2020;40(1):68-74. doi:10.1111/neup.12608

41. Robinson C, Kleinschmidt-DeMasters BK. *IDH1* -Mutation in Diffuse Gliomas in Persons Age 55 Years and Over. *J Neuropathol Exp Neurol*. Published online January 21, 2017:nlw112. doi:10.1093/jnen/nlw112

42. DeWitt JC, Jordan JT, Frosch MP, et al. Cost-effectiveness of IDH testing in diffuse gliomas according to the 2016 WHO classification of tumors of the central nervous system recommendations. *Neuro-Oncology*. 2017;19(12):1640-1650. doi:10.1093/neuonc/nox120

43. Wefel JS, Noll KR, Rao G, Cahill DP. Neurocognitive function varies by IDH1 genetic mutation status in patients with malignant glioma prior to surgical resection. *NEUONC*. 2016;18(12):1656-1663. doi:10.1093/neuonc/now165

44. Benson N, Hulac DM, Kranzler JH. Independent examination of the Wechsler Adult Intelligence Scale—Fourth Edition (WAIS-IV): What does the WAIS-IV measure? *Psychological Assessment*. 2010;22(1):121-130. doi:10.1037/a0017767

45. Tomasino B, De Fraja G, Guarracino I, et al. Cognitive reserve and individual differences in brain tumour patients. *Brain Communications*. 2023;5(4):fcad198. doi:10.1093/braincomms/fcad198

46. Bürkner PC. Advanced Bayesian Multilevel Modeling with the R Package brms. *The R Journal*. 2018;10(1):395-411. doi:10.32614/RJ-2018-017

47. Bürkner PC. Bayesian Item Response Modeling in R with brms and Stan. *Journal of Statistical Software*. 2021;100(5):1-54. doi:10.18637/jss.v100.i05

48. Stan Development Team. Stan Modeling Language Users Guide and Reference Manual. Published online 2023. https://mc-stan.org

49. Buuren SV, Groothuis-Oudshoorn K. **mice**: Multivariate Imputation by Chained Equations in *R*. *J Stat Soft*. 2011;45(3). doi:10.18637/jss.v045.i03

50. Van Erp S, Oberski DL, Mulder J. Shrinkage priors for Bayesian penalized regression. *Journal of Mathematical Psychology*. 2019;89:31-50. doi:10.1016/j.jmp.2018.12.004

51. Piironen J, Vehtari A. On the Hyperprior Choice for the Global Shrinkage Parameter in the Horseshoe Prior. Published online 2017. doi:https://doi.org/10.48550/arXiv.1610.05559

52. Vehtari A, Gelman A, Gabry J. Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Stat Comput*. 2017;27(5):1413-1432. doi:10.1007/s11222-016-9696-4

53. Vehtari A, Simpson D, Gelman A, Yao Y, Gabry J. Pareto Smoothed Importance Sampling. Published online August 4, 2022. Accessed December 6, 2023. http://arxiv.org/abs/1507.02646

54. Vabalas A, Gowen E, Poliakoff E, Casson AJ. Machine learning algorithm validation with a limited sample size. Hernandez-Lemus E, ed. *PLoS ONE*. 2019;14(11):e0224365. doi:10.1371/journal.pone.0224365

55. Kruschke JK. Bayesian Analysis Reporting Guidelines. *Nat Hum Behav*. 2021;5(10):1282-1291. doi:10.1038/s41562-021-01177-7

56. R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing; 2021. https://www.R-project.org/

57. Dallabona M, Sarubbo S, Merler S, et al. Impact of mass effect, tumor location, age, and surgery on the cognitive outcome of patients with high-grade gliomas: a longitudinal study. *Neuro-Oncology Practice*. 2017;4(4):229-240. doi:10.1093/nop/npw030

58. Kocher M, Jockwitz C, Caspers S, et al. Role of the default mode resting-state network for cognitive functioning in malignant glioma patients following multimodal treatment. *NeuroImage: Clinical*. 2020;27:102287. doi:10.1016/j.nicl.2020.102287

59. Moretti R, Caruso P. An Iatrogenic Model of Brain Small-Vessel Disease: Post-Radiation Encephalopathy. *IJMS*. 2020;21(18):6506. doi:10.3390/ijms21186506

60. Wong SS, Case LD, Avis NE, Cummings TL, Cramer CK, Rapp SR. Cognitive functioning following brain irradiation as part of cancer treatment: Characterizing better cognitive performance. *Psycho-Oncology*. 2019;28(11):2166-2173. doi:10.1002/pon.5202

61. Leonetti A, Puglisi G, Rossi M, et al. Factors Influencing Mood Disorders and Health Related Quality of Life in Adults With Glioma: A Longitudinal Study. *Front Oncol*. 2021;11:662039. doi:10.3389/fonc.2021.662039

62. van Kessel E, Berendsen S, Baumfalk AE, et al. Tumor-related molecular determinants of neurocognitive deficits in patients with diffuse glioma. *Neuro-Oncology*. 2022;24(10):1660-1670. doi:10.1093/neuonc/noac036

63. Rekkas A, Rijnbeek PR, Kent DM, Steyerberg EW, Van Klaveren D. Estimating individualized treatment effects from randomized controlled trials: a simulation study to compare risk-based approaches. *BMC Med Res Methodol*. 2023;23(1):74. doi:10.1186/s12874-023-01889-6

64. Prosperi M, Guo Y, Sperrin M, et al. Causal inference and counterfactual prediction in machine learning for actionable healthcare. *Nat Mach Intell*. 2020;2(7):369-375. doi:10.1038/s42256-020-0197-y

65. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects.

66. Doutreligne M, Struja T, Abecassis J, Morgand C, Varoquaux G, Celi LA. Step-by-step causal analysis of Electronic Health Records to ground decision making. Published online August 21, 2023. doi:10.21203/rs.3.rs-3222036/v1

67. Scholkopf B, Locatello F, Bauer S, et al. Toward Causal Representation Learning. *Proc IEEE*. 2021;109(5):612-634. doi:10.1109/JPROC.2021.3058954

68. Murphy K, Bodurka J, Bandettini PA. How long to scan? The relationship between fMRI temporal signal to noise ratio and necessary scan duration. *NeuroImage*. 2007;34(2):565-574. doi:10.1016/j.neuroimage.2006.09.032

69. García-García S, García-Galindo M, Arrese I, Sarabia R, Cepeda S. Current Evidence, Limitations and Future Challenges of Survival Prediction for Glioblastoma Based on Advanced Noninvasive Methods: A Narrative Review. *Medicina*. 2022;58(12):1746. doi:10.3390/medicina58121746

70. Aftab K, Aamir FB, Mallick S, et al. Radiomics for precision medicine in glioblastoma. *J Neurooncol*. 2022;156(2):217-231. doi:10.1007/s11060-021-03933-1

71. Aerts H, Schirner M, Dhollander T, et al. Modeling brain dynamics after tumor resection using The Virtual Brain. *NeuroImage*. 2020;213:116738. doi:10.1016/j.neuroimage.2020.116738