Scientific Research Publishing

World Journal of Neuroscience, 2026, 16(4), 246-297
https://www.scirp.org/journal/wjns
ISSN Online: 2162-2019
ISSN Print: 2162-2000

TRIAGE-SIC: Development and Internal Validation of the First African Emergency Triage Score for Intracranial Suppurative Infections

Faozo Stephane Landry Teti[1,2*], Koffi Yves Soress Dongo[1,2], Yao Bernard Fionko[1,2], Konan Serge Yao[1,2], Kouadio Jean-Baptiste Keke[1,2], Aderehime Haidara[1,2]

[1]Department of Neurosurgery, Bouaké Teaching Hospital, Bouaké, Côte d'Ivoire
[2]Faculty of Medical Sciences, Alassane Ouattara University, Bouaké, Côte d'Ivoire
Email: *landjoss1990@gmail.com

How to cite this paper: Teti, F.S.L., Dongo, K.Y.S., Fionko, Y.B., Yao, K.S., Keke, K.J.-B. and Haidara, A. (2026) TRIAGE-SIC: Development and Internal Validation of the First African Emergency Triage Score for Intracranial Suppurative Infections. *World Journal of Neuroscience*, **16**, 246-297.
https://doi.org/10.4236/wjns.2026.164018

Received: June 21, 2026
Accepted: September 26, 2026
Published: September 29, 2026

Abstract

Background. In sub-Saharan Africa, intracranial suppurative infections present late and triage—who goes to theatre, who needs monitoring, who can be reassessed—is not formalised. No triage rule specific to these infections has been developed in an African population. **Objective.** To develop and internally validate an integer triage score using only variables available at presentation, and to define the targets for which prediction fails. **Methods.** Single-centre cohort, Bouaké Teaching Hospital (2016-2026; 150 admissions, 148 patients). Development endpoint: a composite of in-hospital death, poor functional outcome (Glasgow Outcome Scale ≤ 3) and critical-care use; early surgery was excluded. Twenty-four candidate variables, LASSO selection across ten multiply imputed datasets (60% stability), Firth regression, conversion to integer points, internal validation by 1000 bootstrap resamples of the whole procedure including selection. Pre-specified comparator: continuous Glasgow Coma Scale (DeLong test). **Results.** The composite occurred in 97 admissions (64.7%). Six variables form the score (0 - 14 points): scalp swelling (4), mass effect (3), impaired consciousness (2), immunosuppression (2), hydrocephalus (2), raised intracranial pressure (1). Optimism-corrected area under the curve 0.788 (0.726 - 0.852); observed risks by band 11.5%, 55.3%, 83.6% and 100%. The score outperformed the Glasgow Coma Scale for the composite ($p = 0.002$) and for critical-care use ($p < 0.001$). Pre-specified negative findings: death was not predictable (0.472), and the Glasgow Coma Scale remained superior for functional outcome ($p < 0.001$). **Conclusion.** Six items recorded at presentation stratify neurosurgical resource needs. The instrument is not an individual prognosis and requires African multicentre external validation.

Keywords

Brain Abscess, Côte D'Ivoire, Subdural Empyema, Decision Support Techniques, Triage

1. Introduction

1.1. Burden of Disease

Intracranial suppurative infections—brain abscess, subdural or extradural empyema, ventriculitis, scalp suppuration with intracranial extension—are among the few neurosurgical emergencies whose prognosis has been transformed by a single technological innovation. Since computed tomography came into general use, mortality has fallen from over 40% to under 10% in contemporary high-income series [1]-[4].

Disorders of the nervous system are, moreover, the leading global cause of disability-adjusted life years, with a burden concentrated in low- and middle-income countries [5]. This transformation has not been universal. In sub-Saharan Africa the disease remains common, is seen late, and carries higher case fatality and greater functional morbidity [6]-[10]. Its determinants are documented: prolonged time to consultation, unequal access to cross-sectional imaging, inconsistent microbiological documentation and limited critical-care capacity [6] [8] [11] [12].

The threat to life does not arise from the infection as such. It arises from three mechanical processes: mass effect from the collection and the surrounding oedema, hydrocephalus from obstruction of cerebrospinal fluid pathways, and intraventricular rupture [2] [13]-[15]. All three are identifiable on unenhanced computed tomography. The information that determines how urgent the pathway is, therefore, is precisely the information that remains available where technical resources are limited.

1.2. The Triage Problem in a Resource-Limited Setting

The Lancet Commission on Global Surgery estimated that five billion people lack access to safe, affordable and timely surgery [16]. Deaths avertable by surgery and anaesthesia now exceed the combined total attributable to tuberculosis, HIV and malaria [17]. Applied to neurosurgery, this analysis reveals an annual shortfall of several million essential operations, concentrated in sub-Saharan Africa and South Asia [18] [19]. Resolution WHA68.15 placed essential surgery within universal health coverage, and national surgical, obstetric and anaesthesia plans are its operational instrument [12] [20].

Within this framework, mortality attributable to poor-quality health systems exceeds that attributable to lack of access alone [21]. Intracranial suppurative infections illustrate this mechanism: the required procedure is technically simple, survival depends on how early it is performed, and the obstacle is organisational

rather than technical.

In practical terms, the neurosurgeon on call in a teaching hospital of inland Côte d'Ivoire is not trying to estimate a three-month probability of death. He has to solve a scheduling problem: should this patient occupy the only theatre available tonight? Should one of the few close-monitoring beds be kept for him? Can he be placed on a general ward on antibiotics, with reassessment at twelve hours? This decision is taken several times a week, often by a junior operator, and almost always without any formal support.

1.3. Limitations of Existing Instruments

Published scores do not answer this question. They were developed in populations where resources are not the binding constraint, which shifts the informative content of the variables; they predict endpoints—mortality, late Glasgow Outcome Scale—that do not coincide with the action to be decided; and they frequently include variables unavailable in an emergency in the settings concerned (magnetic resonance imaging, bacteriological documentation, specialised markers). Triage systems validated in Africa, such as the Cape Triage Score and the paediatric triage promoted by the World Health Organization, have demonstrated their value for in-hospital mortality but remain general-purpose and lack neurosurgical granularity [22]-[24]. To our knowledge, no clinical prediction rule for the triage of intracranial suppurative infections has been developed and internally validated in an African population.

1.4. Hypothesis and Objectives

Our hypothesis was that clinical and computed tomography variables available before any therapeutic decision are sufficient to identify, at presentation, the patients whose admission will consume heavy resources or will follow an unfavourable course.

The primary objective was to develop and internally validate an integer triage score, calculable without a calculator, from the variables available at presentation alone.

There were three secondary objectives: to evaluate this score simultaneously against four pre-specified triage targets—in-hospital death, poor functional outcome, surgery within the first 24 hours and use of critical-care therapies; to compare it with a pre-specified comparator, the Glasgow Coma Scale treated as a continuous variable; and to delimit its domain of validity explicitly by reporting in full the targets for which prediction fails.

2. Patients and Methods

2.1. Study Design and Reporting

This is a development and internal validation study of a clinical prediction rule (TRIPOD type 1b: development with internal validation by resampling, without an external dataset) [25] [26]. Reporting follows the TRIPOD checklist together

with the TRIPOD + AI extension for the penalised learning components [27] (Supplementary material S2 and S3); an abridged compliance summary is given in Supplementary **Table S7-4**. The statistical protocol, including the targets, the candidate pool, the stability threshold, the point-conversion unit and the two pre-specified negative analyses, was fixed before any performance figure was examined and is provided in full (S1).

2.2. Population and Data Collection

The cohort comprises all admissions for imaging-documented intracranial suppurative infection managed in the neurosurgery department of Bouaké Teaching Hospital between January 2016 and January 2026. The analytical database holds 150 admissions corresponding to 148 patients; two patients had two distinct episodes, kept as separate admissions, a choice discussed in the limitations section. One hundred and twenty-seven admissions come from a retrospective case-note review (2016-2024) and 23 from structured prospective collection (2025-2026). The flow diagram is shown in Figure 1.

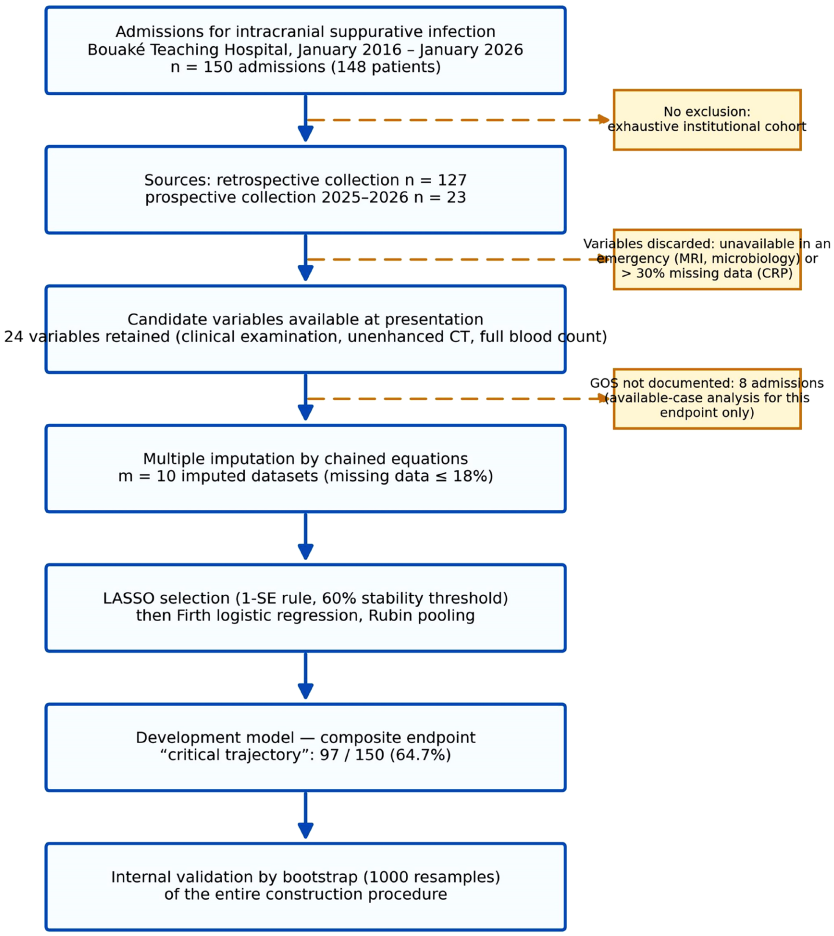

Figure 1. Study flow diagram (TRIPOD format). Admissions included, sources of data collection, candidate variables retained and discarded, multiple imputation, selection and internal validation.

Patients of any age with a brain abscess, a subdural or extradural empyema, a ventriculitis or a scalp suppuration with intracranial repercussion, documented by computed tomography or magnetic resonance imaging, were included. No exclusion criterion was applied a priori, in order to preserve the representativeness of the real patient flow—a necessary condition for the validity of a triage instrument, which must apply to every patient who presents and not to a selected subgroup.

2.3. Triage Targets (Endpoints)

Four targets were pre-specified, each corresponding to a distinct pathway decision:

1) **In-hospital death**, recorded at the end of the admission;

2) **Poor functional outcome**, defined by a Glasgow Outcome Scale score ≤ 3 at the last documented assessment of the index admission, in practice at discharge [28];

3) **Early surgery**, defined by an evacuation procedure performed within 24 hours of admission (time to surgery ≤ 1 day);

4) **Critical-care use**, defined by the administration, at any time during the index admission, of osmotherapy and/or the insertion of an external ventricular drain.

The fourth definition calls for an explicit methodological comment. The database holds no "admission to intensive care" variable, the department having had no dedicated neurosurgical intensive care unit over the whole period. We therefore took as a proxy the use of therapies that, in this setting, mandate critical-care-level monitoring. The proxy is imperfect: it captures therapeutic intensity rather than an administrative destination, and it is liable to be influenced by prescribing habits. It is reported as such, and every result relating to it must be read with that reservation. None of the three components of the composite was time-anchored to surgery: death, poor functional outcome and critical-care use were ascertained over the whole index admission, whether they arose before or after any operative procedure, whereas early surgery—the only time-defined target—was a procedure performed within 24 hours of admission.

2.4. Rationale for the Composite Development Endpoint

The primary development endpoint is a pragmatic composite, the critical trajectory, defined by the occurrence of at least one of the following: death, Glasgow Outcome Scale ≤ 3 or critical-care use. Four arguments support this choice; they are set out in full, with the corresponding methodological references, in the prespecified statistical protocol (S1).

First, the composite matches the decision it informs. An endpoint should reflect the action it is meant to guide [29] [30]. At the moment of triage, the clinician wants to know whether the patient belongs to the category for which a priority pathway is justified. From the standpoint of scheduling, the three components are manifestations of a single underlying state. A composite endpoint is legitimate

when its components share a common causal mechanism and a common decisional implication, a condition met here [31].

Second, each target taken alone is unsuitable for development. Mortality yielded only 16 events, far below recommended thresholds whichever rule is applied [32]-[34]. The Glasgow Outcome Scale is almost entirely explained by the Glasgow Coma Scale, which would have produced an instrument redundant with a variable already measured universally. Critical-care use is a process variable, sensitive to the prescribing habits of the department.

Third, early surgery was deliberately excluded from the composite. It is a therapeutic decision and not an event undergone by the patient. Including it would have introduced circularity between the predictor and the predicted action. It is retained as a secondary evaluation target, in order to document this limitation.

Fourth, the decomposition is reported in full. The main criticism levelled at composite endpoints is that they mask the heterogeneity of their components [29] [31]. Performance of the score is therefore reported for each component separately, including where it is nil.

2.5. Candidate Predictors

Twenty-four candidates were pre-specified under a strict availability rule: only variables obtainable at presentation without magnetic resonance imaging, without microbiology and without specialised assays were retained—that is, from the history, the neurological examination, an unenhanced brain computed tomography scan and a full blood count. These candidates covered four domains: patient background (age ≥ 65 years, age < 5 years, immunosuppression, postoperative origin, referred patient, rural residence, symptom-to-admission interval ≥ 21 days); neurology (Glasgow Coma Scale ≤ 8, Glasgow Coma Scale 9 - 12, impaired consciousness, raised intracranial pressure syndrome, meningeal syndrome, seizures, motor deficit, anisocoria, scalp swelling/cellulitis); imaging (subdural empyema, ventriculitis, multiple collections, maximum size ≥ 30 mm, mass effect, midline shift ≥ 5 mm, hydrocephalus); and laboratory tests (leucocytosis). Immunosuppression was defined operationally as human immunodeficiency virus infection, prolonged corticosteroid therapy, active malignancy or any other cause of immunosuppression documented in the notes; diabetes and sickle cell disease did not fall within this definition. The three items whose bedside assessment is least standardised were, in addition, defined operationally, and these definitions were reproduced verbatim on the score card and in the electronic calculator (S10): impaired consciousness denoted any impairment of the level of consciousness recorded at presentation, whatever its depth; raised intracranial pressure syndrome, the presence of headache, vomiting, papilloedema or a combination of these signs; and mass effect on unenhanced computed tomography, the displacement of adjacent structures, the effacement of the cortical sulci or compression of the ventricular system. Scalp swelling or cellulitis denoted inflammatory involvement of the soft

tissues of the scalp overlying the suppuration, and hydrocephalus on computed tomography an active ventricular dilatation. The full dictionary of definitions and coding rules is given in S4.

C-reactive protein was excluded from the candidate pool despite its apparent clinical relevance: it was missing in 38% of admissions and is not routinely available as an emergency test in a substantial proportion of the settings for which this score is intended. A triage score with an item measurable only once in every two patients is not a triage score.

Dichotomising continuous variables (Glasgow Coma Scale, size, midline shift) is methodologically unfavourable and entails a recognised loss of information [35]. It was nevertheless adopted in a pre-specified manner for the construction of the point scale, the aim being an instrument calculable mentally in the admissions room. The cost of this choice is measured explicitly in the study by the comparison with the Glasgow Coma Scale treated as a continuous variable, and it constitutes one of the two negative findings reported.

2.6. Missing Data

Missing data concerned mainly midline shift (18.0%), white cell count (8.0%) and maximum size of the collection (4.0%). Multiple imputation by chained equations was performed (ten imputed datasets, posterior draw, twenty iterations) [36]-[38], the derived binary variables being recomputed after imputation and then dichotomised. All selection and estimation steps were conducted across the ten datasets, with pooling according to Rubin's rules [36]. A complete-case sensitivity analysis is provided in S6.

2.7. Variable Selection and Estimation

Selection relied on L1-penalised logistic regression (LASSO) [39], the penalty parameter being chosen by five-fold cross-validation under the one-standard-error rule, which deliberately favours parsimony over apparent performance. The procedure was repeated on each of the ten imputed datasets, and only variables selected in at least 60% of the datasets were retained—a stability criterion inspired by stability selection approaches [40], intended to prevent a variable from owing its presence in the score to a particular imputation draw.

The final model was estimated by Firth penalised logistic regression [41] [42], chosen because quasi-complete separation was expected on rare variables (scalp swelling was associated with a critical trajectory in all 12 patients concerned, a configuration that makes ordinary maximum likelihood unusable). Coefficients and standard errors were pooled according to Rubin's rules.

2.8. Construction of the Integer Score and Full Model Equation

The Firth coefficients were divided according to a pre-specified conversion rule—division by the smallest retained coefficient, whose estimated value here is 0.642—then rounded to the nearest integer, with a cap of 4 points per item. This cap,

decided before any performance figure was examined, limits the influence of estimates arising from quasi-separation, whose confidence interval is by construction very wide. The total score ranges from 0 to 14 points.

The score-to-probability mapping was obtained by a Firth regression of the composite endpoint on the score. The model is therefore fully specified by the following three elements, reproduced here so that it can be implemented without recourse to the authors (full detail in S5):

1) Score calculation.

Score = 4 × (scalp swelling or cellulitis) + 3 × (mass effect on computed tomography) + 2 × (impaired consciousness) + 2 × (immunosuppression) + 2 × (hydrocephalus) + 1 × (raised intracranial pressure syndrome)

each item scoring 1 if present and 0 otherwise.

2) Transformation into a probability.

$$\text{logit}(p) = -2.213 + 0.681 \times \textit{Score}$$

that is

$$p = 1/[1 + \text{e}\text{\^{}}-(-2.213 + 0.681 \times \textit{Score})]$$

3) Underlying multivariable model (provided to allow recalibration or reweighting at external validation, and not for routine calculation):

$\text{logit}(p) = -2.096 + 3.270 \times \text{scalp} + 1.795 \times \text{mass effect} + 1.463 \times \text{consciousness} + 1.399 \times \text{immunosuppression} + 1.279 \times \text{hydrocephalus} + 0.642 \times \text{raised intracranial pressure}$

Step-by-step worked examples are given in Supplementary **Table S7-1**.

2.9. Internal Validation

Internal validation followed Harrell's method [43] [44]: 1000 bootstrap resamples with replacement, each replaying the entire construction procedure—fixed imputation, LASSO selection under the one-standard-error rule, Firth estimation, conversion to points, score-to-probability recalibration—the score thus rebuilt being then applied to the original sample. Imputation was carried out once, before the bootstrap, and the ten imputed datasets were then held fixed within the resampling loop rather than regenerated inside each resample; because re-imputing within every resample would propagate imputation uncertainty more completely, holding the imputations fixed is a pragmatic simplification that may render the optimism-corrected intervals slightly narrower than a fully nested imputation-in-bootstrap procedure would yield. Given the low overall fraction of missing data—at most 18.0% for a single variable and 8.0% or less for the others—this effect is expected to be small; it is recorded among the limitations. Optimism was estimated as the mean difference between performance in the resample and performance in the original sample, and subtracted from apparent performance. This approach is more conservative than cross-validation applied after selection. It is also the only one that accounts for the optimism induced by selection itself, in line with recent recommendations on the evaluation of prediction models [45] [46].

Corrected performance measures comprised, in line with current recommen-

dations for prognostic models [47]-[49]:

- **Discrimination**, by the area under the ROC curve;
- **Weak calibration**, by the calibration slope and calibration-in-the-large (intercept estimated with the slope fixed at 1);
- **Moderate calibration**, by a smoothed calibration curve obtained by locally weighted logistic regression (tricube kernel), together with the E_mean, E_90 and E_max indices, defined as the mean, the 90th centile and the maximum of the absolute difference between predicted probability and smoothed observed probability;
- **Overall performance**, by the Brier score and the scaled Brier score (1 − Brier/Brier_max).

Calibration was also tested formally by the Spiegelhalter test and by the Hosmer-Lemeshow test with ten groups, and displayed by deciles and quintiles of predicted risk.

The full bootstrap distribution of the corrected area under the curve and the optimism-corrected calibration curve, with its 95% confidence envelope, are reported in the Results and in Supplementary.

2.10. Pre-Specified Comparator and Comparison of ROC Curves

The Glasgow Coma Scale treated as a continuous variable was pre-specified as the comparator for each target [50] [51]. This choice is deliberately demanding: it poses the question that any new score must answer—does it add anything to the clinical variable the clinician already measures? The areas under the curve of the score and of the comparator were compared for each target by the DeLong test for correlated ROC curves [52], with estimation of the difference in area under the curve and its 95% confidence interval.

2.11. Assessment of Clinical Utility

Clinical utility was assessed by decision curve analysis [53] [54], comparing the net benefit of the score, over a range of probability thresholds from 0.05 to 0.95, with the two reference strategies: directing every patient to the priority pathway, or directing none. Net benefit was converted into a directly interpretable measure, the net reduction in unnecessary priority referrals per 100 patients, obtained by dividing the difference in net benefit by the threshold odds. The range of clinically plausible thresholds (0.15 - 0.60) was discussed a priori with the clinicians of the department and corresponds to the range within which a neurosurgeon on call would accept to commit a scarce resource [55]. A clinical impact curve completes this analysis: for 1000 admissions and at each threshold, it displays the number of patients the score would classify as high risk and, among them, the number who would actually experience the event.

2.12. Pre-Specified Sensitivity Analyses

Three sensitivity analyses were pre-specified: restriction to complete cases; re-

striction to the first admission of each patient; and exclusion of patients operated on within the first 24 hours. This last analysis tests the robustness of the score against the residual circularity linked to the surgical decision: if performance is maintained in the subgroup not operated on immediately, the score does not reduce to a reproduction of the operative indication.

2.13. Patient and Public Involvement

Patients and the public were involved neither in the design, nor in the conduct, nor in the choice of endpoints, nor in the reporting of this research, nor in its dissemination. This absence of involvement is an acknowledged limitation. A prospective impact study, such as the one proposed in the discussion, will however have to involve patients, families and admissions staff in defining the endpoints and the decision thresholds, since the relative value placed on an unnecessary priority referral and on an unanticipated critical trajectory is a trade-off that is not purely clinical.

2.14. Software

Analyses were conducted in Python 3.12 (numpy, pandas, scikit-learn [56], scipy), with an in-house implementation of Firth regression by Newton-Raphson iterations with the score modified by the Jeffreys term, and an in-house implementation of the DeLong test by the method of structural components. The full code, allowing exact reproduction of every figure reported, is provided as supplementary material (S9).

3. Results

3.1. Characteristics of the Population

The flow of admissions is shown in **Figure 1** and the characteristics of the 150 admissions in **Table 1**. Median age was 26 years (IQR 14 - 50), with a male predominance (69.3%) and an almost equal distribution between rural and urban residence. One patient in two had been referred from another facility.

Table 1. Characteristics of the 150 admissions at presentation, on imaging and during management.

Characteristic	Value
Patient background	
Age, years—median (IQR)	26 (14 - 50)
Male sex—n (%)	104 (69.3)
Children < 15 years—n (%)	40 (26.7)
Rural residence—n (%)	78 (52.0)
Referred from another facility—n (%)	77 (51.3)

Continued

Symptom-to-admission interval, days—median (IQR)	19 (15 - 25)
Documented immunosuppression—n (%)	13 (8.7)
Community-acquired/post-traumatic/postoperative origin—n (%)	90 (60.0)/33 (22.0)/27 (18.0)
Clinical findings at presentation	
GCS—median (IQR)	13 (10 - 15)
GCS ≤ 8—n (%)	19 (12.7)
GCS 9 - 12—n (%)	47 (31.3)
Raised intracranial pressure syndrome—n (%)	107 (71.3)
Impaired consciousness—n (%)	73 (48.7)
Motor deficit—n (%)	49 (32.7)
Seizures—n (%)	27 (18.0)
Anisocoria—n (%)	8 (5.3)
Scalp swelling/cellulitis—n (%)	12 (8.0)
Imaging	
CT only (no MRI)—n (%)	101 (67.3)
Brain abscess (pure form)—n (%)	72 (48.0)
Extradural/subdural empyema (pure forms)—n (%)	20 (13.3)/17 (11.3)
Multiple collections—n (%)	57 (38.0)
Maximum size, mm—median (IQR)	36 (30 - 44)
Mass effect—n (%)	110 (73.3)
Midline shift, mm—median (IQR)	2 (0 - 4)
Hydrocephalus—n (%)	16 (10.7)
Laboratory tests and microbiology	
White cell count, $\times 10^9$/L—median (IQR)	15 (13 - 17)
Pus sample obtained—n (%)	116 (77.3)
Positive culture among samples obtained—n (%)	50 (43.1)
Management	
Combined medical and surgical strategy—n (%)	125 (83.3)
Admission-to-surgery interval, days—median (IQR)	2 (1 - 3)
Length of stay, days—median (IQR)	11 (8 - 15)
Triage targets	
Critical trajectory (composite)—n (%)	97 (64.7)
In-hospital death—n (%)	16 (10.7)

Continued

Poor functional outcome (GOS ≤ 3)—n/N (%)	57/142 (40.1)
Surgery ≤ 24 h—n (%)	41 (27.3)
Critical-care use—n (%)	65 (43.3)

IQR: interquartile range; GCS: Glasgow Coma Scale; GOS: Glasgow Outcome Scale; CT: computed tomography; MRI: magnetic resonance imaging. Missing data: midline shift 27/150 (18.0%), white cell count 12/150 (8.0%), maximum size of the collection 6/150 (4.0%), GOS 8/150 (5.3%); all other candidate predictors were complete. Candidate prevalences not listed above were: age ≥ 65 years 11/150, age < 5 years 8/150, symptom-to-admission interval ≥ 21 days 75/150, meningeal syndrome 30/150, ventriculitis 18/150, maximum size ≥ 30 mm 114/150, midline shift ≥ 5 mm 17/150 and leucocytosis 138/150 (TRIPOD item 13c; full dictionary in S4). Documented immunosuppression: human immunodeficiency virus infection, prolonged corticosteroid therapy, active malignancy or any other cause of immunosuppression documented in the notes; diabetes and sickle cell disease did not fall within this definition. So-called pure forms exclude associated locations; the candidate variable "subdural empyema" used in Table 2 and Table 3 denotes any subdural involvement, isolated or associated (27/150).

The median interval between symptom onset and admission reached 19 days (IQR 15 - 25). Median GCS at presentation was 13 (IQR 10 - 15), with 12.7% of patients at GCS ≤ 8. The six items of the future score were present in 71.3% (raised intracranial pressure), 48.7% (impaired consciousness), 73.3% (mass effect), 10.7% (hydrocephalus), 8.7% (immunosuppression) and 8.0% (scalp swelling) of admissions.

Imaging rested on computed tomography alone in 67.3% of cases. The median maximum size of the collections was 36 mm (IQR 30 - 44). Microbiological documentation was attempted in 77.3% of patients, with a positive culture in 43.1% of the samples obtained; sterile cultures accounted for 44.0% of the cohort. A combined medical and surgical strategy was adopted in 83.3% of patients, with a median time to surgery of 2 days (IQR 1 - 3).

The characteristics of admissions with and without a critical trajectory are compared in Supplementary **Table S7-5**. The two groups differed neither in age, nor in sex, nor in place of residence, nor in the symptom-to-admission interval. Differences concerned exclusively the neurological and radiological variables recorded at presentation, which supports the choice of the candidate pool.

3.2. Frequency of the Triage Targets

A critical trajectory was observed in 97 admissions (64.7%). The individual targets were: in-hospital death, 16 (10.7%); poor functional outcome (GOS ≤ 3), 57 of 142 assessable (40.1%); surgery within 24 hours, 41 (27.3%); critical-care use, 65 (43.3%). The three components of the composite overlapped substantially: of the 97 critical trajectories, 2 were defined by death alone, 24 by poor functional outcome alone and 35 by critical-care use alone, whereas 6 combined death and poor functional outcome, 3 death and critical-care use, 22 poor functional outcome and critical-care use, and 5 all three components; death and poor functional outcome

co-occurred in 11 admissions. The number of events per candidate variable was therefore favourable for the composite (97 events for 6 final parameters, about 16 events per variable) but plainly insufficient for death taken alone (16 events), which is a major determinant of the results set out below [32] [34].

3.3. Univariable Associations

Univariable associations with the critical trajectory are given in **Table 2**. The strongest associations concerned mass effect (79.1% versus 25.0%; OR 10.82 (95% CI 4.69 - 24.96); $p < 0.001$), raised intracranial pressure syndrome (75.7% versus 37.2%; OR 5.13 (2.42 - 10.87); $p < 0.001$), impaired consciousness (82.2% versus 48.1%; OR 4.84 (2.31 - 10.13); $p < 0.001$), size ≥ 30 mm (OR 3.52 (1.63 - 7.60); $p = 0.001$), GCS ≤ 8 (OR 8.14 (1.49 - 44.59); $p = 0.004$) and midline shift ≥ 5 mm (OR 7.09 (1.29 - 39.06); $p = 0.006$). Scalp swelling showed a perfect association (12/12 patients concerned). Postoperative origin was associated with a lower risk (44.4% versus 69.1%; OR 0.36 (0.16 - 0.84); $p = 0.025$), a finding consistent with earlier diagnosis in patients already admitted and under follow-up—a pattern also described in series of postoperative neurosurgical infections [57] [58].

Table 2. Univariable associations with the critical trajectory (main candidates).

Predictor	n exposed/N	Events among exposed—n (%)	Events among unexposed—n (%)	OR (95% CI)	p
Mass effect	110/150	87 (79.1)	10 (25.0)	10.82 (4.69 - 24.96)	<0.001
Raised intracranial pressure syndrome	107/150	81 (75.7)	16 (37.2)	5.13 (2.42 - 10.87)	<0.001
Impaired consciousness	73/150	60 (82.2)	37 (48.1)	4.84 (2.31 - 10.13)	<0.001
Size ≥ 30 mm	114/150	82 (71.9)	15 (41.7)	3.52 (1.63 - 7.60)	0.001
GCS ≤ 8	19/150	18 (94.7)	79 (60.3)	8.14 (1.49 - 44.59)	0.004
Midline shift ≥ 5 mm	17/150	16 (94.1)	81 (60.9)	7.09 (1.29 - 39.06)	0.006
Scalp swelling/cellulitis	12/150	12 (100.0)	85 (61.6)	15.64 (0.91 - 269.72)	0.009
Hydrocephalus	16/150	15 (93.8)	82 (61.2)	6.58 (1.19 - 36.41)	0.011
GCS 9 - 12	47/150	37 (78.7)	60 (58.3)	2.57 (1.17 - 5.64)	0.017
Postoperative origin	27/150	12 (44.4)	85 (69.1)	0.36 (0.16 - 0.84)	0.025

Continued

Subdural empyema	27/150	22 (81.5)	75 (61.0)	2.63 (0.97 - 7.14)	0.048
Immunosuppression	13/150	11 (84.6)	86 (62.8)	2.74 (0.67 - 11.22)	0.140
Anisocoria	8/150	7 (87.5)	90 (63.4)	2.90 (0.49 - 17.30)	0.261

OR: odds ratio estimated with a continuity correction (0.5); p: Fisher's exact test. Univariable associations were estimated on the ten imputed datasets, so denominators are out of 150 for every candidate, including midline shift and maximum size. Continuous variables (Mann-Whitney test): median GCS 12 (IQR 9 - 14) in patients with a critical trajectory versus 14 (13 - 15) without ($p < 0.001$); maximum size 39.5 mm (33.0 - 47.2) versus 31.0 mm (23.0 - 36.2) ($p < 0.001$); midline shift 3.1 mm (1.3 - 4.7) versus 0.0 mm (0.0 - 1.1) ($p < 0.001$).

3.4. Variable Selection

LASSO selection repeated across the ten imputed datasets retained six variables above the 60% stability threshold (Table 3, Supplementary Figure S8-1): impaired consciousness, raised intracranial pressure syndrome, mass effect and scalp swelling (stability 100%), immunosuppression and hydrocephalus (80%). No other variable exceeded 20% stability across imputed datasets.

Table 3. Selection stability of the predictors (Endpoint: Critical trajectory).

Predictor	LASSO selection across 10 imputed datasets	Selection across 1000 bootstrap resamples	Retained (60% threshold)
Mass effect	100%	99.6%	yes
Impaired consciousness	100%	94.6%	yes
Scalp swelling/cellulitis	100%	87.6%	yes
Raised intracranial pressure syndrome	100%	75.1%	yes
Hydrocephalus	80%	75.0%	yes
Immunosuppression	80%	64.0%	yes
Midline shift ≥ 5 mm	10%	54.0%	no
GCS ≤ 8	0%	50.2%	no
Referred patient	0%	47.6%	no
Rural residence	0%	46.0%	no
Multiple collections	0%	43.5%	no
Subdural empyema	0%	40.5%	no
Age < 5 years	0%	39.7%	no
Meningeal syndrome	0%	38.0%	no
Other candidates (10)	≤20%	≤37.5%	no

The median number of variables selected per bootstrap resample was 11 (out of 24 candidates), the final selection resting on the stability threshold and not on penalisation alone.

The selection frequency across the 1000 bootstrap resamples, reported here for the first time, confirms this hierarchy while showing the true uncertainty of the procedure: mass effect 99.6%, impaired consciousness 94.6%, scalp swelling 87.6%, raised intracranial pressure syndrome 75.1%, hydrocephalus 75.0%, immunosuppression 64.0%. The six items retained are therefore also the six variables most frequently selected under resampling, with a ten-point margin over the seventh candidate (midline shift ≥ 5 mm, 54.0%). The median number of variables selected per resample was 11, which indicates that penalisation alone would have produced a markedly less parsimonious model and underlines the role of the stability threshold.

Notably, the dichotomised GCS was not retained: in this dataset, clinically recorded impaired consciousness captures most of the information carried by the GCS for this composite, while being more robust to variation in scoring between operators—inter-observer variability being well documented for the GCS, including among trained assessors [50].

3.5. Final Model, Point Scale and Equation

The pooled Firth model and the corresponding point scale are given in **Table 4**, which reports for each item, on a single line, the coefficient β, its pooled standard error, the odds ratio and its confidence interval, the p value and the points assigned. Adjusted odds ratios, with profile penalised-likelihood confidence intervals and penalised likelihood-ratio test p values, were 26.32 (95% CI 1.6 to >1000; p = 0.010) for scalp swelling, 6.02 (2.1 - 18.0; p < 0.001) for mass effect, 4.32 (1.8 - 10.5; p < 0.001) for impaired consciousness, 4.05 (0.7 - 25.0; p = 0.12) for immunosuppression, 3.59 (0.6 - 23.0; p = 0.17) for hydrocephalus and 1.90 (0.7 - 5.3; p = 0.21) for raised intracranial pressure syndrome. Once the quasi-complete separation is handled by penalised likelihood, the confidence interval for scalp swelling no longer includes 1 and this item reaches significance.

Table 4. Final firth model (Rubin pooling) and point scale of the TRIAGE-SIC score.

Variable	β	Standard error	Adjusted OR	95% CI	p	Points
Intercept	−2.096	0.513	-	-	<0.001	-
Scalp swelling/cellulitis	3.270	1.818	26.32	1.6 to >1000	0.010	4
Mass effect (CT)	1.795	0.547	6.02	2.1 - 18.0	<0.001	3
Impaired consciousness	1.463	0.437	4.32	1.8 - 10.5	<0.001	2
Immunosuppression	1.399	0.917	4.05	0.7 - 25.0	0.12	2
Hydrocephalus	1.279	0.960	3.59	0.6 - 23.0	0.17	2
Raised intracranial pressure syndrome	0.642	0.517	1.90	0.7 - 5.3	0.21	1
Total						**0 - 14**

Confidence intervals are profile penalised-likelihood intervals and p values are from the penalised likelihood-ratio test (Firth method [26]); under the quasi-complete separation of scalp swelling (12/12) the Wald statistic is invalid and was therefore not used. Points = β/0.642, rounded to the nearest integer, capped at 4. Standard error pooled according to Rubin's rules (within variance + between variance × [1 + 1/m]). Score-to-probability mapping: logit(p) = −2.213 + 0.681 × score. Full equation in Supplementary **Table S7-1**.

Three of the six items—immunosuppression, hydrocephalus and raised intracranial pressure syndrome—are not significant taken in isolation on the basis of the penalised likelihood-ratio test; they were retained because selection rested on the stability of the penalised contribution and not on a significance threshold—selection by p value being explicitly discouraged in the development of prognostic models [43] [44]—and because removing them degraded calibration in the upper bands.

The resulting TRIAGE-SIC point scale (0 - 14 points) assigns 4 points to scalp swelling/cellulitis, 3 points to mass effect, 2 points to impaired consciousness, to immunosuppression and to hydrocephalus, and 1 point to raised intracranial pressure syndrome. The corresponding nomogram is presented in **Figure 2**, the paper version of the score in **Table 5**, and the full model equation in Supplementary **Table S7-1**. The distribution of the score in the cohort is shown in Supplementary **Figure S8-2**: it covers the whole 0 - 14 range, with a main mode at 6 points (n = 44), a secondary mode at 4 points (n = 32) and a distinct group of 20 patients at 0 points. The extreme values are sparsely populated (12 points, n = 2; 14 points, n = 1), which justifies not interpreting predicted probabilities individually beyond 10 points.

Table 5. TRIAGE-SIC score—paper version (for display in the admissions room).

	Item (tick if present)	Points
☐	Scalp swelling or cellulitis	4
☐	Mass effect on computed tomography	3
☐	Impaired consciousness	2
☐	Known immunosuppression: human immunodeficiency virus infection, prolonged corticosteroid therapy, active malignancy, or other documented immunosuppression	2
☐	Hydrocephalus on computed tomography	2
☐	Raised intracranial pressure syndrome	1
	TOTAL	**/14**

Total	Band (colour code)	Estimated probability	Observed critical trajectory	Observed critical-care use	Suggested management
0 - 1	Low (green)	10% - 18%	11.5%	0%	Antibiotics, planned reassessment
2 - 4	Intermediate (yellow)	30% - 63%	55.3%	40.4%	Clinical decision; re-score at 12 h
5 - 6	High (orange)	77% - 87%	83.6%	49.1%	Anticipate theatre and close monitoring

Continued

≥7	Very high (red)	≥93%	100%	86.4%	Immediate priority pathway

This score predicts neither death nor individual functional outcome. For functional prognosis, use the Glasgow Coma Scale. The estimated probabilities are the apparent probabilities of the derivation cohort; as the optimism-corrected calibration slope is 0.640, they are slightly too widely dispersed at both ends of the scale and call for local recalibration before any numerical use. **Operational definitions of the items.** Scalp swelling or cellulitis: inflammatory involvement of the soft tissues of the scalp overlying the suppuration. Mass effect: displacement of adjacent structures, effacement of the cortical sulci or ventricular compression on unenhanced computed tomography. Impaired consciousness: any impairment of the level of consciousness at presentation, whatever its depth. Hydrocephalus: active ventricular dilatation on computed tomography. Raised intracranial pressure syndrome: headache, vomiting, papilloedema or a combination of these signs.

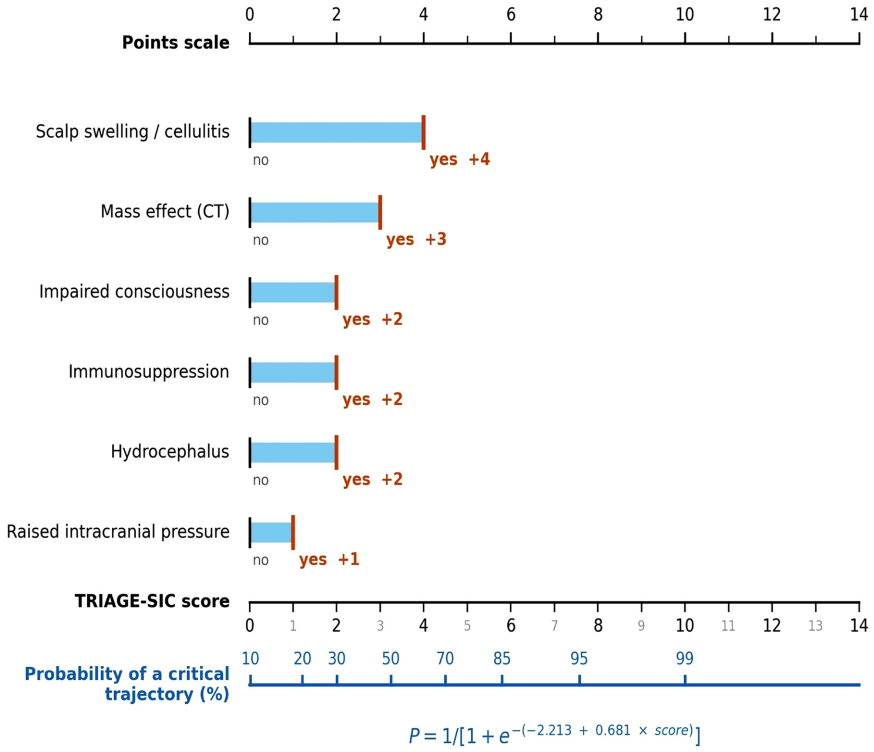

Figure 2. TRIAGE-SIC nomogram. Points assigned to each item, points scale, total score and corresponding probability of a critical trajectory. The associated logistic equation is restated below the probability axis: p = 1/[1 + e^−(−2.213 + 0.681 × score)].

3.6. Performance and Internal Validation

For the critical trajectory, the apparent area under the curve was 0.858. The bootstrap encompassing selection estimated an optimism of 0.070, giving a corrected area under the curve of 0.788 (95% CI 0.726 - 0.852). The full bootstrap distribution of the corrected area under the curve is shown in Supplementary: it is approximately symmetrical and centred on 0.788; the apparent value of 0.858 lies beyond the upper bound of the confidence interval of the corrected area under the

curve, which gives the measure of the optimism that a naive validation would have let through.

The calibration indices, reported in full in Table 6 and in Supplementary **Table S7-6**, were as follows. Corrected calibration-in-the-large was 0.011 (95% CI −0.548 to 0.604). The corrected calibration slope was 0.640 (0.367 - 0.977). The corrected Brier score was 0.185 for a maximum Brier score of 0.229. The corrected scaled Brier score was 0.196 (95% CI −0.010 to 0.372); it was corrected for optimism as a quantity in its own right and is therefore not exactly the transformation of the two preceding values. This interval contains zero: at this sample size, and unlike discrimination and net benefit, overall performance is not distinguishable from that of an uninformative strategy. The corrected moderate calibration indices were E_mean 0.058, E_90 0.099 and E_max 0.131.

Table 6. Internal validation by bootstrap (1000 resamples of the entire procedure)—endpoint: critical trajectory.

Performance index	Apparent value	Optimism	Corrected value	95% CI of the corrected value
Discrimination—AUC	0.858	0.070	**0.788**	0.726 - 0.852
Calibration—slope	1.000	0.360	**0.640**	0.367 - 0.977
Calibration-in-the-large (intercept)	0.004	−0.007	**0.011**	−0.548 - 0.604
Brier score	0.140	−0.044	**0.185**	0.143 - 0.231
Scaled Brier score	0.385	0.190	**0.196**	−0.010 - 0.372
E_mean	0.020	−0.038	**0.058**	0.007 - 0.115
E_90	0.025	−0.073	**0.099**	0.000 - 0.221
E_max	0.050	−0.081	**0.131**	0.000 - 0.323

Maximum Brier score (Brier score of the uninformative strategy): 0.229. Scaled Brier score = 1 − Brier/Brier_max. E_mean, E_90 and E_max: mean, 90th centile and maximum of the absolute difference between predicted probability and observed probability smoothed by local logistic regression. Optimism is the mean difference between performance in the resample and performance of the resampled model applied to the original cohort; for indices where a lower value is better (Brier, E), a negative optimism means that the corrected value is higher than the apparent value. The apparent calibration slope equals 1.000 by construction, the score-to-probability relation being estimated on the very data used to evaluate it; it therefore carries no information, and the corrected value of 0.640 is determined entirely by the estimation of optimism. The lower bounds of the intervals for E_90 and E_max are truncated at 0.000: the corresponding bootstrap percentiles were negative, which is impossible for absolute differences.

The formal calibration tests applied to the apparent model showed no significant departure from perfect calibration: Spiegelhalter test $z = -0.094$ ($p = 0.925$), Hosmer-Lemeshow test $C = 5.13$ on 8 degrees of freedom ($p = 0.744$). The latter must be read with caution: the score takes only fifteen distinct values and twenty admissions share a score of 0, so that the deciles separate patients with identical predicted probabilities; it is reported as secondary to the calibration curve. Calibration by deciles of predicted risk is shown in Figure 3(b) and detailed in Supplementary **Table S7-6**: the observed proportion of events rose from 20.0% in the first decile (mean predicted probability 0.099) to 100% in the tenth (0.977), and no decile departed from the diagonal beyond its confidence interval.

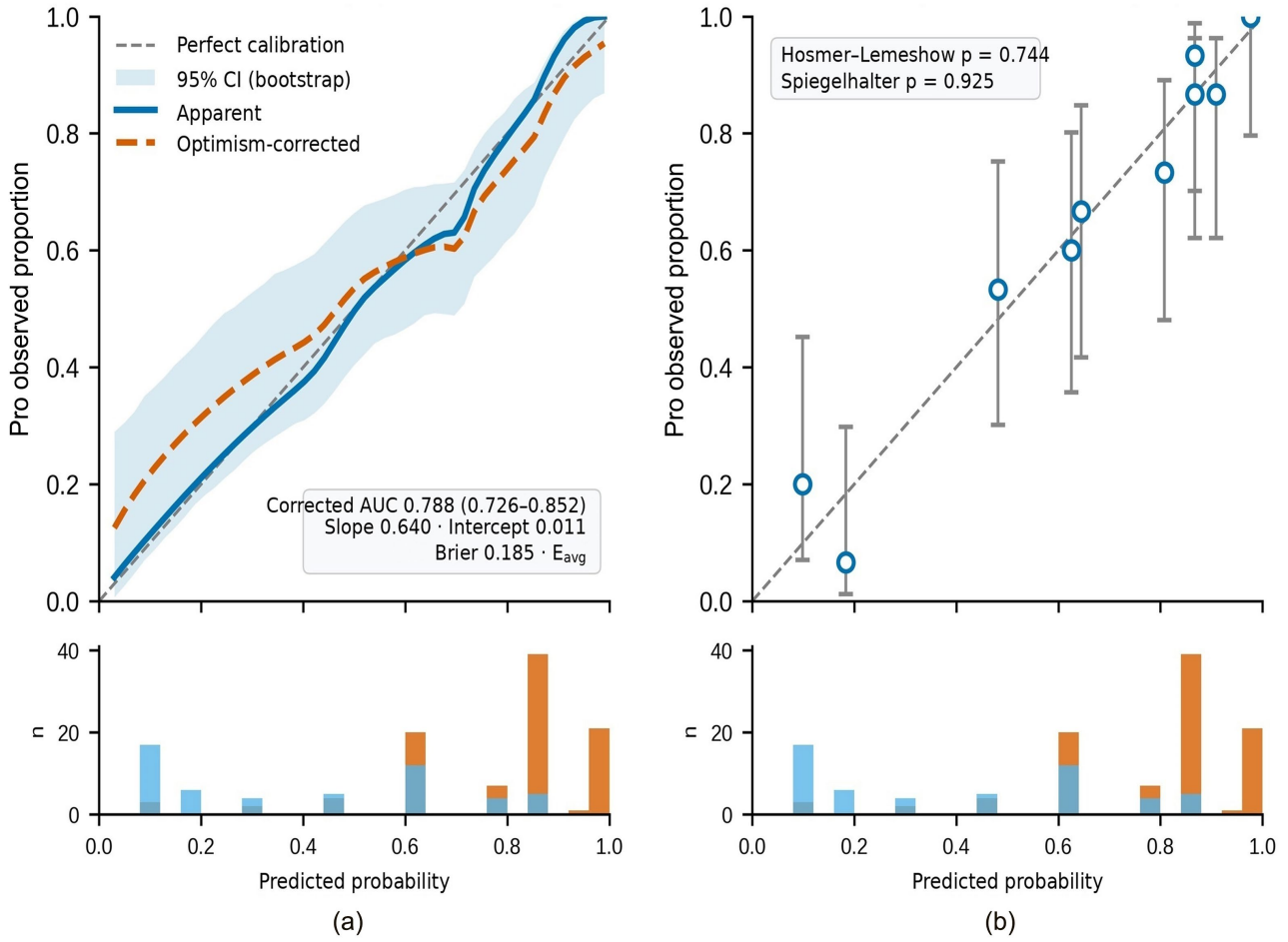

Figure 3. Calibration of the score on the composite endpoint. (a) Calibration curve smoothed by local logistic regression: apparent curve (solid line), optimism-corrected curve (dashed), 95% confidence envelope and line of identity. (b) Calibration by deciles of predicted risk, with Wilson confidence intervals and the results of the Spiegelhalter and Hosmer-Lemeshow tests. The lower panels show the distribution of predicted probabilities according to occurrence of the event.

The apparent calibration curve and the optimism-corrected curve, with the line of identity and the 95% confidence envelope, are shown in Figure 3, together with the histogram of predicted probabilities by status. Correction for optimism shifts the curve upwards in the lower half of the range, which means that the score, transposed to an unobserved population, would slightly underestimate the risk of low-scoring patients—the clinically preferable direction of error for a triage instrument, whose most costly error is underestimation of high risk.

3.7. Performance by Target and Comparison with the Glasgow Coma Scale

Applied to the individual targets (Table 7, Figure 4), the score retained useful discrimination for critical-care use (AUC 0.787; 95% CI 0.714 - 0.851), moderate discrimination for poor functional outcome (0.711; 0.628 - 0.791), weak discrimination for early surgery (0.588; 0.494 - 0.683) and none for death (0.472; 0.336 - 0.608).

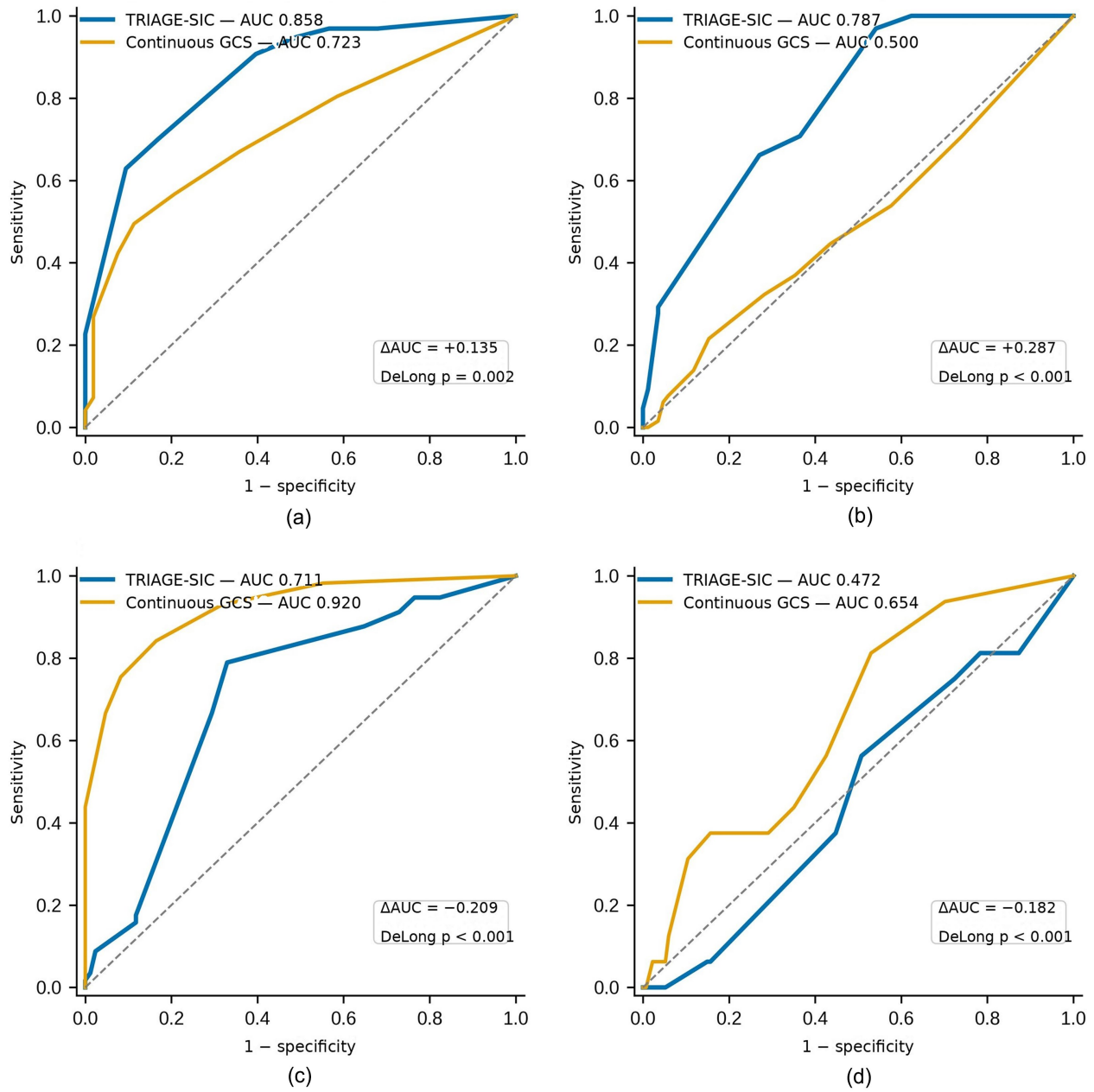

Figure 4. Discrimination of the score by triage target, compared with the continuous Glasgow Coma Scale. (a) Critical trajectory (composite); (b) critical-care use; (c) poor functional outcome (GOS ≤ 3); (d) in-hospital death. Each panel reports the difference in area under the curve and the p value of the DeLong test for correlated ROC curves.

Table 7. Performance of the TRIAGE-SIC score by triage target, and comparison with the continuous GCS (DeLong test).

Target	Events/N	AUC of the score (95% CI)	AUC of the continuous GCS	ΔAUC (95% CI)	p (DeLong)	Interpretation
Critical trajectory (composite)	97/150	0.858 (0.795 - 0.914)*	0.723 (0.642 - 0.804)	+0.135 (0.050 - 0.221)	0.002	The score is **superior** to the GCS
Critical-care use	65/150	0.787 (0.714 - 0.851)	0.500 (0.406 - 0.594)	+0.287 (0.203 - 0.372)	<0.001	The score is **superior**; the GCS adds nothing

Continued

Surgery ≤ 24 h	41/150	0.588 (0.494 - 0.683)	0.685 (0.584 - 0.786)	−0.097 (−0.187 to −0.006)	0.037	Weakly predictable; GCS superior
Poor outcome (GOS ≤ 3)	57/142	0.711 (0.628 - 0.791)	0.920 (0.869 - 0.971)	−0.209 (−0.289 to −0.130)	<0.001	**The continuous GCS is superior**
In-hospital death	16/150	0.472 (0.336 - 0.608)	0.654 (0.501 - 0.807)	−0.182 (−0.284 to −0.080)	<0.001	**Not predictable by the score**

*Apparent AUC; the value corrected for optimism by 1000 bootstrap resamples encompassing selection is 0.788 (0.726 - 0.852). ΔAUC = AUC of the score−AUC of the continuous GCS. A positive value favours the score. The 95% confidence intervals of the AUC of the continuous GCS are given in the table. For critical-care use the AUC of the GCS is exactly 0.500 (95% CI 0.406 - 0.594), computed on all 150 admissions with the conventional orientation of the scale; the interval includes 0.5, confirming the absence of any discrimination and supporting one of the two negative findings.

Direct comparison with the pre-specified comparator by the DeLong test gives these figures their meaning. For the two resource-allocation targets, TRIAGE-SIC is significantly superior to the continuous Glasgow Coma Scale: critical trajectory, 0.858 versus 0.723, ΔAUC + 0.135 (95% CI 0.050 - 0.221; p = 0.002); critical-care use, 0.787 versus 0.500, ΔAUC + 0.287 (0.203 - 0.372; p < 0.001). This second result deserves emphasis: the Glasgow Coma Scale, taken alone, does not discriminate at all for the use of critical-care therapies in this cohort (AUC 0.500), whereas the triage score reaches 0.787. The information that determines the need for osmotherapy or ventricular drainage is therefore essentially morphological—mass effect, hydrocephalus—and not related to the depth of coma.

Conversely, for the other three targets the comparator is significantly superior to the score: poor functional outcome, ΔAUC −0.209 (−0.289 to −0.130; p < 0.001); death, ΔAUC −0.182 (−0.284 to −0.080; p < 0.001); early surgery, ΔAUC −0.097 (−0.187 to −0.006; p = 0.037). The two instruments therefore do not answer the same question, and this divergence is not a sampling artefact but a statistically established result across all five comparisons.

3.8. Pre-Specified Negative Findings

Two pre-specified analyses produced negative results, reported here on the same footing as the positive ones.

3.8.1. Negative Finding No. 1: In-Hospital Mortality Is Not Predictable from Variables Available at Presentation Alone

The selection procedure applied specifically to death retained no variable. The three most frequently selected candidates—age ≥ 65 years, GCS ≤ 8 and subdural empyema—reached only 40% stability, below the pre-specified threshold of 60%. The continuous Glasgow Coma Scale discriminated death only weakly (AUC 0.654) and the TRIAGE-SIC score not at all (0.472; 95% CI 0.336 - 0.608, an interval containing 0.500).

3.8.2. Negative Finding No. 2: For Functional Outcome, the Glasgow Coma Scale Outperforms the Point Score

The pre-specified comparator reached an area under the curve of 0.920 for GOS

≤ 3, against 0.711 for the point score, a difference of 0.209 that is significant on the DeLong test ($p < 0.001$). A model specific to this target, selected by the same procedure (GCS ≤ 8, GCS 9 - 12, impaired consciousness, subdural empyema, midline shift ≥ 5 mm), reached 0.889, still below the continuous GCS. Conversely, for the composite endpoint the score outperformed the continuous GCS (0.858 versus 0.723; $p = 0.002$). This dissociation can be read directly in the distribution by band (Table 8(a)): the proportion of poor functional outcomes peaks at 66.0% in the 5 - 6 band and then falls back to 50.0% in the ≥7 band, and mortality, also non-monotonic, falls from 14.5% to 4.5% between these same two bands. The score ranks resource needs, not individual neurological severity.

Table 8. Observed risk by score band, operating characteristics of the thresholds and likelihood ratios. (a) Observed risk by band, (b) Operating characteristics of the thresholds (target: critical trajectory; prevalence 64.7%).

(a)

Band	n (%)	Critical trajectory (95% CI, Wilson)	Critical care	GOS ≤ 3	Death	Surgery ≤ 24 h
0 - 1 point	26 (17.3)	3/26—11.5% (4.0 - 29.0)	0/26 (0%)	3/23 (13.0%)	3/26 (11.5%)	3/26 (11.5%)
2 - 4 points	47 (31.3)	26/47—55.3% (41.2 - 68.6)	19/47 (40.4%)	9/46 (19.6%)	4/47 (8.5%)	11/47 (23.4%)
5 - 6 points	55 (36.7)	46/55—83.6% (71.7 - 91.1)	27/55 (49.1%)	35/53 (66.0%)	8/55 (14.5%)	21/55 (38.2%)
≥7 points	22 (14.7)	22/22—100% (85.1 - 100)	19/22 (86.4%)	10/20 (50.0%)	1/22 (4.5%)	6/22 (27.3%)

(b)

Threshold	Sn	Sp	PPV	NPV	LR+ (95% CI)	LR− (95% CI)	Post-test probability if positive	Post-test probability if negative	Use
≥2	96.9%	43.4%	75.8%	88.5%	1.71 (1.35 - 2.17)	0.07 (0.02 - 0.23)	75.8%	11.5%	**Rule-out**—allow a non-priority pathway
≥3	94.8%	50.9%	78.0%	84.4%	1.93 (1.46 - 2.55)	0.10 (0.04 - 0.25)	78.0%	15.6%	Intermediate
≥4	90.7%	60.4%	80.7%	78.0%	2.29 (1.63 - 3.21)	0.15 (0.08 - 0.30)	80.7%	22.0%	Intermediate
≥5	70.1%	83.0%	88.3%	60.3%	4.13 (2.24 - 7.59)	0.36 (0.26 - 0.50)	88.3%	39.7%	**Rule-in**—trigger the priority pathway
≥6	62.9%	90.6%	92.4%	57.1%	6.67 (2.85 - 15.57)	0.41 (0.31 - 0.54)	92.4%	42.9%	Strict rule-in

Sn: sensitivity; Sp: specificity; PPV/NPV: positive and negative predictive values; LR +/LR−: positive and negative likelihood ratios. Post-test probabilities computed from the observed prevalence of 64.7%. At the ≥ 7 threshold, specificity reaches 100% and LR + is not estimable (no false positives); this threshold is therefore not proposed as a decision rule.

3.9. Risk Bands and Likelihood Ratios

Four bands were defined a priori on the distribution of points (Table 8, Figure 5). The observed risk of a critical trajectory was 11.5% (3/26; Wilson 95% CI 4.0 - 29.0) for 0 - 1 point, 55.3% (26/47; 41.2 - 68.6) for 2 - 4 points, 83.6% (46/55; 71.7 - 91.1) for 5 - 6 points and 100% (22/22; 85.1 - 100) for ≥7 points. Critical-care use followed the same gradient: 0%, 40.4%, 49.1% and 86.4%. No patient in the lowest band required osmotherapy or an external ventricular drain.

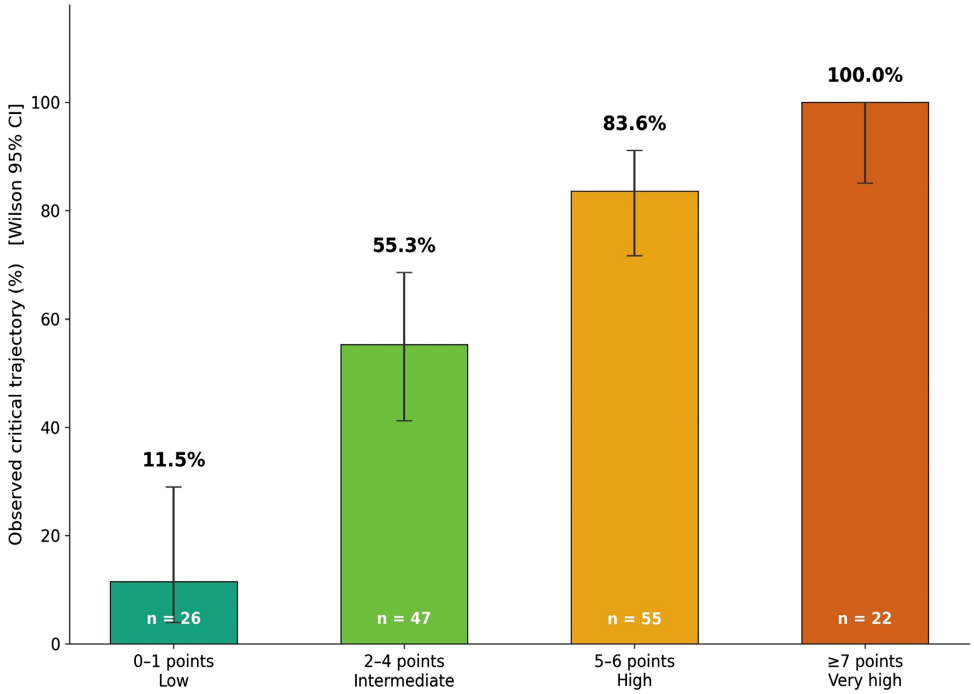

Figure 5. Observed risk of a critical trajectory by score band. Observed proportions with Wilson 95% confidence intervals, numbers per band and a graded colour code corresponding to the four priority levels.

At the operating thresholds, a score ≥ 2 identified the critical trajectory with a sensitivity of 96.9% and a negative predictive value of 88.5%, for a negative likelihood ratio of 0.07 (95% CI 0.02 - 0.23)—a configuration suited to rule-out use, that is, to allow a non-priority referral: a patient scoring below 2 sees his probability of a critical trajectory fall from 64.7% (prevalence) to 11.5%. A score ≥ 5 gave a sensitivity of 70.1%, a specificity of 83.0%, a positive predictive value of 88.3% and a positive likelihood ratio of 4.13 (2.24 - 7.59), raising the post-test probability to 88.3%—a configuration suited to rule-in use, that is, to trigger the priority pathway. A score ≥ 6 (LR + 6.67; 2.85 - 15.57) raises the post-test probability to 92.4%.

These two thresholds correspond to the two distinct triage decisions and are proposed jointly rather than as a single cut-off, a choice consistent with the logic of an instrument whose rule-out use and rule-in use do not carry the same consequence when they err.

3.10. Clinical Utility

Decision curve analysis (**Figure 6(a)**) showed a net benefit superior to both reference strategies across the whole range of clinically plausible thresholds. Expressed as a net reduction in unnecessary priority referrals per 100 patients admitted, the gain was 7 at the 0.20 threshold, 15 at the 0.50 threshold and 19 at the 0.65 threshold. At this last threshold, the "refer everyone" strategy became harmful (negative net benefit) whereas the score retained a net benefit of 0.342.

The clinical impact curve (**Figure 6(b)**) expresses the same result in absolute numbers. For 1000 admissions and at the 0.20 threshold, the score would classify 827 patients as high risk, of whom 627 would actually experience a critical trajectory. At the 0.50 threshold, these figures become 727 and 587 respectively.

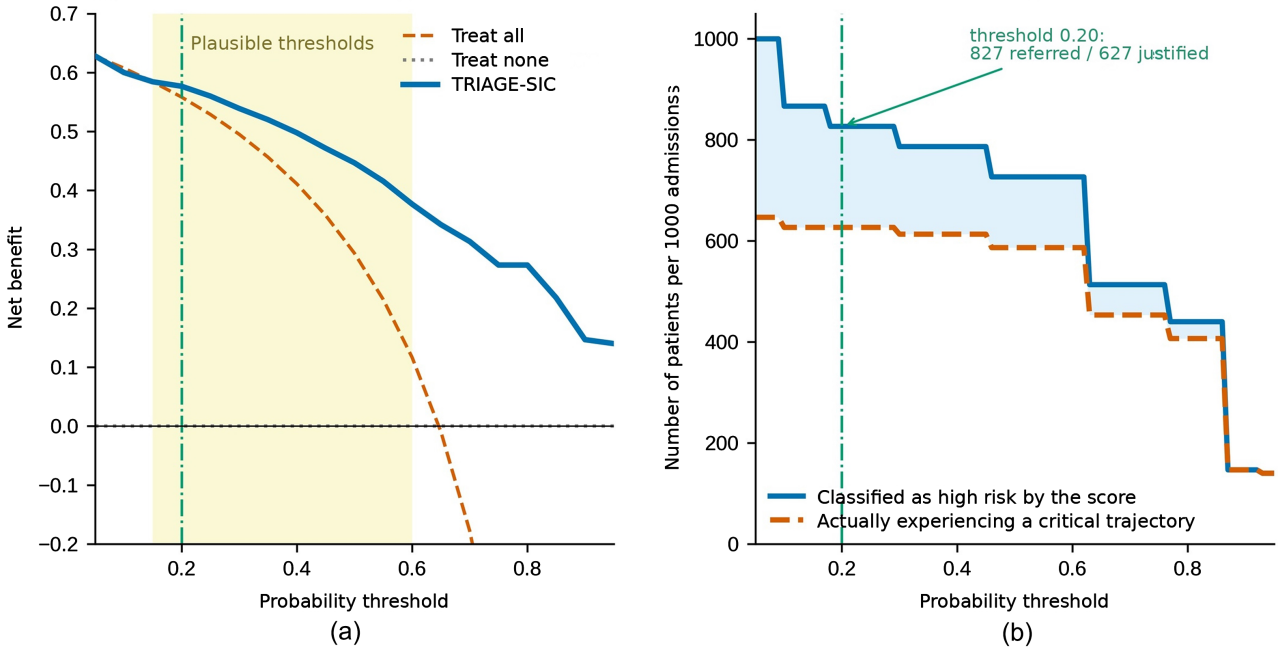

Figure 6. Clinical utility. (a) Clinical decision curve: net benefit of the score compared with the "refer everyone" and "refer no one" strategies. The shaded area delimits the range of clinically plausible thresholds for a triage decision (0.15 - 0.60); the reference threshold of 0.20 is indicated. (b) Clinical impact curve: for 1000 admissions and at each threshold, the number of patients the score would classify as high risk and, among them, the number who would actually experience a critical trajectory.

3.11. Sensitivity Analyses

Excluding the 41 patients operated on within the first 24 hours left 109 admissions and 65 events. Discrimination of the score there was 0.852, against 0.858 in the whole cohort. Calibration-in-the-large was 0.002 and the calibration slope 1.000; both values result from re-estimation of the score-to-probability relation in the 109 admissions and therefore do not constitute a test of the transportability of the derivation equation: the calibration that would have been obtained by applying that equation to this subgroup without re-estimation was not computed. The gradient between bands was preserved: 13.0%, 52.8%, 79.4% and 100% for 0 - 1, 2 - 4, 5 - 6 and ≥7 points. A LASSO selection conducted de novo in this subgroup

retained seven variables, including five of the six items of the final score; immunosuppression was no longer retained there, whereas GCS ≤ 8 and midline shift ≥ 5 mm were.

The complete-case analysis and the analysis restricted to each patient's first admission did not change the variables selected (S6).

Finally, a post-hoc sensitivity analysis, requested to gauge the dependence of the score on its treatment-dependent component, restricted the development endpoint to death or poor functional outcome (GOS ≤ 3), excluding critical-care use. Sixty-two of 150 admissions (41.3%) met this restricted endpoint. Repeating the imputation and LASSO selection procedure retained four variables above the 60% stability threshold: GCS ≤ 8, GCS 9 - 12, impaired consciousness and mass effect. The resulting Firth model had an apparent AUC of 0.831; bootstrap correction for optimism yielded an AUC of 0.817 (bootstrap 95% CI approximately 0.763 - 0.892), with an optimism-corrected calibration slope of 0.99. Removing the process-dependent critical-care component therefore reduced the number of selected variables and shifted selection towards neurological severity, while preserving useful discrimination.

3.12. Implementation Tools

Four supports are provided: a paper version fitting on half a page, with colour coding of the bands and the suggested management (**Table 5**); the nomogram, intended for teaching and for reading the relative weight of each item (**Figure 2**); the clinical algorithm (**Figure 7**); and a standalone calculator contained in a single HTML file, running offline on a mobile phone with no data transmission, which returns the total, the estimated probability, the risk band and an explicit reminder that the score predicts neither death nor individual functional outcome (S10). None of these supports requires a network connection or a licence.

4. Discussion

4.1. Main Findings and Positioning

We have developed and internally validated a triage score for intracranial suppurative infections in an Ivorian cohort of 150 admissions. Six variables available at presentation, without magnetic resonance imaging or microbiology, stratify the risk of an unfavourable in-hospital course with an optimism-corrected area under the curve of 0.788, and critical-care use with an area under the curve of 0.787—in both cases significantly better than the continuous Glasgow Coma Scale. The score predicts neither death nor individual functional outcome. These results matter because they provide, for the first time in this setting, a reproducible instrument for a decision that is currently taken without any formal support.

One point of vocabulary conditions the reading of everything that follows. TRIAGE-SIC is not a prognostic score: it is a resource-allocation instrument intended to structure the pathway decision in a resource-limited setting**.** It supports clinical judgement and does not replace it.

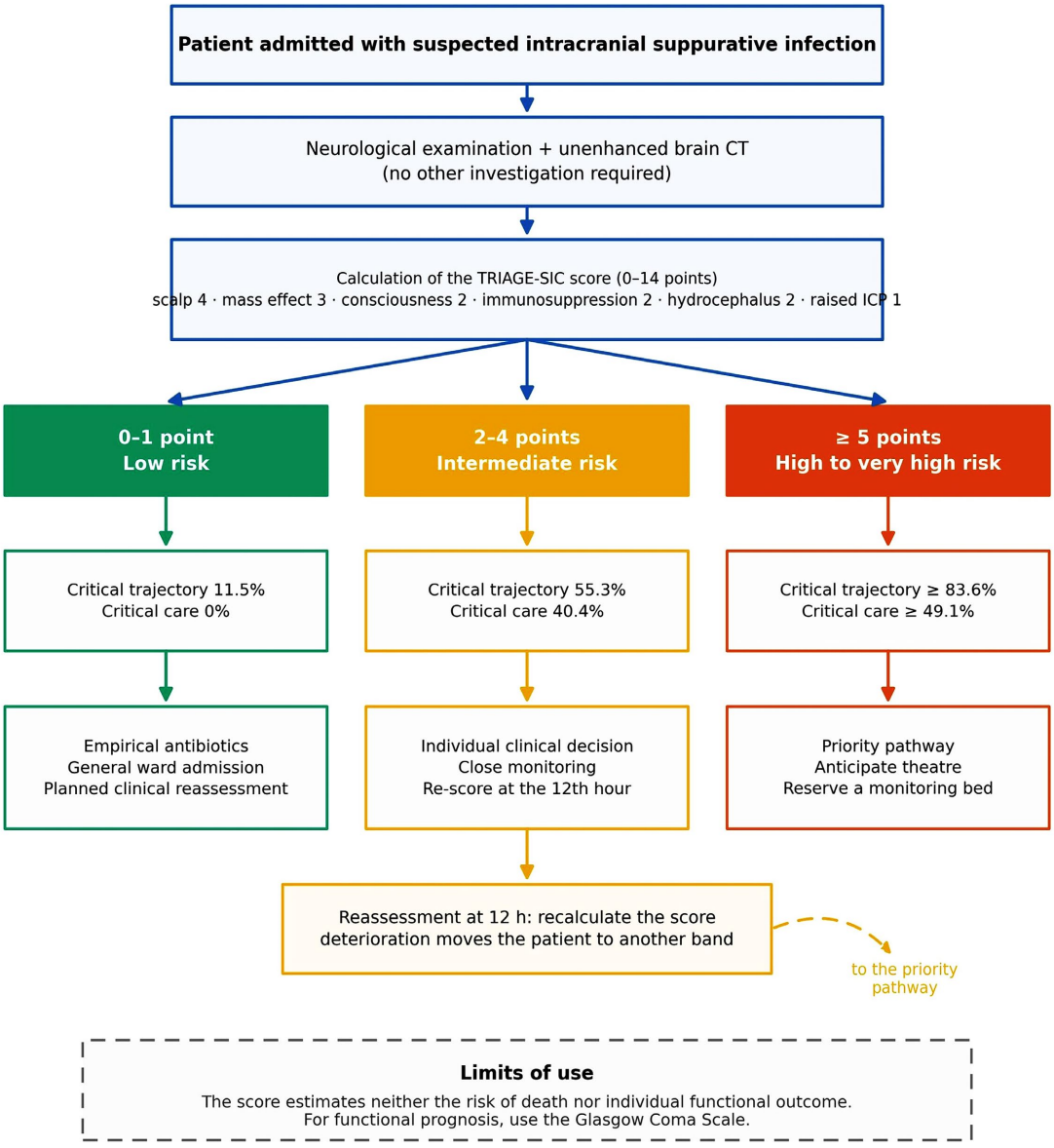

Figure 7. Clinical algorithm for use of the TRIAGE-SIC score. Sequence of steps from admission to referral, with the three decision bands, the corresponding observed risks, the reassessment loop at the twelfth hour and a reminder of the limits of use of the instrument.

4.2. Statistical Implications

4.2.1. Rationale for the Methodological Choices

Four methodological choices structure this work and deserve to be made explicit, because they determine the credibility of the model more than its apparent performance does.

LASSO was chosen because stepwise selection based on p values produces, at this sample size, unstable and overfitted models [43] [44]. L1 penalisation performs selection and shrinkage of the coefficients simultaneously; the one-standard-error rule favours parsimony over apparent performance; and the 60% stability threshold applied across the ten imputed datasets ensures that a variable does

not owe its inclusion to a particular draw [40].

Firth regression answers a specific difficulty: scalp swelling was associated with a critical trajectory in all 12 patients concerned, a quasi-complete separation configuration that makes ordinary maximum likelihood diverge. Penalisation by the Jeffreys term makes the estimate finite and corrects the second-order bias of classical estimators, a bias that is substantial when events are rare [41] [42].

The bootstrap encompassing selection is more conservative than cross-validation applied after selection. Replaying the entire procedure on each resample is the only way to quantify the optimism induced by selection itself [43]-[45]; a validation applied to the final model alone would have underestimated it.

Rubin pooling combines the estimates from the ten imputed datasets while incorporating imputation uncertainty into the variance [36] [38]. Analysing a single dataset would have produced spuriously narrow intervals; analysing complete cases would have discarded 18% of admissions because of midline shift.

4.2.2. Interpretation of Calibration

The corrected calibration slope, below 1, reflects moderate residual overfitting. This is the expected behaviour of a first derivation cohort. It implies that extreme predictions are too widely dispersed: high risks are slightly overestimated and low risks slightly underestimated. The practical consequence is not a revision of the coefficients but recalibration whenever the score is transposed [49] [59].

Calibration-in-the-large close to zero indicates that the model neither systematically overestimates nor systematically underestimates the mean probability of the event. The Spiegelhalter test and the Hosmer-Lemeshow test show no significant departure from perfect calibration, and calibration by deciles remains within the expected confidence intervals across the whole range of risk.

4.2.3. Discrimination, Prevalence and Transportability

An area under the curve approaching 0.80 is usually regarded as good discrimination. It must nevertheless be recalled that, for an implementation decision, calibration and clinical utility are more informative than discrimination alone [49] [54] [60]. That is why we report all three.

The prevalence of a critical trajectory reaches 64.7% in our cohort, a high value reflecting the recruitment of a referral centre receiving late-presenting patients. The positive and negative predictive values reported depend directly on this prevalence and are not transposable as they stand. Likelihood ratios, by contrast, are independent of prevalence and are the quantity to be carried across to an external validation. That is why they are reported with their confidence intervals, and why their application to any local prevalence is facilitated by the Fagan nomogram (Supplementary **Figure S8-3**).

Three levels of transportability must be distinguished. Statistical transportability concerns recalibration of the model: it is almost always necessary and technically simple, the intercept being the first parameter to readjust. Clinical transportability concerns the preservation of discrimination in a population whose case-

mix differs. Organisational transportability, finally, concerns adaptation to hospitals whose capacities differ: a centre with an identified intensive care unit does not share our operational definition of "critical care". All three must be assessed separately at external validation [59] [61] [62].

4.3. Clinical Implications

4.3.1. Plausibility of the Point Scale

A score is adopted only if its content is recognisable to the clinician who has to use it. The six items retained by a wholly data-driven procedure describe the three mechanisms of decompensation set out in the introduction, which amounts to a form of construct validation.

Mass effect (3 points) reflects the first mechanism. It aggregates the volume of the collection and the surrounding oedema, whose contribution often exceeds that of the collection itself [13] [14], and proved more informative than size alone: two 30 mm collections do not have the same consequence depending on their location. **Hydrocephalus** (2 points) represents the second mechanism; it predicts the critical trajectory beyond the mere need for drainage, which suggests that it marks a more extensive involvement of cerebrospinal fluid pathways.

Impaired consciousness (2 points) reflects decompensation, whatever its cause. That the procedure retained it in preference to the dichotomised GCS deserves comment. The GCS remains the reference instrument for quantifying the depth of coma [51], but its inter-observer reliability is imperfect, particularly for the verbal component and in the aphasic, intubated or non-French-speaking patient [50]. The observation "this patient is not normally alert" is cruder but more reproducible. For an instrument intended for the least experienced operator in the chain, this robustness has a value of its own—at the cost, as our results show, of a loss of prognostic information.

Immunosuppression (2 points) modulates the response to infection, delays encapsulation and broadens the microbiological spectrum [15] [63] [64]; in a region where the prevalence of human immunodeficiency virus infection remains substantial, it is a demographic determinant of the patient flow rather than a rare variable. The definition adopted is deliberately restrictive: diabetes and sickle cell disease are not included, and the item concerned 8.7% of admissions. An external validation that included them would mechanically raise the prevalence of the item and shift the calibration of the score; the definition must therefore be taken over unchanged. **Raised intracranial pressure syndrome** (1 point) makes a weak but stable contribution, its modest weighting being consistent with its prevalence of 71.3%: a sign present in three patients out of four necessarily discriminates little.

4.3.2. The "Scalp Swelling" Item

All twelve patients concerned had a severe in-hospital course. This complete separation produces a very high but very imprecise estimate, which Firth penalisation renders estimable without rendering it reliable [41] [42]. Its weighting was capped at 4 points.

Clinically, the sign is coherent. Scalp suppuration with cellulitis in a patient with an intracranial suppurative infection generally reflects either an infection spreading through tissue planes by contiguity, with intercurrent osteitis, or a surgical site infection after craniotomy [57] [58]. In both cases, management requires wide debridement, often removal of a bone flap or of implanted material, and prolonged antibiotic therapy. The sign therefore marks surgical complexity rather than neurological severity—precisely the information a triage instrument needs.

This finding should be regarded as a hypothesis to be confirmed and is the priority target of external validation. The decision rule is stated in advance: if the association is not reproduced in an independent cohort, the item will have to be removed and the scale recalibrated over 10 points.

4.3.3. Domain of Validity

The scope of the instrument deserves to be delimited precisely. A poorly framed triage score is quickly read as an individual prognosis, and it is by this mechanism that a methodologically sound instrument can cause clinical harm.

The score does not predict death. An area under the curve of 0.472 indicates a complete absence of discrimination. With 16 deaths, the available power is far below that required to identify a stable combination of predictors [32]-[34]: the result therefore does not mean that no variable recorded at presentation is associated with death, but that no stable combination is identifiable at this sample size. This distinction matters, because failing to make it probably contributes to the publication of fragile mortality scores that subsequently fail at external validation. An explicit restriction of use follows: TRIAGE-SIC should not be presented to a patient or to a family as an estimate of the risk of death.

A hypothesis also follows. In a cohort where death occurs once therapeutic resources have already been committed, mortality probably depends more on events occurring after triage—the actual time to theatre, the real availability of monitoring, antibiotic stock-outs, secondary complications—than on the state at presentation. This reasoning is consistent with the analysis according to which mortality attributable to the quality of health systems exceeds that attributable to lack of access [21].

The score does not replace the GCS for functional prognosis, and this gap is statistically established. The practical conclusion is simple: to discuss functional outcome with a family, the clinician uses the GCS; to organise the pathway, he uses the score.

Lastly, the score predicts early surgery poorly. This result was foreseeable: time to surgery is a decision variable, determined as much by the state of the patient as by the availability of the theatre and of the anaesthetist. A score that predicted current time to surgery perfectly would in fact be a poor triage score, since it would merely reproduce existing decisions, biases included.

The sensitivity analysis excluding patients operated on immediately supports this reasoning. In this subgroup, where the immediate surgical decision can no longer contribute to the endpoint, discrimination of the score remains 0.852 and

calibration remains excellent. The score therefore does not reduce to a reproduction of the operative indication. The disappearance of immunosuppression in favour of GCS ≤ 8 and midline shift in a selection conducted de novo in this subgroup does, however, recall the instability of selection expected at this sample size, and underlines that the point scale must not be read as identifying the "true" causal factors.

4.3.4. Clinical Relevance and Use

TRIAGE-SIC is designed to support immediate triage decisions in resource-limited hospitals. It does not replace clinical judgement: it provides a standardised estimate of the risk of an unfavourable course, allowing a more rational allocation of neurosurgical resources and explicit communication between neurosurgeons, emergency physicians, anaesthetists and intensive care physicians.

Use proceeds in five steps: perform the unenhanced computed tomography scan, tick the six items, add up the total, identify the band, apply the suggested management (Figure 7, Table 5). Three clinical vignettes illustrate application of the score in Table 9.

Table 9. Three clinical vignettes illustrating use of the score.

	Patient A	Patient B	Patient C
Presentation	Man aged 32, headache and vomiting for three weeks, conscious and orientated	Woman aged 41, headache, drowsiness, right hemiparesis	Man aged 27, previous craniotomy, headache, inflammatory parietal swelling
Scalp swelling/cellulitis	no	no	**yes (+4)**
Mass effect (CT)	no	**yes (+3)**	**yes (+3)**
Impaired consciousness	no	**yes (+2)**	no
Immunosuppression	no	no	no
Hydrocephalus	no	no	no
Raised intracranial pressure syndrome	**yes (+1)**	no	**yes (+1)**
Total score	**1**	**5**	**8**
Estimated probability of a critical trajectory	18%	77%	96%
Band	Low (green)	High (orange)	Very high (red)
Suggested management	Empirical antibiotics, general ward admission, planned reassessment	Anticipate theatre and reserve a monitoring bed	Immediate priority pathway; plan wide debridement and possible removal of implanted material

These vignettes are constructed for teaching purposes and do not correspond to real patients. They illustrate the mechanics of the point scale, not a therapeutic recommendation: the suggested management remains subordinate to clinical judgement.

Three uses are conceivable without waiting for external validation, because they carry no risk for the patient: documenting the score on admission as a variable describing severity; using it as a formalised argument when negotiating an operating slot; and using it in teaching, including with the doctors of the peripheral facilities that refer the patients.

The expected organisational consequences—better theatre scheduling, anticipation of monitoring beds, fewer unjustified transfers—were not measured here and constitute the endpoints of a future impact study [65].

4.4. Implications for Global Neurosurgery

Three publications from 2015 structured the field of global surgery: the *Essential Surgery* volume of the Disease Control Priorities [66], the Lancet Commission on Global Surgery [16] [67] and Resolution WHA68.15 [12]. They put an end to the view of surgery as the "neglected stepchild of global health" [68]. Applied to neurosurgery, this analysis quantified a major shortfall in essential operations [11] [18] [19] [69], whose translation into public policy runs through national surgical, obstetric and anaesthesia plans [20] [70]—in which neurosurgery remains underrepresented [71]. In this respect, intracranial suppurative infections are a sentinel condition: preventable upstream, curable downstream, and dependent on nothing but the speed of access to a technically simple procedure.

The score fits into this framework along three lines.

Resource allocation. Decision curve analysis shows that the advantage of the score over a policy of systematic referral grows with the scarcity of the resource. The instrument is therefore most useful where it is most needed. In a department sharing a theatre with all the surgical specialties, it provides an objectifiable justification for the priority requested—an argument that, in practice, carries more weight than a subjective appraisal. Beyond the bedside, describing a cohort by its distribution of scores rather than by its mortality alone allows the resource burden to be compared between centres independently of their capacities, and hence a centre receiving more severely ill patients to be distinguished from a centre obtaining poorer results at equal severity. This descriptive function would give national plans a specifically neurosurgical indicator, where the indicators of the Lancet Commission remain general-purpose [16] [20].

Equity of access. Triage that is not formalised is not neutral triage. In the absence of an explicit criterion, priority is set by factors that are not clinical: time of arrival, insistence of the accompanying relative, familiarity of the referring doctor with the team, ability to pay up front. This last point is not anecdotal in a system where out-of-pocket payment remains dominant and where catastrophic health expenditure is one of the indicators of the Lancet Commission [16] [67]. An explicit score, displayed and documented, makes the decision auditable and shifts the discussion towards clinical criteria. This argument of distributive justice and decisional transparency is independent of the statistical performance of the instrument.

Essential neurosurgery. The availability constraint was imposed before the analysis rather than observed afterwards. The six items require a neurological examination and an unenhanced computed tomography scan; none requires magnetic resonance imaging, microbiology, a biomarker, a computer algorithm or a network connection. This parsimony is a design specification, aligned with the definition of essential neurosurgery [19] [69]. As the six items belong to the common clinical vocabulary of the subregion, a West African multicentre external validation is technically feasible from data already collected.

4.5. Comparison with the Literature

European series are the methodological reference in this field. The meta-analysis by Brouwer and colleagues places mortality at around 10% in series published after 2000, with a favourable outcome in about two thirds of survivors [1]. The prognostic determinants identified differ, however, from one cohort to another. The British series identifies impaired consciousness on admission [58]; the prognostic analysis of the Danish national cohort—whose incidence and mortality were described separately [72]—retains intraventricular rupture, immunosuppression, age over 65 years and a collection diameter over 30 mm, without impaired consciousness appearing among them [73]. Asian and Turkish series place impaired consciousness first and add intraventricular rupture [74]-[77]. This heterogeneity deserves note: three of the four factors of the Danish series describe the patient's background or the mechanical repercussion of the collection rather than the depth of coma, which chimes with the dissociation we observe between resource needs and individual neurological prognosis. North American series historically established the transformative effect of computed tomography and remain the reference for the choice of therapeutic modality [3] [4] [78].

African series are the most relevant comparator. They report mortality of between 10% and 25% [6]-[10] [79]. The South African series of Nathoo and colleagues, the largest published in the computed tomography era on the continent, shows that preoperative neurological deterioration governs outcome more than the organism or the location does [10] [79]. Series from French-speaking West Africa describe a younger population, a predominance of otogenic and post-traumatic forms, and decisive delays in access to imaging [8] [9]; Ghanaian and Nigerian work stresses the frequency of presentation with established raised intracranial pressure [6].

Three observations emerge from this comparison. The clinical picture on admission is remarkably homogeneous from one African country to another, which argues in favour of the transportability of the score. Mortality, by contrast, varies by a factor of two and a half without initial severity fully explaining it, which points to a system effect. Finally, none of these series produced a decision instrument, which leaves each team to rebuild its prioritisation rules empirically; the positioning of TRIAGE-SIC relative to existing instruments is detailed in Supplementary **Table S7-3**.

The differences from high-income series concern delays and microbiology more than mortality. Our median interval of 19 days between first symptoms and admission is far longer than European medians, which are counted in days [72] [73] [80]. Symmetrically, our proportion of sterile cultures far exceeds that of series where microbiological yield reaches 70% when the sample precedes antibiotic therapy [63] [64] [77]. These two differences define the context of use of the score and empirically justify the availability rule imposed on the predictors.

4.6. Comparison with International Guidelines

Several normative texts frame the management of intracranial suppurative infections: reference syntheses [2] [15] and the European ESCMID guidelines for brain abscess [63], the Infectious Diseases Society of America having published a dedicated guideline only for healthcare-associated ventriculitis and meningitis [81]. All converge on a three-stage architecture: cross-sectional imaging without delay, drainage or excision of collections larger than 2.5 cm or responsible for mass effect, then prolonged antibiotic therapy guided by the intraoperative sample.

Three observations follow from comparing the score with this framework.

Our point scale is consistent with the recommended surgical threshold but expresses it differently. Our procedure retained only the mechanical criterion, size ≥ 30 mm failing to reach the stability threshold despite a strong univariable association. In a late-presenting population where median size already reaches 36 mm, almost every patient crosses the dimensional criterion, which thereby loses its discriminating power. A threshold derived from a population with short delays may thus fail in a population with long delays.

The guidelines then presuppose a resource that our setting does not guarantee. They state what should be done, not in what order to do it when two eligible patients present on the same night for a single theatre. TRIAGE-SIC therefore does not replace them: it operates in a decision space they do not cover.

The place of unenhanced computed tomography deserves, finally, to be made explicit. The guidelines favour diffusion-weighted magnetic resonance imaging for the differential diagnosis between abscess and necrotic tumour, an undisputed superiority [2] [15]. But the triage question is not the differential diagnosis question: once suppuration has been established, the three pieces of information that determine urgency are accessible without contrast. Designing the score around this constraint is not a degraded compromise but the matching of the instrument to the decision it serves.

4.7. The Place of Machine Learning Approaches

A more complex model—random forest, gradient boosting, neural network—could achieve better discrimination on these data. This comparison was conducted separately in the same cohort and is the subject of a distinct paper: at this sample size, algorithmic complexity does not improve prediction. We deliberately chose here an instrument calculable in the head. The systematic review by War-

man and colleagues shows that most machine learning models published in neurosurgery are single-centre and that fewer than one in six undergo external validation [82]. An opaque model, dependent on computing infrastructure and hard to verify at the bedside, would meet the stated constraint less well than a six-item point scale calculable in the head. Transparency and interpretability are not concessions here: they are conditions for adoption and for external validation. Future work may compare TRIAGE-SIC with these approaches, provided that discrimination, calibration and clinical utility are assessed jointly.

4.8. Strengths

This work brings together four features that, to our knowledge, are not found together in the literature of this field: a neurosurgical triage score developed in an African population; an endpoint centred on resource allocation rather than on survival; internal validation encompassing the entire construction procedure, selection included; and reporting compliant with TRIPOD and TRIPOD + AI, together with publication of the full equation and of the code.

Three methodological points may be added. The availability constraint on the predictors was imposed before the analysis, guaranteeing that the score can actually be used in the settings concerned. Internal validation replays the entire procedure, selection included, which avoids the substantial optimism introduced by a validation applied after selection [43]-[45]. Calibration is reported at every recommended level, including by formal tests and by deciles [47]-[49]. The negative analyses and the sensitivity analyses were pre-specified, and the former are reported on the same footing as the positive results. Reporting follows TRIPOD and TRIPOD + AI [25]-[27]. Finally, the protocol, the full code, the complete equation and the implementation tools are provided: the whole analytical pipeline—preprocessing, multiple imputation, selection, estimation, bootstrap validation and implementation of the score—is reproducible as it stands.

4.9. Limitations

Sample size and power. One hundred and fifty admissions is a modest sample. The composite endpoint produced 97 events, about 16 events per retained predictor, which exceeds contemporary minimum recommendations for penalised modelling [33] [34] [83]. The sample nevertheless remains insufficient for the individual targets, and plainly insufficient for death (16 events). External validation remains indispensable before any clinical implementation.

Single centre and absence of external validation. This is a single-centre, predominantly retrospective study. The performance reported, even corrected for optimism, will probably be lower in another centre.

Absence of internal temporal validation. The cohort spans ten years, over which service provision and practice have changed. We did not carry out development on the earlier period followed by evaluation on the recent period, the sample size not allowing the cohort to be split without compromising the stability

of selection. Performance on the most recent admissions, which correspond to the future context of use, therefore remains unknown.

Data drift over ten years. Beyond the absence of temporal validation, a gradual shift in the distribution of the predictors and in the predictor-endpoint relation is plausible over such a period. This phenomenon is the main mechanism of failure of prediction models deployed in clinical practice [84] and could not be quantified here.

Absence of geographical validation. No evaluation in a separate centre. Bouaké Teaching Hospital receives a predominantly rural and referred population, with long delays. A centre in Abidjan, Dakar or Johannesburg would have a different profile, with a probably lower prevalence of the composite, which would mechanically shift the predictive values.

Selection bias. The cohort consists of patients who reached the neurosurgery department of a teaching hospital. Patients who died before transfer, were directed elsewhere, or never obtained imaging do not appear in it. The score therefore describes risk conditional on admission to a referral centre, not the risk of the disease in the population.

Treatment indication bias. The most severely ill patients received the heaviest treatments, and critical-care use—a component of the primary endpoint—is partly a consequence of the medical decision rather than an independent event. This bias is intrinsic to a composite endpoint incorporating a process variable and cannot be neutralised in a retrospective dataset. A sensitivity analysis restricted to the more patient-centred endpoint of death or poor functional outcome, excluding critical-care use, preserved useful discrimination (optimism-corrected AUC 0.817), although selection then shifted towards GCS-defined neurological severity—which supports reading TRIAGE-SIC as a triage rather than a purely prognostic instrument.

Inter-observer reproducibility not assessed. The six items, particularly impaired consciousness and mass effect, rest on an appraisal whose agreement between assessors was not measured. A dedicated agreement study is a prerequisite for the impact study.

Proxy for the "critical care" target. A therapeutic definition in the absence of an intensive care admission variable, influenced by prescribing habits and not transposable as it stands to a centre with an identified unit.

Fixed imputation within the bootstrap. The imputations were generated once and held fixed across bootstrap resamples rather than regenerated within each resample. This pragmatic choice, adopted because the fraction of missing data was low, may make the optimism-corrected confidence intervals marginally narrower than those a fully nested imputation-in-bootstrap procedure would yield; the reported intervals should be read with that caveat.

Other limitations. The "early surgery" target is a decision and not an event. Microbiological sampling was not performed in every patient and culture conditions varied. Two patients each contribute two admissions; an analysis restricted

to the first admission did not change the variables selected. Finally, the composite endpoint must not be broken down into specific predictions of each of its components, as the failure to predict death demonstrates.

4.10. Research Programme

We propose a seven-step programme, the first three steps of which are feasible in the short term:

1) **Ivorian external validation**—an independent cohort from one or two national centres, with recalibration of the intercept.

2) **West African validation**—five to ten centres, with a sample size calculated a priori according to the methods dedicated to models with a binary endpoint rather than set for convenience [85]-[87]; as an indication, about 500 patients would allow the calibration slope to be estimated with useful precision. The evaluation will address discrimination, calibration, the recalibration required and clinical utility jointly.

3) **Validation in French-speaking and then sub-Saharan Africa**—geographical extension and assessment of organisational transportability.

4) **Inter-observer agreement study** of the six items, a prerequisite for any impact study.

5) **Prospective impact study**, with the admission-to-theatre interval of high-risk patients as the primary endpoint.

6) **Cluster randomised trial** comparing formalised triage with usual practice.

7) **Comparison with machine learning models** and digital deployment, once external validity has been established.

Prospective data collection including process variables will also have to test the hypothesis that mortality depends more on the system than on the state at presentation. TRIAGE-SIC should be regarded as version 1.0 of an instrument that is meant to evolve.

5. Conclusions

TRIAGE-SIC is, to our knowledge, the first African neurosurgical triage instrument for intracranial suppurative infections built to the contemporary methodological standards of clinical prediction rule development—pre-specification of the protocol, LASSO selection by stability, Firth penalised regression, internal validation by bootstrap encompassing selection, full assessment of calibration, decision curves and TRIPOD-compliant reporting. Six items, readable in under a minute and requiring only a neurological examination and an unenhanced computed tomography scan, usefully stratify the risk of a critical in-hospital trajectory and of critical-care use, with significant superiority over the Glasgow Coma Scale for these two targets.

The instrument does not predict death and does not improve on the Glasgow Coma Scale for functional prognosis. These two negative findings, pre-specified and reported in full, delimit its domain of validity and constitute, as much as the

positive results, the contribution of this work: they indicate that, at this sample size and in this setting, in-hospital mortality from intracranial suppurative infections is not predictable from admission variables alone, and they direct research towards the process variables of the health system.

Positioned as a resource-allocation instrument and not as an individual prognosis, TRIAGE-SIC provides a working basis for structuring neurosurgical triage in a resource-limited setting. Prospective multicentre validation in several African countries, followed by an impact study, is now the indispensable step before any clinical implementation.

Beyond the prediction of clinical deterioration, TRIAGE-SIC offers a pragmatic framework for the equitable allocation of scarce neurosurgical resources in a low-resource setting, and could serve as a prototype for context-adapted neurosurgical triage tools across sub-Saharan Africa.

Author Contributions

All authors made a substantial contribution to this work, whether to the design of the study, the collection of the data, the analysis and interpretation of the results, or the drafting and critical revision of the manuscript. All have read and approved the submitted version and take collective responsibility for it.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Ethical Approval

The study was conducted in accordance with the principles of the Declaration of Helsinki and approved internally by the Department of Neurosurgery of Bouaké Teaching Hospital, where it was carried out.

Consent

Informed consent was obtained for the prospective inclusions; a waiver of consent was granted for the retrospective phase, the data having been anonymised.

Availability of Data and Code

The pre-specified statistical protocol (S1), the completed TRIPOD checklist (S2), the TRIPOD + AI checklist (S3), the variable dictionary (S4), the full model equation (S5), the sensitivity analyses (S6), the supplementary tables and figures (S7 and S8), the full analysis code in Python 3.12 (S9), the standalone web calculator (S10) and the graphical abstract (S11) are provided in a separate supplementary material document accompanying this submission. They will also be deposited before publication in a public repository with a persistent identifier (DOI), to which the manuscript will refer. De-identified individual data are available from the corresponding author on reasonable request and subject to a data sharing agreement.

Funding

No specific funding.

Overlap with Previous Work

This analysis uses the same institutional cohort as several papers from the same group: a prognostic score for poor functional outcome (SIC-4 score), the epidemiological and prognostic description of the cohort, an explainable machine learning analysis of the same 142 assessable episodes and the same GOS ≤ 3 endpoint, and an analysis of antibiotic stewardship. The machine learning paper concludes, as the present manuscript does, that the admission Glasgow Coma Scale concentrates most of the measurable predictive signal for functional outcome. The question asked here is a different one—a multi-target triage instrument usable from presentation—and the results presented in no way constitute an external validation of the earlier score, nor the reverse. The reader should regard the two papers as two non-independent analyses of a single population.

Transparency Statement

This work rests on four commitments, all verifiable in the supplementary material: a pre-specified statistical protocol established before any performance figure was examined (S1); publication of the full model equation, intercept included (S5); release of the complete analysis code, allowing every value reported to be reproduced (S9); and a reproducible internal validation encompassing the entire construction procedure, variable selection included.

References

[1] Brouwer, M.C., Coutinho, J.M. and van de Beek, D. (2014) Clinical Characteristics and Outcome of Brain Abscess: Systematic Review and Meta-Analysis. *Neurology*, **82**, 806-813. https://doi.org/10.1212/wnl.0000000000000172

[2] Brouwer, M.C., Tunkel, A.R., McKhann, G.M. and van de Beek, D. (2014) Brain Abscess. *The New England Journal of Medicine*, **371**, 447-456. https://doi.org/10.1056/nejmra1301635

[3] Rosenblum, M.L., Hoff, J.T., Norman, D., Weinstein, P.R. and Pitts, L. (1978) Decreased Mortality from Brain Abscesses since Advent of Computerized Tomography. *Journal of Neurosurgery*, **49**, 658-668. https://doi.org/10.3171/jns.1978.49.5.0658

[4] Yang, S.Y. (1981) Brain Abscess: A Review of 400 Cases. *Journal of Neurosurgery*, **55**, 794-799. https://doi.org/10.3171/jns.1981.55.5.0794

[5] GBD 2021 Nervous System Disorders Collaborators (2024) Global, Regional, and National Burden of Disorders Affecting the Nervous System, 1990-2021: A Systematic Analysis for the Global Burden of Disease Study 2021. *The Lancet Neurology*, **23**, 344-381.

[6] Dakurah, T.K., Iddrissu, M., Wepede, G. and Nuamah, I. (2006) Hemispheric Brain Abscess: A Review of 46 Cases. *West African Journal of Medicine*, **25**, 126-129. https://doi.org/10.4314/wajm.v25i2.28262

[7] Hassani, F.D., Fatemi, N.E., Moufid, F., Yassad Oudrhiri, M., Gana, R., Maaqili, R.E.,

et al. (2014) Abcès encéphaliques: Prise en charge, à propos d'une série de 82 cas. *The Pan African Medical Journal*, **18**, 110. https://doi.org/10.11604/pamj.2014.18.110.2247

[8] Kabré, A., Zabsonré, S., Diallo, O. and Cissé, R. (2014) Prise en charge médico-chirurgicale des abcès du cerveau à l'ère du scanner en Afrique sub-saharienne: À propos de 112 cas. *Neurochirurgie*, **60**, 249-253. https://doi.org/10.1016/j.neuchi.2014.06.011

[9] Loembe, P.M., Okome-Kouakou, M. and Alliez, B. (1997) Les suppurations col-lectées intra-crâniennes en milieu africain. *Médecine Tropicale*, **57**, 186-194. https://pubmed.ncbi.nlm.nih.gov/9304016/

[10] Nathoo, N., Nadvi, S.S., Narotam, P.K. and van Dellen, J.R. (2011) Brain Abscess: Management and Outcome Analysis of a Computed Tomography Era Experience with 973 Patients. *World Neurosurgery*, **75**, 716-726. https://doi.org/10.1016/j.wneu.2010.11.043

[11] Punchak, M., Mukhopadhyay, S., Sachdev, S., Hung, Y., Peeters, S., Rattani, A., *et al.* (2018) Neurosurgical Care: Availability and Access in Low-Income and Middle-Income Countries. *World Neurosurgery*, **112**, e240-e254. https://doi.org/10.1016/j.wneu.2018.01.029

[12] World Health Organization (2015) Resolution WHA68.15: Strengthening Emergency and Essential Surgical Care and Anaesthesia as a Component of Universal Health Coverage. World Health Organization. https://apps.who.int/gb/ebwha/pdf_files/WHA68/A68_R15-en.pdf

[13] Mathisen, G.E. and Johnson, J.P. (1997) Brain Abscess. *Clinical Infectious Diseases*, **25**, 763-779. https://doi.org/10.1086/515541

[14] Muzumdar, D., Jhawar, S. and Goel, A. (2011) Brain Abscess: An Overview. *International Journal of Surgery*, **9**, 136-144. https://doi.org/10.1016/j.ijsu.2010.11.005

[15] Sonneville, R., Ruimy, R., Benzonana, N., Riffaud, L., Carsin, A., Tadié, J., *et al.* (2017) An Update on Bacterial Brain Abscess in Immunocompetent Patients. *Clinical Microbiology and Infection*, **23**, 614-620. https://doi.org/10.1016/j.cmi.2017.05.004

[16] Meara, J.G., Leather, A.J.M., Hagander, L., Alkire, B.C., Alonso, N., Ameh, E.A., *et al.* (2015) Global Surgery 2030: Evidence and Solutions for Achieving Health, Welfare, and Economic Development. *The Lancet*, **386**, 569-624. https://doi.org/10.1016/s0140-6736(15)60160-x

[17] Shrime, M.G., Bickler, S.W., Alkire, B.C. and Mock, C. (2015) Global Burden of Surgical Disease: An Estimation from the Provider Perspective. *The Lancet Global Health*, **3**, S8-S9. https://doi.org/10.1016/s2214-109x(14)70384-5

[18] Dewan, M.C., Rattani, A., Fieggen, G., Arraez, M.A., Servadei, F., Boop, F.A., *et al.* (2019) Global Neurosurgery: The Current Capacity and Deficit in the Provision of Essential Neurosurgical Care. Executive Summary of the Global Neurosurgery Initiative at the Program in Global Surgery and Social Change. *Journal of Neurosurgery*, **130**, 1055-1064. https://doi.org/10.3171/2017.11.jns171500

[19] Park, K.B., Johnson, W.D. and Dempsey, R.J. (2016) Global Neurosurgery: The Unmet Need. *World Neurosurgery*, **88**, 32-35. https://doi.org/10.1016/j.wneu.2015.12.048

[20] Truché, P., Shoman, H., Reddy, C.L., Jumbam, D.T., Ashby, J., Mazhiqi, A., *et al.* (2020) Globalization of National Surgical, Obstetric and Anesthesia Plans: The Critical Link between Health Policy and Action in Global Surgery. *Globalization and Health*, **16**, Article No. 1. https://doi.org/10.1186/s12992-019-0531-5

[21] Kruk, M.E., Gage, A.D., Joseph, N.T., Danaei, G., García-Saisó, S. and Salomon, J.A.

(2018) Mortality Due to Low-Quality Health Systems in the Universal Health Coverage Era: A Systematic Analysis of Amenable Deaths in 137 Countries. *The Lancet*, **392**, 2203-2212. https://doi.org/10.1016/s0140-6736(18)31668-4

[22] Gottschalk, S.B., Wood, D., Devries, S., Wallis, L.A. and Bruijns, S. (2006) The Cape Triage Score: A New Triage System South Africa. Proposal from the Cape Triage Group. *Emergency Medicine Journal*, **23**, 149-153. https://doi.org/10.1136/emj.2005.028332

[23] Molyneux, E., Ahmad, S. and Robertson, A. (2006) Improving Triage and Emergency Care for Children Reduces Inpatient Mortality in a Resource-Constrained Setting. *Bulletin of the World Health Organization*, **84**, 314-319. https://doi.org/10.2471/blt.04.019505

[24] Twomey, M., Wallis, L.A. and Myers, J.E. (2007) Limitations in Validating Emergency Department Triage Scales. *Emergency Medicine Journal*, **24**, 477-479. https://doi.org/10.1136/emj.2007.046383

[25] Collins, G.S., Reitsma, J.B., Altman, D.G. and Moons, K.G.M. (2015) Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD): The TRIPOD Statement. *BMJ*, **350**, g7594. https://doi.org/10.1136/bmj.g7594

[26] Moons, K.G.M., Altman, D.G., Reitsma, J.B., Ioannidis, J.P.A., Macaskill, P., Steyerberg, E.W., *et al.* (2015) Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD): Explanation and Elaboration. *Annals of Internal Medicine*, **162**, W1-W73. https://doi.org/10.7326/m14-0698

[27] Collins, G.S., Moons, K.G.M., Dhiman, P., Riley, R.D., Beam, A.L., Van Calster, B., *et al.* (2024) TRIPOD + AI Statement: Updated Guidance for Reporting Clinical Prediction Models That Use Regression or Machine Learning Methods. *BMJ*, **385**, e078378. https://doi.org/10.1136/bmj-2023-078378

[28] Jennett, B. and Bond, M. (1975) Assessment of Outcome after Severe Brain Damage: A Practical Scale. *The Lancet*, **305**, 480-484. https://doi.org/10.1016/s0140-6736(75)92830-5

[29] Cordoba, G., Schwartz, L., Woloshin, S., Bae, H. and Gotzsche, P.C. (2010) Definition, Reporting, and Interpretation of Composite Outcomes in Clinical Trials: Systematic Review. *BMJ*, **341**, c3920. https://doi.org/10.1136/bmj.c3920

[30] Freemantle, N., Calvert, M., Wood, J., Eastaugh, J. and Griffin, C. (2003) Composite Outcomes in Randomized Trials: Greater Precision but with Greater Uncertainty? *JAMA*, **289**, 2554-2559. https://doi.org/10.1001/jama.289.19.2554

[31] Ferreira-González, I., Permanyer-Miralda, G., Domingo-Salvany, A., Busse, J.W., Heels-Ansdell, D., Montori, V.M., *et al.* (2007) Problems with Use of Composite End Points in Cardiovascular Trials: Systematic Review of Randomised Controlled Trials. *BMJ*, **334**, Article No. 786. https://doi.org/10.1136/bmj.39136.682083.ae

[32] Peduzzi, P., Concato, J., Kemper, E., Holford, T.R. and Feinstein, A.R. (1996) A Simulation Study of the Number of Events per Variable in Logistic Regression Analysis. *Journal of Clinical Epidemiology*, **49**, 1373-1379. https://doi.org/10.1016/s0895-4356(96)00236-3

[33] Riley, R.D., Ensor, J., Snell, K.I.E., Harrell, F.E., Martin, G.P., Reitsma, J.B., *et al.* (2020) Calculating the Sample Size Required for Developing a Clinical Prediction Model. *BMJ*, **368**, m441. https://doi.org/10.1136/bmj.m441

[34] van Smeden, M., Moons, K.G., de Groot, J.A., Collins, G.S., Altman, D.G., Eijkemans, M.J., *et al.* (2019) Sample Size for Binary Logistic Prediction Models: Beyond Events

per Variable Criteria. *Statistical Methods in Medical Research*, **28**, 2455-2474. https://doi.org/10.1177/0962280218784726

[35] Royston, P., Altman, D.G. and Sauerbrei, W. (2006) Dichotomizing Continuous Predictors in Multiple Regression: A Bad Idea. *Statistics in Medicine*, **25**, 127-141. https://doi.org/10.1002/sim.2331

[36] Rubin, D.B. (1987) Multiple Imputation for Nonresponse in Surveys. Wiley. https://doi.org/10.1002/9780470316696

[37] van Buuren, S. and Groothuis-Oudshoorn, K. (2011) MICE: Multivariate Imputation by Chained Equations in *R*. *Journal of Statistical Software*, **45**, 1-67. https://doi.org/10.18637/jss.v045.i03

[38] White, I.R., Royston, P. and Wood, A.M. (2011) Multiple Imputation Using Chained Equations: Issues and Guidance for Practice. *Statistics in Medicine*, **30**, 377-399. https://doi.org/10.1002/sim.4067

[39] Tibshirani, R. (1996) Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society Series B*: *Statistical Methodology*, **58**, 267-288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

[40] Meinshausen, N. and Bühlmann, P. (2010) Stability Selection. *Journal of the Royal Statistical Society Series B*: *Statistical Methodology*, **72**, 417-473. https://doi.org/10.1111/j.1467-9868.2010.00740.x

[41] Firth, D. (1993) Bias Reduction of Maximum Likelihood Estimates. *Biometrika*, **80**, 27-38. https://doi.org/10.1093/biomet/80.1.27

[42] Heinze, G. and Schemper, M. (2002) A Solution to the Problem of Separation in Logistic Regression. *Statistics in Medicine*, **21**, 2409-2419. https://doi.org/10.1002/sim.1047

[43] Harrell, F.E., Lee, K.L. and Mark, D.B. (1996) Multivariable Prognostic Models: Issues in Developing Models, Evaluating Assumptions and Adequacy, and Measuring and Reducing Errors. *Statistics in Medicine*, **15**, 361-387. https://doi.org/10.1002/(sici)1097-0258(19960229)15:4<361::aid-sim168>3.0.co;2-4

[44] Harrell, F.E. Jr. (2015) Regression Modeling Strategies: With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis. 2nd Edition, Springer.

[45] Collins, G.S., Dhiman, P., Ma, J., Schlussel, M.M., Archer, L., Van Calster, B., *et al.* (2024) Evaluation of Clinical Prediction Models (Part 1): From Development to External Validation. *BMJ*, **384**, e074819. https://doi.org/10.1136/bmj-2023-074819

[46] Steyerberg, E.W. and Vergouwe, Y. (2014) Towards Better Clinical Prediction Models: Seven Steps for Development and an ABCD for Validation. *European Heart Journal*, **35**, 1925-1931. https://doi.org/10.1093/eurheartj/ehu207

[47] Austin, P.C. and Steyerberg, E.W. (2014) Graphical Assessment of Internal and External Calibration of Logistic Regression Models by Using Loess Smoothers. *Statistics in Medicine*, **33**, 517-535. https://doi.org/10.1002/sim.5941

[48] Steyerberg, E.W. (2019) Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating. 2nd Edition, Springer.

[49] Van Calster, B., McLernon, D.J., van Smeden, M., Wynants, L. and Steyerberg, E.W. (2019) Calibration: The Achilles Heel of Predictive Analytics. *BMC Medicine*, **17**, Article No. 230. https://doi.org/10.1186/s12916-019-1466-7

[50] Reith, F.C.M., Van den Brande, R., Synnot, A., Gruen, R. and Maas, A.I.R. (2016) The Reliability of the Glasgow Coma Scale: A Systematic Review. *Intensive Care Medicine*, **42**, 3-15. https://doi.org/10.1007/s00134-015-4124-3

[51] Teasdale, G. and Jennett, B. (1974) Assessment of Coma and Impaired Consciousness: A Practical Scale. *The Lancet*, **304**, 81-84. https://doi.org/10.1016/s0140-6736(74)91639-0

[52] DeLong, E.R., DeLong, D.M. and Clarke-Pearson, D.L. (1988) Comparing the Areas under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach. *Biometrics*, **44**, 837-845. https://doi.org/10.2307/2531595

[53] Vickers, A.J. and Elkin, E.B. (2006) Decision Curve Analysis: A Novel Method for Evaluating Prediction Models. *Medical Decision Making*, **26**, 565-574. https://doi.org/10.1177/0272989x06295361

[54] Vickers, A.J., Van Calster, B. and Steyerberg, E.W. (2016) Net Benefit Approaches to the Evaluation of Prediction Models, Molecular Markers, and Diagnostic Tests. *BMJ*, **352**, i6. https://doi.org/10.1136/bmj.i6

[55] Wynants, L., van Smeden, M., McLernon, D.J., Timmerman, D., Steyerberg, E.W. and Van Calster, B. (2019) Three Myths about Risk Thresholds for Prediction Models. *BMC Medicine*, **17**, Article No. 192. https://doi.org/10.1186/s12916-019-1425-3

[56] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., *et al.* (2011) Scikit-Learn: Machine Learning in Python. *Journal of Machine Learning Research*, **12**, 2825-2830. https://jmlr.org/papers/v12/pedregosa11a.html

[57] Corsini Campioli, C., Castillo Almeida, N.E., O'Horo, J.C., Esquer Garrigos, Z., Wilson, W.R., Cano, E., *et al.* (2021) Bacterial Brain Abscess: An Outline for Diagnosis and Management. *The American Journal of Medicine*, **134**, 1210-1217.e2. https://doi.org/10.1016/j.amjmed.2021.05.027

[58] Widdrington, J.D., Bond, H., Schwab, U., Price, D.A., Schmid, M.L., McCarron, B., *et al.* (2018) Pyogenic Brain Abscess and Subdural Empyema: Presentation, Management, and Factors Predicting Outcome. *Infection*, **46**, 785-792. https://doi.org/10.1007/s15010-018-1182-9

[59] Debray, T.P.A., Vergouwe, Y., Koffijberg, H., Nieboer, D., Steyerberg, E.W. and Moons, K.G.M. (2015) A New Framework to Enhance the Interpretation of External Validation Studies of Clinical Prediction Models. *Journal of Clinical Epidemiology*, **68**, 279-289. https://doi.org/10.1016/j.jclinepi.2014.06.018

[60] Van Calster, B., Wynants, L., Timmerman, D., Steyerberg, E.W. and Collins, G.S. (2019) Predictive Analytics in Health Care: How Can We Know It Works? *Journal of the American Medical Informatics Association*, **26**, 1651-1654. https://doi.org/10.1093/jamia/ocz130

[61] Riley, R.D., Ensor, J., Snell, K.I.E., Debray, T.P.A., Altman, D.G., Moons, K.G.M., *et al.* (2016) External Validation of Clinical Prediction Models Using Big Datasets from E-Health Records or IPD Meta-Analysis: Opportunities and Challenges. *BMJ*, **353**, i3140. https://doi.org/10.1136/bmj.i3140

[62] Sperrin, M., Riley, R.D., Collins, G.S. and Martin, G.P. (2022) Targeted Validation: Validating Clinical Prediction Models in Their Intended Population and Setting. *Diagnostic and Prognostic Research*, **6**, Article No. 24. https://doi.org/10.1186/s41512-022-00136-8

[63] Bodilsen, J., D'Alessandris, Q.G., Humphreys, H., Iro, M.A., Klein, M., Last, K., *et al.* (2024) European Society of Clinical Microbiology and Infectious Diseases Guidelines on Diagnosis and Treatment of Brain Abscess in Children and Adults. *Clinical Microbiology and Infection*, **30**, 66-89. https://doi.org/10.1016/j.cmi.2023.08.016

[64] Brook, I. (2017) Microbiology and Treatment of Brain Abscess. *Journal of Clinical Neuroscience*, **38**, 8-12. https://doi.org/10.1016/j.jocn.2016.12.035

[65] Reilly, B.M. and Evans, A.T. (2006) Translating Clinical Research into Clinical Prac-

tice: Impact of Using Prediction Rules to Make Decisions. *Annals of Internal Medicine*, **144**, 201-209. https://doi.org/10.7326/0003-4819-144-3-200602070-00009

[66] Debas, H.T., Donkor, P., Gawande, A., Jamison, D.T., Kruk, M.E. and Mock, C.N. (2015) Essential Surgery. Disease Control Priorities, Vol. 1, 3rd Edition, The World Bank.

[67] Alkire, B.C., Raykar, N.P., Shrime, M.G., Weiser, T.G., Bickler, S.W., Rose, J.A., *et al.* (2015) Global Access to Surgical Care: A Modelling Study. *The Lancet Global Health*, **3**, e316-e323. https://doi.org/10.1016/s2214-109x(15)70115-4

[68] Farmer, P.E. and Kim, J.Y. (2008) Surgery and Global Health: A View from Beyond the OR. *World Journal of Surgery*, **32**, 533-536. https://doi.org/10.1007/s00268-008-9525-9

[69] Corley, J., Lepard, J., Barthélemy, E., Ashby, J.L. and Park, K.B. (2019) Essential Neurosurgical Workforce Needed to Address Neurotrauma in Low- and Middle-Income Countries. *World Neurosurgery*, **123**, 295-299. https://doi.org/10.1016/j.wneu.2018.12.042

[70] Sonderman, K.A., Citron, I. and Meara, J.G. (2018) National Surgical, Obstetric, and Anesthesia Planning in the Context of Global Surgery: The Way Forward. *JAMA Surgery*, **153**, 959-960. https://doi.org/10.1001/jamasurg.2018.2440

[71] Citron, I., Chokotho, L. and Lavy, C. (2015) Prioritisation of Surgery in the National Health Strategic Plans of Africa: A Systematic Review. *The Lancet*, **385**, S53. https://doi.org/10.1016/s0140-6736(15)60848-0

[72] Bodilsen, J., Dalager-Pedersen, M., van de Beek, D., Brouwer, M.C. and Nielsen, H. (2020) Incidence and Mortality of Brain Abscess in Denmark: A Nationwide Population-Based Study. *Clinical Microbiology and Infection*, **26**, 95-100. https://doi.org/10.1016/j.cmi.2019.05.016

[73] Bodilsen, J., Duerlund, L.S., Mariager, T., Brandt, C.T., Petersen, P.T., Larsen, L., *et al.* (2023) Clinical Features and Prognostic Factors in Adults with Brain Abscess. *Brain*, **146**, 1637-1647. https://doi.org/10.1093/brain/awac312

[74] Hakan, T., Ceran, N., Erdem, İ., Berkman, M.Z. and Göktaş, P. (2006) Bacterial Brain Abscesses: An Evaluation of 96 Cases. *Journal of Infection*, **52**, 359-366. https://doi.org/10.1016/j.jinf.2005.07.019

[75] Tseng, J.H. and Tseng, M.Y. (2006) Brain Abscess in 142 Patients: Factors Influencing Outcome and Mortality. *Surgical Neurology*, **65**, 557-562. https://doi.org/10.1016/j.surneu.2005.09.029

[76] Xiao, F., Tseng, M.Y., Teng, L.J., Tseng, H.M. and Tsai, J.C. (2005) Brain Abscess: Clinical Experience and Analysis of Prognostic Factors. *Surgical Neurology*, **63**, 442-449. https://doi.org/10.1016/j.surneu.2004.08.093

[77] Zhang, C., Hu, L., Wu, X., Hu, G., Ding, X. and Lu, Y. (2014) A Retrospective Study on the Aetiology, Management, and Outcome of Brain Abscess in an 11-Year, Single-Centre Study from China. *BMC Infectious Diseases*, **14**, Article No. 311. https://doi.org/10.1186/1471-2334-14-311

[78] Mamelak, A.N., Mampalam, T.J., Obana, W.G. and Rosenblum, M.L. (1995) Improved Management of Multiple Brain Abscesses: A Combined Surgical and Medical Approach. *Neurosurgery*, **36**, 76-86. https://doi.org/10.1227/00006123-199501000-00010

[79] Nathoo, N., Nadvi, S.S., van Dellen, J.R. and Gouws, E. (1999) Intracranial Subdural Empyemas in the Era of Computed Tomography: A Review of 699 Cases. *Neurosurgery*, **44**, 529-535. https://doi.org/10.1097/00006123-199903000-00055

[80] Helweg-Larsen, J., Astradsson, A., Richhall, H., Erdal, J., Laursen, A. and Brennum, J. (2012) Pyogenic Brain Abscess, a 15 Year Survey. *BMC Infectious Diseases*, **12**, Article No. 332. https://doi.org/10.1186/1471-2334-12-332

[81] Tunkel, A.R., Hasbun, R., Bhimraj, A., Byers, K., Kaplan, S.L., Scheld, W.M., *et al.* (2017) 2017 Infectious Diseases Society of America's Clinical Practice Guidelines for Healthcare-Associated Ventriculitis and Meningitis. *Clinical Infectious Diseases*, **64**, e34-e65. https://doi.org/10.1093/cid/ciw861

[82] Warman, A., Kalluri, A.L. and Azad, T.D. (2023) Machine Learning Predictive Models in Neurosurgery: An Appraisal Based on the TRIPOD Guidelines. Systematic Review. *Neurosurgical Focus*, **54**, E8. https://doi.org/10.3171/2023.3.focus2386

[83] Riley, R.D., van Calster, B. and Collins, G.S. (2024) Developing and Validating Clinical Prediction Models. *BMJ*, **385**, e078379.

[84] Finlayson, S.G., Subbaswamy, A., Singh, K., Bowers, J., Kupke, A., Zittrain, J., *et al.* (2021) The Clinician and Dataset Shift in Artificial Intelligence. *The New England Journal of Medicine*, **385**, 283-286. https://doi.org/10.1056/nejmc2104626

[85] Riley, R.D., Archer, L., Snell, K.I.E., Ensor, J., Dhiman, P., Martin, G.P., *et al.* (2024) Evaluation of Clinical Prediction Models (Part 2): How to Undertake an External Validation Study. *BMJ*, **384**, e074820. https://doi.org/10.1136/bmj-2023-074820

[86] Riley, R.D., Debray, T.P.A., Collins, G.S., Archer, L., Ensor, J., van Smeden, M., *et al.* (2021) Minimum Sample Size for External Validation of a Clinical Prediction Model with a Binary Outcome. *Statistics in Medicine*, **40**, 4230-4251. https://doi.org/10.1002/sim.9025

[87] Riley, R.D., Snell, K.I.E., Archer, L., Ensor, J., Debray, T.P.A., van Calster, B., *et al.* (2024) Evaluation of Clinical Prediction Models (Part 3): Calculating the Sample Size Required for an External Validation Study. *BMJ*, **384**, e074821. https://doi.org/10.1136/bmj-2023-074821

TRIAGE-SIC—Consolidated Supplementary Material (S1 - S11)

Development and Internal Validation of the First African Emergency Triage Score for Intracranial Suppurative Infections

*Single consolidated file. The numbering below (S1 - S11, **Tables S7-1-S7-6**, **Figures S8-1-S8-3**) matches the in-text citations of the main manuscript exactly, so that every callout resolves to an item here.*

To preserve the integrity of the manuscript, nothing in this file was invented. Each item carries one of the following labels:

- —computed deterministically from the coefficients, tables or equation already reported in the main text; to be confirmed by the authors against the original analysis.
- —a reporting checklist filled by mapping each item to where it is addressed; page numbers to be verified by the authors.
- —requires the original protocol document, the anonymised dataset, the per-resample outputs, the analysis code or a design deliverable; it cannot be reconstructed from published values without fabricating data.

S1. Pre-specified statistical protocol

The protocol below restates the analysis plan as described in the manuscript. The manuscript states that this plan was fixed before any performance figure was examined; the authors alone can attest to its pre-specification and should substitute the original document.

1) Targets. One composite primary target (critical trajectory) and four secondary targets analysed separately: critical-care use, surgery within 24 h, poor functional outcome (GOS ≤ 3) and in-hospital death.

2) Candidate pool. Twenty-four candidate predictors pre-specified under a strict availability rule (variables obtainable in the admissions room without specialised resources). See S4.

3) Missing data. Multiple imputation by chained equations, 10 imputed datasets; univariable and selection analyses run across the imputations (Rubin pooling).

4) Variable selection. LASSO applied across the 10 imputed datasets and across 1000 bootstrap resamples; a predictor was retained if selected in ≥60% of resamples (stability threshold fixed a priori).

5) Final model. Firth penalised logistic regression on the retained predictors; coefficients pooled by Rubin's rules.

6) Point conversion. Points = β/0.642, rounded to the nearest integer, capped at 4 (unit fixed a priori).

7) Internal validation. 1000 bootstrap resamples encompassing the whole selection process; optimism-corrected AUC, calibration-in-the-large, calibration slope, Brier and scaled Brier.

8) Calibration. Smoothed calibration curve with 95% envelope, calibration by deciles, Spiegelhalter and Hosmer-Lemeshow tests (the latter secondary).

9) Comparator. Continuous Glasgow Coma Scale, pre-specified for every tar-

get; ROC comparison by the DeLong test.

10) Two pre-specified negative analyses. Absence of discrimination of the score for death, and absence of discrimination of the continuous GCS for critical-care use—reported on the same footing as the positive results.

S2. TRIPOD checklist (2015)

Adherence mapping; *item topics are paraphrased* (*the official checklist wording is not reproduced here*).

TRIPOD area	Addressed	Where in the manuscript
Title—identifies study as developing a prediction model	Yes	Title
Abstract—objectives, design, setting, participants, outcome, results	Yes	Abstract
Background and objectives	Yes	Introduction
Source of data (design)	Yes	§2.1 (TRIPOD type 1b)
Participants (setting, eligibility, dates)	Yes	§2.2
Outcome definition	Yes	§2.3 (critical trajectory + secondary targets)
Predictors (definitions, blinding)	Yes	§2.4, Table 5 note, S4
Sample size	Partial	§2.2 (150 admissions/148 patients)
Missing data handling	Yes	§2.6 (multiple imputation, 10 sets)
Statistical analysis—selection, model, validation	Yes	§2.7 - §2.9
Risk groups (bands)	Yes	Table 5, §3.8
Development vs validation	Yes	§2.1 (internal validation by resampling)
Participants—flow, characteristics	Yes	Figure 1, Table 1
Model specification (full model, coefficients)	Yes	Table 4, S5
Model performance (discrimination, calibration)	Yes	§3.6 - §3.7, Table 6 and Table 7, Figure 3 and Figure 4
Model updating	Not applicable	-
Limitations	Yes	§Discussion
Interpretation/implications	Yes	§Discussion
Supplementary information/data availability	Yes	"Availability of data and code" (S1 - S11)

S3. TRIPOD + AI checklist (2024)

The AI/penalised-learning components that this extension asks to report, and where they already appear, are: penalised (LASSO/Firth) methods and hyper-pa-

rameter handling (§2.7 - §2.8, Table 4); resampling-based internal validation with optimism correction (§2.9, Table 6); absence of external validation stated explicitly (§2.1, Discussion); code availability (S9); fairness/subgroup considerations (§Discussion). The authors should transcribe these onto the official TRIPOD + AI checklist with item numbers and page references.

S4. Variable dictionary

Final score items (operational definitions, from §3.5):

- **Scalp swelling or cellulitis**—inflammatory involvement of the soft tissues of the scalp overlying the suppuration.
- **Mass effect**—displacement of adjacent structures, effacement of the cortical sulci or ventricular compression on unenhanced CT.
- **Impaired consciousness**—any impairment of the level of consciousness at presentation, whatever its depth.
- **Known immunosuppression**—HIV infection, prolonged corticosteroid therapy, active malignancy, or other documented immunosuppression.
- **Hydrocephalus**—active ventricular dilatation on CT.
- **Raised intracranial pressure syndrome**—headache, vomiting, papilloedema or a combination of these signs.

Pre-specified candidate pool (24 candidates). Individually named in Table 2 and Table 3: mass effect, raised ICP syndrome, impaired consciousness, size ≥ 30 mm, GCS ≤ 8, midline shift ≥ 5 mm, scalp swelling/cellulitis, hydrocephalus, GCS 9 - 12, postoperative origin, subdural empyema, immunosuppression, anisocoria, referred patient, rural residence, multiple collections, age < 5 years, meningeal syndrome.

S5. Final model, point scale and score-to-probability equation

Penalised (Firth) logistic model; endpoint = critical trajectory (composite); n = 148 patients/150 admissions; event proportion 97/150 (64.7%).

Model coefficients (Table 4): intercept −2.096; scalp swelling/cellulitis β = 3.270 (4 pts); mass effect β = 1.795 (3 pts); impaired consciousness β = 1.463 (2 pts); immunosuppression β = 1.399 (2 pts); hydrocephalus β = 1.279 (2 pts); raised ICP syndrome β = 0.642 (1 pt).

Point scale: Points = β/0.642, rounded to the nearest integer, capped at 4.

Score-to-probability mapping: logit(p) = −2.213 + 0.681 × Score, so p = 1/[1 + e^−(−2.213 + 0.681 × Score)]. Total score range: 0 - 14 points.

S6. Sensitivity analyses

- **Exclusion of patients operated on within 24 h** (removes 41 admissions → 109 admissions, 65 events): apparent discrimination 0.852 (vs 0.858 in the whole cohort). Calibration-in-the-large 0.002 and calibration slope 1.000, both from re-estimation of the score-to-probability relation in the 109 admissions (therefore not a transportability test). The band gradient was preserved: 13.0%, 52.8%, 79.4% and 100% for 0 - 1, 2 - 4, 5 - 6 and ≥7 points. A de novo LASSO in this subgroup retained seven variables, including five of the six final items; immunosuppression was no longer retained, whereas GCS ≤ 8 and midline shift ≥ 5 mm were.

- **Complete-case analysis** and **analysis restricted to each patient's first admission**: did not change the variables selected.

Supplementary Table S7-1. Worked score-to-probability examples

Predicted probability of a critical trajectory from logit(p) = −2.213 + 0.681 × Score. *The full table for all attainable totals is given in Table S7-2.*

Worked examples. A patient scoring 3 points (impaired consciousness [2] + raised ICP [1]) has a predicted risk of 45.8% (Intermediate/yellow). A patient scoring 6 points (mass effect [3] + impaired consciousness [2] + raised ICP [1]) has a predicted risk of 86.7% (High/orange).

Supplementary Table S7-2. Full score-to-probability table (all attainable totals)

Total score	logit(p)	Predicted probability	Band (colour code)
0	−2.213	9.9%	Low (green)
1	−1.532	17.8%	Low (green)
2	−0.851	29.9%	Intermediate (yellow)
3	−0.170	45.8%	Intermediate (yellow)
4	+0.511	62.5%	Intermediate (yellow)
5	+1.192	76.7%	High (orange)
6	+1.873	86.7%	High (orange)
7	+2.554	92.8%	Very high (red)
8	+3.235	96.2%	Very high (red)
9	+3.916	98.0%	Very high (red)
10	+4.597	99.0%	Very high (red)
11	+5.278	99.5%	Very high (red)
12	+5.959	99.7%	Very high (red)
13	+6.640	99.9%	Very high (red)
14	+7.321	99.9%	Very high (red)

Supplementary Table S7-3. Comparison with existing instruments

The manuscript's stated positioning is that no triage score specific to intracranial suppurative infection exists, and that the pre-specified comparator is therefore the continuous Glasgow Coma Scale; the head-to-head comparison by target is reported in Table 7 (ΔAUC and DeLong test). *A formal comparison table against other published neurosurgical/sepsis triage instruments, with their references, is to be supplied by the authors* (*specific instruments and citations are not reconstructable from published values*).

Supplementary Table S7-4. TRIPOD compliance summary

The reporting follows the TRIPOD 2015 checklist for a type-1b development-with-internal-validation study (S2), with the TRIPOD + AI extension for the penalised-learning components (S3). No external validation was performed; this is stated explicitly. Full item-by-item adherence with page numbers is given in S2 and to be finalised for S3.

Supplementary Table S7-5. Comparison by outcome (with vs without critical trajectory)

Continuous variables (Mann-Whitney), critical trajectory vs no critical trajectory: median GCS 12 (IQR 9 - 14) vs 14 (13 - 15), $p < 0.001$; maximum collection size 39.5 mm (33.0 - 47.2) vs 31.0 mm (23.0 - 36.2), $p < 0.001$; midline shift 3.1 mm (1.3 - 4.7) vs 0.0 mm (0.0 - 1.1), $p < 0.001$. Univariable associations for the candidate predictors (events among exposed vs unexposed, OR) are in Table 2. *A complete baseline table stratified by outcome, for all variables, is to be supplied by the authors.*

Supplementary Table S7-6. Calibration by deciles of predicted risk

Decile of predicted risk	Mean predicted probability	Observed proportion of events
1 (lowest)	0.099	20.0%
2 - 9	[*TO BE SUPPLIED BY THE AUTHORS*]	[*TO BE SUPPLIED*]
10 (highest)	0.977	100%

Formal tests (on the apparent model): Hosmer-Lemeshow $C = 5.13$ on 8 df, $p = 0.744$; Spiegelhalter $z = -0.094$, $p = 0.925$. No decile departed from the diagonal beyond its confidence interval.

Supplementary Figure S8-1. LASSO selection stability of the predictors

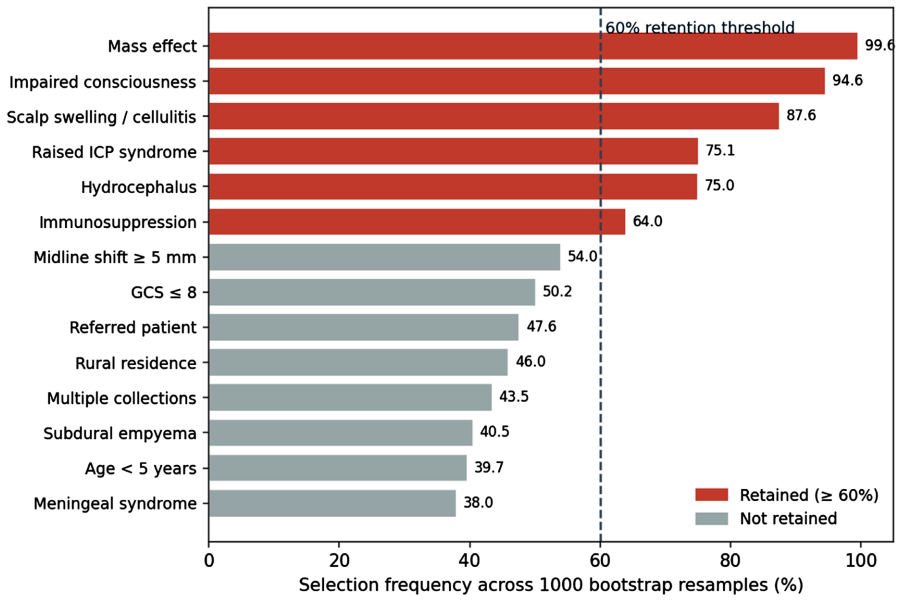

Figure S8-1. LASSO selection stability of the predictors.

Selection frequency of each candidate across 1000 bootstrap resamples; the dashed line marks the pre-specified 60% retention threshold. The six predictors above the threshold form the final score.

Supplementary Figure S8-2. Distribution of the score in the cohort

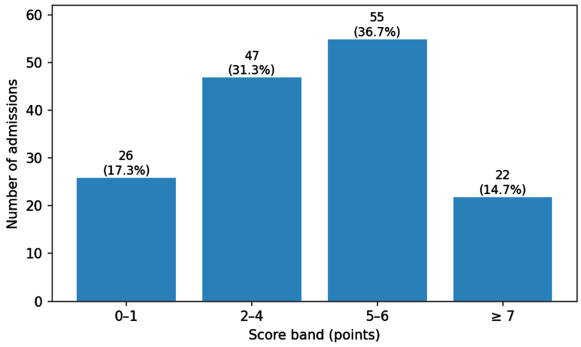

Figure S8-2. Distribution of the score by band (n = 150).

Distribution of the 150 admissions across the four score bands. *A per-score histogram (15 bins, 0 - 14) requires the individual scores and is to be supplied by the authors; the main text reports the modes (6 points, n = 44; 4 points, n = 32; 0 points, n = 20).*

Supplementary. Bootstrap distribution of the corrected AUC

The corrected AUC is 0.788 (95% CI 0.726 - 0.852) and the apparent AUC 0.858; the distribution is described in §3.6 as approximately symmetrical and centred on 0.788, with the apparent value beyond the upper bound of the corrected interval. The histogram itself cannot be reconstructed from published summary values.

Supplementary Figure S8-3. Fagan nomogram

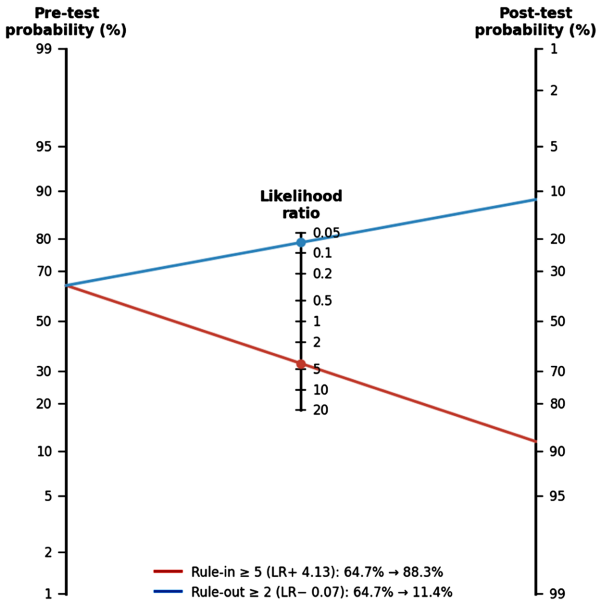

Figure S8-3. Fagan nomogram (prevalence 64.7%).

Fagan nomogram at the cohort prevalence of 64.7%. The rule-in threshold (score ≥ 5; LR+ 4.13) raises the probability of a critical trajectory to 88.3%; the rule-out threshold (score ≥ 2; LR− 0.07) lowers it to 11.4%.

S9. Analysis code

The score and its probability are deterministic and can be reproduced by:

```
POINTS = {"scalp": 4, "mass_effect": 3, "impaired_consciousness": 2,
          "immunosuppression": 2, "hydrocephalus": 2, "raised_icp": 1}

def triage_sic(items: dict) -> dict:
    score = sum(POINTS[k] for k, present in items.items() if present)
    import math
    p = 1 / (1 + math.exp(-(-2.213 + 0.681 * score)))
    band = ("Low" if score <= 1 else "Intermediate" if score <= 4
            else "High" if score <= 6 else "Very high")
    return {"score": score, "probability": round(p, 3), "band": band}
```

The full pipeline that produced the reported estimates (*multiple imputation by chained equations*, *LASSO selection across imputations and 1000 bootstraps*, *Firth penalised regression with Rubin pooling, and optimism-corrected internal validation*) *is the authors' original code and must be attached here* (*Python 3.12*), *with the random seeds, to allow exact reproduction.*

S10. Standalone web calculator

A self-contained, offline calculator (no network connection required) that applies the point scale and the score-to-probability equation and returns the band and suggested management. Delivered alongside this document as *TRIAGE-SIC_calculator.html*.

S11. Graphical abstract

A one-page visual summary (design deliverable). It can be built from the score card (Table 5), the nomogram (Figure 2) and the decision algorithm (Figure 7); the authors should provide or approve the final version.

Abbreviations

Abbreviation	Definition
AUC	Area under the ROC curve
CI	Confidence interval
CITL	Calibration-in-the-large (calibration intercept)
CT	Computed tomography
EVD	External ventricular drain
GCS	Glasgow Coma Scale
GOS	Glasgow Outcome Scale
ICP	Intracranial pressure
IQR	Interquartile range
ISI	Intracranial suppurative infection

Continued

LASSO	*Least absolute shrinkage and selection operator*
LR+/LR−	Positive/negative likelihood ratio
MRI	Magnetic resonance imaging
NSOAP	National surgical, obstetric and anaesthesia plan
OR	Odds ratio