Online Resource 1

Preoperative Machine-Learning Models for Predicting Postoperative Morbidity after Anterior Clinoidal Meningioma Surgery: Development and Multinational Internal–External Validation

Acta Neurochirurgica

[1]Noa Ben Dor, MD; [1]Gal Ziv, BA; [1]Filippo Friso, MD; [1]Carlo Bortolotti, MD; [1,2]Diego Mazzatenta, MD, PhD; [1,2]Matteo Zoli, MD, PhD; [1]Matteo Martinoni, MD; [1]Filippo Badaloni, MD; [1]Giulia Casulli, MD; [1]Vittoria Rosetti, MD; [1]Luca Zanuttini, MD; [2]Lorenzo Faccin, MD; [5]Arish Mustafa; [3]Francesco Travaglini, MD; [3]Luigino Tosatto, MD; [4]He-ming Wu, MD; [5]Catherine Elizabeth Gilkes, MD; [5]Cathal John Hannan, MD, PhD; [5,6]Michael D. Jenkinson, MD, PhD; [3]Marcello D'Andrea, MD; [1,2]Alfredo Conti, MD, PhD.

1 Department of Neurosurgery; IRCCS Institute of Neurological Sciences of Bologna (ISNB), Bologna, Italy
2 Department of Biomedical and Neuromotor Sciences (DIBINEM), Alma Mater Studiorum – University of Bologna, Bologna, Italy
3 Department of Neurosurgery, Maurizio Bufalini Hospital, AUSL Romagna, Cesena, Italy
4 Department of Neurosurgery, Nanjing First Hospital, Nanjing Medical University, Nanjing, Jiangsu, China
5 Department of Neurosurgery, The Walton Centre NHS Foundation Trust, Liverpool L9 7LJ, UK
6 Institute of Systems, Molecular and Integrative Biology, University of Liverpool, Liverpool, UK

Corresponding author:
Noa Ben Dor, MD
Department of Neurosurgery, IRCCS Institute of Neurological Sciences (ISNB), Bologna, Italy
Via Altura 3, 40139 Bologna, Italy
E-mail: noa.bendorziv@studio.unibo.it
ORCID: 0009-0009-8608-6418

Supplementary methods: machine-learning methodology (sections S1–S8, Tables S1–S7, Fig. S1)

This document expands the Machine Learning Workflow section of the Methods. It describes the modelling population, predictor representation, model specification, validation, performance measures, sample size and software in enough detail for the analyses to be reproduced. Results are reported in the main text and in Online Resource 3.

S1. Modelling population and analysis design

Two prediction analyses were performed independently from the same database: histogram gradient boosting and logistic regression with ridge (L2) penalization. Of the 241 patients, the ten patients from the Chinese center were excluded from modelling because no complication of any type had been recorded and their records were otherwise not comparable with the cohort (see below and Table S1). Both analyses therefore used the same modelling population of 231 patients: Bologna (n = 115), Cesena (n = 68) and Liverpool (n = 48).

Three validation designs were applied to every endpoint: internal validation in all 231 patients; internal–external cross-validation, in which each center was held out in turn and the model refitted on the remaining centers; and a geographical split, with development in Bologna and Cesena (n = 183) and external validation in Liverpool (n = 48). Held-out data were withheld from predictor specification, penalty selection and all model fitting. Postoperative cognitive decline was not recorded in Cesena, so the cognitive-decline model was developed in Bologna (n = 115) and validated in Liverpool (n = 48). Because the two algorithms used different predictor sets and representations,

their estimates are reported side by side and are not a formal comparison of algorithms. The workflow is shown in Fig. S1.

Excluded center

The ten patients from the Chinese center differed systematically from the remainder of the cohort: no postoperative complication of any type was recorded (0 of 10, against 40.3% elsewhere; Fisher exact p = 0.008; the exact probability of observing no event in ten patients at the cohort rate is 0.006); every patient was discharged with a Karnofsky score of 90 or 100, including one admitted at 50; all had a Charlson Comorbidity Index of 0 or 1 (cohort mean 2.4); 30-day, 6-month and 12-month follow-up status was complete for all ten but missing for 56%, 76% and 76% of the other patients; and operative duration was recorded for these ten patients only. Whether this reflects case selection or a difference in how outcomes were recorded could not be determined; in either case the records were not comparable. The ten patients were excluded from all modelling analyses and retained in the descriptive tables (Tables 1 and 2 of the main text; Online Resource 2, Table S8). Model performance with the ten patients included is given in Online Resource 3, Table S9 and is marginally higher for every endpoint.

Table S1 compares the development, external and excluded groups. The withheld Liverpool cohort had larger tumors with more frequent cavernous sinus and major cerebral artery involvement, but a lower comorbidity burden.

S2. Outcomes, predictors and encoding

Outcomes were defined as in the Methods of the main text and ascertained at discharge: medical complications (systemic infection, deep vein thrombosis, acute respiratory insufficiency, heart failure), surgical complications (postoperative hematoma, seizure, wound infection, CSF leak, hydrocephalus), the composite of any new or worsened neurological deficit and its motor, cognitive and visual components. New ischemic stroke was assigned to the neurological domain. The non-neurological composite (any medical or surgical complication or new ischemic stroke) was analyzed as a sensitivity endpoint.

All predictors were recorded preoperatively (Table S2). The gradient-boosting analysis used 25 candidate variables. Sex, ASA class, tumor side and vessel-encasement category were one-hot encoded, each with an explicit missing-category indicator, giving 39 encoded inputs; the remaining variables were entered on their original scale. The ridge analysis used two fixed sets. The non-neurological set (age, Charlson Comorbidity Index, cardiorespiratory comorbidity, preoperative KPS, maximum axial tumor diameter, vessel encasement grade and ACA/MCA/PCA involvement; seven variables) was used for the medical, surgical and non-neurological composite endpoints. ASA class was part of the pre-specified non-neurological set and was removed before the final models were fitted because its adjusted association with medical complications was inverse and clinically implausible, reflecting its overlap with age and the Charlson index and inconsistent grading across centers (Online Resource 3, Table S16); the pre-specified eight-variable medical model is reported there as a sensitivity analysis. The neurological set (age, Charlson Comorbidity Index, preoperative KPS, preoperative motor symptoms, maximum axial tumor diameter, vessel encasement grade, ICA/carotid siphon involvement and cavernous sinus involvement; eight variables) was fixed for the composite neurological endpoint and applied unchanged to its components. In the ridge models vessel encasement was entered as an ordinal grade from 0 (none) to 4 (> 270°).

There were no missing values in any predictor or outcome variable of the modelling population, so no imputation was needed. Cognitive decline in Cesena was not recorded rather than absent; Cesena was therefore excluded from every analysis of that endpoint.

S3. Histogram gradient boosting

Models used scikit-learn's HistGradientBoostingClassifier with its default settings (logarithmic loss, learning rate 0.1, up to 100 boosting iterations, 31 leaf nodes, at least 20 samples per leaf, unrestricted depth, no L2 regularization) and early stopping on a 10% internal validation fraction after 10 iterations without improvement of the loss. No hyperparameter search and no class weighting were used. Internal performance was the mean AUC

over five repeats of stratified five-fold cross-validation. A high-sensitivity operating threshold was estimated within the development cohort (Bologna and Cesena) from repeated cross-validated predictions, as the largest threshold on a grid from 0.01 to 0.60 (steps of 0.01) that reached at least 90% sensitivity; it was then frozen and applied to the external predictions. Because a single boosted model depends on the random early-stopping split, the external AUC was also recomputed over 20 random seeds. Settings are listed in Table S3.

S4. Ridge-penalized logistic regression

Each model was a scikit-learn pipeline of StandardScaler and LogisticRegressionCV with an L2 penalty, the lbfgs solver, up to 2,000 iterations and no class weighting. The inverse penalty C was selected from eleven log-spaced values between 10^{-3} and 10^{2} by stratified five-fold cross-validation minimizing the log-loss, and was re-selected within every training set of every validation procedure. Predictors were standardized within each training set. For the final models fitted on all eligible patients, the selected values were C = 0.1 (medical), C = 0.032 (motor) and C = 0.01 (cognitive decline, Bologna). Coefficients are reported as standardized odds ratios per 1-SD increase, which describe relative contribution within the penalized model; confidence intervals are not given because penalized coefficients have no valid closed-form standard errors. Raw-scale coefficients and intercepts, which the online calculator implements, are given in Online Resource 3 (Tables S12 and S15). Settings are listed in Table S4.

S5. Validation procedures

Ridge models were validated internally by nested 10-fold stratified cross-validation repeated five times, with the penalty selected in the inner loop; the reported AUC is the mean over repeats, with a 95% percentile bootstrap interval (2,000 replicates) computed on the out-of-fold predictions averaged over repeats. Optimism was estimated with 200 bootstrap replicates that repeated the whole procedure, including penalty selection; the optimism-corrected AUC is the apparent AUC minus the mean optimism.

In internal–external cross-validation, the held-out AUC of a center was reported when it contained at least five events and five non-events, with a 1,000-replicate percentile bootstrap interval. Per-center AUCs were pooled on the logit scale by DerSimonian–Laird random-effects meta-analysis, with standard errors derived from the bootstrap interval width (upper minus lower logit bound divided by 3.92). With three contributing centers the between-center variance is estimated imprecisely. External AUCs in the geographical split used 2,000 bootstrap replicates. The cognitive-decline model followed the same procedures within Bologna before a single test in Liverpool.

Sensitivity analyses comprised the full cohort of 241 patients, the non-neurological composite endpoint, the motor model without the preoperative motor-symptom predictor, the pre-specified medical model with ASA class and, for cognitive decline, a model with preoperative cognitive or psychiatric symptoms in place of motor symptoms, nested cross-validation with Bologna and Liverpool pooled, and the reverse direction (development in Liverpool, test in Bologna). The designs of the two analyses are compared in Table S5.

S6. Performance measures

Discrimination was measured by the AUC. Calibration was summarized by the calibration slope (logistic regression of the outcome on the logit of the predicted probability) and by calibration-in-the-large (the intercept of that regression with the slope fixed at 1; negative values indicate over-prediction), and overall accuracy by the Brier score, shown with the Brier score of a model predicting the event rate for everyone. Sensitivity, specificity, positive and negative predictive values and net benefit were computed across probability thresholds from the held-out predictions of internal–external cross-validation, in which every patient is predicted by a model that has not seen his or her center. These thresholds describe trade-offs and were not validated as cut-points. Observed event rates were also summarized across tertiles of held-out predicted risk.

S7. Sample size

No a priori sample-size calculation was performed; the sample was the available multicenter cohort. Events per predictor are given in Table S6. The low ratios motivated the fixed parsimonious predictor sets and penalization of the ridge models, preset settings with early stopping for gradient boosting, strict separation of development and validation data, and resampling-based estimates of optimism.

Supplementary tables

Table S1 Composition of the development, external and excluded cohorts

Variable	Development (Bologna + Cesena)	External (Liverpool)	Excluded (n = 10)	p, development vs external
Patients	183	48	10	
Age, years (mean ± SD)	60.8 ± 11.9	59.8 ± 13.9	55.0 ± 12.4	0.675
Male sex, n (%)	55 (30.1)	12 (25.0)	6 (60.0)	0.593
Charlson Comorbidity Index (mean ± SD)	2.6 ± 1.7	2.1 ± 1.4	0.2 ± 0.4	0.043
ASA class III–IV, n (%)	105 (57.4)	15 (31.2)	0 (0.0)	0.002
Any comorbidity, n (%)	127 (69.4)	31 (64.6)	8 (80.0)	0.601
Cardiorespiratory disease n (%)	94 (51.4)	21 (43.8)	4 (40.0)	0.418
Preoperative KPS, median (IQR)	80 (70–90)	80 (70–80)	90 (82–98)	0.016
Preoperative motor deficit, n (%)	44 (24.0)	12 (25.0)	0 (0.0)	0.852
Preoperative visual deficit, n (%)	45 (24.6)	9 (18.8)	0 (0.0)	0.449
Maximum axial diameter, mm (mean ± SD)	35.1 ± 13.8	40.2 ± 14.9	38.0 ± 13.2	0.036
Vessel encasement grade 3–4, n (%)	57 (31.1)	22 (45.8)	0 (0.0)	0.062
ICA/carotid siphon involvement, n (%)	108 (59.0)	23 (47.9)	0 (0.0)	0.192
Cavernous sinus involvement, n (%)	29 (15.8)	16 (33.3)	0 (0.0)	0.013
ACA/MCA/PCA involvement, n (%)	102 (55.7)	35 (72.9)	0 (0.0)	0.033
Optic canal invasion, n (%)	23 (12.6)	8 (16.7)	0 (0.0)	0.478
Medical complication, n (%)	29 (15.8)	7 (14.6)	0 (0.0)	1.000
Surgical complication, n (%)	15 (8.2)	6 (12.5)	0 (0.0)	0.397
New ischemic stroke, n (%)	5 (2.7)	2 (4.2)	0 (0.0)	0.637
Non-neurological composite, n (%)	39 (21.3)	11 (22.9)	0 (0.0)	0.845
New or worsened motor deficit, n (%)	20 (10.9)	10 (20.8)	0 (0.0)	0.090
Cognitive decline, n (%) [a]	16 (8.7)	12 (25.0)	0 (0.0)	0.005
Visual deterioration, n (%)	14 (7.7)	5 (10.4)	0 (0.0)	0.557
Any new neurological deficit, n (%)	48 (26.2)	20 (41.7)	0 (0.0)	0.050
Any complication, n (%)	71 (38.8)	22 (45.8)	0 (0.0)	0.411

Values are mean ± SD, median (IQR) or n (%). p-values: Welch t-test (means), Mann–Whitney U test (medians) or Fisher exact test (proportions), development versus external cohort; the excluded group is shown for completeness only. [a] Cognitive decline was not recorded in Cesena; the development value is 16 of 115 (13.9%) in Bologna. The model predictor 'cardiorespiratory comorbidity' used the variable as recorded (development 81 of 183, 44.3%; external 7 of 48, 14.6%). ASA, American Society of Anesthesiologists; ICA, internal carotid artery; KPS, Karnofsky Performance Status.

Table S2 Candidate and model-specific predictor sets

Predictor	Gradient boosting	Ridge – non-neurological set	Ridge – neurological set
Demographic and comorbidity			
Age	✓	✓	✓
Sex	✓	–	–
Charlson Comorbidity Index	✓	✓	✓
ASA physical status	✓	–	–
Any comorbidity	✓	–	–
Cardiorespiratory comorbidity	✓	✓	–
Hypertension	✓	–	–
Hyperlipidemia	✓	–	–
Preoperative clinical status			

Predictor	Gradient boosting	Ridge – non-neurological set	Ridge – neurological set
Karnofsky Performance Status	✓	✓	✓
Headache	✓	–	–
Motor deficits	✓	–	✓
Visual deficits	✓	–	–
Radiological			
Maximum axial tumor diameter	✓	✓	✓
Tumor side	✓	–	–
Vessel encasement	✓ (categorical)	✓ (ordinal 0–4)	✓ (ordinal 0–4)
Midline shift	✓	–	–
Dural tail	✓	–	–
Hyperostosis	✓	–	–
Optic canal invasion	✓	–	–
Optic nerve or chiasm compression	✓	–	–
ICA compression	✓	–	–
Carotid siphon involvement	✓	–	✓
ICA encasement	✓	–	–
ACA/MCA/PCA involvement	✓	✓	–
Cavernous sinus involvement	✓	–	✓
Total predictors	25 (39 encoded inputs)	7	8

All predictors were recorded preoperatively and were available at the time of surgical counselling. The non-neurological set was used for the medical, surgical and non-neurological composite endpoints; the neurological set for the neurological composite, motor, cognitive and visual endpoints. ASA physical status was part of the pre-specified non-neurological set and was removed before the final models were fitted (Online Resource 3, Table S16). The count of 39 encoded inputs reflects one-hot encoding of sex, ASA class, tumor side and vessel-encasement category, each with an explicit missing-category indicator. ACA, anterior cerebral artery; ASA, American Society of Anesthesiologists; ICA, internal carotid artery; MCA, middle cerebral artery; PCA, posterior cerebral artery.

Table S3 Histogram gradient boosting: specification

Setting	Value
Estimator	scikit-learn HistGradientBoostingClassifier
Hyperparameter search	None; scikit-learn default settings
Loss	Logarithmic loss
Learning rate	0.1
Maximum boosting iterations	100
Maximum leaf nodes	31
Minimum samples per leaf	20
Maximum depth	Unrestricted
L2 regularization	None
Class weighting	None
Early stopping	10% internal validation fraction; stop after 10 iterations without improvement of the loss
Inputs	25 candidate variables, 39 encoded inputs (Table S2)
Internal validation	Stratified five-fold cross-validation, five repeats; AUC = mean over repeats
Operating threshold	Largest threshold (grid 0.01–0.60, step 0.01) with ≥ 90% sensitivity in repeated cross-validation within the development cohort; frozen before external evaluation
Seed sensitivity	External AUC recomputed over 20 random seeds
Random seed	2026

AUC, area under the receiver operating characteristic curve.

Table S4 Ridge-penalized logistic regression: specification

Setting	Value
Pipeline	scikit-learn Pipeline: StandardScaler, then LogisticRegressionCV
Penalty	L2 (ridge); lbfgs solver; up to 2,000 iterations; no class weighting
Penalty grid (C)	11 log-spaced values from 10^{-3} to 10^{2}
Penalty selection	Stratified five-fold cross-validation minimizing log-loss, repeated within every training set
Selected penalty, final models	Medical C = 0.1; motor C = 0.032; cognitive decline C = 0.01
Predictors	Two fixed sets of seven (non-neurological) and eight (neurological) variables (Table S2); standardized within each training set
Vessel encasement	Ordinal grade 0 (none) to 4 (> 270°)
Missing values	None in the modelling population
Coefficient reporting	Standardized odds ratios per 1 SD; raw-scale coefficients for the calculator (Online Resource 3, Tables S12 and S15)
Random seed	2026

C, inverse of the penalty strength (λ = 1/C); larger penalties (smaller C) shrink coefficients more strongly toward zero.

Table S5 Validation design and statistical inference, by analysis

Element	Gradient boosting	Ridge-penalized regression
Modelling population	231 patients (Bologna, Cesena, Liverpool)	231 patients (Bologna, Cesena, Liverpool)
Predictors	25 candidate variables (39 encoded inputs)	Fixed set per domain: seven (non-neurological) or eight (neurological) variables
Tuning	None (default settings with early stopping)	Penalty selected by inner five-fold cross-validation (log-loss)
Internal validation	Stratified five-fold cross-validation, five repeats	Nested 10-fold cross-validation, five repeats
Optimism	Not estimated separately	200-replicate bootstrap repeating the full procedure
Internal–external cross-validation	Each center held out in turn	Each center held out in turn; DerSimonian–Laird pooling
Geographical validation	Bologna + Cesena (n = 183) → Liverpool (n = 48)	Bologna + Cesena (n = 183) → Liverpool (n = 48)
Cognitive decline	Bologna (n = 115) → Liverpool (n = 48)	Bologna (n = 115) → Liverpool (n = 48)
AUC confidence intervals	Percentile bootstrap: 2,000 replicates; 1,000 per held-out center	Percentile bootstrap: 2,000 replicates; 1,000 per held-out center
Calibration	Calibration slope; Brier score	Calibration slope and calibration-in-the-large; Brier score
Thresholds	High-sensitivity threshold frozen in development, applied externally	Descriptive trade-offs across thresholds; decision-curve analysis

AUC, area under the receiver operating characteristic curve.

Table S6 Sample size and events per predictor

Endpoint	Patients / events, modelling population	Patients / events, development cohort	Events per predictor, ridge (231)	Events per predictor, ridge (development)	Events per input, gradient boosting
Medical complication	231 / 36	183 / 29	5.1	4.1	0.92
Surgical complication	231 / 21	183 / 15	3.0	2.1	0.54
Non-neurological composite	231 / 50	183 / 39	7.1	5.6	1.28
New or worsened motor deficit	231 / 30	183 / 20	3.8	2.5	0.77
Any new or worsened neurological deficit	231 / 68	183 / 48	8.5	6.0	1.74
Visual deterioration	231 / 19	183 / 14	2.4	1.8	0.49
Cognitive decline (Bologna)	–	115 / 16	–	2.0	0.41

Ridge models have seven (non-neurological set) or eight (neurological set) predictors; gradient-boosting models have 39 encoded inputs (events per input shown for the modelling population, and for Bologna for cognitive decline). Development cohort: Bologna and Cesena (n = 183), or Bologna alone (n = 115) for cognitive decline.

Table S7 Software environment

Software	Version	Use
Python	3.11.15	Programming environment
scikit-learn	1.8.0	Model fitting, cross-validation, metrics
NumPy	2.4.4	Numerical computing, bootstrap resampling
pandas	3.0.2	Data handling
SciPy	1.17.1	Statistical tests (Welch t, Mann–Whitney U, Fisher exact)
matplotlib	3.10.9	Figures

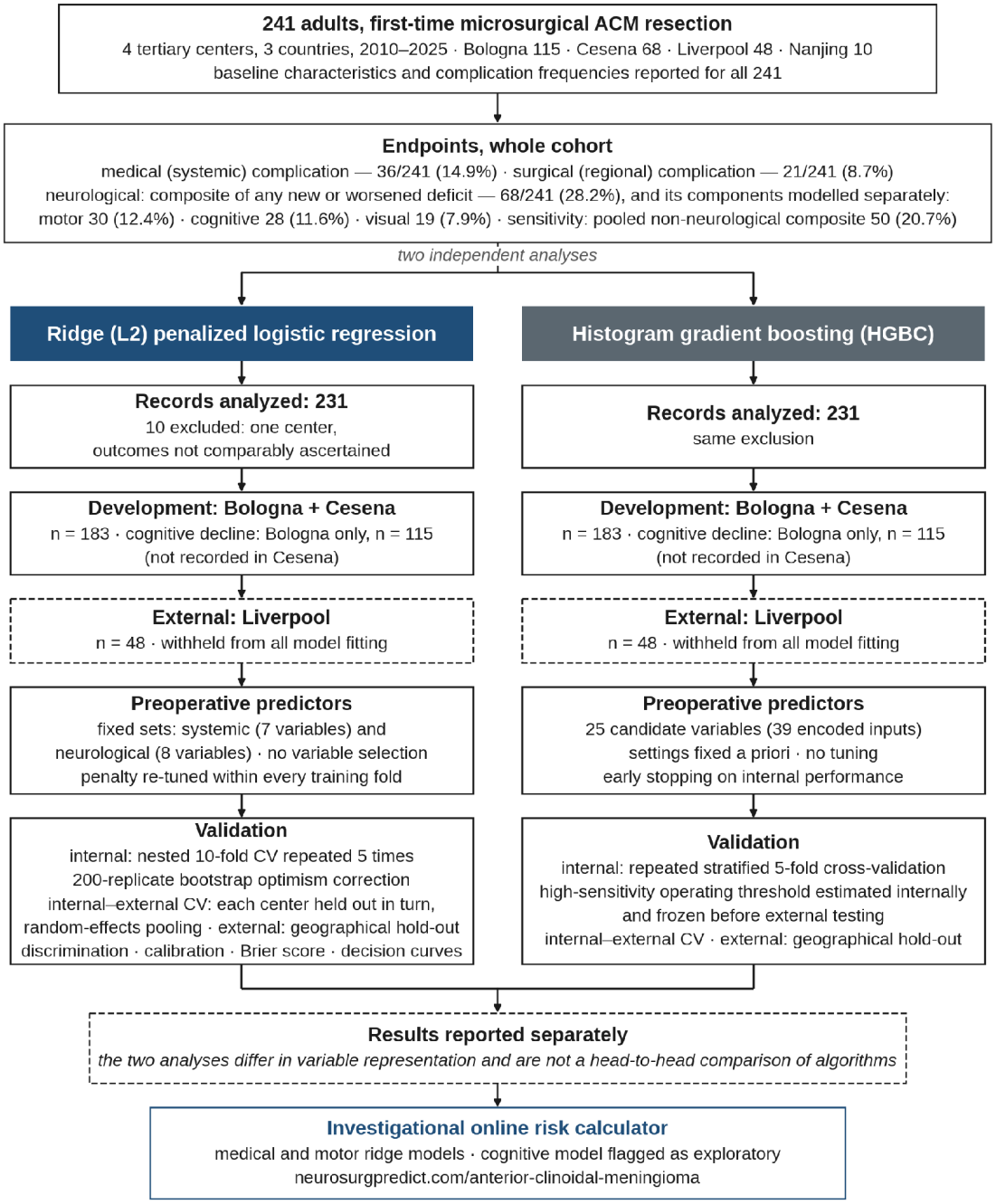


Fig. S1 Study and machine-learning workflow. Each complication category (medical, surgical, and the neurological composite with its motor, cognitive and visual components) was modeled from preoperative data alone in 231 patients; because cognitive decline was not recorded in Cesena, the cognitive model was developed in Bologna and validated in Liverpool. Gradient boosting and ridge analyses were conducted independently and reported separately. ACM, anterior clinoidal meningioma; CV, cross-validation; HGBC, histogram gradient boosting classifier; KPS, Karnofsky Performance Status.