International Journal of Scientific Research and Modern Technology (IJSRMT) *ijsrmt.com*
Volume 5, Issue 9, 2026
DOI: https://doi.org/10.38124/ijsrmt.v5i9.1664

Admission Glasgow Coma Scale Captures Most of the Measurable Prognostic Signal in Intracranial Suppurative Infections: An Explainable Machine-Learning Cohort Study

Faozo Stéphane Landry Teti[1]; Kouadio Jean-Baptiste Keke[2]; Konan Serge Yao[3]; Koffi Yves Soress Dongo[4]; Yao Bernard Fionko[5]; Aderehime Haidara[6]

[1,2,3,4,5,6] Department of Neurosurgery, Bouaké University Hospital, Bouaké, Côte d'Ivoire
[1,2,3,4,5,6] Faculty of Medical Sciences, Alassane Ouattara University, Bouaké, Côte d'Ivoire

Publication Date : 2026/09/26

Abstract

➢ *Background.*
Prognostic studies rank factors by the p value, which measures the compatibility of an association with the null hypothesis rather than the contribution of a variable to prediction. The actual added value of machine-learning models is rarely assessed against the best single clinical predictor.

➢ *Objective.*
To quantify, using several complementary definitions of importance, the concentration of the predictive signal among the variables available at admission, and to assess the added value of multivariable models relative to a prespecified clinical comparator.

➢ *Methods.*
Retrospective cohort of 150 admission episodes corresponding to 148 patients managed for intracranial suppurative infection at Bouaké University Hospital between 1 January 2016 and 31 January 2026; 142 episodes had an evaluable Glasgow Outcome Scale at discharge. Primary outcome: unfavourable functional outcome at discharge (GOS ≤ 3). Twenty-seven admission variables were entered into four prespecified algorithms — penalised logistic regression, random forest, XGBoost, LightGBM — compared with a model containing the Glasgow Coma Scale (GCS) alone, prespecified as the clinical comparator. The episode was the unit of prediction and the patient the unit of partitioning and resampling. Out-of-fold probabilities from 20 repetitions of grouped stratified 5-fold cross-validation were averaged; uncertainty was estimated by patient-level paired percentile bootstrap (1000 replications), the principal comparative inference. Discrimination, probabilistic accuracy, calibration and net benefit were assessed jointly, and importance was quantified by SHAP values, permutation importance and mutual information.

➢ *Results.*
Fifty-seven of the 142 episodes (40.1%) had an unfavourable outcome, including 16 deaths. GCS alone achieved an AUC of 0.896 (95% bootstrap CI 0.834–0.945) and a Brier score of 0.110, versus 0.849 to 0.869 and 0.134 to 0.146 for the four multivariable models. No model showed superior performance: paired Brier score differences excluded zero for all four comparisons, and AUC differences excluded zero versus penalised logistic regression (+0.047; +0.007 to +0.093) and LightGBM (+0.036; +0.003 to +0.073). GCS alone offered the highest net benefit across the range of thresholds explored. All three importance approaches ranked GCS first, but the estimated magnitude of concentration varied markedly with the

Teti, F. S. L., Dongo, K. Y. S., Fionko, Y. B., Yao, K. S., Keke, K. J.-B., & Haidara, A. (2026). Admission Glasgow Coma Scale Captures Most of the Measurable Prognostic Signal in Intracranial Suppurative Infections: An Explainable Machine- Learning Cohort Study. *International Journal of Scientific Research and Modern Technology*, *5*(9), 111–149. https://doi.org/10.38124/ijsrmt.v5i9.1664

metric. GCS ranked first in all 1000 internal resamples; beyond that, ranks were unstable. After selection recalculated within each training fold, the AUC fell from 0.903 with one variable to 0.854 with twenty-seven. The advantage of GCS persisted after exclusion of episodes with GCS 3–8 (0.851 versus 0.797–0.821) and within the range of clinical uncertainty, GCS 9–14 (0.769 versus 0.671–0.736). A prespecified parsimonious clinical model combining GCS, collection size and midline shift reached 0.911, without the difference from GCS alone being distinguishable from zero.

➢ *Conclusion.*

In this single-centre cohort, admission GCS captured most of the measurable predictive signal contained in the baseline variables, and additional algorithmic complexity did not improve validated performance. These observations support systematically comparing any complex model with a prespecified simple clinical predictor. External evaluation is required before any clinical implementation or methodological generalisation.

Keywords : *Intracranial Suppurative Infections; Brain Abscess; Empyema; Explainable Machine Learning; SHAP Values; Mutual Information; Calibration; Decision Curve Analysis; Glasgow Coma Scale; Global Neurosurgery.*

I. INTRODUCTION

➢ *What the p value does not tell us*

For forty years, the prognostic literature on intracranial suppurative infections has followed a stable grammar: a series is described, a univariable and then a multivariable analysis is performed, and the article concludes with a list of independent prognostic factors ordered by p value or by the magnitude of the odds ratio [1–5]. This grammar answers a precise question — is the association distinguishable from chance? — but it does not answer the one the clinician asks at the bedside: among everything I can observe at admission, what helps me anticipate this patient's outcome, and what merely repeats what I already know?

These two questions are distinct and may receive divergent answers [6,7]. A variable may be significantly associated with an outcome without improving out-of-sample prediction — because its effect is weak, unstable, or already contained in another variable. Since significance also depends on sample size, ranking by p value tells us little about predictive contribution.

Interpretable machine learning has made other metrics operational: SHAP values distribute a model's prediction across its inputs [8,9], permutation importance measures what performance loses when the link between a variable and outcome is destroyed [10,11], and mutual information quantifies the reduction in outcome entropy without recourse to any model [12,13]. These tools allow a shift from a discourse on association to a discourse on predictive contribution. Yet they are rarely deployed for the question that follows naturally from them: does the complex model do better than its dominant variable taken alone?

➢ *A Question of Particular Acuity in Resource-Limited Settings*

Intracranial suppurative infections remain serious everywhere: mortality from brain abscess still ranges between 10 and 20% in contemporary cohorts from high-income countries, and functional prognosis depends largely on initial neurological status [14,15]. International guidelines frame antibiotic therapy but leave prognostic stratification to clinical judgement [16–18]. In sub-Saharan Africa, presentation is later, mortality higher and diagnostic resources constrained.

Extending data collection is not neutral there: every additional variable costs staff time, consumables, sometimes a transfer. Knowing which ones deserve to be measured rigorously is therefore a question of resource allocation as much as of method.

➢ *Objectives*

This study pursues five objectives: (1) to compare, under a strictly identical protocol, four prespecified algorithms and a single-variable clinical comparator; (2) to quantify the concentration of the predictive signal using three complementary approaches, one of which relies on no fitted model; (3) to describe the functional form of the learned relationships; (4) to determine the minimum number of variables required, with selection recalculated within each training fold; (5) to verify that the conclusions withstand calibration, net benefit, optimism correction, hyperparameter optimisation and subgroup analyses.

We state explicitly the hypothesis we set out to confirm: that of a predictive signal distributed across many determinants, whose aggregation by a non-linear model would improve prediction. The results do not support it.

II. METHODS

➢ *Study Design, Setting and Population*

Retrospective single-centre analytical study conducted in the department of neurosurgery of Bouaké University Hospital (Côte d'Ivoire), a regional referral hospital. All patients managed for an imaging-documented intracranial suppurative infection between January 2016 and January 2026 were included: brain abscess, subdural empyema, extradural empyema, ventriculitis, pre-suppurative cerebritis and scalp suppuration with intracranial extension. The database comprises 150 admissions corresponding to 148 patients, two patients having had two admissions for recurrence.

Reporting follows the TRIPOD+AI recommendations [19,20]. Risk of bias was examined using the PROBAST domains [21], for which an updated version adapted to artificial-intelligence-based models has since been published [85]; in line with section 4.11, we do

not claim full compliance, the "Analysis" domain remaining at risk because of the small number of events and the absence of external validation. The flow diagram of observation selection and the analytical pipeline are shown in Figure 1.

➢ *Hierarchy of Analyses*

The primary analyses compare, on the same out-of-sample predictions, the discrimination, probabilistic accuracy, calibration and net benefit of the model based on the Glasgow Coma Scale (GCS) alone and of the four multivariable models. The variable importance analyses and the parsimony analysis are secondary. Hyperparameter optimisation, temporal validation, the ordinal model, reclassification indices, interactions, mortality prediction and subgroup analyses are exploratory. This hierarchy was fixed before the results were written up and is restated in Table S10.

The model based solely on the GCS was defined before analysis as the reference clinical comparator, because of its availability at first contact and its established prognostic association in intracranial infections [14,15,17]. The importance analyses subsequently confirmed its empirical first rank in this cohort; the comparator was therefore not chosen after inspection of the results.

➢ *Unit of Analysis and Prediction Horizon*

The admission episode was the unit of prediction; the patient was the unit of partitioning and resampling. This distinction allows readmissions to contribute as distinct clinical episodes while preventing any information leakage and accounting for within-patient dependence. Two patients each contributed two episodes; when a patient is drawn in a bootstrap resample, both of their episodes are drawn together.

Time zero corresponds to the first documented assessment at admission to neurosurgery, before definitive surgical treatment. The outcome is assessed at discharge, a non-uniform time horizon determined by length of stay; this limitation is discussed in section 4.11.

➢ *Outcomes*

The primary outcome is unfavourable outcome at discharge, defined by a Glasgow Outcome Scale (GOS) of 3 or less — death, vegetative state or severe disability with dependence [22]. The GOS was recorded for 142 of the 150 observations; the 8 observations without an evaluable GOS were excluded. The secondary outcome is in-hospital death (16 events), treated as an exploratory analysis of predictive feasibility and not as a model.

Dichotomising an ordinal scale entails a loss of information. A sensitivity analysis using an ordinal logistic model on the five GOS levels was therefore performed.

➢ *Candidate Variables*

Twenty-seven variables were retained. The selection principle is temporal and clinical, not statistical: only variables available at the time of the initial decision were retained. Therapeutic and course-related variables were excluded because they postdate the decision and may be mediators or consequences of prognosis. The bivariable comparisons in Table 1 are descriptive and exploratory; no correction for multiple comparisons was applied, since candidate variables were not selected on the basis of their p value.

The variables cover six domains: demographic, organisational, background, clinical, lesional and biological. The complete operational dictionary — definition, coding, unit, source and time of availability of each variable — is provided in the variable dictionary of Appendix S1. The variable measured in millimetres corresponds to midline shift and not to cerebral herniation in the anatomical sense; it is designated as such throughout the manuscript. Several candidate variables were expected to capture overlapping dimensions of intracranial mass effect; they were deliberately retained in order to quantify predictive redundancy rather than discarded on the sole basis of their correlation. Two non-measurement indicators were added, on the grounds that failure to perform an investigation may reflect resource constraints or severity-dependent diagnostic pathways; a sensitivity analysis compares models with and without these indicators.

One reservation must be stated regarding the horizon of availability. Bacteriological culture is not in practice available at the time of the initial decision: its result follows the specimen, sometimes operative, by several days. The variable set therefore corresponds to the initial inpatient work-up rather than to information strictly available at the bedside at first assessment. A sensitivity analysis excluding positive culture and its non-measurement indicator — that is, a restricted set of 25 variables genuinely available at the bedside — was conducted and is reported in section 3.13.5. The exclusion of therapeutic variables is aimed at availability at the time of prediction and prevention of temporal leakage, not at building a causal model.

➢ *Missing Data and Preprocessing*

Missing data are detailed in Table 2. Imputation was performed using the median learned in the training fold, applied to continuous variables as well as to binary variables coded 0/1; for the latter, the median coincides with the majority category of the fold, so that imputation is equivalent to mode imputation. It was computed exclusively within each training fold and applied to the test fold. No imputation was applied to the outcome: the 8 episodes without an evaluable GOS were excluded, which constitutes a complete-case analysis for the outcome. Standardisation, applied to the linear models only, covered all variables in the set, binary variables being coded 0/1; it was likewise learned in the training fold. Imputation and standardisation are encapsulated in the same pipeline object for all models, which guarantees the absence of information leakage and strict identity of preprocessing across compared models.

➢ *Models and Hyperparameters*

Four models were developed on the same variable set and the same partitions: penalised logistic regression (L2, C = 0.5), random forest (600 trees, depth 4, 5 observations minimum per leaf), XGBoost (300 trees, depth 2, learning rate 0.05, L2 regularisation = 3) and LightGBM (300 trees, 7 leaves, depth 3). A fifth model, containing the GCS alone, serves as the clinical comparator and follows exactly the same pipeline.

The values retained come from the strongly regularised ranges recommended for small samples rather than from a pilot study: shallow depth and low learning rate to limit variance, a high number of trees to compensate for that rate, explicit L2 regularisation. The complete parameterisation of each algorithm, including solver, subsampling, feature fraction, minimum leaf size and seeds, is given in Table S11. No class rebalancing was applied, the prevalence of 40.1% not warranting it, and no early stopping procedure was used. Hyperparameters were fixed a priori. This strategy aimed to limit optimisation variance and the risk of overfitting, at the possible cost of slightly suboptimal performance. To address the objection that this choice might disadvantage the complex models, two complementary analyses were conducted: a nested cross-validation with hyperparameter optimisation by random search in the inner loop, and a sensitivity analysis exploring several regularisation and depth values for each family.

The events-per-variable ratio is 2.1 for the primary outcome, well below usual recommendations [23,24]. This constraint is acknowledged and structures all the methodological choices.

➢ *Estimation of Performance and Uncertainty*

Validation relies on stratified 5-fold cross-validation grouped by patient identifier, repeated 20 times. Choosing 5 folds rather than 10 preserves a sufficient training sample size while leaving about 11 events per test fold; twentyfold repetition brings the partitioning variability of the aggregated AUC below 0.005. The figure of 1000 bootstrap replications was chosen so that the 2.5th and 97.5th percentiles rest on about 25 ordered values on each side. Each episode receives one out-of-sample prediction per repetition, that is 20 probabilities averaged for the primary analysis. Grouping by patient guarantees that the two admissions of the same patient can never be split between training and test.

Two modes of aggregation coexist in the literature and do not answer the same question; we report both. The first computes the metric separately at each repetition and then averages the values: it describes the variability related to partitioning, and the corresponding 2.5–97.5 percentiles do not constitute sampling confidence intervals. The second, retained as the primary one, averages for each observation the out-of-sample probabilities obtained across the 20 repetitions, then computes the metrics on these aggregated probabilities; uncertainty is then estimated by patient-level bootstrap, with 1000 replications.

The estimates derived from repeated cross-validation, nested cross-validation, temporal validation and optimism correction correspond to distinct estimands and are interpreted within each procedure, never compared across procedures. The metrics reported are the AUC [25], the Brier score [26], the Brier Skill Score relative to the null model predicting the prevalence for all patients, the area under the precision–recall curve [27] interpreted relative to the prevalence reference, the calibration slope and intercept [28,29], the expected calibration error, and the Spiegelhalter test applied to the aggregated out-of-sample probabilities. Calibration is assessed principally by the curve, the slope and the recalibration intercept; formal goodness-of-fit tests are considered secondary because of their limited power and their dependence on grouping. The expected calibration error was computed over five equal-sized bins and interpreted descriptively, its value depending on the number of bins. The Brier Skill Score is defined as 1 − (model Brier / null model Brier), the null model predicting the prevalence of the analysis sample. Clinical utility was assessed by decision curve analysis [30].

Performance differences between GCS alone and each multivariable model were estimated by patient-level paired bootstrap, the predictions bearing on the same observations: we report the difference in AUC and Brier score, their 95% percentile intervals, the proportion of resamples in which the AUC difference is positive, and the DeLong test for correlated areas [31]. The patient-level paired bootstrap intervals constitute the principal comparative inference; the DeLong test is reported as a secondary conventional analysis, conditional on the scores obtained and not accounting for the uncertainty related to repeated model training. Two distinct bootstrap procedures are used and must not be confused: a patient-level bootstrap of the metrics, which resamples the episode–prediction pairs already obtained without retraining the models, and a refitting bootstrap for optimism estimation (section 2.9).

➢ *Optimism Correction*

A bootstrap validation following Harrell's procedure was conducted: the model is refitted on each bootstrap sample, performance is measured on that sample and then on the original sample, and the mean difference provides the optimism, which is subtracted from the apparent performance [32,33].

➢ *Quantifying Importance: Three Complementary Approaches*

- *General Framework and Terminology*

No single importance measure answers the question posed on its own: each defines importance differently, and their divergences are interpretable [34]. We therefore use three metrics and systematically compare their results. For the sake of accuracy, we distinguish throughout the text between the relative share of global importance normalised according to absolute SHAP values, the relative share of the positive AUC loss attributed to each variable by permutation, and mutual information

expressed in bits and as a percentage of the outcome entropy. The term information is retained as a narrative concept, its operational definition being restated at each use.

- *SHAP Values*

SHAP values attribute to each variable, for each patient, its contribution to the gap between the prediction and the mean prediction, according to an allocation based on Shapley values satisfying the usual axioms of additive explanation for a given value function and background distribution [8]. This uniqueness is relative to those choices: several formulations coexist depending on whether a conditional or an interventional distribution is adopted, and on the reference retained [35,36]. SHAP values quantify a contribution to a model's prediction and must receive no causal interpretation. We used the linear explainer for logistic regression and the tree explainer in path-dependent mode (tree_path_dependent) for the three tree-based models, with the observations of the fold concerned as background distribution [9]. This mode does not rest on a fully independent background distribution, but its attributions remain specific to the model and to the dependence assumptions retained; in the presence of correlated variables, attribution remains shared, a property we explicitly exploit in section 3.6. The SHAP hierarchy reported in Table 5 comes from the XGBoost model prespecified as the primary explainability model; rank concordance across the four algorithms is reported as a sensitivity analysis. SHAP values explain the predictive function learned by a given model, not the clinical phenomenon itself.

Values were computed on episodes excluded from training, from models fitted exclusively on the corresponding training folds. The share attributed to a variable is the mean of the absolute value of its contribution, divided by the sum of these means, expressed as a percentage. Its distribution was obtained by 1000 patient-level bootstrap resamples, the model being refitted and SHAP values recomputed at each replication; we report the median, the 95% percentile interval, the median rank, the rank interval and the probability of belonging to the top three variables. Bootstrap median shares do not sum exactly to 100% by construction.

- *Permutation Importance*

For each variable, the observed values were randomly permuted within the test fold and the AUC loss measured, the operation being repeated thirty times per variable and per fold. We report the mean loss, its standard deviation and the proportion of permutations producing a positive loss, the latter making it possible to distinguish a weak but stable contribution from an estimate oscillating around zero. Relative contributions were computed after truncating negative losses to zero, then normalising by the sum of positive losses, and averaged across the four algorithms. This convention is made explicit because it mechanically increases the relative share of variables with a positive loss: the value obtained for the rank-1 variable depends strongly on it, and results without truncation are provided in the replication archive. Permutation importance measures a performance loss of a given model and therefore remains, like SHAP values, specific to the model to which it is applied.

- *Mutual Information*

Mutual information between each variable, discretised into quartiles, and the outcome was estimated by direct counting, with a significance test using 500 permutations of the outcome. Conditional mutual information given the GCS, computed as a weighted average over the Glasgow strata, measures what each variable adds once the GCS is known. This approach depends on no fitted predictive model and therefore provides a complementary check on the two preceding ones [12,13]; it nevertheless depends on the discretisation retained, on the estimator and on the handling of missing data, and does not constitute a neutral arbiter. Quartile discretisation was chosen a priori as a compromise between resolution and stability of cell counts, and computed on the whole analysis sample: this is a global descriptive analysis and not an out-of-sample estimate, which distinguishes it from the performance measures. Count-based estimation is positively biased in small samples; significance was assessed by 500 permutations of the outcome, which bounds the attainable p value at about 0.002. Two denominators are used and systematically specified: the percentage of the outcome entropy and the percentage of the sum of univariable mutual information estimates, this sum not constituting a total theoretical quantity of information because of double counting across correlated variables.

- *Marginal Effects and Interactions*

Partial dependence plots describe the mean predicted probability as a function of a variable [37]. Their known limitation — extrapolation towards unobserved combinations — is mitigated by displaying the marginal distribution beneath each curve and by restricting interpretation to populated ranges; accumulated local effects, less sensitive to correlation, were computed in addition [38]. SHAP interaction values were estimated to verify that no substantial interaction had been overlooked.

➢ *Parsimony Analysis with Nested Selection*

The ranking of variables was recalculated within each training fold, from the SHAP values of a model fitted on that fold alone, after which the top k variables were used to fit a model evaluated on the corresponding test fold. This nesting eliminates the optimistic bias of a selection performed on the whole dataset. The marginal gain of each variable added to GCS alone was estimated using the same procedure, with a bootstrap interval and the proportion of resamples in which the gain is positive.

➢ *Functional Form of the Clinical Comparator*

Three specifications of the GCS were compared with the pipeline held identical: linear on the log-odds scale, categorised into three strata, and modelled by a restricted cubic spline with three knots. Uncertainty of the prognostic threshold was assessed by estimating, in each of 1000 resamples, the threshold maximising the Youden index, and by reporting the resulting distribution. Any

interpretation of a data-derived threshold is presented as exploratory.

➢ *Sensitivity and Subgroup Analyses*

Three prespecified subgroup analyses aim to determine whether the advantage of the clinical comparator depends on the episodes whose outcome is most obvious: exclusion of episodes with GCS 3–8, where complete separation is observed; restriction to the GCS 9–14 range, chosen before analysis as the zone of clinical uncertainty because it excludes both extremes; and restriction to survivors, in order to check whether the signal also bears on severe disability and not on mortality alone. A prespecified parsimonious clinical model combining the GCS, the maximum size of the collection and midline shift was compared with GCS alone. The following sensitivity analyses were also conducted: mortality outcome with bootstrap intervals; ordinal logistic model on the full GOS; sensitivity to hyperparameters; models with and without non-measurement indicators; temporal cross-validation between the 2016–2020 and 2021–2026 periods; comparison of the information-based hierarchy with the p-value-based hierarchy; net reclassification and integrated discrimination indices, reported for exploratory purposes given their known limitations [39].

➢ *Tools and Reproducibility*

Analyses were conducted under Python 3.12 (scikit-learn 1.8.0, xgboost 3.3.0, lightgbm 4.7.0, shap 0.52.0, statsmodels 0.14.6, matplotlib 3.10.8), with fixed random seeds. Appendix S1 provides the complete code, a versioned dependency file, a containerised environment, the variable dictionary and an execution notebook reproducing all figures, tables and plots. The equation of the single-variable model is given in section 3.10.

➢ *Ethics and Patient Involvement*

Retrospective study of records anonymised at source, approved by the institutional ethics committee of Bouaké University Hospital (CEI/CHU-BKE/126/2026, favourable opinion of 14 July 2026) with waiver of individual consent. Neither patients nor the public were involved in the design of this methodological strand; the main result is intended to be fed back to front-line teams in the form of a single instruction rather than a form.

III. RESULTS

➢ *Population and Missing Data*

Of 150 admission episodes corresponding to 148 patients, 142 episodes — corresponding to 140 unique patients, including the two patients who each contributed two episodes — had an evaluable GOS and constitute the analysis sample (Figure 1A). The 8 excluded episodes did not differ markedly from the 142 included ones in age (median 28 versus 26 years; p = 0.72), admission GCS (12 versus 13; p = 0.67) or symptom-to-admission delay (18 versus 19 days; p = 0.38) (Table S16), but with eight observations these comparisons are devoid of power and their exclusion cannot be presumed non-informative. Fifty-seven (40.1%) had an unfavourable outcome, including 16 deaths. Median age was 26 years (interquartile range 14–51), median symptom-to-admission delay 19 days (15–25), median GCS 13 (10–15); 17 patients (12.0%) were admitted with a Glasgow score between 3 and 8.

Table 1 presents the characteristics compared according to outcome. Only three variables differ markedly between the groups: the GCS, the maximum size of the collection and midline shift ($p < 0.001$ for all three). Neither symptom-to-admission delay ($p = 0.175$), nor age ($p = 0.708$), nor referral status ($p = 1.000$) shows a significant bivariable association. Table 2 details the missing data, which concern four variables and at most 20.4% of observations.

➢ *The Single Clinical Predictor is not Outperformed by the Multivariable Models*

Table 3 and Figure 2 report performance. GCS alone achieves an AUC of 0.896 (95% patient-level bootstrap interval 0.834–0.945), a Brier score of 0.110 and a Brier Skill Score of 0.54. The four multivariable models achieve AUCs of 0.849 to 0.869, Brier scores of 0.134 to 0.146 and Brier Skill Scores of 0.39 to 0.44. Relative to the null model predicting prevalence (Brier score 0.240), all approaches improve prediction markedly; GCS alone, however, achieves the best probabilistic accuracy.

All areas under the precision–recall curve far exceed the prevalence reference of 0.401, GCS alone again achieving the highest value. The two aggregation modes give close but not identical values for the GCS (0.896 aggregated, 0.911 averaged across repetitions), an expected discrepancy reported in Table 3 so that no reading is open to confusion.

➢ *Paired Performance Differences*

Table 4 and Figure 3 report the paired comparisons, the only statistically admissible basis for the central result.

The Brier score differences exclude zero for all four models: from −0.024 (95% CI −0.040 to −0.009) versus XGBoost to −0.036 (−0.056 to −0.016) versus logistic regression, all in favour of the Glasgow score. The AUC differences exclude zero versus penalised logistic regression (+0.0473; +0.0071 to +0.0925; DeLong p = 0.027) and, more narrowly, versus LightGBM (+0.036; +0.0030 to +0.073; p = 0.039), but not versus XGBoost (+0.0351; −0.0058 to +0.0751) or random forest (+0.0271; −0.0241 to +0.0724). The lower bound of the comparison with LightGBM is reported to four decimal places (+0.0030) because it is on this bound that the exclusion of zero rests and usual rounding to three decimals would mask it. The proportion of resamples in which the AUC difference is positive ranges from 87% to 99% depending on the model; this is a descriptive measure of comparative stability and not a probability of superiority.

The statement these results permit is therefore precise: none of the multivariable models evaluated showed performance superior to the model based on GCS alone; all showed inferior probabilistic accuracy, and two

inferior discrimination. The statistical evidence is strongest against penalised logistic regression and less conclusive for the tree-based methods. We do not claim that the GCS outperforms every possible model.

➢ *Calibration, Clinical Utility and Optimism*

Figure 4 presents the out-of-sample calibration curves. GCS alone shows a slope of 1.27 (95% CI 0.89–1.66) and a null intercept; XGBoost a slope of 0.86 and the lowest expected calibration error (0.039). The random forest shows a slope of 1.87, indicating underfitting of extreme probabilities, with a significant Spiegelhalter test ($p = 0.002$); LightGBM a slope of 0.69 ($p = 0.009$). We refrain from attributing the advantage observed on the Brier score to the calibration component alone, since the Brier score aggregates discrimination and calibration.

Decision curve analysis (Figure 5) shows that GCS alone yields the highest net benefit across the whole range of clinically plausible thresholds: 0.296 at the 30% threshold, versus 0.280 for XGBoost and 0.257 for LightGBM. No multivariable model exceeds the GCS at any threshold examined. This result directly addresses the objection that predictive redundancy would not amount to an absence of clinical utility: here, the two converge.

Optimism correction by refitting bootstrap (Table S1) calls for cautious reading. The AUC optimism estimated for GCS alone is negligible and rounds to 0.000; the optimism of its Brier score is small but not null (0.108 apparent, 0.111 corrected). The multivariable models show an AUC optimism of 0.042 to 0.060 and a Brier optimism of 0.033 to 0.055; LightGBM reaches an apparent AUC of 1.000 with 142 episodes and 27 variables, a clear signal of overfitting. The larger gap between apparent and validated performance observed for the multivariable models is consistent with greater susceptibility to overfitting given the available sample size and dimensionality, without all of it being attributable to algorithmic complexity alone. The AUCs corrected by this procedure (up to 0.955 for LightGBM) remain markedly higher than the out-of-sample values from cross-validation (0.860): optimism correction estimates the performance of a different development process and itself remains optimistic in small samples. It should therefore not be read as an estimate of external performance, and this table illustrates the limitations of apparent evaluation rather than a high corrected performance.

➢ *Three Complementary Metrics Converge*

Table 6 and Figure 6 report the three importance metrics.

According to the SHAP values of the prespecified XGBoost model, the GCS accounts for 37.2% of normalised global importance (bootstrap interval 27.3–50.2), ahead of the maximum size of the collection (10.5%), midline shift (8.2%), white cell count (6.7%) and age (4.7%). According to permutation importance averaged across the four algorithms, it accounts for 86.6% of the normalised sum of positive AUC losses — a value that depends strongly on the prespecified truncation of negative losses and must not be read as a share of information, with a mean loss of 0.293 (standard deviation 0.052) positive in 100% of permutations; no other variable exceeds 3.8%. According to mutual information, a metric resting on no fitted predictive model, it carries 0.428 bit, that is 44.1% of the outcome entropy (0.971 bit) and 49.9% of the sum of univariable mutual information estimates between candidate variables and outcome (0.857 bit; $p \leq 0.002$, the resolution bound of 500 permutations). These two denominators answer different questions and are systematically specified; the sum of univariable mutual information values is not a total quantity of information, because of double counting across correlated variables.

The proportion of permutations producing a positive loss clearly distinguishes two regimes: 100% for the GCS, 67% for size, 64% for midline shift, but 48% for the raised intracranial pressure syndrome and 45% for the white cell count. For these latter variables, the estimate oscillates around zero: we conclude that no detectable contribution of their own exists in this cohort and with this procedure, not that noise has been demonstrated.

➢ *The Gap Between Metrics Measures Redundancy*

The gap between 37% (SHAP) and 87% (permutation) is not an inconsistency: it quantifies the redundancy of the system. SHAP values share credit across correlated variables; permutation attributes to each only what no other can replace. Now the GCS is correlated with the size of the collection (Spearman −0.53; 95% CI −0.64 to −0.40) and with midline shift (−0.33; −0.47 to −0.17), correlations reported together with the full matrix in Figure S4.

Conditional mutual information confirms this reading: once the GCS is known, size adds no more than 0.048 bit and midline shift 0.056 bit, that is less than 6% of the outcome entropy each.

We therefore state the finding as follows: the GCS concentrates a large share of predictive importance and renders largely substitutable the information carried by several correlated variables. This is the formulation the data support.

➢ *An Information Concentration Index*

The phenomenon observed can usefully be summarised by a simple descriptive statistic. We introduce the Information Concentration Index (ICI), defined as the share of total importance attributed to the rank-1 variable, expressed as a percentage. Since importance depends both on the metric and, for two of the three metrics, on the model being explained, the index is denoted ICI(M,A), where M designates the metric and A the algorithm concerned; for permutation importance, the truncation convention forms part of the definition. Under an equal distribution across 27 variables, the index would be 3.7%: this is a uniform-distribution benchmark and not a sampling null hypothesis.

In this cohort, the ICI is 37.2% (CI 27.3–50.2) according to SHAP values, 86.6% according to

permutation importance and 49.9% (CI 29.7–52.3) according to mutual information (Figure 6B). These three values, arising from distinct definitions, are all an order of magnitude above the uniform benchmark. They do not constitute three interchangeable estimates of the same latent quantity and must be neither aggregated nor numerically compared with one another: their convergence is appreciated qualitatively, through the fact that they all designate the same variable as rank 1. The index could facilitate comparison between cohorts using the same metric and the same type of model, provided that subsequent external evaluations confirm its usefulness.

➢ *Functional Form and Threshold Uncertainty*

Table 6 and Figure 9A compare three specifications of the Glasgow score. The linear form on the log-odds scale achieves an AUC of 0.896 (0.840–0.942) and the restricted cubic spline with three knots placed at GCS 8, 12 and 14 achieves 0.895 (0.828–0.948): no measurable predictive improvement is observed with the additional flexibility, which does not demonstrate that the relationship is strictly linear on the logit scale. Categorisation into three strata, by contrast, lowers the AUC to 0.762 (0.669–0.855).

This result substantially qualifies the step-function reading suggested by the partial dependence plots: the relationship does include a transition zone, but dichotomising or categorising the score destroys a substantial part of its predictive value. The continuous form must be retained.

The bootstrap distribution of the threshold maximising the Youden index (Figure 9B) places 88.0% of the estimates at 12 or 13 (12: 40.9%; 13: 47.1%), with 3.2% at 11 and 8.8% at 14. The transition zone around 12–13 is therefore reproducible, but the exact threshold is not: any decisional use of a single threshold remains exploratory and requires independent validation.

The observed risk by stratum (Figure 9C) is 100% for a Glasgow score of 3 to 8 (17/17; exact Clopper–Pearson interval 80.5–100%), 68.9% between 9 and 12 (53.4–81.8%) and 11.2% between 13 and 15 (5.3–20.3%). The complete separation of the most severe group explains the high discrimination and precludes estimation of a finite conventional odds ratio; the exact interval makes the uncertainty visible despite the proportion of 100%.

➢ *Parsimony Withstands Nested Selection*

Table 7 and Figure 7 report the parsimony analysis, with the ranking recalculated within each training fold. With logistic regression, the AUC is 0.903 (0.847–0.950) with one variable, 0.907 with three, 0.897 with eight, 0.871 with twenty and 0.854 (0.786–0.913) with twenty-seven. With XGBoost, it falls from 0.887 to 0.870.

This result matters because the earlier version of this analysis used a ranking established on the whole dataset, which made it optimistic. Nesting the selection does not alter the shape of the curve: performance plateaus from the very first variables onwards and then declines slowly, a classic dilution phenomenon.

The marginal gain analysis (Table S2) is more direct still. Added to GCS alone, the most favourable variable — subdural empyema — yields +0.0103 of AUC, with an interval from −0.0074 to +0.0368 and a positive gain in only 79% of resamples. None of the twenty-six variables yields a gain whose interval excludes zero.

➢ *Stability of the Hierarchy and Model Equation*

Across 1000 patient-level bootstrap resamples (Figure 8), the GCS holds first rank in 100% of cases, with a rank interval reduced to [1; 1] and a probability of belonging to the top three variables of 1.00. The following variables are poorly separated: maximum size has a median rank of 3 (interval 2–10) and a probability of appearing in the top three of 0.63; midline shift, 0.43; white cell count, 0.29. No reliable ranking can be claimed beyond first place, and we refrain from doing so.

The single-variable model is written, on unstandardised data:

$$\mathbf{logit(p) = 11.118 - 0.949 \times GCS}$$

That is, p = 1 / (1 + e^−(11.118 − 0.949 × GCS)). The intercept is 11.118 (standard error 1.812; 95% CI 7.567–14.669) and the GCS coefficient −0.949 (standard error 0.147; 95% CI −1.237 to −0.661; p < 0.001), that is an odds ratio of 0.387 (95% CI 0.290–0.516) per additional GCS point: each point gained is associated with a 61.3% reduction in the odds — not the probability — of unfavourable outcome in the fitted model. The patient-level bootstrap intervals are concordant (intercept 8.621–15.711; coefficient −1.315 to −0.738). The corresponding predicted probabilities, with bootstrap interval, are about 4% at a GCS of 15 (1.3–8.7), 23% at 13 (12.3–34.0), 66% at 11 (52.2–81.3) and 93% at 9 (86.1–98.1).

Two clarifications are indispensable. These coefficients come from a final refitting of the single-variable model on all 142 episodes, with unstandardised GCS values; this refitting was not used to estimate performance, which rests exclusively on the out-of-sample predictions. A penalised estimation, conducted because of the complete separation observed at the lowest GCS values, gives attenuated coefficients (intercept 9.576; slope −0.825) without changing the direction or the order of magnitude. The intervals moreover quantify the uncertainty of the estimated mean risk function and do not represent the individual prognostic uncertainty of a given patient.

The equation is provided for replication and external evaluation purposes, not for decisional use at the bedside as it stands.

➢ *Interactions and Marginal Effects*

SHAP interaction values (Figure 12B, Table S3) attribute 76.6% of total importance to main effects and 23.4% to interactions as a whole. The strongest, age ×

Glasgow, represents only 2.6% of the total, and no other exceeds 2.2%. No interaction therefore dominates the prediction, although interactions as a whole represent 23.4% of the attribution under this decomposition. This result provides a partial and complementary explanation for the absence of gain from non-linear methods; it is not on its own sufficient to establish it, since a non-linear model may also capture marginal non-linearities or local effects. The absence of empirical gain is established by the performance comparison, not by this decomposition.

The partial dependence plots (Figure S8) and the accumulated local effects agree: the Glasgow curve describes a steady decline with a marked transition between 11 and 13, that of maximum size a monotonic rise plateauing beyond 45 mm, and that of symptom-to-admission delay a practically flat curve.

➢ *Sensitivity Analyses*

- *Hyperparameter Optimisation and Robustness of the Parameterisation*

In nested cross-validation with optimisation by random search in the inner loop (Table S4, Figure S5D), the AUCs are 0.832 for logistic regression, 0.881 for random forest, 0.873 for XGBoost and 0.877 for LightGBM, versus 0.919 for GCS alone. The maximum gain from optimisation is +0.013. The sensitivity analysis of the parameterisation gives AUCs between 0.848 and 0.871 across all configurations tested. The objection that the complex models had been disadvantaged by an overly regularised parameterisation is therefore not supported by the data.

- *Non-Measurement Indicators*

Removing the two non-measurement indicators does not degrade performance and slightly improves it for logistic regression (0.865 versus 0.855) as well as for XGBoost (0.869 versus 0.868). These indicators, retained on the grounds that the absence of an investigation might be informative in a constrained system, therefore add nothing measurable here. We report this negative result and retain the indicators in the primary analysis for transparency, the conclusion being identical under both configurations.

- *Temporal Cross-Validation*

The cohort shows no detectable drift between 2016–2020 and 2021–2026: unfavourable outcome 38.2% versus 42.4% ($p = 0.612$), median Glasgow score 14 versus 13 ($p = 0.236$), median delay 19 days in both periods ($p = 0.572$). Trained on the earlier period and tested on the more recent one, GCS alone achieves an AUC of 0.944 and the full model 0.828; in the reverse direction, 0.903 versus 0.812 (Figure S5C). The simple predictor transports better over time than the complex model.

- *Ordinal Analysis*

An ordinal logistic model on the five GOS levels confirms the association of the Glasgow score with outcome on the full scale, without dichotomisation (coefficient 0.582; proportional odds ratio 1.79 per Glasgow point; $p < 0.001$). Dichotomisation therefore does not create the result.

- *Mortality: An Inconclusive Exploratory Analysis*

With 16 deaths, the intervals are too wide to draw conclusions. An audit of this analysis became necessary, since the earlier version of this work reported an out-of-sample AUC of 0.436 for GCS alone whereas the AUC of the raw score is 0.659. This is not a reversal of orientation — patients who died had a lower mean GCS (10.94 versus 12.35) — but an artefact of the model used: with three deaths per test fold, the tree ensemble produces probabilities that are almost indistinguishable between those who died and those who survived (0.118 versus 0.114) and its out-of-sample AUC becomes noise. Reported with the logistic model, the out-of-sample AUC of GCS alone is 0.623 (95% CI 0.472–0.755) and that of the 27-variable model 0.476 (0.320–0.633). Neither of these values can be distinguished from chance, but neither can the absence of signal be asserted. We report this episode because it directly illustrates the point of this work: with so few events, an ensemble method can produce a meaningless estimate that an elementary descriptive inspection suffices to unmask.

We therefore present this analysis as an exploratory assessment of feasibility and not as a result, and draw no clinical conclusion from it: with this number of events, no generalisable multivariable model can be built, which is in itself a warning for a literature that regularly publishes such models on comparable sample sizes. Deaths are moreover distributed across the whole GCS scale, four of them occurring in patients admitted with a GCS of 13 (Table S14), which sheds light on the difficulty of predicting them.

- *Reclassification*

The integrated discrimination indices favour GCS alone versus logistic regression (IDI +0.057) and, more weakly, versus XGBoost (+0.011). The continuous net reclassification index is +0.017 versus logistic regression but −0.101 versus XGBoost. These indices, whose limitations are documented [39], therefore do not fully converge and are reported for exploratory purposes; they do not alter the conclusion based on the primary metrics.

➢ *Robustness of the Main Result*

The seventeen episodes admitted with a GCS of 3 to 8 all had an unfavourable outcome. This complete separation probably contributes substantially to the high apparent discrimination and may reduce the stability of coefficients in resamples containing few severe episodes. Three prespecified analyses test whether the main result depends on it (Table 9).

After exclusion of episodes with GCS 3–8, 125 episodes and 40 events remain. GCS alone achieves an AUC of 0.851 (95% CI 0.769–0.919), versus 0.797 for logistic regression, 0.804 for LightGBM, 0.814 for XGBoost and 0.821 for random forest. The hierarchy is therefore preserved, with difference intervals that include zero owing to the reduced sample size.

Within the zone of clinical uncertainty, GCS 9–14 — 86 episodes, 39 events, prevalence 45.3% — where the contribution of a model would a priori be greatest since both prognostic extremes are set aside, GCS alone achieves 0.769 (0.663–0.862) versus 0.671 for logistic regression, 0.719 for LightGBM, 0.724 for XGBoost and 0.736 for random forest. The paired difference versus logistic regression is more marked there than in the whole cohort and excludes zero (+0.098; +0.029 to +0.168).

Restricted to the 126 survivors, the analysis contrasts severe disability (41 episodes, 32.5%) with favourable outcome. GCS alone achieves an AUC of 0.944 (0.899–0.978), the four multivariable models lying between 0.932 and 0.940. The signal therefore does not bear on mortality alone: it also discriminates severe disability among survivors.

A prespecified parsimonious clinical model combining the GCS, the maximum size of the collection and midline shift reaches an AUC of 0.911, versus 0.896 for GCS alone. This difference of +0.015 is numerically in favour of the three-variable model, but its interval includes zero (−0.003 to +0.036 in favour of the trio). We report this result without playing it down: it indicates that a very parsimonious clinical enrichment might yield a gain that this cohort lacks the power to distinguish from zero. It qualifies the reading that a single variable would suffice, while leaving intact the finding that the 27-variable models add nothing.

A final analysis addresses the most serious reservation about the composition of the variable set. Since bacteriological culture is not available at the time of the initial decision, the models were refitted on a restricted set of 25 variables genuinely available at the bedside, excluding positive culture and its non-measurement indicator. Removal slightly improves all four models — 0.858 for logistic regression, 0.871 for random forest, 0.863 for XGBoost and 0.861 for LightGBM, versus 0.849 to 0.869 with the full set — without ever reaching GCS alone (0.896). The four paired Brier score differences continue to exclude zero. The AUC differences, by contrast, now exclude zero only for LightGBM and in a henceforth marginal way (+0.0004 as lower bound), that versus logistic regression becoming −0.0028.

Two consequences follow. The main result does not depend on the inclusion of a variable unavailable at the bedside, and the phrasing "variables available at admission" can be maintained by restricting the set. But the statistical support for the discrimination comparison weakens on this restricted set: the robust finding is the probabilistic inferiority of the multivariable models, the discrimination superiority of the GCS being less clearly established. Culture moreover carried only 0.64% of SHAP importance, and its non-measurement indicator an undetectable share.

➢ *Hierarchy by Importance and Hierarchy by P Value*

Table 8 contrasts the two rankings. They agree at the top and diverge thereafter. Symptom-to-admission delay, non-significant in bivariable analysis (p = 0.175), ranks seventh by SHAP importance. Subdural empyema, at the limit of significance (p = 0.066), is the only variable whose median marginal gain exceeds 0.01 AUC points. Conversely, midline shift, highly significant (p < 0.001), contributes on its own only 3.0% of the permutation AUC loss: its significance stems largely from its correlation with level of consciousness.

IV. DISCUSSION

- *Main Result*

Among 142 evaluable admission episodes, drawn from a cohort of 148 patients with intracranial suppurative infection, the measurable predictive signal contained in the baseline variables appears strongly concentrated on a single one of them. The GCS accounts for 37% of normalised SHAP importance, 87% of the permutation AUC loss and 44% of the outcome entropy according to mutual information; it holds first rank in all 1000 internal resamples, which reflects strong ranking stability within the sample without establishing its transportability; and a model containing it alone is outperformed by none of the four multivariable models, neither in discrimination, nor in probabilistic accuracy, nor in net benefit.

Three prespecified analyses rule out the most natural objection, namely that this advantage would stem solely from identifying the most severe neurological presentations: it persists after exclusion of episodes with GCS 3–8, it is most marked within the zone of clinical uncertainty GCS 9–14, and it is also observed among survivors in discriminating severe disability. This result contradicts the hypothesis stated in the introduction. We expected importance distributed across lesional, biological and organisational factors, which a non-linear model would have been able to aggregate. That is not what the data show, and the whole set of robustness analyses — nested selection, hyperparameter optimisation, optimism correction, temporal validation, ordinal analysis — leaves this finding unchanged.

- *Why Does A Single Variable Concentrate Predictive Importance?*

Three non-exclusive mechanisms may be invoked, and our data allow them to be partly distinguished.

The first is mediation. The GCS is not one variable among others: it measures the result of all lesional mechanisms on brain function at a given moment. A large collection, herniation, extensive ventriculitis or severe sepsis compromise prognosis insofar as they impair consciousness. The GCS lies downstream and might constitute their common final pathway. This pattern is compatible with mediation through neurological severity, but no causal mediation analysis was conducted and none of our analyses allows it to be established. Conditional mutual information supports this reading: once the GCS is known, the size of the collection adds no more than 0.048 bit and midline shift 0.056 bit.

The second is collinearity. The Spearman correlations between the GCS and the lesional variables (−0.53 with size, −0.33 with midline shift) render the latter largely substitutable. This is exactly the pattern revealed by the gap between shared attribution (SHAP) and contribution of its own (permutation).

The third is its status as an integrative measure. The GCS may be viewed as a pragmatic indicator of overall neurological severity, of which the other variables would be only partial and noisy manifestations; no latent-variable modelling was, however, performed. In this framework, a model that adds these manifestations to their integrator mainly adds measurement error.

This integrative property explains the robustness of the Glasgow scale across very different conditions since its initial description [40,41]. The large prognostic models in head injury, IMPACT and CRASH, developed on tens of thousands of patients, arrive at a comparable architecture: a few dominant variables and a weak marginal gain from those that follow [42–44].

- *Redundancy is not Uselessness*

One distinction deserves to be stated explicitly, because it is little articulated in the clinical literature. A variable may be clinically indispensable, causally important, and yet redundant in predictive terms. The size of the collection and midline shift guide the operative decision; they add no detectable discrimination once the level of consciousness is known. These two propositions are not contradictory: the first concerns what should be done, the second the separation of outcomes. The result reported here in no way diminishes the diagnostic and therapeutic importance of imaging, microbiology and lesional assessment.

The cautious statement is therefore that several lesional variables appear largely redundant for predicting outcome once the admission GCS is available, in this cohort and with these measurements.

- *Functional Form: A Transition Zone, Not A Dichotomy*

Our analyses correct an appealing but inaccurate reading. The partial dependence plots suggest a step-function relationship around 12–13, and the observed risk by stratum is strikingly contrasted. But formal comparison of the specifications shows that categorisation into three strata loses 0.13 AUC points relative to the continuous form, while the spline adds nothing.

The clinical conclusion is twofold. On the one hand, the transition zone around 12–13 is real and reproducible in 88% of resamples: a patient referred for intracranial suppurative infection with a Glasgow score of 12 is not a mildly affected patient, since two-thirds of that group have an unfavourable outcome. On the other hand, this observation must not be turned into a binary threshold: the score should be used in its continuous form, every point counting. The thresholds inherited from head trauma, where 13–15 defines mild injury, lead to underestimating severity in this condition.

- *Causal Effect And Predictive Importance: Two Different Questions*

Symptom-to-admission delay carries only 4.3% of SHAP importance and an undetectable contribution of its own. This result might seem to contradict earlier strands of this series, which identified prehospital delay as a central determinant and estimated by g-computation a substantial reduction in expected mortality under shortening scenarios.

There is no contradiction, but two distinct questions. Causal estimation asks what would happen if the delay were modified for all patients; importance quantification asks whether knowing a patient's delay helps distinguish their outcome from another's. A variable with a narrow distribution may have a strong causal effect and weak discriminating value. That is the case here: the interquartile range of the delay extends from 15 to 25 days, in other words almost all patients arrive late. A variable that does not vary does not separate.

This distinction has opposite practical consequences: delay is a public health intervention target of the first order and a poor individual triage tool; the GCS is the reverse.

- *What The Complex Models Did Not Add Here*

In this cohort, ensemble methods added no measurable gain in discrimination, calibration or net benefit relative to a penalised logistic regression, itself no better than the Glasgow score alone. This finding is in line with a convergent methodological literature [45,46].

Our study adds a mechanistic explanation rather than a mere observation. Non-linear methods can yield a gain only if there are interactions or non-linearities to capture. Yet SHAP interaction values attribute 76.6% of importance to main effects and the strongest interaction reaches only 2.6%; the spline does not improve on the linear form. There is, in these data, almost nothing to capture.

The corollary matters more than the observation. Since the complex models add no further performance, their use would be justified by no benefit while adding an implementation cost and decisional opacity. The argument that explainability techniques would make opaque models acceptable should be reversed: when they reveal that a model rests essentially on one variable, the reasonable conclusion is not to deploy the model with its explanations, but to evaluate the variable alone [47,48].

- *Why Did Machine Learning Not Do Better Here?*

Four conditions must be met for a flexible model to outperform a simple predictor. None is met in the data analysed, and our analyses make it possible to verify this one by one rather than assert it.

- There must be interactions or non-linearities to capture. Main effects represent 76.6% of SHAP attribution, the strongest interaction 2.6%, and a spline does not improve on the linear form of the GCS.

- There must be a signal distributed across several variables. Three distinct metrics rank the GCS first, and conditional mutual information shows that once the GCS is known, the size of the collection adds no more than 0.048 bit and midline shift 0.056 bit.
- There must be a sufficient number of events to estimate the additional parameters. With 57 events for 27 variables, the ratio is 2.1, well below usual recommendations [23,24]; the bootstrap optimism of 0.042 to 0.060 is its direct trace.
- The variables added must be measured precisely. Several lesional variables are correlated with the GCS and affected by measurement error: adding them to their integrator mainly adds noise.

These four explanations are complementary and not competing. The first and second concern the structure of the data, the third their volume, the fourth their quality. None allows the conclusion that machine learning would be useless in intracranial suppurative infections in general: they describe the conditions under which it could not yield a gain here, and provide a framework of interpretation applicable to other cohorts.

- *Two Methodological Proposals*

- *The Information Concentration Index*

We suggest reporting, alongside a model's performance, the share of total importance carried by its rank-1 variable. It is useful to specify what this index is and what it is not. It is neither a causal measure, nor a universal quantity, nor a prognostic score, nor a variable selection criterion, nor a decision threshold: a high index does not on its own establish that the other variables are useless, added value always having to be assessed empirically. It is a descriptive statistic, specific to the metric and to the model whose predictions it explains, which summarises the concentration of the predictive signal in a given cohort and can be compared across studies using the same metric and the same type of model. Its principal value is to flag situations in which a multivariable model might merely restate a single variable.

- *The Principle of Predictive Sufficiency*

We suggest stating it as follows: when a variable concentrates most of the predictive signal, it is useful for a more complex model to demonstrate explicitly, on out-of-sample data and by paired comparison, that it makes a contribution of its own in discrimination, calibration or net benefit. We present this formulation as a reporting heuristic calling for external evaluation, not as an established rule. The term sufficiency is used here in the pragmatic sense of reporting and not in the formal sense of the Fisher–Neyman sufficient statistic; concentration, dominance and sufficiency are not the same thing, and it is the conjunction of a high index and an absence of paired gain that supports parsimony.

This principle is not a matter-of-principle hostility towards machine learning: on the contrary, it makes the most rigorous use of it, by requiring that complexity justify its cost. It leads to a minimum reporting requirement which we summarise in the box in section 6, in the form of a list of seven items.

- *A five-Step Framework*

Figure 10 formalises the progression that structures this work: statistical association, predictive capacity, contribution of its own, clinical utility, decision. Each step can fail independently of the preceding ones. A significant variable may not be predictive; a predictive variable may make no contribution of its own; an informative model may bring no net benefit. The clinical literature often crosses these steps implicitly, treating "prognostic factor", "risk factor" and "predictor" as synonyms.

- *Scientific Frugality*

One practical consequence deserves to be stated directly: before developing a machine-learning model, it is useful to check that there is genuinely information left to learn. This check is inexpensive — it requires an importance metric, a paired comparison and a parsimony analysis — and it avoids building tools whose contribution is nil.

In resource-limited systems, this frugality has an immediate operational translation: an excellent clinical measurement is better than a sophisticated but poorly fed model. This formulation is not a slogan but the consequence of the figures reported here, where the effort of collecting twenty-six additional variables translates into no decisional gain.

- *The Value of Negative Results in Clinical Machine Learning*

This work reports a negative result: four algorithms did not improve on a single clinical predictor. Such a result is informative only when obtained through a rigorous validation procedure, failing which it cannot be distinguished from an implementation flaw. This is why internal validation, calibration, net benefit, nested selection, hyperparameter optimisation and subgroup analyses were conducted: they serve less to demonstrate superiority than to make an absence of gain credible.

This study also illustrates a symmetrical risk. LightGBM reaches an apparent AUC of 1.000 on 142 episodes; published without validation, such a value would constitute a methodological false positive. The distance between apparent and validated performance is, in this field, the principal source of excessive conclusions.

- *Algorithmic Frugality*

The frugality we describe is not opposed to sophisticated algorithms: it gives priority to the simplest validated solution achieving equivalent clinical performance. It presupposes considering that every variable has a cost — financial, human, logistical — and that this cost must be weighed against the expected benefit. We did not conduct a health-economic analysis and therefore do not quantify what, in this department, the reliable collection of the GCS would represent compared with that of twenty-six additional variables; the order of magnitude nonetheless remains asymmetrical, the first

measurement requiring neither consumables, nor equipment, nor transfer.

- *Comparison with the Literature*

- *Reference Series*

The Danish national cohort of Bodilsen and colleagues, several analyses of which have explored oral aetiology and treatment strategy [49,50], identifies age, comorbidities, impaired consciousness and intraventricular rupture as determinants of unfavourable outcome [15]. The meta-analysis of Brouwer and colleagues arrives at a similar hierarchy [51], as do series combining abscesses and empyemas [17]. These works converge with ours on the rank of impaired consciousness but also retain age and comorbidities, which our analysis ranks low.

Two explanations are plausible. The age structure differs radically: our population has a median age of 26 years versus more than 50 years in European cohorts, and comorbidities are rare there. Moreover, a variable whose effect is real but modest becomes detectable across several hundred patients and undetectable across one hundred and forty-two. We therefore do not conclude that age is unimportant in intracranial suppurative infections, but that it makes no detectable contribution in this young population.

- *Explainable Models in Neurosurgery and Beyond*

Applications of explainable machine learning have proliferated in head trauma, intracerebral haemorrhage, ischaemic disease, neuro-oncology and hydrocephalus [52–57], as well as in critical care and sepsis prediction [69–71]. A pattern recurs: the models achieve respectable performance, SHAP values designate one to three dominant variables, and comparison with the best single variable is almost never reported. Our contribution consists less in adding a model than in carrying the analysis through to that comparison. The phenomenon described here has no reason to be specific to intracranial suppurative infections, and the framework proposed applies to any prognostic study based on tabular clinical data.

- *Reporting Recommendations*

The ecosystem of recommendations for clinical artificial intelligence has structured itself rapidly: TRIPOD+AI for prognostic models [20], PROBAST for risk of bias [21], DECIDE-AI for early clinical evaluation [58], CONSORT-AI and SPIRIT-AI for trials [59,60], MI-CLAIM, CLAIM and MINIMAR for reporting transparency [61–63]. Our proposal does not replace them: it adds an element that these recommendations do not yet explicitly require, the comparison of the model with its best single variable. This requirement echoes recent calls for real clinical evaluation of artificial intelligence models, beyond discrimination metrics alone [72–74].

- *Explainability, Trust and Fairness*

The growing use of explainability methods is accompanied by a critical literature warning against the illusion of understanding they can produce: a plausible explanation is not a correct explanation, and two attribution methods can designate different variables for the same model [75–77]. Our approach partly addresses this criticism by comparing three metrics and adding a model-free measure, but it does not escape it entirely.

The question of algorithmic fairness arises here in a particular form. A model developed on a young, predominantly rural and late-presenting population is not transportable to a different population without reassessment, and the under-representation of African cohorts in the datasets used to develop clinical models is documented [78–80]. The result reported here — the sufficiency of a universally available variable — partly mitigates this problem, since it depends on no technological resource.

- *Implications for Global Neurosurgery*

The main result has direct operational relevance in resource-limited systems, and it is counter-intuitive: the priority is neither to extend data collection nor to acquire algorithmic capacity, but to ensure that the variable carrying most of the predictive importance is measured correctly, systematically and by everyone.

- *Training.*

The inter-observer reproducibility of the GCS is modest without structured training. Investing in the reliability of its measurement improves the prognostic capacity of the system more than adding any further investigation.

- *Traceability.*

A Glasgow score not recorded at admission is an irreversible loss of information; its presence in the record should constitute a quality indicator.

- *Referral.*

Since the GCS is available before imaging, it can support risk communication and the prioritisation of transfers; its use as a formal transfer rule would require prospective evaluation. Imaging remains indispensable for diagnosis and treatment planning, even if the variables it provides added only limited incremental prognostic value beyond the GCS here.

This reading is in keeping with the priorities of the Lancet Commission on Global Surgery and the essential neurosurgery movement [64–67,81,82], as well as with the goal of universal health coverage: the most informative variable is also the one available earliest and furthest upstream in the chain of care. It carries an implication that extends beyond this department. In resource-limited systems, access to data is often more constrained than access to algorithms: improving the quality of what is already measured may produce more benefit than collecting more variables. A district hospital without cross-sectional imaging nonetheless has the predictor that, in this cohort, carries most of the signal — which makes conceivable a prognostic stratification upstream of the referral chain, where it is today absent.

- *Strengths*

This study has eight methodological strengths: partitions and resamples grouped by patient; a strictly identical pipeline across compared models, preprocessing included; SHAP values computed outside the training sample; three independent importance metrics, one of them model-free; paired comparisons with patient-level bootstrap intervals and the DeLong test; joint assessment of discrimination, calibration and net benefit; a parsimony analysis with nested selection; and full reporting of negative results, including the failure of mortality prediction and the uselessness of the non-measurement indicators.

- *Limitations*

The first limitation is the absence of external validation. The results describe the structure of predictive importance in a single-centre Ivorian cohort; the generalisability of a clinical model is not postulated, it is demonstrated [68]. The concentration observed might be smaller in an older, more comorbid or earlier-presenting population.

The second is sample size. With 57 events for 27 variables, the events-per-variable ratio is 2.1. This constraint structurally limits what can be established: we can assert that one variable dominates, not rank those that follow. It also makes it possible for a genuinely informative but rare variable — anisocoria, present in 7 patients — to be wrongly classified as non-contributory.

The third is the retrospective nature of the data collection. The GCS is extracted from clinical records and the quality of its measurement was not monitored. One possible bias deserves to be named: if the GCS influenced the therapeutic decision, part of its predictive value would reflect management and not intrinsic prognosis. The exclusion of therapeutic variables preserves the clinical use of the model but does not neutralise this mechanism.

The fourth relates to the outcome. The GOS at discharge is a coarse measure, sensitive to length of stay, and does not capture long-term outcome; its dichotomisation entails a loss of information, mitigated but not eliminated by the complementary ordinal analysis. The 8 observations without an evaluable GOS were excluded rather than imputed, and their exclusion is not necessarily random.

The fifth relates to residual risk of bias. We used the PROBAST domains as a structured self-assessment grid for risk of bias, and not as a formal independent assessment, which normally falls to third parties. According to the PROBAST domains, the "Analysis" domain remains at risk because of the small number of events, the number of candidate variables, simple imputation and the absence of external validation. We therefore do not claim full compliance.

The sixth relates to the nature of the methods used. SHAP values explain a model, not a phenomenon; permutation importance evaluates combinations of variables that are sometimes unobserved; partial dependence plots extrapolate beyond the support of the data [11,36,37]. These limitations are mitigated by the convergence of three approaches and by the addition of a model-free metric, but they preclude any causal reading: none of the figures reported here should be interpreted as an effect [83,84].

The seventh relates to transportability. A population differing in age, comorbidities, time to care or access to imaging would alter the distribution of the variables; a different technical platform would alter the measurement of several of them; and a different definition or horizon of the outcome would alter the target itself. Each of these shifts may alter the concentration of signal observed here, which can only be verified by external evaluation. No prospective evaluation, no impact study on decisions or outcomes, and no health-economic analysis were moreover conducted: this work describes a signal structure, not a validated tool. The eighth relates to the measurement of the dominant predictor itself. The inter-observer reliability and the exact timing of GCS assessment were not standardised prospectively; part of the estimated predictive signal might reflect documentation practices and measurement error. The outcome itself is assessed at a non-uniform horizon, determined by length of stay, and may partly reflect discharge practices, access to rehabilitation and socio-economic constraints, in addition to neurological recovery. The ninth relates to the threshold. The transition zone around 12–13 is data-derived and remains exploratory; it requires independent validation before any decisional use.

- *Future Directions*

Four extensions are under way. A multicentre external validation in two to three West African centres would test the reproducibility of the importance structure more than that of the model, and would allow a decision curve analysis on independent data to be added [30]. An inter-observer reliability study of Glasgow score measurement at the different levels of the chain of care would quantify the share of information lost to measurement error — the variable most worth improving if it is poorly measured. Extending the outcome to a three- or six-month horizon would shed light on the stability of the hierarchy. Finally, the prospective aggregation of several African cohorts would make it possible to reach the number of events required to pose seriously the question of mortality prediction, which remains unanswered here.

- *Key Messages*

- A p value measures the compatibility of an association with the null hypothesis, not the contribution of a variable to predicting a patient's outcome.
- In this cohort, admission GCS captured most of the measurable predictive signal contained in the baseline variables.
- None of the four multivariable models evaluated yielded a measurable gain in discrimination, calibration or net benefit.

- An explainability analysis gains from being accompanied by a paired comparison with the best single clinical predictor.
- In resource-limited settings, improving the measurement quality of a simple predictor may have more impact than developing a complex model.

V. CONCLUSION

In this single-centre cohort, the admission Glasgow Coma Scale captured most of the measurable predictive signal available for functional outcome at discharge. None of the four multivariable models evaluated showed superior discrimination, probabilistic accuracy or net benefit, and this advantage persisted after exclusion of the most severe presentations as well as within the zone of clinical uncertainty.

This result does not diminish the value of the other variables: it shows that their prognostic contribution passes largely through impaired consciousness, of which the GCS constitutes the integrated measure. It does, however, shift the operational priority. In a resource-constrained system, the most worthwhile effort is neither extending data collection nor algorithmic sophistication: it is the reliability, earliness and traceability of a measurement that any health worker can perform without equipment.

Finally, we offer a methodological suggestion. A study reporting a machine-learning-based prognostic model would gain from reporting, alongside its performance, the concentration of the predictive signal and the performance of its dominant predictor taken in isolation, compared with the model by a paired procedure. When a complex model does not improve on its dominant predictor, reporting the simplest predictor may constitute the most transparent and most directly applicable conclusion. External evaluation remains necessary before any clinical implementation or methodological generalisation.

MINIMUM REPORTING ITEMS FOR EXPLAINABLE CLINICAL MACHINE LEARNING

We propose that any study reporting an explainable prognostic model document the following seven items, none of which requires additional data:

Table 1 Study reporting an Explainable Prognostic Model Document the Following Seven Items

#	Item	What must be reported
1	Comparison with the best single variable	Model performance and performance of its dominant variable alone, on the same partitions and the same pipeline
2	Paired comparison	Performance difference with confidence interval obtained by resampling at the level of the unit of analysis
3	Calibration	Out-of-sample calibration slope, intercept and curve, in addition to discrimination
4	Clinical utility	Net benefit across the relevant range of thresholds, compared with reference strategies
5	Concentration of importance	Concentration index, with the metric used specified, with resampling interval
6	Parsimony	Performance as a function of the number of variables, with selection nested within validation
7	Reproducibility	Equation or serialised model, code, versioned environment, random seeds, variable dictionary

➢ *Tables*

Table 2 Admission Characteristics According to Functional Outcome at Discharge (n = 142)

Variable	Overall (n = 142)	GOS 4–5 (n = 85)	GOS ≤ 3 (n = 57)	p
Age, years — median [IQR]	26 [14–51]	30 [14–50]	25 [15–51]	0.708
Symptom-to-admission delay, days	19 [15–25]	18 [15–24]	19 [16–25]	0.175
Glasgow Coma Scale	13 [10–15]	14 [13–15]	10 [8–11]	< 0.001
Maximum size, mm	36 [30–44]	32 [29–38]	43 [38–50]	< 0.001
Midline shift, mm	1.8 [0.0–4.1]	1.1 [0.0–2.7]	3.7 [1.7–5.0]	< 0.001
White cell count, G/L	14.7 [12.6–17.0]	14.4 [12.6–16.9]	14.9 [12.6–17.3]	0.610
Male sex — n (%)	99 (69.7)	57 (67.1)	42 (73.7)	0.459
Rural residence — n (%)	75 (52.8)	43 (50.6)	32 (56.1)	0.607
Referred patient — n (%)	72 (50.7)	43 (50.6)	29 (50.9)	1.000
Immunosuppression — n (%)	13 (9.2)	5 (5.9)	8 (14.0)	0.138
Diabetes — n (%)	10 (7.0)	5 (5.9)	5 (8.8)	0.522
Postoperative origin — n (%)	25 (17.6)	18 (21.2)	7 (12.3)	0.187
Post-traumatic origin — n (%)	32 (22.5)	18 (21.2)	14 (24.6)	0.685
Fever — n (%)	89 (62.7)	48 (56.5)	41 (71.9)	0.077
Seizures — n (%)	27 (19.0)	18 (21.2)	9 (15.8)	0.515
Motor deficit — n (%)	46 (32.4)	27 (31.8)	19 (33.3)	0.857
Anisocoria — n (%)	7 (4.9)	2 (2.4)	5 (8.8)	0.117
Raised ICP syndrome — n (%)	103 (72.5)	60 (70.6)	43 (75.4)	0.570
Brain abscess — n (%)	89 (62.7)	58 (68.2)	31 (54.4)	0.112
Subdural empyema — n (%)	24 (16.9)	10 (11.8)	14 (24.6)	0.066
Ventriculitis — n (%)	16 (11.3)	6 (7.1)	10 (17.5)	0.062
Multiple collections — n (%)	54 (38.0)	30 (35.3)	24 (42.1)	0.481
Mass effect — n (%)	106 (74.6)	59 (69.4)	47 (82.5)	0.115
Hydrocephalus — n (%)	14 (9.9)	10 (11.8)	4 (7.0)	0.404
Positive culture — n (%)	50/113 (44.2)	29/65 (44.6)	21/48 (43.8)	1.000

IQR: interquartile range. Mann–Whitney tests (continuous variables) and Fisher's exact tests (binary variables). These comparisons are descriptive and exploratory; no correction for multiple comparisons was applied, since candidate variables were not selected on the basis of their p value.

Table 3 Missing Data in the Analysis Sample (n = 142)

Variable	Missing observations	Proportion
Maximum size of the collection (mm)	6	4.2%
Herniation / midline shift (mm)	26	18.3%
White cell count (G/L)	11	7.7%
Positive culture	29	20.4%

The other 23 variables are complete. Median imputation was computed exclusively within each training fold.

Table 4 Out-of-Sample Performance of the Four Models and of the Single Clinical Predictor

Model	AUC	95% CI (patient bootstrap)	Brier	BSS	AUPRC	Cal. slope	Intercept	ECE
Glasgow Coma Scale alone	0.896	0.834 – 0.945	0.110	0.54	0.880	1.27	-0.00	0.049
Penalised logistic regression	0.849	0.780 – 0.914	0.146	0.39	0.811	0.79	-0.04	0.090
Random forest	0.869	0.803 – 0.929	0.145	0.39	0.849	1.87	-0.00	0.129
XGBoost	0.861	0.793 – 0.923	0.134	0.44	0.853	0.86	-0.00	0.039
LightGBM	0.860	0.792 – 0.917	0.144	0.40	0.854	0.69	0.01	0.101

The AUC values differ from one analysis to another because they arise from distinct resampling estimands; they must not be compared across procedures. Out-of-sample probabilities averaged over 20 repetitions of patient-grouped stratified 5-fold cross-validation, with metrics then computed on these aggregated probabilities; intervals by patient-level bootstrap (1000 replications). BSS: Brier Skill Score relative to the null model (Brier 0.240). AUPRC: prevalence reference 0.401. ECE: expected calibration error. Under the alternative aggregation mode (metric computed at each repetition and then averaged), the AUCs are 0.911 for GCS alone, 0.836 for logistic regression, 0.867 for random forest, 0.856 for XGBoost and 0.848 for LightGBM; the corresponding percentiles describe partitioning variability and are not confidence intervals.

Table 5 Paired Comparisons: GCS Alone Versus Each Multivariable Model. A Positive ΔAUC and a Negative ΔBrier Favour GCS Alone

Comparison (GCS – model)	ΔAUC	95% CI	ΔBrier	95% CI	ΔAUC > 0 (bootstraps)	DeLong p
Penalised logistic regression	+0.0473	0.0071 ; 0.0925	-0.036	-0.056 ; -0.016	99%	0.027
Random forest	+0.0271	-0.0241 ; 0.0724	-0.035	-0.050 ; -0.018	87%	0.232
XGBoost	+0.0351	-0.0058 ; 0.0751	-0.024	-0.040 ; -0.009	95%	0.083
LightGBM	+0.0364	0.0030 ; 0.0733	-0.034	-0.052 ; -0.017	98%	0.039

Patient-level paired bootstrap, 1000 replications, predictions bearing on the same observations. A positive ΔAUC and a negative ΔBrier favour GCS alone. The "ΔAUC > 0" column gives the proportion of resamples in which the difference is positive; this is a descriptive measure of comparative stability, not a probability of superiority. The principal comparative inference rests on the paired bootstrap intervals, the DeLong test being secondary. All four Brier score differences exclude zero.

Table 6 Relative Share of Predictive Importance According to Three Complementary Approaches (Top Ten Variables; Full Table in 21)

Variable	SHAP (%)	95% CI	Median rank	P(top 3)	ΔAUC perm.	Positive perm.	Perm. share (%)	MI (bits)	MI (%)
Admission Glasgow Coma Scale	**37.2**	**27.3 – 50.2**	**1 [1–1]**	**1.00**	**0.2927**	**100%**	**86.6**	**0.428**	**49.9**
Maximum size of the collection (mm)	10.5	2.5 – 20.2	3 [2–10]	0.63	0.0130	67%	3.8	0.172	20.1
Herniation / midline shift (mm)	8.2	1.8 – 15.3	4 [2–11]	0.43	0.0101	64%	3.0	0.079	9.2
White cell count (G/L)	6.7	1.6 – 13.7	5 [2–12]	0.29	0.0002	45%	0.6	0.003	0.4
Age (years)	4.7	1.1 – 10.9	7 [2–12]	0.11	-0.0042	32%	0.1	0.006	0.7
Brain abscess (vs other form)	4.3	0.2 – 12.1	7 [2–17]	0.14	0.0065	61%	1.9	0.014	1.6
Symptom-to-admission delay (days)	4.3	1.0 – 10.6	7 [3–13]	0.08	-0.0040	30%	0.0	0.012	1.3

Variable	SHAP (%)	95% CI	Median rank	P(top 3)	ΔAUC perm.	Positive perm.	Perm. share (%)	MI (bits)	MI (%)
Fever	3.6	0.0 – 11.7	8 [2–19]	0.11	0.0048	56%	1.5	0.018	2.1
Raised intracranial pressure syndrome	2.6	0.0 – 11.7	10 [3–19]	0.09	0.0026	48%	0.8	0.002	0.2
Multiple collections	2.4	0.1 – 11.1	10 [3–19]	0.06	-0.0027	29%	0.1	0.003	0.4

SHAP: share of global importance normalised according to the means of absolute values, derived from the prespecified XGBoost model and computed on episodes excluded from training; bootstrap median shares do not sum exactly to 100% by construction; interval, rank and probability of belonging to the top three variables over 1000 patient-level bootstraps. Permutation: mean AUC loss over 30 permutations per variable and per fold; "positive perm." denotes the proportion of permutations producing a positive loss; the relative share is computed after truncating negative losses to zero and then normalising by the sum of positive losses. Permutation: mean of the four algorithms; this metric remains specific to the models to which it is applied. MI: mutual information with the outcome, in bits and as a percentage of the sum of univariable mutual information estimates (0.857 bit); relative to the outcome entropy (0.971 bit), the GCS share is 44.1%.

Table 7 Functional form of the GCS and Observed Risk by Stratum

Specification	AUC	95% CI	Brier
linear	0.896	0.840 – 0.942	0.110
categorised (3-8/9-12/13-15)	0.762	0.669 – 0.855	0.130
restricted cubic spline	0.895	0.828 – 0.948	0.112

Glasgow stratum	n	Events	Observed risk	Exact 95% CI
3-8	17	17	100.0%	80.5 – 100.0
9-12	45	31	68.9%	53.4 – 81.8
13-15	80	9	11.2%	5.3 – 20.3

Clopper–Pearson intervals. Categorisation clearly degrades performance relative to the continuous form; the restricted cubic spline yields no gain. The complete separation of the 3–8 group precludes estimation of a finite conventional odds ratio.

Table 8 Parsimony Analysis With Selection Nested Within Validation

Number of variables (k)	AUC — PLR	95% CI	Brier — PLR	AUC — XGBoost	95% CI
1	**0.903**	**0.847 – 0.950**	**0.110**	**0.887**	**0.828 – 0.944**
2	0.909	0.859 – 0.955	0.113	0.884	0.820 – 0.940
3	0.907	0.847 – 0.957	0.115	0.879	0.814 – 0.932
5	0.907	0.856 – 0.948	0.115	0.876	0.808 – 0.934
8	0.897	0.838 – 0.944	0.121	0.873	0.802 – 0.936
12	0.888	0.826 – 0.944	0.125	0.868	0.803 – 0.931
20	0.871	0.804 – 0.924	0.136	0.869	0.796 – 0.931
27	0.854	0.786 – 0.913	0.147	0.870	0.802 – 0.928

PLR: penalised logistic regression. The ranking of variables is recalculated within each training fold from the SHAP values of a model fitted on that fold alone, then applied to the corresponding test fold. This nesting eliminates the optimistic bias of a selection performed on the whole dataset.

Table 9 Two Hierarchies: Ranking by p Value and Ranking by Predictive Importance

Variable	Rank by p	p	SHAP rank	SHAP (%)	Contribution of its own (perm., %)
Glasgow Coma Scale	1	< 0.001	1	37.2	86.6
Maximum size of the collection	2	< 0.001	2	10.5	3.8
Midline shift	3	< 0.001	3	8.2	3.0
Subdural empyema	5	0.066	≈ 13	1.6	0.3
Ventriculitis	4	0.062	≈ 24	0.8	0.1
Fever	6	0.077	8	3.6	1.5
Brain abscess	8	0.112	7	4.3	1.9
Symptom-to-admission delay	13	0.175	7	4.3	0.1
Age	23	0.708	7	4.7	0.1

The two rankings agree at the top and diverge thereafter. Midline shift, highly significant, contributes on its own only 3.0% of the AUC loss: its significance stems largely from its correlation with level of consciousness (Spearman −0.33).

Table 10 Prespecified Subgroup Analyses: Out-of-Sample AUC (Bootstrap CI for GCS Alone)

Population	n	Events	GCS alone	PLR	Random forest	XGBoost	LightGBM
Whole cohort	142	57	0.896 [0.840–0.948]	0.849	0.869	0.861	0.860
GCS 9–15 (3–8 excluded)	125	40	0.851 [0.769–0.919]	0.797	0.821	0.814	0.804
GCS 9–14 (zone of uncertainty)	86	39	0.769 [0.663–0.862]	0.671	0.736	0.724	0.719
Survivors (severe disability)	126	41	0.944 [0.899–0.978]	0.940	0.937	0.933	0.932

PLR: penalised logistic regression. The three restrictions test whether the advantage of the clinical comparator depends on the episodes whose outcome is most obvious. Within the GCS 9–14 range, the paired difference versus logistic regression is +0.098 (95% CI +0.029 to +0.168). Among survivors, the outcome contrasts severe disability (GOS 2–3) with favourable outcome (GOS 4–5).

Table 11 Final Single-Variable Model: Coefficients and Uncertainty

Parameter	Coefficient	Standard error	95% CI (Wald)	95% CI (patient bootstrap)
Intercept	11.118	1.812	7.567 – 14.669	8.621 – 15.711
Glasgow Coma Scale	-0.949	0.147	-1.237 – -0.661	-1.315 – -0.738

Glasgow Coma Scale	Predicted probability	95% CI (bootstrap)
15	4.2%	1.3 – 8.7
13	22.8%	12.3 – 34.0
11	66.4%	52.2 – 81.3
9	92.9%	86.1 – 98.1

Unpenalised logistic regression refitted on the 142 episodes with unstandardised GCS values; this refitting was not used to estimate performance. The intervals quantify the uncertainty of the estimated mean risk function and not individual prognostic uncertainty. A penalised estimation, motivated by the complete separation at the lowest GCS values, gives an intercept of 9.576 and a slope of −0.825.

➢ *Figures*

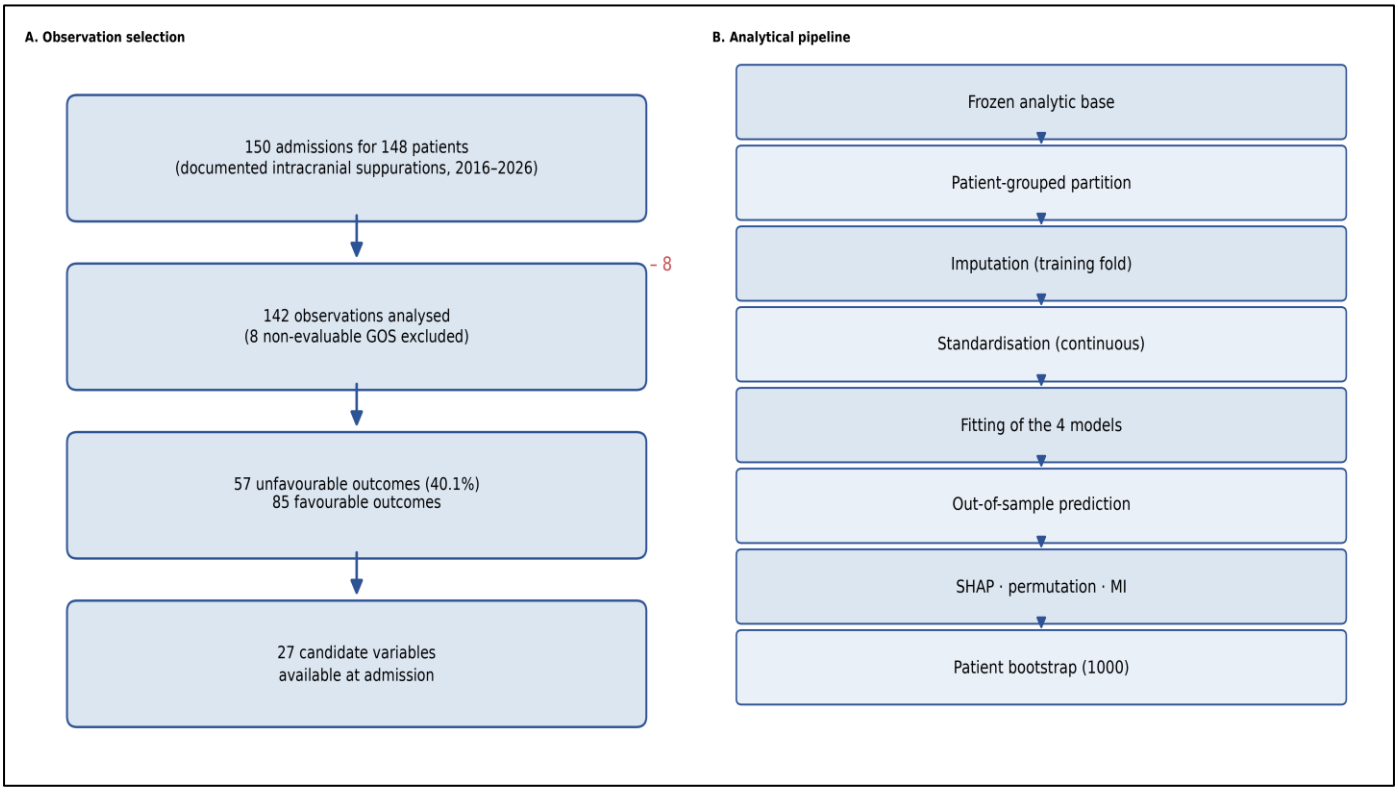

Fig 1 Selection of Observations and Analytical Pipeline.
A. Flow diagram. B. Analytical sequence, from database freeze to bootstrap resampling. All preprocessing steps are encapsulated in the pipeline and learned within the training fold.

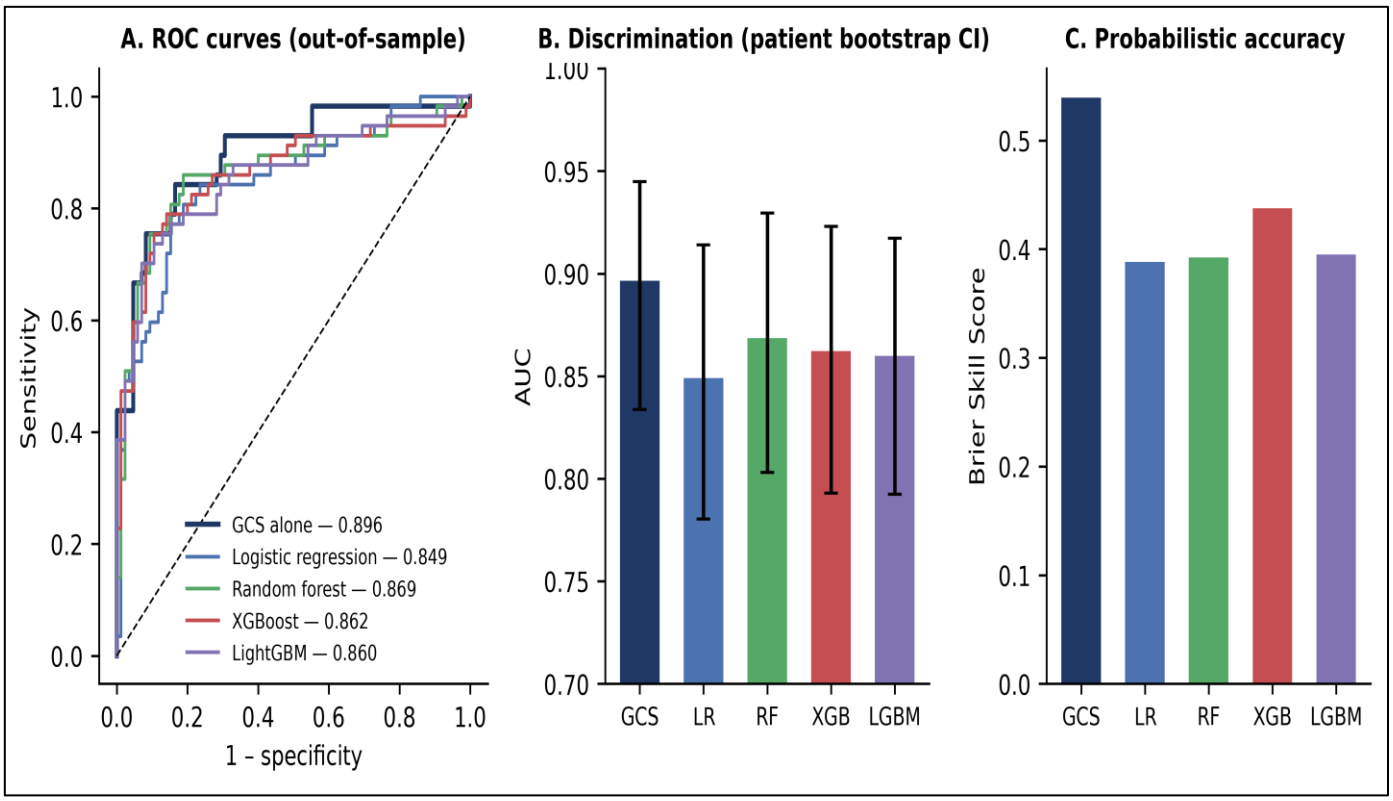

Fig 2 Out-of-Sample Performance of the Four Models and of GCS Alone.
A. ROC curves on the aggregated out-of-sample probabilities. B. AUC and patient-level bootstrap intervals. C. Brier Skill Score, relative to the null model predicting prevalence.

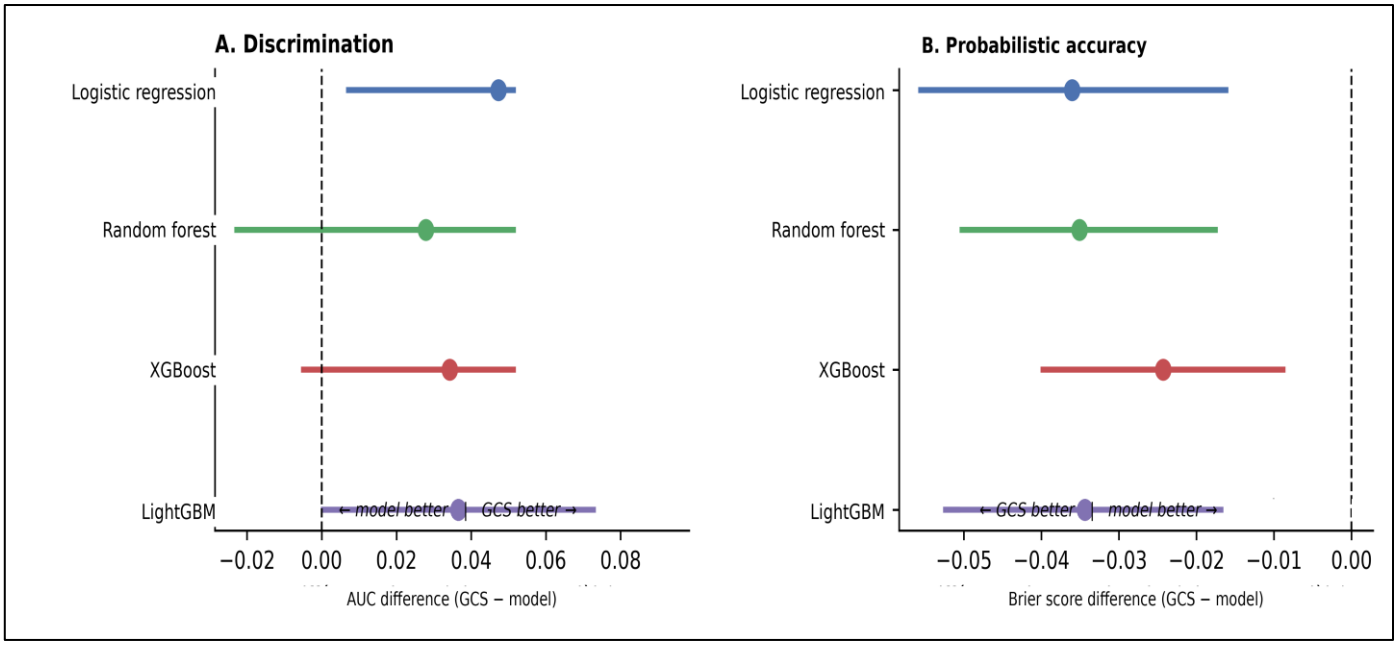

Fig 3 Paired Comparisons Between GCS Alone and Each Multivariable Model.
A. AUC difference. B. Brier score difference. Points: point estimate; bars: 95% patient-level paired bootstrap interval. All four Brier score differences exclude zero.

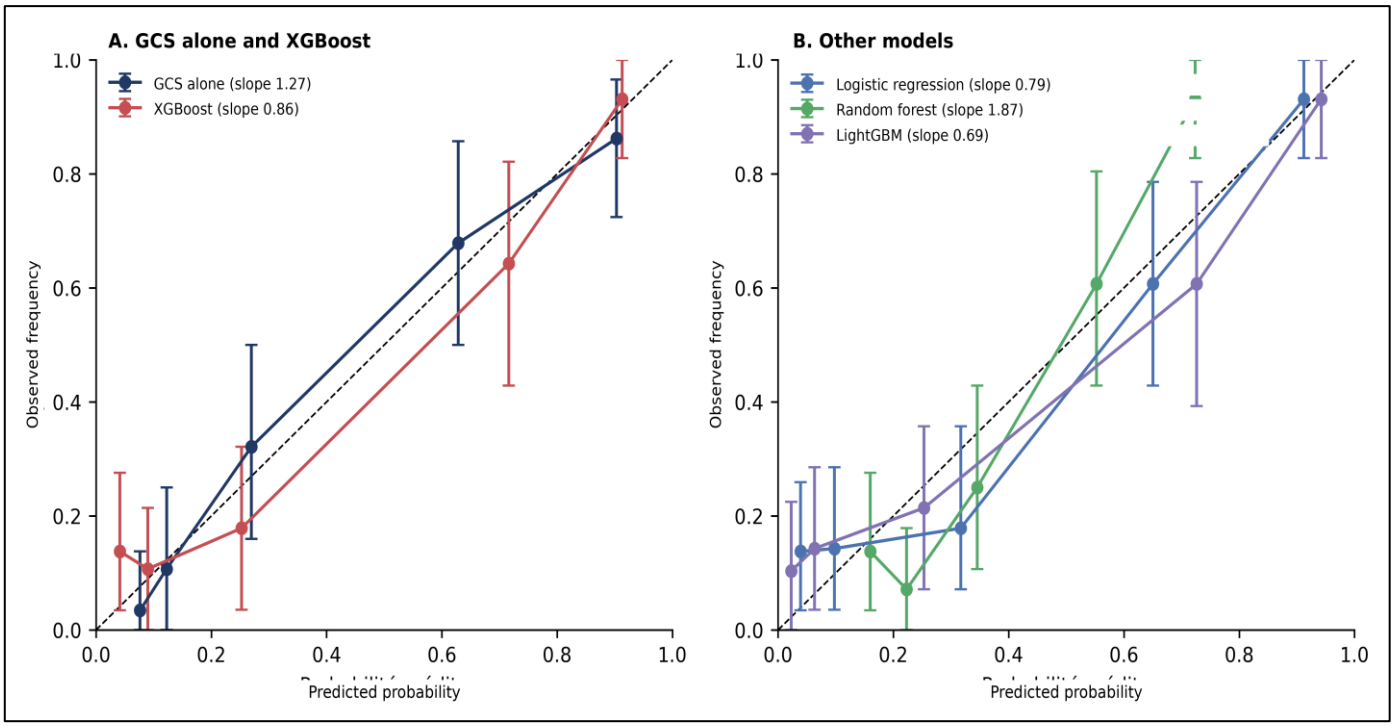

Fig 4 Calibration of out-of-Sample Predicted Probabilities.

Quintiles of predicted risk with 95% bootstrap intervals, identity line and calibration slope in the legend. A. Glasgow alone and XGBoost. B. Logistic regression, random forest and LightGBM. The random forest shows a slope of 1.87, indicating underfitting of extreme probabilities.

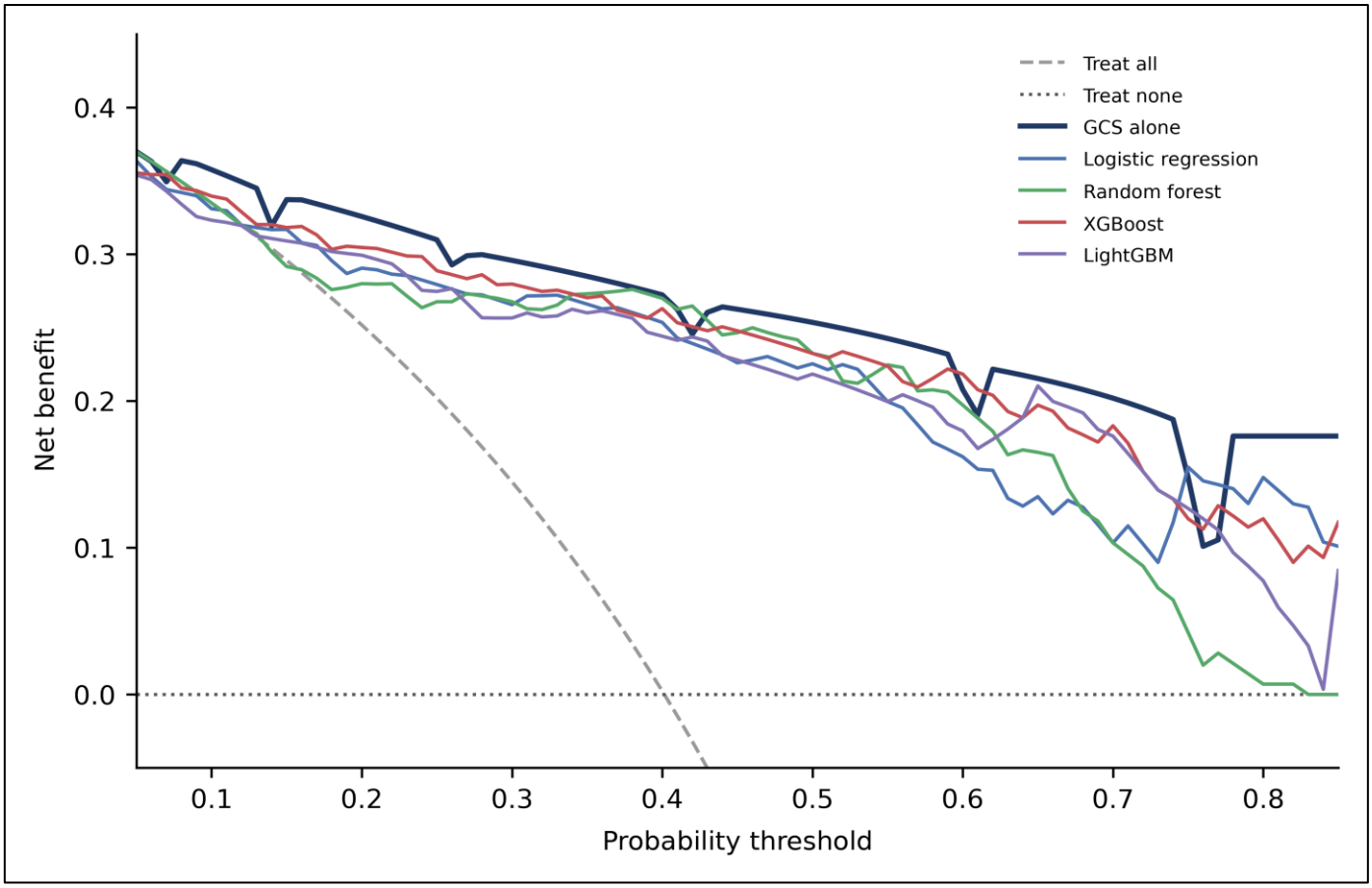

Fig 5 Decision Curve Analysis.

Out-of-sample net benefit as a function of the probability threshold, compared with the "treat all" and "treat none" strategies. GCS alone yields the highest net benefit across the whole range of clinically plausible thresholds.

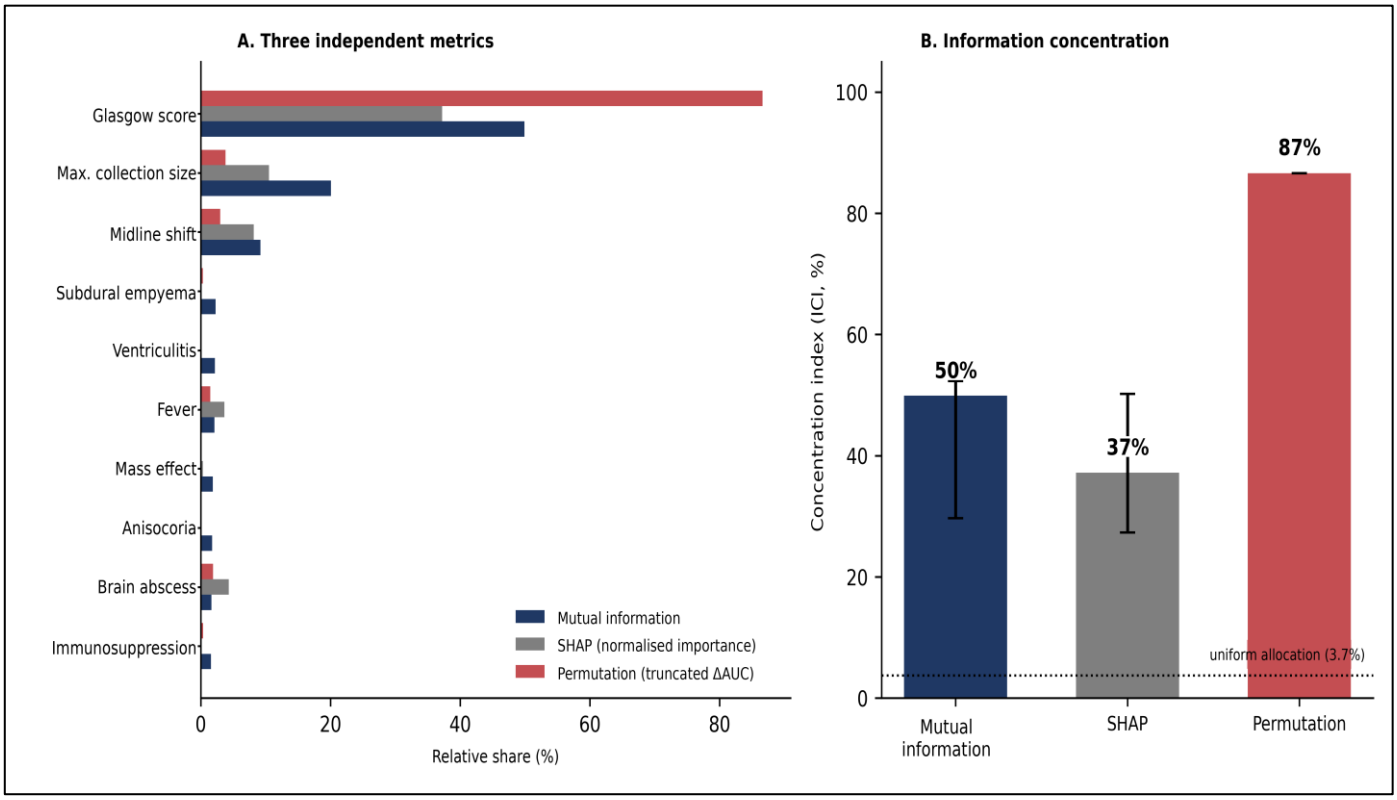

Fig 6 Three Importance Metrics and the Concentration Index.
A. Variable-by-variable comparison of mutual information, normalised SHAP importance and permutation importance. B. Information concentration index according to each of the three metrics, compared with the value expected under a uniform distribution across 27 variables (3.7%).

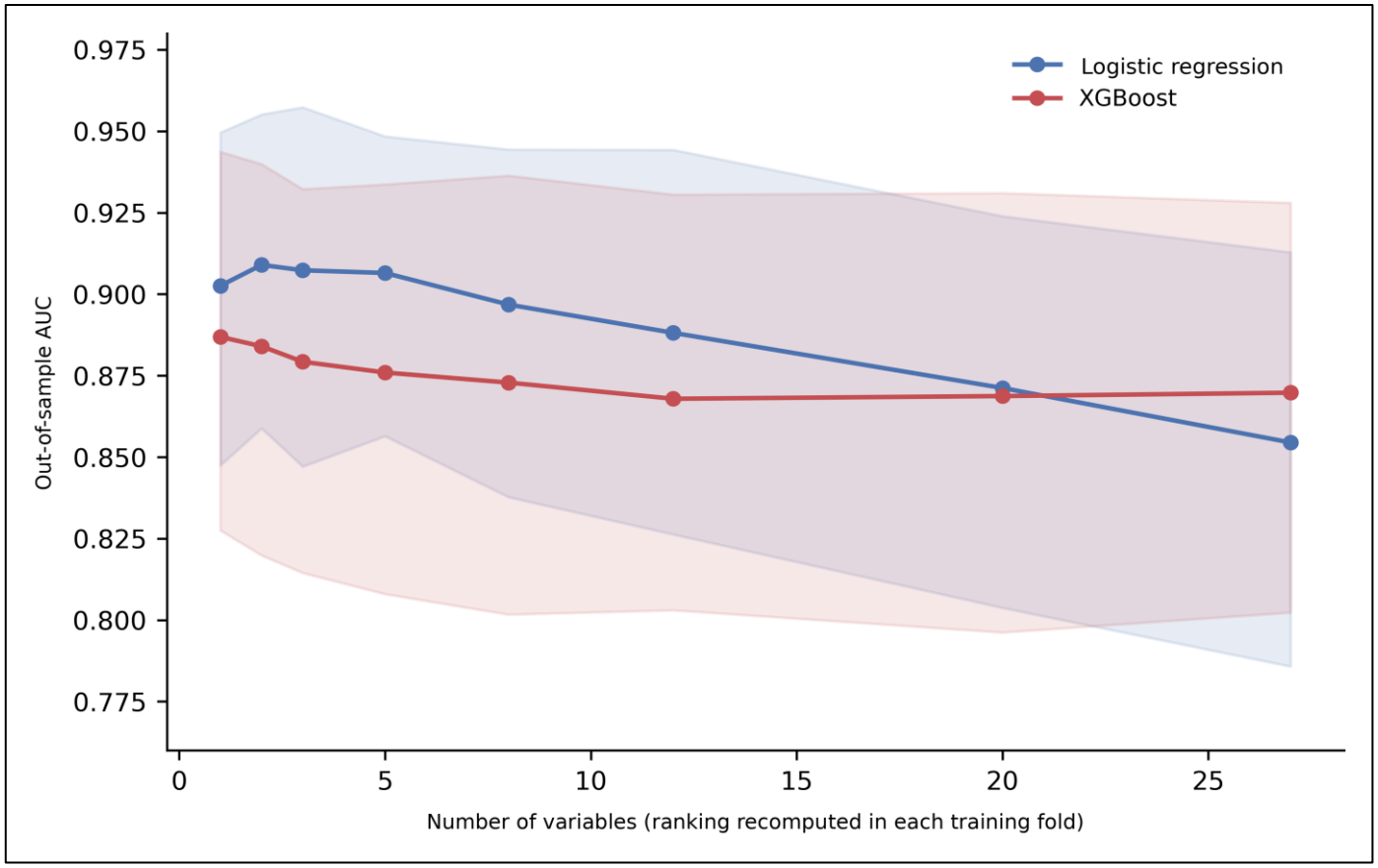

Fig 7 Parsimony Analysis with Nested Selection.
Variables are added in decreasing order of importance, the ranking being recalculated within each training fold. Bands: 95% bootstrap intervals. Performance plateaus from the first variables onwards and then declines.

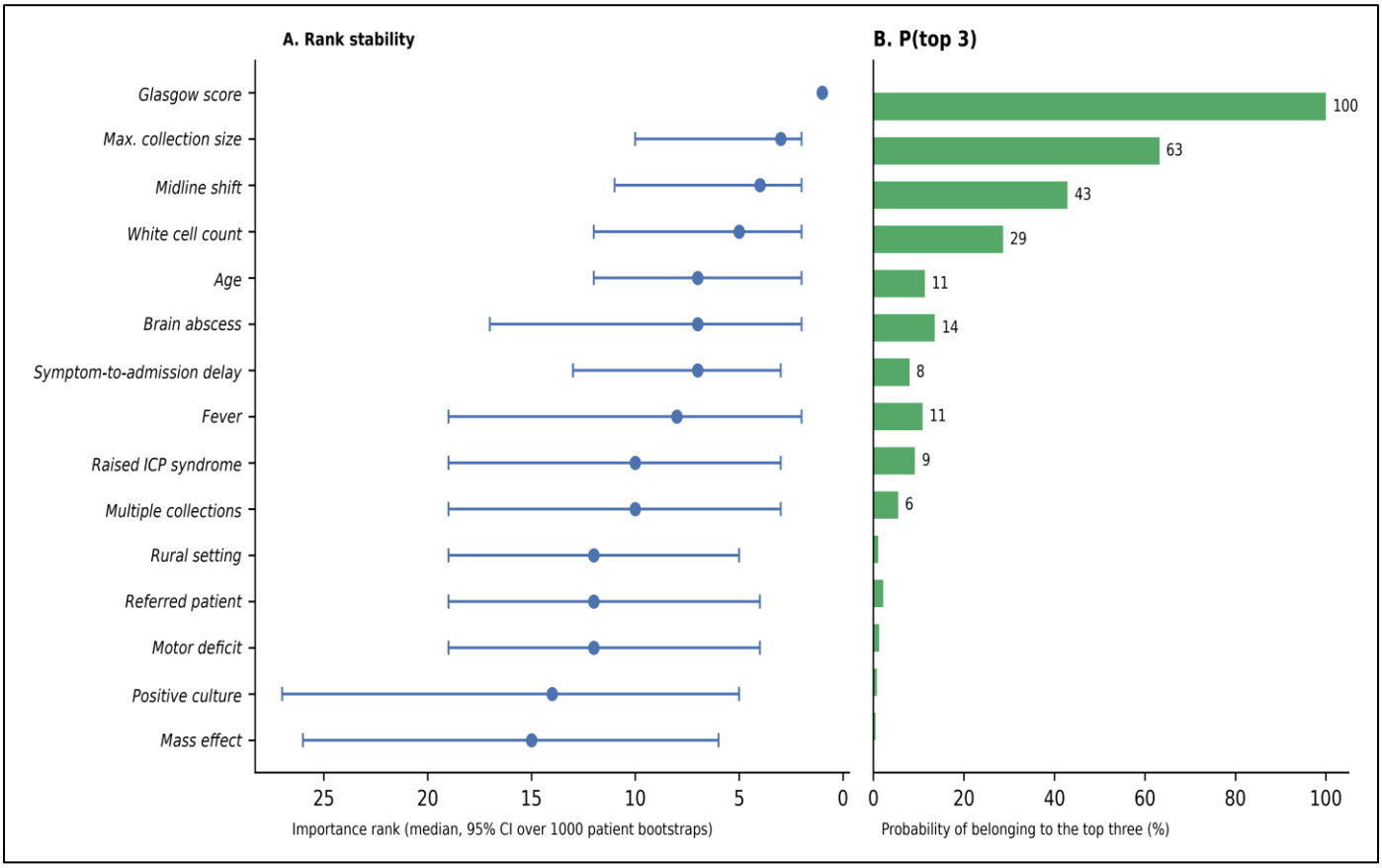

Fig 8 Stability of the Hierarchy Across 1000 Patient-Level Bootstrap Resamples.
A. Median rank and 95% interval. B. Probability of belonging to the top three variables. The GCS holds first rank in all resamples; beyond that, the intervals overlap widely and no ranking is claimed.

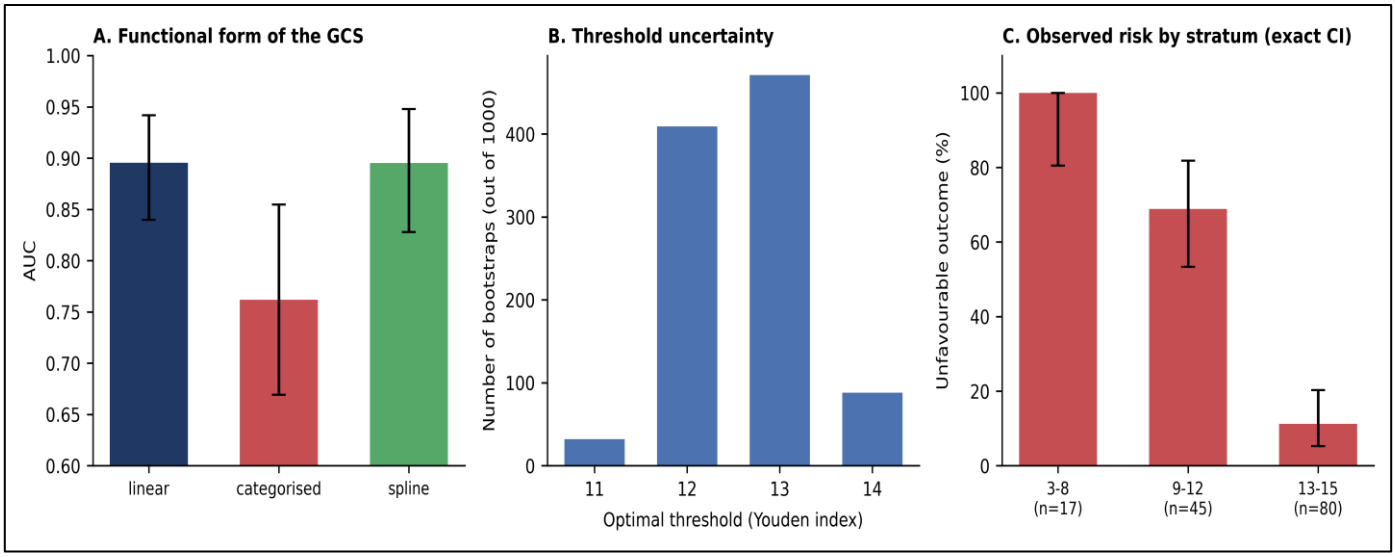

Fig 9 Functional form of the GCS, Threshold Uncertainty and Risk by Stratum.
A. Comparison of the linear, categorised and restricted cubic spline specifications. B. Bootstrap distribution of the threshold maximising the Youden index. C. Observed risk by stratum with exact Clopper–Pearson intervals.

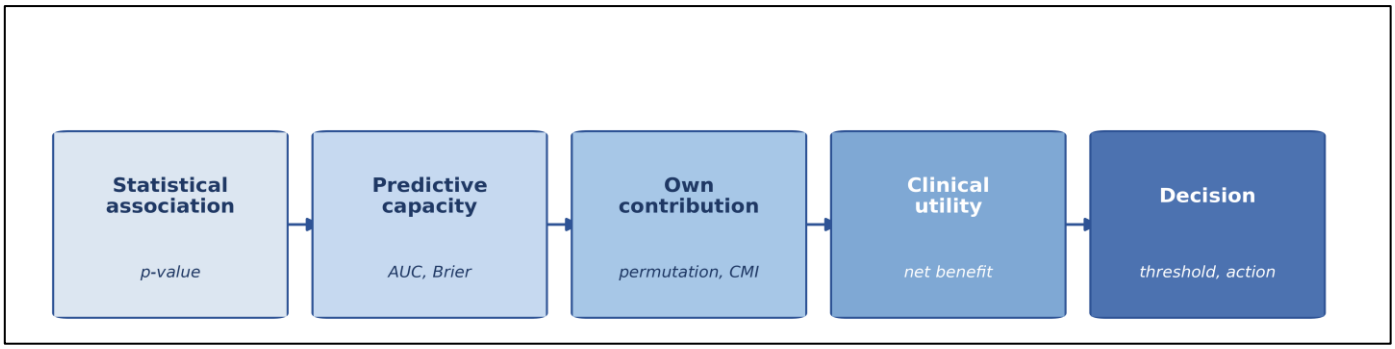

Fig 10 Five-Step Framework: from Statistical Association to Clinical Decision.

Each step can fail independently of the preceding ones. This framework structures the distinction, central to this work, between significance, predictive capacity, contribution of its own and decisional utility.

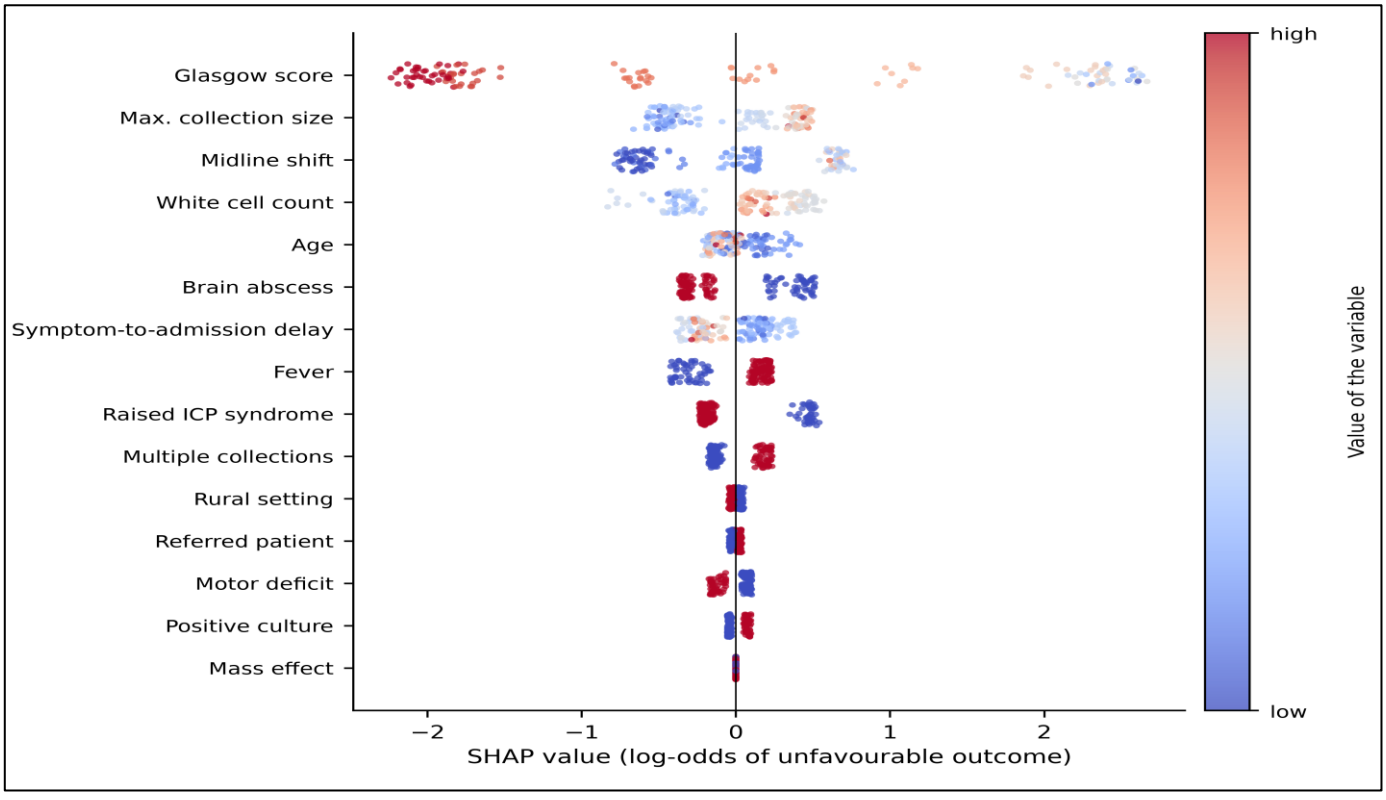

Fig 11 Individual SHAP Contributions (Top Fifteen Variables, XGBoost).

Each point represents one observation; the x-axis indicates the effect on the predicted log-odds, the colour codes the value of the variable. The spread of the GCS, out of all proportion to that of the other variables, makes the concentration of importance visible.

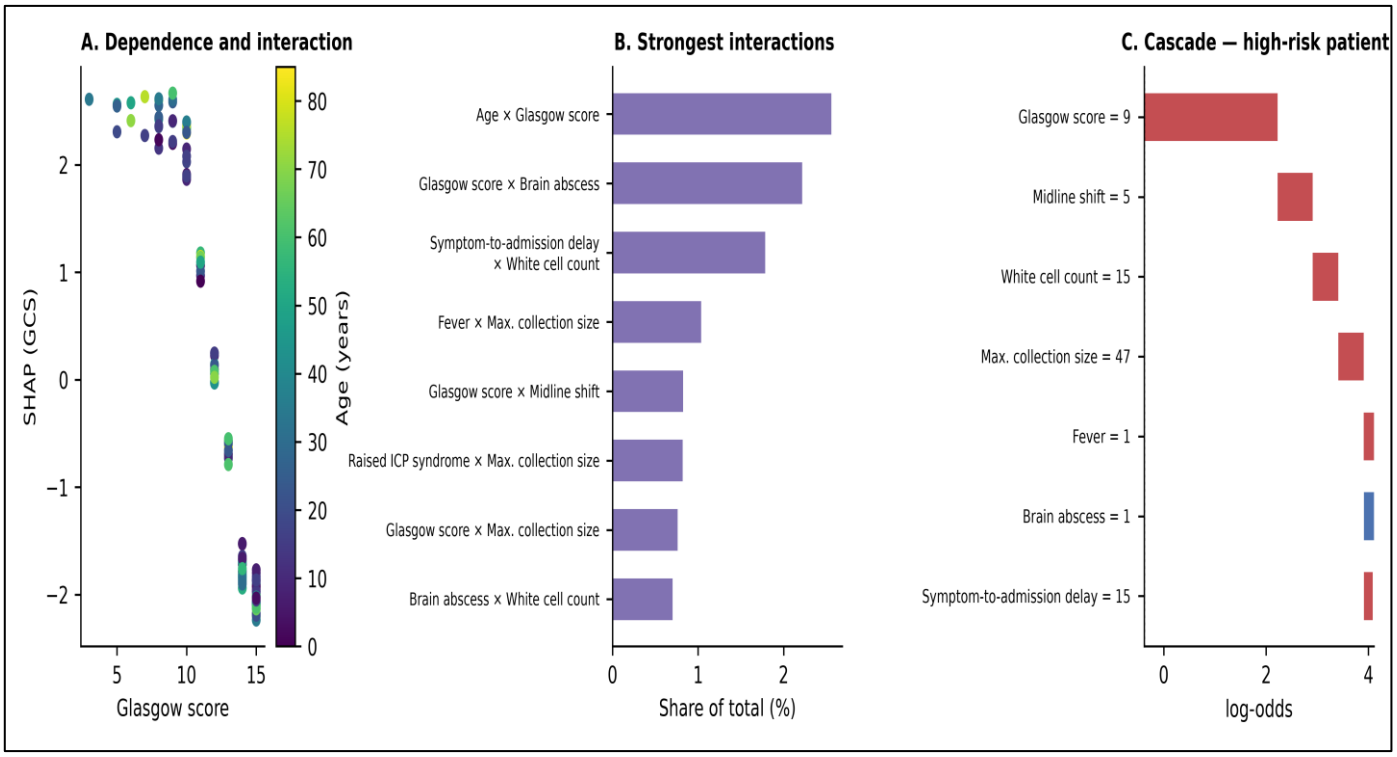

Fig 12 Interactions and individual attribution.
A. SHAP dependence for the Glasgow score, coloured by age. B. Strongest interactions according to SHAP interaction values; main effects represent 76.6% of total importance. C. Waterfall decomposition of the prediction for a high-risk patient.

DECLARATIONS

➢ *Funding.*

No specific funding was received for this work.

➢ *Conflicts Of Interest.*

The authors declare that they have no conflicts of interest.

➢ *Ethical Approval.*

Retrospective study conducted in accordance with the Declaration of Helsinki, approved by the Institutional Ethics Committee of Bouaké University Hospital under number CEI/CHU-BKE/126/2026 on 14 July 2026. Waiver of individual consent was granted on the grounds that the study concerns retrospectively collected and de-identified clinical data. Data were pseudonymised at source; since exact dates and internal identifiers were retained in the working database, re-identification by a third party holding the source record cannot be formally excluded.

➢ *Code Availability.*

The analysis code, the versioned dependency file, the definition of the containerised environment, the variable dictionary, the master file of out-of-sample predictions and the execution notebook reproducing all results are provided in Appendix S1 and will be deposited in a frozen public archive with a persistent identifier at the time of submission.

➢ *Data Availability.*

Individual-level data cannot be shared publicly for confidentiality reasons. A de-identified dataset may be considered after institutional and ethical approval, subject to applicable data protection requirements.

Table 12 Author Contributions (CRediT)

Role	Authors
Conceptualisation and methodology	All authors
Data collection	TETI Faozo Stephane Landry
Formal analysis	FIONKO Yao Bernard
Writing — original draft	DONGO Koffi Yves Soress
Writing — review and editing	All authors
Supervision	HAIDARA Aderehime

➢ *Acknowledgements.*

The authors thank the care teams of the department of neurosurgery of Bouaké University Hospital.

REFERENCES

[1]. Nathoo N, Nadvi SS, Narotam PK, van Dellen JR. Brain abscess: management and outcome analysis of a computed tomography era experience with 973 patients. World Neurosurg. 2011;75(5-6):716-26.

[2]. Xiao F, Tseng MY, Teng LJ, Tseng HM, Tsai JC. Brain abscess: clinical experience and analysis of prognostic factors. Surg Neurol. 2005;63(5):442-9.

[3]. Carpenter J, Stapleton S, Holliman R. Retrospective analysis of 49 cases of brain abscess and review of the literature. Eur J Clin Microbiol Infect Dis. 2007;26(1):1-11.

[4]. Landriel F, Ajler P, Hem S, Bendersky D, Goldschmidt E, Garategui L, et al. Supratentorial and infratentorial brain abscesses: surgical treatment, complications and outcomes — a 10-year single-centre study. Acta Neurochir (Wien). 2012;154(5):903-11.

[5]. Muzumdar D, Jhawar S, Goel A. Brain abscess: an overview. Int J Surg. 2011;9(2):136-44.

[6]. Shmueli G. To explain or to predict? Stat Sci. 2010;25(3):289-310.

[7]. Lo A, Chernoff H, Zheng T, Lo SH. Why significant variables aren't automatically good predictors. Proc Natl Acad Sci U S A. 2015;112(45):13892-7.

[8]. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. Adv Neural Inf Process Syst. 2017;30:4765-74.

[9]. Lundberg SM, Erion G, Chen H, DeGrave A, Prutkin JM, Nair B, et al. From local explanations to global understanding with explainable AI for trees. Nat Mach Intell. 2020;2(1):56-67.

[10]. Breiman L. Random forests. Mach Learn. 2001;45(1):5-32.

[11]. Hooker G, Mentch L, Zhou S. Unrestricted permutation forces extrapolation: variable importance requires at least one more model, or there is no free variable importance. Stat Comput. 2021;31(6):82.

[12]. Shannon CE. A mathematical theory of communication. Bell Syst Tech J. 1948;27(3):379-423.

[13]. Cover TM, Thomas JA. Elements of information theory. 2nd ed. Hoboken: Wiley-Interscience; 2006.

[14]. Brouwer MC, Tunkel AR, McKhann GM, van de Beek D. Brain abscess. N Engl J Med. 2014;371(5):447-56.

[15]. Bodilsen J, Duerlund LS, Mariager T, Brandt CT, Petersen PT, Larsen L, et al. Clinical features and prognostic factors in adults with brain abscess. Brain. 2023;146(4):1637-47.

[16]. Tunkel AR, Hasbun R, Bhimraj A, Byers K, Kaplan SL, Scheld WM, et al. 2017 Infectious Diseases Society of America's clinical practice guidelines for healthcare-associated ventriculitis and meningitis. Clin Infect Dis. 2017;64(6):e34-65.

[17]. Bodilsen J, D'Alessandris QG, Humphreys H, Iro MA, Klein M, Last K, et al. European Society of Clinical Microbiology and Infectious Diseases guidelines on diagnosis and treatment of brain abscess in children and adults. Clin Microbiol Infect. 2024;30(1):66-89.

[18]. van de Beek D, Brouwer MC, Koedel U, Wall EC. Community-acquired bacterial meningitis. Lancet. 2021;398(10306):1171-83.

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

[20]. Collins GS, Moons KGM, Dhiman P, Riley RD, Beam AL, Van Calster B, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ. 2024;385:e078378.

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

[22]. Jennett B, Bond M. Assessment of outcome after severe brain damage: a practical scale. Lancet. 1975;1(7905):480-4.

[23]. Peduzzi P, Concato J, Kemper E, Holford TR, Feinstein AR. A simulation study of the number of events per variable in logistic regression analysis. J Clin Epidemiol. 1996;49(12):1373-9.

[24]. Riley RD, Snell KIE, Ensor J, Burke DL, Harrell FE Jr, Moons KGM, et al. Minimum sample size for developing a multivariable prediction model: PART II — binary and time-to-event outcomes. Stat Med. 2019;38(7):1276-96.

[25]. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology. 1982;143(1):29-36.

[26]. Brier GW. Verification of forecasts expressed in terms of probability. Mon Weather Rev. 1950;78(1):1-3.

[27]. Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLoS One. 2015;10(3):e0118432.

[28]. Van Calster B, McLernon DJ, van Smeden M, Wynants L, Steyerberg EW. Calibration: the Achilles heel of predictive analytics. BMC Med. 2019;17(1):230.

[29]. Steyerberg EW, Vickers AJ, Cook NR, Gerds T, Gonen M, Obuchowski N, et al. Assessing the performance of prediction models: a framework for traditional and novel measures. Epidemiology. 2010;21(1):128-38.

[30]. Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. Med Decis Making. 2006;26(6):565-74.

[31]. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a

nonparametric approach. Biometrics. 1988;44(3):837-45.
[32]. Harrell FE Jr, Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Stat Med. 1996;15(4):361-87.
[33]. Steyerberg EW, Harrell FE Jr, Borsboom GJ, Eijkemans MJ, Vergouwe Y, Habbema JD. Internal validation of predictive models: efficiency of some procedures for logistic regression analysis. J Clin Epidemiol. 2001;54(8):774-81.
[34]. Fisher A, Rudin C, Dominici F. All models are wrong, but many are useful: learning a variable's importance by studying an entire class of prediction models simultaneously. J Mach Learn Res. 2019;20(177):1-81.
[35]. Aas K, Jullum M, Løland A. Explaining individual predictions when features are dependent: more accurate approximations to Shapley values. Artif Intell. 2021;298:103502.
[36]. Sundararajan M, Najmi A. The many Shapley values for model explanation. Proc Int Conf Mach Learn. 2020;119:9269-78.
[37]. Friedman JH. Greedy function approximation: a gradient boosting machine. Ann Stat. 2001;29(5):1189-232.
[38]. Apley DW, Zhu J. Visualizing the effects of predictor variables in black box supervised learning models. J R Stat Soc Series B Stat Methodol. 2020;82(4):1059-86.
[39]. Kerr KF, Wang Z, Janes H, McClelland RL, Psaty BM, Pepe MS. Net reclassification indices for evaluating risk prediction instruments: a critical review. Epidemiology. 2014;25(1):114-21.
[40]. Teasdale G, Jennett B. Assessment of coma and impaired consciousness: a practical scale. Lancet. 1974;2(7872):81-4.
[41]. Teasdale G, Maas A, Lecky F, Manley G, Stocchetti N, Murray G. The Glasgow Coma Scale at 40 years: standing the test of time. Lancet Neurol. 2014;13(8):844-54.
[42]. Steyerberg EW, Mushkudiani N, Perel P, Butcher I, Lu J, McHugh GS, et al. Predicting outcome after traumatic brain injury: development and international validation of prognostic scores based on admission characteristics. PLoS Med. 2008;5(8):e165.
[43]. MRC CRASH Trial Collaborators. Predicting outcome after traumatic brain injury: practical prognostic models based on large cohort of international patients. BMJ. 2008;336(7641):425-9.
[44]. Roozenbeek B, Lingsma HF, Lecky FE, Lu J, Weir J, Butcher I, et al. Prediction of outcome after moderate and severe traumatic brain injury: external validation of the IMPACT and CRASH prognostic models. Crit Care Med. 2012;40(5):1609-17.
[45]. Christodoulou E, Ma J, Collins GS, Steyerberg EW, Verbakel JY, Van Calster B. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. J Clin Epidemiol. 2019;110:12-22.
[46]. van der Ploeg T, Austin PC, Steyerberg EW. Modern modelling techniques are data hungry: a simulation study for predicting dichotomous endpoints. BMC Med Res Methodol. 2014;14:137.
[47]. Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat Mach Intell. 2019;1(5):206-15.
[48]. Ghassemi M, Oakden-Rayner L, Beam AL. The false hope of current approaches to explainable artificial intelligence in health care. Lancet Digit Health. 2021;3(11):e745-50.
[49]. Bodilsen J, Mariager T, Duerlund LS, Storgaard M, Larsen L, Brandt CT, et al. Brain abscess caused by oral cavity bacteria: a nationwide, population-based cohort study. Clin Infect Dis. 2024;78(3):544-53.
[50]. Eriksen EM, Larsen L, Storgaard M, Mens H, Wiese L, Jepsen MPG, et al. Nonoperative versus neurosurgical treatment of brain abscess: an emulated trial nested within a nationwide, population-based cohort. Clin Infect Dis. 2026;82(2):219-27.
[51]. Brouwer MC, Coutinho JM, van de Beek D. Clinical characteristics and outcome of brain abscess: systematic review and meta-analysis. Neurology. 2014;82(9):806-13.
[52]. Senders JT, Staples PC, Karhade AV, Zaki MM, Gormley WB, Broekman MLD, et al. Machine learning and neurosurgical outcome prediction: a systematic review. World Neurosurg. 2018;109:476-86.
[53]. Feng Y, Wang Y, Zhang M, Zhou W, Zhou Y, Liu H. Interpretable machine learning models for predicting outcome after traumatic brain injury. Front Neurol. 2023;14:1198042.
[54]. Wang HL, Hsu WY, Lee MH, Weng HH, Chang SW, Yang JT, et al. Automatic machine learning-based classification and outcome prediction in spontaneous intracerebral hemorrhage. Front Neurol. 2019;10:910.
[55]. van Os HJA, Ramos LA, Hilbert A, van Leeuwen M, van Walderveen MAA, Kruyt ND, et al. Predicting outcome of endovascular treatment for acute ischemic stroke: potential value of machine learning algorithms. Front Neurol. 2018;9:784.
[56]. Karhade AV, Thio QCBS, Ogink PT, Shah AA, Bono CM, Oh KS, et al. Development of machine learning algorithms for prediction of survival in spinal metastatic disease. Neurosurgery. 2019;85(4):E671-81.
[57]. Muscas G, Matteuzzi T, Becattini E, Orlandini S, Battista F, Laiso A, et al. Development of machine learning models to prognosticate chronic shunt-dependent hydrocephalus after aneurysmal subarachnoid hemorrhage. Acta Neurochir (Wien). 2020;162(12):3093-105.
[58]. Vasey B, Nagendran M, Campbell B, Clifton DA, Collins GS, Denaxas S, et al. Reporting guideline for the early-stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. BMJ. 2022;377:e070904.

[59]. Liu X, Cruz Rivera S, Moher D, Calvert MJ, Denniston AK; SPIRIT-AI and CONSORT-AI Working Group. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. Nat Med. 2020;26(9):1364-74.
[60]. Cruz Rivera S, Liu X, Chan AW, Denniston AK, Calvert MJ; SPIRIT-AI and CONSORT-AI Working Group. Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension. Nat Med. 2020;26(9):1351-63.
[61]. Norgeot B, Quer G, Beaulieu-Jones BK, Torkamani A, Dias R, Gianfrancesco M, et al. Minimum information about clinical artificial intelligence modeling: the MI-CLAIM checklist. Nat Med. 2020;26(9):1320-4.
[62]. Mongan J, Moy L, Kahn CE Jr. Checklist for Artificial Intelligence in Medical Imaging (CLAIM): a guide for authors and reviewers. Radiol Artif Intell. 2020;2(2):e200029.
[63]. Hernandez-Boussard T, Bozkurt S, Ioannidis JPA, Shah NH. MINIMAR (MINimum Information for Medical AI Reporting): developing reporting standards for artificial intelligence in health care. J Am Med Inform Assoc. 2020;27(12):2011-5.
[64]. Meara JG, Leather AJM, Hagander L, Alkire BC, Alonso N, Ameh EA, et al. Global Surgery 2030: evidence and solutions for achieving health, welfare, and economic development. Lancet. 2015;386(9993):569-624.
[65]. Dewan MC, Rattani A, Fieggen G, Arraez MA, Servadei F, Boop FA, et al. Global neurosurgery: the current capacity and deficit in the provision of essential neurosurgical care. J Neurosurg. 2019;130(4):1055-64.
[66]. Park KB, Johnson WD, Dempsey RJ. Global neurosurgery: the unmet need. World Neurosurg. 2016;88:32-5.
[67]. Kanmounye US, Robertson FC, Thango NS, Doe AN, Bankole NDA, Ginette PA, et al. Needs of young African neurosurgeons and residents: a cross-sectional study. Front Surg. 2021;8:647279.
[68]. Ramspek CL, Jager KJ, Dekker FW, Zoccali C, van Diepen M. External validation of prognostic models: what, why, how, when and where? Clin Kidney J. 2021;14(1):49-58.
[69]. Fleuren LM, Klausch TLT, Zwager CL, Schoonmade LJ, Guo T, Roggeveen LF, et al. Machine learning for the prediction of sepsis: a systematic review and meta-analysis of diagnostic test accuracy. Intensive Care Med. 2020;46(3):383-400.
[70]. Shillan D, Sterne JAC, Champneys A, Gibbison B. Use of machine learning to analyse routinely collected intensive care unit data: a systematic review. Crit Care. 2019;23(1):284.
[71]. Moor M, Rieck B, Horn M, Jutzeler CR, Borgwardt K. Early prediction of sepsis in the ICU using machine learning: a systematic review. Front Med (Lausanne). 2021;8:607952.
[72]. Wong A, Otles E, Donnelly JP, Krumm A, McCullough J, DeTroyer-Cooley O, et al. External validation of a widely implemented proprietary sepsis prediction model in hospitalized patients. JAMA Intern Med. 2021;181(8):1065-70.
[73]. Finlayson SG, Subbaswamy A, Singh K, Bowers J, Kupke A, Zittrain J, et al. The clinician and dataset shift in artificial intelligence. N Engl J Med. 2021;385(3):283-6.
[74]. Youssef A, Pencina M, Thakur A, Zhu T, Clifton D, Shah NH. External validation of AI models in health should be replaced with recurring local validation. Nat Med. 2023;29(11):2686-7.
[75]. Lipton ZC. The mythos of model interpretability. Commun ACM. 2018;61(10):36-43.
[76]. Krishna S, Han T, Gu A, Wu S, Jabbari S, Lakkaraju H. The disagreement problem in explainable machine learning: a practitioner's perspective. arXiv:2202.01602 [cs.LG]. 2022.
[77]. Molnar C. Interpretable machine learning: a guide for making black box models explainable. 2nd ed. Munich: Independently published; 2022.
[78]. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. Science. 2019;366(6464):447-53.
[79]. Celi LA, Cellini J, Charpignon ML, Dee EC, Dernoncourt F, Eber R, et al. Sources of bias in artificial intelligence that perpetuate healthcare disparities — a global review. PLOS Digit Health. 2022;1(3):e0000022.
[80]. Chen IY, Pierson E, Rose S, Joshi S, Ferryman K, Ghassemi M. Ethical machine learning in healthcare. Annu Rev Biomed Data Sci. 2021;4:123-44.
[81]. Rudolfson N, Dewan MC, Park KB, Shrime MG, Meara JG, Alkire BC. The economic consequences of neurosurgical disease in low- and middle-income countries. J Neurosurg. 2018;130(4):1149-56.
[82]. Corley J, Lepard J, Barthélemy E, Ashby JL, Park KB. Essential neurosurgical workforce needed to address neurotrauma in low- and middle-income countries. World Neurosurg. 2019;123:295-9.
[83]. Hernán MA, Hsu J, Healy B. A second chance to get causal inference right: a classification of data science tasks. Chance. 2019;32(1):42-9.
[84]. Prosperi M, Guo Y, Sperrin M, Koopman JS, Min JS, He X, et al. Causal inference and counterfactual prediction in machine learning for actionable healthcare. Nat Mach Intell. 2020;2(7):369-75.
[85]. Moons KGM, Damen JAA, Kaul T, Hooft L, Andaur Navarro C, Dhiman P, et al. PROBAST+AI: an updated quality, risk of bias, and applicability assessment tool for prediction models using regression or artificial intelligence methods. BMJ. 2025;388:e082505.

SUPPLEMENTARY MATERIAL

Table 12 Internal Bootstrap Validation: Optimism and Corrected Performance

Model	Apparent AUC	Optimism	Corrected AUC	Apparent Brier	Corrected Brier
Glasgow Coma Scale alone	0.920	0.000	0.920	0.108	0.111
Penalised logistic regression	0.950	0.060	0.890	0.088	0.135
Random forest	0.961	0.042	0.919	0.100	0.133
XGBoost	0.979	0.052	0.927	0.060	0.106
LightGBM	1.000	0.045	0.955	0.024	0.079

Harrell's procedure: refitting of the model and of all preprocessing on each bootstrap sample, evaluation on that sample and then on the original sample; the mean optimism is subtracted from the apparent performance. The AUC optimism of GCS alone rounds to 0.000, but the optimism of its Brier score is small and not null. The corrected AUCs remain markedly higher than the out-of-sample values from cross-validation: this procedure estimates the performance of a different development process, itself remains optimistic in small samples, and must not be interpreted as an expected external performance. This table illustrates susceptibility to overfitting rather than a corrected performance.

Table 13 Marginal AUC Gain of Each Variable Added to GCS Alone

Variable added	ΔAUC	95% CI	Positive gain
Subdural empyema	+0.0103	-0.0074 ; 0.0368	79%
Fever	+0.0054	-0.0115 ; 0.0242	73%
Raised intracranial pressure syndrome	+0.0052	-0.0069 ; 0.0165	79%
Herniation / midline shift (mm)	+0.0047	-0.0195 ; 0.0368	61%
Maximum size of the collection (mm)	+0.0004	-0.0272 ; 0.0282	51%
Immunosuppression	−0.0014	-0.0054 ; 0.0012	20%
Diabetes	−0.0014	-0.0057 ; 0.0013	18%
Anisocoria	−0.0014	-0.0059 ; 0.0014	17%
Hydrocephalus	−0.0014	-0.0057 ; 0.0012	19%
Motor deficit	−0.0029	-0.0154 ; 0.0077	30%
Ventriculitis	−0.0029	-0.0086 ; 0.0008	11%
Symptom-to-admission delay (days)	−0.0037	-0.0197 ; 0.0107	30%
Midline shift not measured (indicator)	−0.0039	-0.0126 ; 0.0036	15%
Seizures	−0.0043	-0.0113 ; 0.0010	6%
Post-traumatic origin	−0.0052	-0.0163 ; 0.0040	18%
Brain abscess (vs other form)	−0.0056	-0.0324 ; 0.0193	32%
White cell count (G/L)	−0.0058	-0.0295 ; 0.0136	34%
Postoperative origin	−0.0062	-0.0248 ; 0.0054	24%
Rural residence	−0.0080	-0.0209 ; 0.0006	4%
Mass effect	−0.0095	-0.0292 ; 0.0035	8%
Culture not performed (indicator)	−0.0101	-0.0268 ; 0.0004	3%
Referred patient	−0.0105	-0.0220 ; -0.0014	0%
Multiple collections	−0.0107	-0.0271 ; 0.0011	4%
Positive culture	−0.0147	-0.0301 ; -0.0034	0%
Male sex	−0.0159	-0.0334 ; -0.0028	1%
Age (years)	−0.0215	-0.0454 ; 0.0009	3%

Reference AUC of GCS alone: 0.887 in this procedure. Intervals by patient-level bootstrap (500 replications). "Positive gain": proportion of resamples in which the gain is greater than zero. No interval excludes zero.

Table 14 Strongest Interactions According to SHAP Interaction Values

Variable pair	Share of total importance (%)
Age (years) × Admission Glasgow Coma Scale	2.6
Admission Glasgow Coma Scale × Brain abscess (vs other form)	2.2
Symptom-to-admission delay (days) × White cell count (G/L)	1.8
Fever × Maximum size of the collection (mm)	1.0

Variable pair	Share of total importance (%)
Admission Glasgow Coma Scale × Herniation / midline shift (mm)	0.8
Raised intracranial pressure syndrome × Maximum size of the collection (mm)	0.8
Admission Glasgow Coma Scale × Maximum size of the collection (mm)	0.8
Brain abscess (vs other form) × White cell count (G/L)	0.7

Main effects represent 76.6% of total importance and interactions as a whole 23.4%. No interaction exceeds 2.6%.

Table 15 Nested Cross-Validation with Hyperparameter Optimisation

Model	AUC (fixed hyperparameters)	AUC (optimised)	Difference
PLR	0.839	0.832	−0.008
RF	0.877	0.881	+0.004
XGB	0.868	0.873	+0.005
LGBM	0.864	0.877	+0.013
Glasgow Coma Scale alone	0.919	0.919	+0.000

Optimisation by random search in the inner loop. The maximum gain from optimisation is +0.013 and does not alter the hierarchy.

Table 16 Sensitivity Analysis of Model Parameterisation

Configuration	AUC
PLR C=0.1	0.866
PLR C=0.5	0.855
PLR C=1.0	0.848
XGB depth=1	0.870
XGB depth=2	0.868
XGB depth=3	0.868
RF depth=3	0.865
RF depth=4	0.868
RF depth=5	0.871
LGBM 5 leaves	0.858
LGBM 7 leaves	0.860
LGBM 10 leaves	0.860

All configurations tested give AUCs between 0.848 and 0.871, lower than that of GCS alone (0.896).

Table 17 Net Benefit by Probability Threshold (Decision Curve Analysis)

Model	Threshold 20%	Threshold 30%	Threshold 40%	Threshold 50%
Glasgow Coma Scale alone	0.326	0.296	0.272	0.254
Penalised logistic regression	0.290	0.266	0.254	0.225
Random forest	0.280	0.268	0.270	0.232
XGBoost	0.305	0.280	0.263	0.232
LightGBM	0.299	0.257	0.244	0.218

Table 18 Additional Sensitivity Analyses

Analysis (logistic model)	AUC	95% CI
GCS alone	0.623	0.472 – 0.755
27 variables	0.476	0.320 – 0.633

Configuration (non-measurement indicators)	AUC
with indicators | PLR	0.855
with indicators | XGB	0.868
without indicators | PLR	0.865
without indicators | XGB	0.869

Table 19 Sensitivity Analysis: Set Restricted to Variables Available at the Bedside (Without Culture)

Model	AUC — full set (27 var.)	AUC — bedside set (25 var.)	95% CI of ΔAUC (GCS − model)	95% CI of ΔBrier
Penalised logistic regression	0.849	0.858	−0.0028 ; +0.0773	−0.0532 ; −0.0130
Random forest	0.869	0.871	−0.0225 ; +0.0711	−0.0458 ; −0.0154
XGBoost	0.861	0.863	−0.0092 ; +0.0737	−0.0396 ; −0.0084
LightGBM	0.860	0.861	+0.0004 ; +0.0729	−0.0545 ; −0.0158

Removing positive culture and its non-measurement indicator slightly improves all four models without any of them reaching GCS alone (0.896). The four Brier score differences continue to exclude zero; the AUC differences now exclude zero only for LightGBM, marginally. The main result therefore does not depend on the inclusion of a variable unavailable at the time of the initial decision, but the statistical support for the discrimination comparison weakens on this restricted set.

Table 20 Observed Risk by Exact GCS Value

GCS	n	Unfavourable outcomes	Observed risk	Exact 95% CI	of which deaths
3	1	1	100.0%	2.5 – 100.0	0
5	3	3	100.0%	29.2 – 100.0	1
6	2	2	100.0%	15.8 – 100.0	0
7	2	2	100.0%	15.8 – 100.0	1
8	9	9	100.0%	66.4 – 100.0	3
9	8	8	100.0%	63.1 – 100.0	1
10	17	13	76.5%	50.1 – 93.2	0
11	8	5	62.5%	24.5 – 91.5	1
12	12	5	41.7%	15.2 – 72.3	2
13	18	5	27.8%	9.7 – 53.5	4
14	23	3	13.0%	2.8 – 33.6	2
15	39	1	2.6%	0.1 – 13.5	1

Clopper–Pearson intervals. Complete separation extends up to a GCS of 9. Deaths are distributed across the whole scale, four occurring in patients admitted with a GCS of 13, which sheds light on the difficulty of predicting them. The numbers per value are small for the lowest GCS values: interpretation of the fitted curve must take this uneven density into account.

Table 21 Relative Share of Predictive Importance — All 27 Candidate Variables

Variable	SHAP (%)	95% CI	Median rank	P(top 3)	ΔAUC perm.	Positive perm.	Perm. share (%)	MI (bits)	MI (%)
Admission Glasgow Coma Scale	**37.2**	**27.3 – 50.2**	**1 [1–1]**	**1.00**	**0.2927**	**100%**	**86.6**	**0.428**	**49.9**
Maximum size of the collection (mm)	10.5	2.5 – 20.2	3 [2–10]	0.63	0.0130	67%	3.8	0.172	20.1
Herniation / midline shift (mm)	8.2	1.8 – 15.3	4 [2–11]	0.43	0.0101	64%	3.0	0.079	9.2
White cell count (G/L)	6.7	1.6 – 13.7	5 [2–12]	0.29	0.0002	45%	0.6	0.003	0.4
Age (years)	4.7	1.1 – 10.9	7 [2–12]	0.11	-0.0042	32%	0.1	0.006	0.7
Brain abscess (vs other form)	4.3	0.2 – 12.1	7 [2–17]	0.14	0.0065	61%	1.9	0.014	1.6
Symptom-to-admission delay (days)	4.3	1.0 – 10.6	7 [3–13]	0.08	-0.0040	30%	0.0	0.012	1.3
Fever	3.6	0.0 – 11.7	8 [2–19]	0.11	0.0048	56%	1.5	0.018	2.1
Raised intracranial pressure syndrome	2.6	0.0 – 11.7	10 [3–19]	0.09	0.0026	48%	0.8	0.002	0.2
Multiple collections	2.4	0.1 – 11.1	10 [3–19]	0.06	-0.0027	29%	0.1	0.003	0.4
Rural residence	1.3	0.1 – 6.6	12 [5–19]	0.01	-0.0011	28%	0.2	0.002	0.3
Referred patient	1.2	0.1 – 7.4	12 [4–19]	0.02	-0.0017	19%	0.0	0.000	0.0
Motor deficit	1.1	0.0 – 7.3	12 [4–19]	0.01	-0.0006	29%	0.2	0.000	0.0
Positive culture	0.6	0.0 – 7.1	14 [5–27]	0.01	-0.0023	18%	0.0	0.001	0.1
Mass effect	0.3	0.0 – 5.6	15 [6–26]	0.01	-0.0018	20%	0.0	0.016	1.9
Male sex	0.2	0.0 – 4.7	16 [6–22]	0.00	-0.0023	13%	0.0	0.004	0.4
Midline shift not measured (indicator)	0.2	0.0 – 5.2	19 [6–27]	0.00	-0.0027	26%	0.1	0.002	0.3
Seizures	0.0	0.0 – 3.0	18 [9–23]	0.00	-0.0034	14%	0.0	0.003	0.4
Anisocoria	0.0	0.0 – 0.0	19 [17–25]	0.00	-0.0015	6%	0.0	0.015	1.7
Postoperative origin	0.0	0.0 – 1.2	20 [13–24]	0.00	-0.0001	23%	0.1	0.010	1.1
Diabetes	0.0	0.0 – 0.0	21 [16–24]	0.00	-0.0006	6%	0.0	0.002	0.3
Immunosuppression	0.0	0.0 – 0.0	22 [15–23]	0.00	0.0013	20%	0.4	0.014	1.6
Post-traumatic origin	0.0	0.0 – 4.0	19 [7–23]	0.00	-0.0000	22%	0.1	0.001	0.1
Ventriculitis	0.0	0.0 – 0.3	24 [15–27]	0.00	-0.0011	14%	0.1	0.019	2.2
Hydrocephalus	0.0	0.0 – 0.0	25 [24–27]	0.00	-0.0002	13%	0.0	0.005	0.5
Subdural empyema	0.0	0.0 – 3.7	22 [9–26]	0.00	0.0009	31%	0.3	0.020	2.3
Culture not performed (indicator)	0.0	0.0 – 1.4	26 [11–27]	0.00	-0.0010	17%	0.1	0.007	0.8

Full version of Table 6. Values below 0.05% are displayed as 0.0 by rounding and are not necessarily exactly null.

The mortality outcome (16 events) permits no conclusion: both intervals encompass 0.500, over 1000 valid resamples. The values are reported with the logistic model; with this outcome the tree ensemble produces a meaningless out-of-sample estimate (0.436), the predicted probabilities being almost identical between those who died and those who survived. Removing the non-measurement indicators does not degrade performance.

Table 22 Spearman Correlations Between the GCS and the Main Variables

Variable	Spearman ρ	95% CI
Maximum size of the collection (mm)	-0.53	-0.64 – -0.40
Herniation / midline shift (mm)	-0.33	-0.47 – -0.17
Mass effect	-0.14	-0.30 – 0.03
Symptom-to-admission delay (days)	-0.22	-0.37 – -0.06
Age (years)	-0.07	-0.23 – 0.09

The full correlation matrix, reordered by hierarchical agglomerative clustering, is shown in Figure 13.

Table 23 Bootstrap Distribution of the Glasgow Threshold Maximising the Youden Index

Threshold	Number of resamples	Proportion
11	32	3.2%
12	409	40.9%
13	471	47.1%
14	88	8.8%

Over 1000 patient-level resamples. The 12–13 transition zone concentrates 88.0% of the estimates, but the exact threshold is not reproducible.

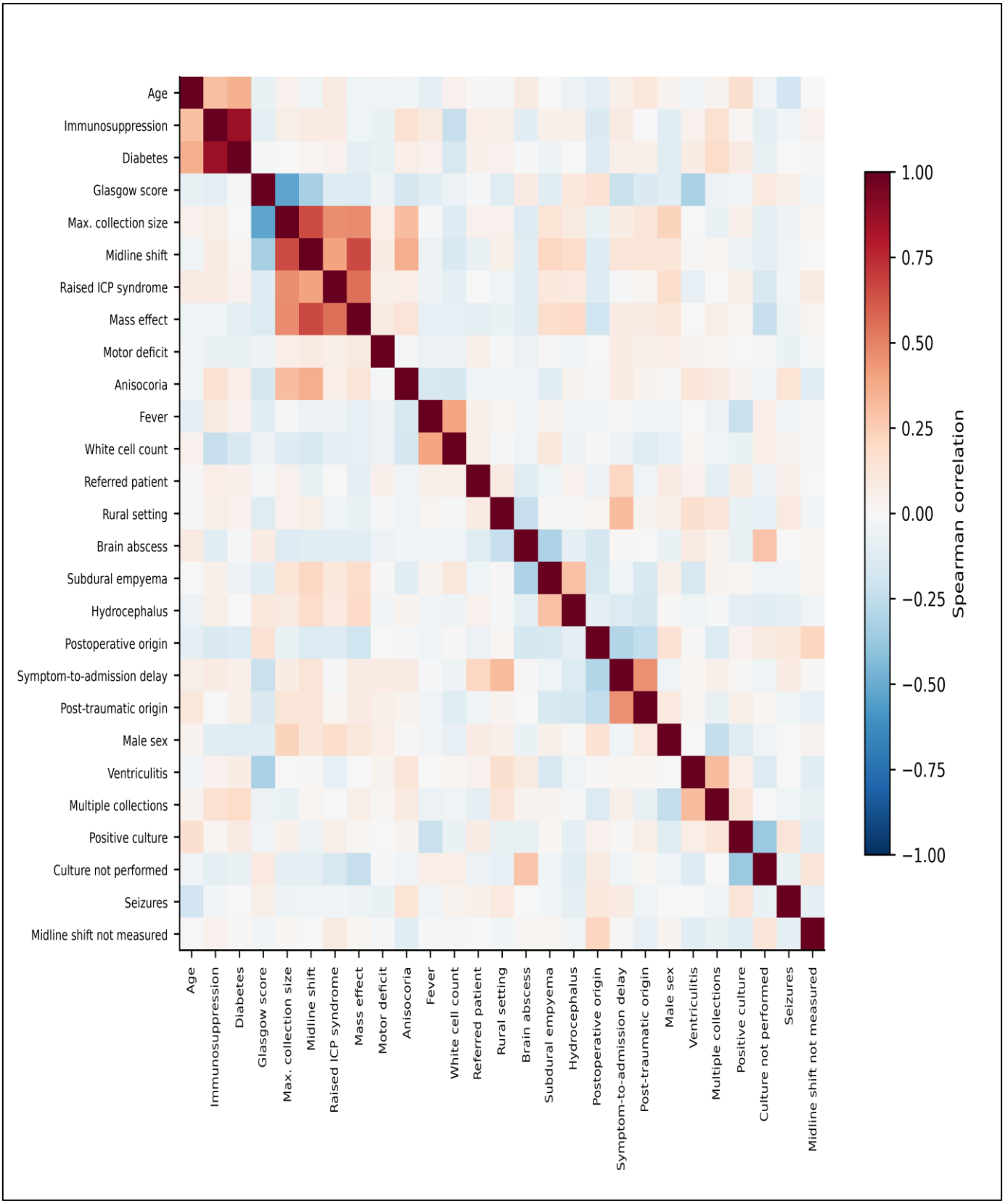

Fig 13 Matrix of Spearman Correlations Between the 27 Candidate Variables.

Reordered by hierarchical agglomerative clustering on the distance $1 - |\rho|$. The blocks of correlated variables explain the gap between shared attribution (SHAP) and contribution of its own (permutation).

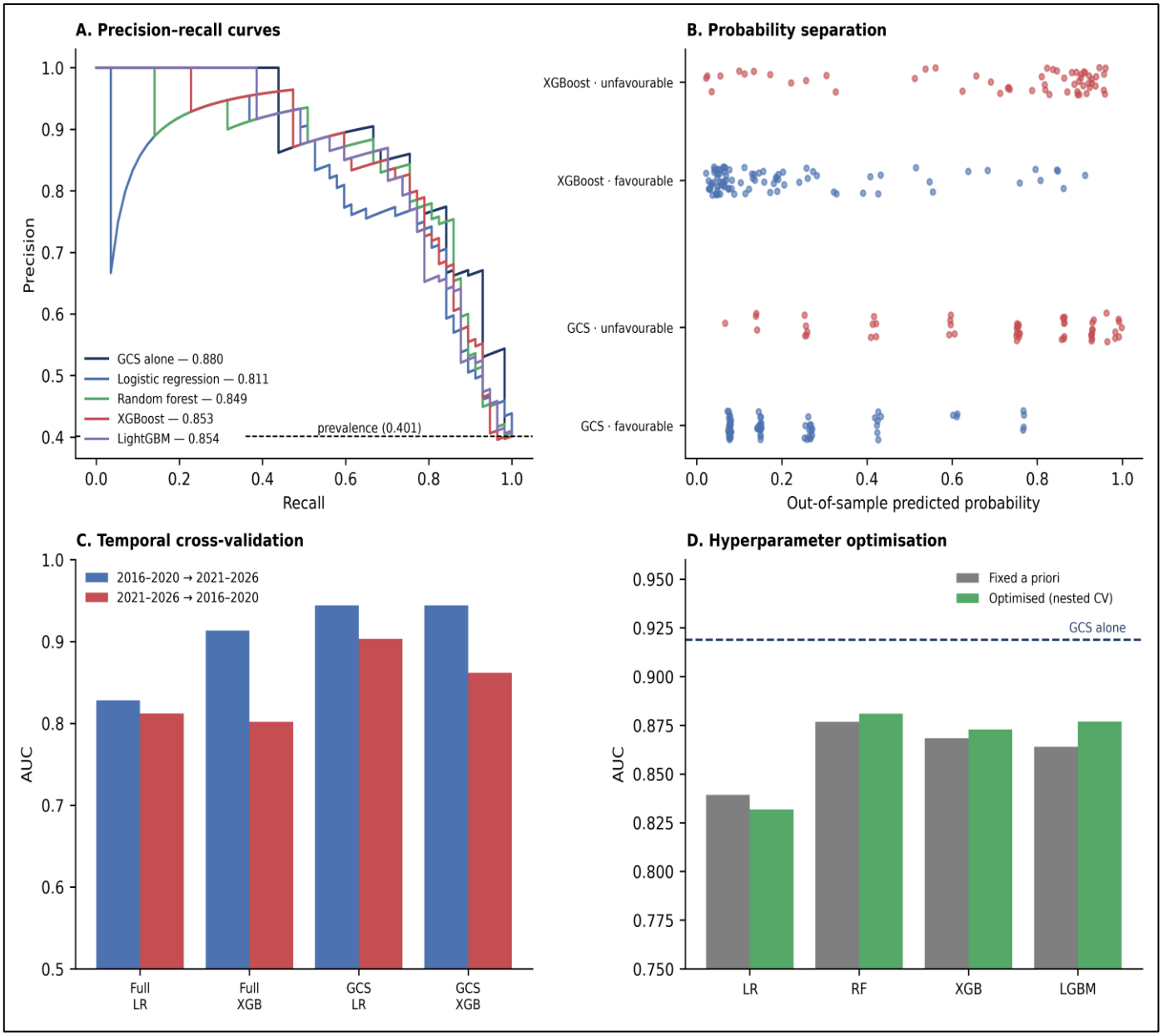

Fig 14 Additional Performance Analyses.
A. Precision–recall curves with prevalence reference. B. Separation of predicted probabilities according to observed outcome. C. Temporal cross-validation between the 2016–2020 and 2021–2026 periods. D. Effect of hyperparameter optimisation in nested cross-validation.

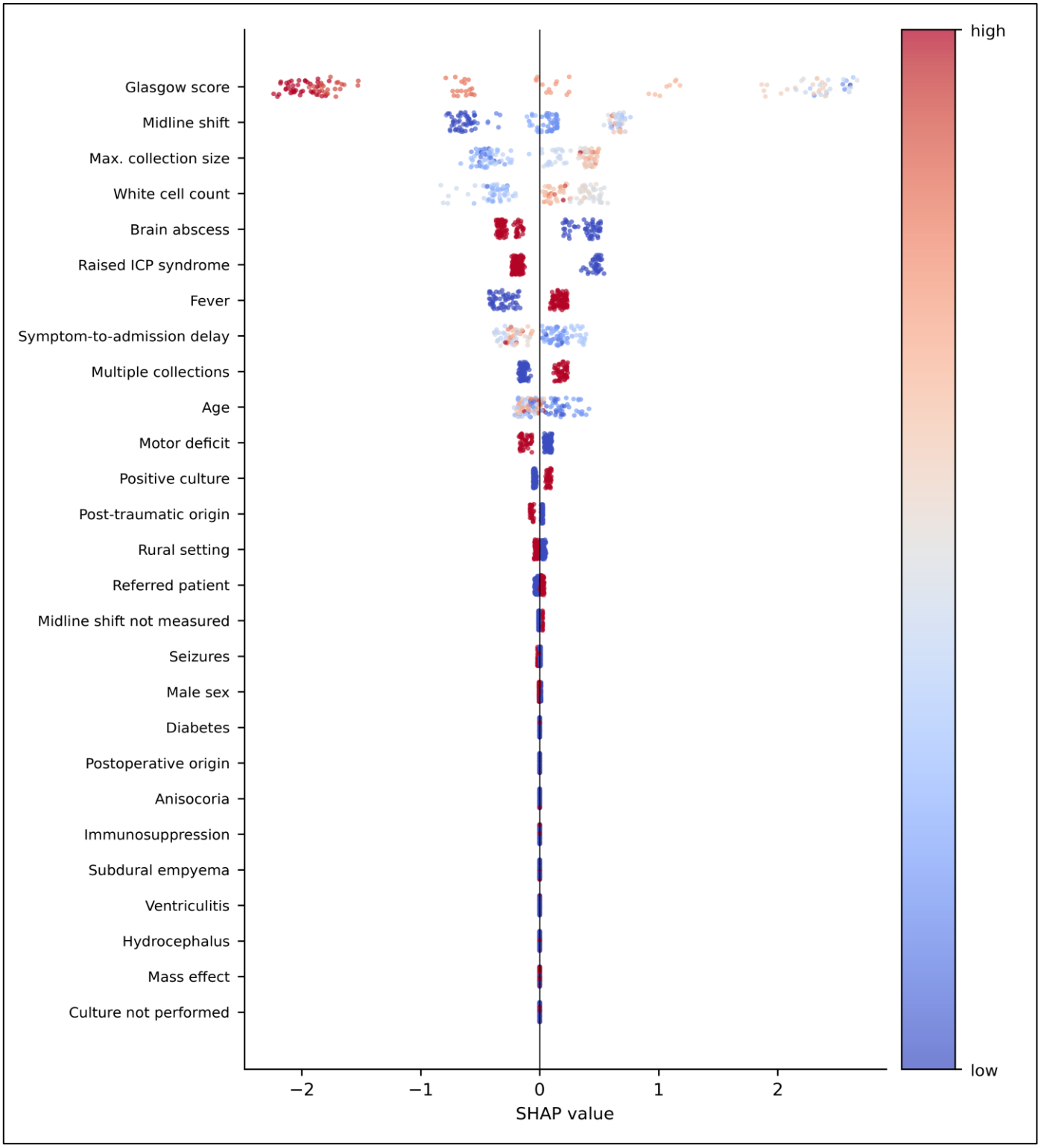

Fig 15 Individual SHAP Contributions for All 27 Variables.

Full version of Figure 11. The spread decreases very rapidly beyond the first variables.

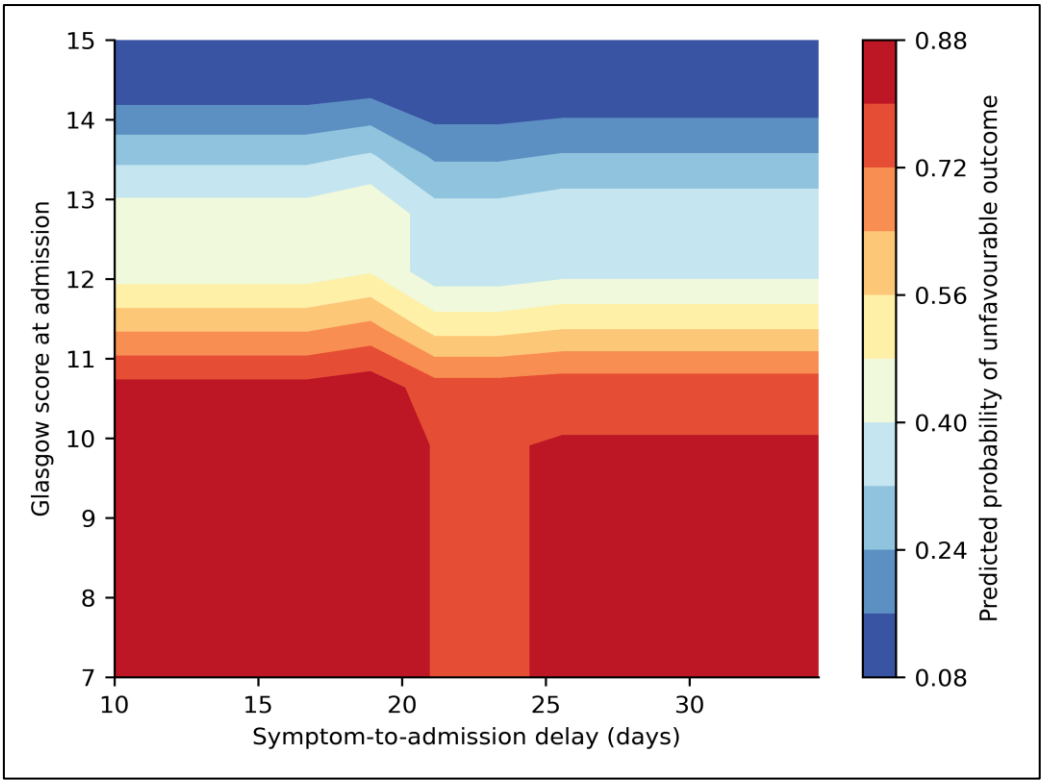

Fig 16 Two-Dimensional Partial Dependence Plots.

Interactions between the GCS and, respectively, age, maximum size of the collection and symptom-to-admission delay. The surfaces are practically additive, confirming the absence of substantial interaction.

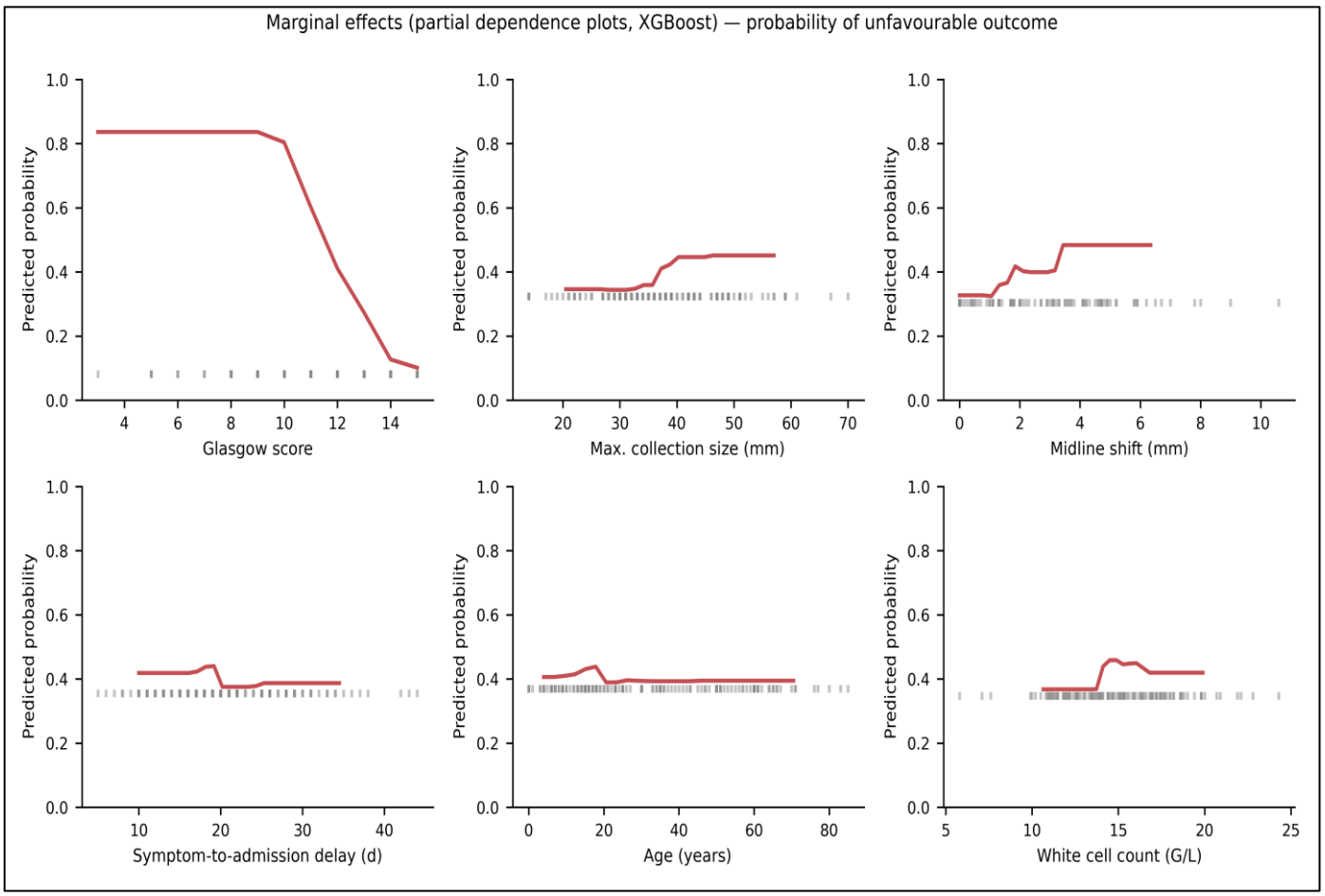

Fig 17 Univariable Partial Dependence Plots and Accumulated Local Effects.

The marginal distribution of each variable is displayed beneath the curve; interpretation is restricted to the ranges actually populated.

Table 24 Hierarchy of Analyses

Analysis	Status	Question addressed	Criterion
GCS alone versus multivariable models	Primary	Added predictive value	AUC, Brier, paired comparison
Calibration and decision curve	Primary	Reliability of probabilities and utility	slope, intercept, net benefit
SHAP, permutation, mutual information	Secondary	Concentration of the signal	rank, relative share
Parsimony with nested selection	Secondary	Minimum number of variables	AUC as a function of k
Subgroups (GCS 9–15, 9–14, survivors)	Prespecified exploratory	Dependence on extreme cases	AUC, paired difference
Hyperparameter optimisation	Exploratory	Does the parameterisation disadvantage the models?	Nested AUC
Temporal validation	Exploratory	Stability across periods	AUC by direction
Ordinal model, reclassification, interactions	Exploratory	Robustness and mechanism	OR, IDI, decomposition
Mortality prediction	Exploratory	Feasibility	AUC and interval

This hierarchy was fixed before the results were written up. The exploratory analyses ground no conclusion and are reported for their descriptive value.

Table 25 Complete Parameterisation of the Algorithms

Model	Parameters fixed a priori	Range explored in nested validation
Logistic regression	L2 penalty, C = 0.5, lbfgs solver, 5000 iterations maximum, unpenalised intercept, standardisation within the pipeline	$C \in [0.01; 10]$
Random forest	600 trees, depth 4, 5 observations minimum per leaf, $\sqrt{p}$ variables per split, Gini criterion, bootstrap enabled	depth 3–6, leaf 3–10
XGBoost	300 trees, depth 2, learning rate 0.05, subsampling 0.8, feature fraction 0.8, $\lambda = 3$, minimum child weight 3, logistic objective	depth 1–4, λ 1–10, rate 0.01–0.1
LightGBM	300 trees, 7 leaves, depth 3, learning rate 0.05, 10 observations minimum per leaf, subsampling 0.8, feature fraction 0.8, $\lambda = 3$	leaves 5–15, λ 1–10

Main seed 20260720, derived deterministically for each repetition. No class rebalancing, no early stopping. The probabilities reported are the raw model outputs, without recalibration. The random search of the inner loop comprised 30 draws, an internal AUC metric and an internal 5-fold validation likewise grouped by patient; the outer fold was used only once, for evaluation.

Table 25 Comparison of Episodes Excluded for Non-Evaluable GOS and Included Episodes

Variable	Excluded (n = 8)	Included (n = 142)	p
Age, years — median [IQR]	28 [16–47]	26 [14–51]	0.722
Glasgow Coma Scale — median [IQR]	12 [9–14]	13 [10–15]	0.671
Symptom-to-admission delay, days	18 [14–23]	19 [15–25]	0.377

Mann–Whitney tests. With eight observations, these comparisons are devoid of power: the absence of a detected difference does not demonstrate that the exclusion is non-informative. It is reported for transparency, not as validation of the missing-at-random assumption.

Table 26 Attempted Estimation of the Integrated Calibration Index: Documented Instability

Smoothing parameter (frac)	Index estimated for GCS alone
0.20	0.324
0.30	0.190
0.50	0.401
0.75	0.401

The integrated calibration index based on loess smoothing was computed at the request of a methodological review. On 142 episodes, the estimate varies by a factor of two according to the smoothing parameter alone, and the values obtained at high parameter settings correspond to smoothing towards the sample prevalence (0.401), that is an artefact and not a measure of calibration. These values are incompatible with a calibration slope of 1.27 and a Brier score of 0.110. We therefore do not report this index and confine ourselves to the calibration curve, slope and intercept, whose small-sample behaviour is better characterised.

APPENDIX S1. REPRODUCIBILITY

The replication archive accompanying this article contains the following items. All the primary metrics are recomputed from a single master file of out-of-sample predictions, which guarantees that Table 3, the ROC curves, calibration, the Brier score, the decision curve and the paired comparisons bear on exactly the same predictions. A control script automatically checks the sample sizes, prevalence, primary metrics and equation coefficients against the values reported in the manuscript.

Table 27 Reproducibility

File	Content
README.md	Step-by-step execution instructions and mapping between each script and each table or figure produced
dictionnaire_variables.csv	Definition, type, coding, unit and provenance of the 27 candidate variables and of the outcome
requirements.txt	Exact versions of all Python dependencies
environment.yml	Equivalent conda environment
Dockerfile	Container reproducing the complete execution environment
core.py	Data loading, pipeline definition, patient-grouped cross-validation, patient bootstrap
A_perf.py	Performance, calibration, paired comparisons, DeLong test
B_importance.py	Permutation importance (30 permutations) and SHAP values with bootstrap
C_parsimony.py	Parsimony analysis with nested selection and marginal gains
D_robustesse.py	Functional form, thresholds, hyperparameter sensitivity, additional analyses
infotheory.py	Mutual information and conditional mutual information
validation2.py	Bootstrap optimism, decision curve analysis, calibration tests
figs5.py, figs5b.py	Generation of all figures
E_subgroups.py	Prespecified subgroup analyses and risk by GCS value
F_final.py	Final model, coefficient uncertainty, mortality by logistic regression
audit.py	Checks of orientation, exact bounds and normalisation sums
oof_predictions.csv	Master file: identifiers, outcome, GCS and out-of-sample probabilities of the five models
test_reproducibility.py	Automated checks of the manuscript's figures

Main random seed: 20260720. All stochastic procedures are initialised deterministically; the full run reproduces the reported values identically.