

Measuring the Quality of Neurosurgical Care Through Intracranial Suppurations: A Twenty-Indicator Framework Tested on 150 Cases at Bouaké University Hospital

Serge Konan Yao^{1*}, Landry Stéphane Faozo Teti², Djoko Traoré³, Adama Kouadio⁴, Koffi Jean-Jacques Kouassi⁵, Aderehime Haidara⁶

^{1,2,3,4,5,6}Department of Neurosurgery, Bouaké University Hospital (CHU de Bouaké), Bouaké, Côte d'Ivoire

^{1,2,3,4,5,6}Faculty of Medical Sciences (UFR des Sciences Médicales), Alassane Ouattara University, Bouaké, Côte d'Ivoire

¹ORCID: 0009-0004-6417-4066

²ORCID: 0009-0007-8602-9881

³ORCID: 0009-0008-4138-254X

⁴ORCID: 0009-0008-5081-4826

⁵ORCID: 0009-0000-7433-6701

⁶ORCID: 0009-0001-0412-9982

ABSTRACT: Background. Global surgery indicators measure a country's capacity, not what happens to a patient between the first symptoms and discharge. Intracranial suppurations (ICS) test every link in the chain of care. We used them as a tracer condition to build and test a framework for measuring quality in neurosurgery. Methods. Single-center cohort study, retrospective then prospective, of 150 observations (148 patients) of ICS managed at Bouaké University Hospital (Côte d'Ivoire) from 2016 to 2026. This is a secondary analysis of a cohort whose main study is the subject of a separate manuscript. The ISQF framework comprises twenty indicators distributed according to the Donabedian model: six of structure, eight of process, and six of outcomes. A composite compliance score (ISI-QI) was calculated for each patient. Two Firth penalized logistic regression models (in-hospital mortality; unfavorable outcome, GOS 1–3) combined the ISI-QI, a Glasgow Coma Scale (GCS) score below 9, and an age of at least 65 years. Internal validation relied on 2,000 bootstrap samples with optimism correction. No protocol was registered. The report follows STROBE, SQUIRE 2.0, and TRIPOD+AI; no machine learning method was used. Results. Overall process compliance was 89.0% (878/987 care opportunities; 95% CI: 86.9–90.8). Clinical indicators met or approached their targets. Timeliness indicators fell short: 67.0% of patients with a surgical indication were operated on within 48 hours, and 2.0% were admitted within seven days of symptom onset. The median delay to surgery was 20 days, of which about 95% elapsed before reaching the hospital. Only 34 of 150 patients (22.7%) received antibiotic therapy guided by the antibiogram. The ISI-QI was not associated with mortality (aOR 1.18 per 10-point increment; 95% CI: 0.72–2.06) but was associated with unfavorable outcome (aOR 1.52; 1.04–2.29), solely because of the early-surgery indicator (aOR 3.72; 1.31–11.41): patients operated on early were more severe (GCS < 9 in 18.2% of cases versus 2.6%). Without this indicator, the association disappeared (aOR 1.15; 0.80–1.69). After optimism correction, the ISI-QI added nothing to initial severity for mortality (C statistic 0.646 versus 0.651), and for functional outcome the apparent gain (0.676 to 0.739) disappeared without the early-surgery indicator (0.685). Conclusion. ICS lend themselves to quality measurement from ordinary records. The failure lies before the hospital and in the microbiological chain, not in the operating room. The ISI-QI describes the compliance of an organization; it does not predict the outcome of a patient. From this we derive a ten-indicator dashboard, the ISI-QI-10. Level of evidence. III. Protocol registration. None (secondary analysis of an observational cohort).

KEYWORDS: quality of care; tracer condition; Donabedian model; intracranial suppurations; brain abscess; global neurosurgery; sub-Saharan Africa; performance indicators; confounding by indication

I. INTRODUCTION

In resource-limited health systems, quality of care has long come second to access. The Lancet Global Health Commission on High-Quality Health Systems reversed that order: in low- and middle-income countries, poor-quality care kills more people than lack of access [1]. Neurosurgery illustrates this clearly. Five billion people lack access to safe and timely surgery [2], and more than five million essential neurosurgical procedures go unperformed each year, mostly in sub-Saharan Africa and South-East Asia [3]. Global

Measuring the Quality of Neurosurgical Care Through Intracranial Suppurations: A Twenty-Indicator Framework Tested on 150 Cases at Bouaké University Hospital

surgery indicators track a country's capacity: access within two hours, surgeon density, operative volume, perioperative mortality. They say nothing about what happens to a given patient between the first symptoms and discharge.

The Donabedian model remains the reference framework for assessing quality: structure (the resources), process (what is done to the patient), and outcomes (what becomes of the patient) [4,5]. Its limitation is well known. Process indicators often predict outcomes poorly, and outcomes depend heavily on patient case mix, so that rankings of institutions based on crude rates can be misleading [6,7]. The difficulty can be sidestepped by measuring the system through a single condition whose management draws on many links: this is the tracer condition, proposed by Kessner in 1973 [8]. Hernia repair, for example, has been used to assess access to and quality of elective care in 83 countries [9].

Intracranial suppurations (ICS)—brain abscesses, empyemas, ventriculitis—meet these criteria, with urgency in addition. Prognosis depends on delay. Diagnosis requires cross-sectional imaging. Treatment combines surgery, microbiology, and prolonged antibiotic therapy, and follow-up lasts months [10,11]. Primary care, transport, emergency services, radiology, laboratory, operating theater, intensive care, and rehabilitation each come into play in turn, and the failure of a single link compromises the whole. African series describe long presentation delays and a low microbiological yield [12–17], with an in-hospital mortality of 13 to 17% when reported [13,15,18]. To our knowledge, none has turned these findings into measurable indicators with explicit denominators and targets, or subjected such a system to the methodological requirements of prognostic models [19,20].

A note on vocabulary. This work uses no machine learning method: all models are penalized logistic regressions. We follow TRIPOD+AI because this guideline replaced TRIPOD in 2024 and also covers regression models [20]. We chose a simple model, written out in full and recalculable by hand: with a small sample and limited resources, interpretability and transportability take precedence over algorithmic flexibility.

We pursue four objectives: to build a framework of structure, process, and outcome indicators specific to ICS in resource-limited settings; to apply it to 150 observations to test its feasibility and discriminative power; to examine its measurement properties (relation to clinical outcomes, discrimination, calibration, clinical utility) through a complete internal validation; and to derive a minimal set of indicators for a regional registry.

II. PATIENTS AND METHODS

A. Study Design, Setting, and Population

This is an observational, single-center cohort study with retrospective then prospective recruitment, conducted in the neurosurgery department of Bouaké University Hospital (Côte d'Ivoire) from 2016 to 2026. Bouaké University Hospital is one of the two public neurosurgical centers of the country. It serves several million inhabitants of the center and north. It has a continuously operating CT scanner, a dedicated operating theater, and a bacteriology laboratory that performs aerobic cultures. During the study period, it had neither anaerobic media nor molecular amplification techniques.

We included all patients managed for an ICS documented by clinical findings and imaging: brain abscess, subdural empyema, extradural empyema, ventriculitis, or extracranial suppuration with intracranial involvement or risk of extension. We excluded records whose clinical or imaging data did not allow the diagnosis to be characterized. This work is a secondary analysis: the selection of observations and the STROBE flow diagram appear in the main report on the same cohort. The cohort comprises 150 observations (episodes) for 148 patients. Two patients contributed two episodes each and both survived. The episode is the unit of analysis, without accounting for within-patient correlation; mortality is identical by episode (16/150) and by patient (16/148).

The report follows STROBE for observational studies [21], SQUIRE 2.0 for quality measurement [22], and TRIPOD+AI for the development and internal validation of prognostic models [20].

B. Construction of the Indicator Framework

The framework was built in four steps. A targeted literature review on ICS (guidelines of the European Society of Clinical Microbiology and Infectious Diseases [11], meta-analyses and large cohorts [23–25]) first identified the elements of care on which there is consensus. We then classified them according to Donabedian's three domains, following the usual principles of indicator construction [26,27]. Each candidate indicator had to be clinically relevant, measurable in existing records without additional data collection, equipped with an unambiguous denominator, and sensitive to organizational failures. Finally, we set a target for each process indicator, drawn from guidelines where they exist and, failing that, from an internal consensus within the department.

Content validity therefore rests on the literature and on an internal expert consensus, without a Delphi procedure or external panel. This is an acknowledged limitation at this stage of development. On the other hand, no indicator was retained or discarded on the basis of cohort data, which protects the construction from a selection bias driven by results. The guiding principle was feasibility: we discarded any indicator that required equipment that was absent, dedicated prospective data collection, or a computerized information system, however theoretically interesting.

The framework comprises twenty indicators (Table I and Fig. 1): six structure markers, eight process indicators (P1 to P8), and six outcome indicators. Complete operational definitions appear in Table S1.

Measuring the Quality of Neurosurgical Care Through Intracranial Suppurations: A Twenty-Indicator Framework Tested on 150 Cases at Bouaké University Hospital

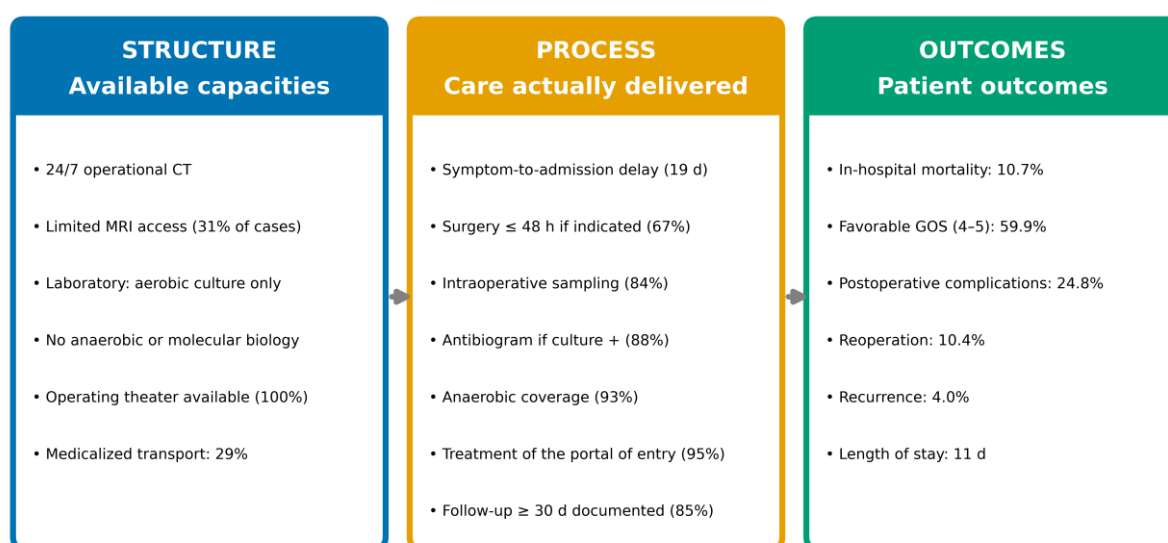


Fig. 1. The Donabedian model applied to intracranial suppurations at Bouaké University Hospital. Values shown are those observed in the cohort (N = 150).

TABLE I. ISQF INDICATOR FRAMEWORK: DOMAINS, INDICATORS, DENOMINATORS, AND TARGETS

Domain	Indicator	Denominator	Proposed Target
Structure	Operational cross-sectional imaging	All patients	24/7 availability
	MRI access	All patients	No reference norm
	Microbiological capacity (aerobic, anaerobic, molecular)	Facility	Anaerobic available
	Access to the operating theater when an indication is established	Patients with a surgical indication	100%
	Medicalized transport to the referral center	Patients with documented transport	No reference norm
	Health facilities passed through before admission	Patients with a documented pathway	≤ 2
Process	P1. Cross-sectional imaging before the therapeutic decision	All patients	$\geq 95\%$
	P2. Surgery within 48 h of admission	Patients with a surgical indication and a documented delay	$\geq 90\%$
	P3. Pus sample sent to the laboratory	Operated patients	$\geq 95\%$
	P4. Antibigram reported	Positive cultures	$\geq 90\%$
	P5. Empirical antibiotic therapy covering anaerobes	All patients	$\geq 95\%$
	P6. Total antibiotic duration ≥ 6 weeks	Survivors at discharge	$\geq 90\%$
	P7. Treatment of the portal of entry	Patients with an identified portal of entry	$\geq 90\%$
	P8. Documented follow-up ≥ 30 days	Survivors at discharge	$\geq 80\%$
Outcomes	All-cause in-hospital mortality	All patients	Risk-adjusted (O:E)
	Favorable outcome (GOS 4-5)	Patients with an evaluable GOS	Risk-adjusted
	Postoperative complications	Operated patients	Surveillance
	Reoperation	Operated patients	Surveillance
	Recurrence	All patients	Surveillance

Measuring the Quality of Neurosurgical Care Through Intracranial Suppurations: A Twenty-Indicator Framework Tested on 150 Cases at Bouaké University Hospital

Domain	Indicator	Denominator	Proposed Target
	Length of stay	All patients	Surveillance

GOS: Glasgow Outcome Scale; O/E: observed/expected ratio. A “no reference norm” target means that the indicator is followed descriptively, no reference value having been established. Operational definitions: Table S1.

C. ISI-QI Score, Outcomes, and Predictors

For each patient, we classified each of the eight process indicators as met, not met, or not applicable. Applicability matters: P4 is assessed only if the culture is positive, P6 and P8 only in survivors. The ISI-QI score of patient i is written:

$$\text{ISI-QI}(i) = 100 \times \left[\frac{\sum I(i,j), j \in A(i)}{|A(i)|} \right]$$

$A(i)$ is the set of indicators applicable to patient i , and $I(i,j)$ equals 1 if indicator j is met, 0 otherwise. The score ranges from 0 to 100%. Writing it in terms of “care opportunities” avoids penalizing a patient for a care that was not due [28]. The overall compliance of the department is the ratio of the sum of indicators met to the sum of applicable indicators over the whole cohort.

The ISI-QI is a descriptive instrument: its value does not depend on the patient’s outcome, and it requires neither calibration nor clinical utility assessment. The models described in Section II-E are, for their part, predictive tools that use it as an explanatory variable; it is they that are subjected to the analyses of discrimination, calibration, and clinical utility.

The outcomes are all-cause in-hospital mortality and functional outcome at the last assessment, measured by the Glasgow Outcome Scale (GOS): favorable if GOS 4–5, unfavorable if GOS 1–3. The GOS was not collected independently: it was assigned from sequelae and level of autonomy, according to a mapping rule, without a standardized time point. The predictors are the ISI-QI (per 10-point increment), a GCS below 9 at admission, and an age of at least 65 years. Delays are expressed in calendar days. Indicator definitions appear in Table S1.

D. Missing Data and Data Quality

We imputed no data. Three situations were distinguished. A non-applicable indicator (the antibiogram in a patient whose culture is sterile, for example) enters neither the numerator nor the denominator. A care not delivered counts as a failure. Missing data make the indicator non-evaluable in that patient: this is the case for the ten operated patients whose surgical delay was not documented, excluded from the denominator of P2. This choice avoids confusing a care failure with a traceability failure. The proportion of missing data is reported separately.

The missing-data mechanism is probably not entirely random [29]. The operative delay is missing mainly from the oldest and least complete records, which suggests a mechanism dependent on observed variables. The GOS is missing in 8 survivors (5.3% of the cohort) whose follow-up did not allow the outcome to be rated; this missingness might depend on the outcome itself, which is a limitation. A sensitivity analysis with extreme bounds gives a proportion of favorable outcome between 56.7% (85/150, if the 8 patients all had an unfavorable outcome) and 62.0% (93/150, if they all had a favorable outcome), against 59.9% (85/142) in complete cases. We extended it to the unfavorable-outcome model (Section III-H). Because the variables of the analytic core (age, sex, admission delay, GCS, lesion type, therapeutic strategy, vital status, length of hospital stay) are recorded in 100% of cases, we favored complete-case analysis [30]. Multiple imputation would have added more uncertainty than it removed in a sample of this size. The numbers analyzed are indicated for each estimate.

A single investigator collected the data from the paper records of the department. Definitions are thus homogeneous, but inter-rater agreement cannot be estimated. Three checks were applied: an internal consistency check (no operative delay in a non-operated patient, no antibiogram without a positive culture, vital status compatible with the duration of follow-up), a systematic return to the source record for any value outside plausible bounds, and a time-stamped correction log retaining the initial value, the corrected value, and its justification. There was neither independent double entry nor external audit. Partly retrospective recruitment exposes the study to information bias: a care that was not recorded cannot be distinguished from a care that was not delivered.

E. Statistical Analysis

Quantitative variables are described by the median and interquartile range (IQR), qualitative variables by counts and percentages. The 95% confidence intervals of proportions come from the Jeffreys method, better suited than the normal approximation to extreme proportions. We compared groups with Fisher’s exact test (qualitative variables) and the Mann–Whitney test (quantitative variables).

Two prognostic models were specified, one per outcome:

$$\text{logit}(p_{\text{death}}) = \beta_0 + \beta_1 \cdot (\text{ISI-QI}/10) + \beta_2 \cdot I(\text{GCS} < 9) + \beta_3 \cdot I(\text{age} \geq 65)$$

$$\text{logit}(p_{\text{GOS 1-3}}) = \gamma_0 + \gamma_1 \cdot (\text{ISI-QI}/10) + \gamma_2 \cdot I(\text{GCS} < 9) + \gamma_3 \cdot I(\text{age} \geq 65)$$

GCS and age were retained because they were the only two prognostic determinants identified in the main analysis of the same cohort. This choice, made on the same data, precedes resampling. We deliberately limited the number of covariates given the small number of events. Estimation relies on Firth penalized logistic regression, retained because of the small number of events and the

Measuring the Quality of Neurosurgical Care Through Intracranial Suppurations: A Twenty-Indicator Framework Tested on 150 Cases at Bouaké University Hospital

risk of quasi-complete separation: no patient with a GCS below 9 had a favorable outcome, which would prevent maximizing the likelihood in conventional regression [31,32]. Confidence intervals and p-values come from the penalized profile likelihood, more reliable than the Wald approximation in this situation. The coefficients, their intervals, and the intercepts are reported in full (Table V) so that any prediction can be recalculated. These are the unshrunk Firth coefficients.

We compared four specifications for each outcome: severity alone ($GCS < 9$ and $age \geq 65$ years), the ISI-QI alone, the full model, and a sensitivity model in which the score is calculated without the surgical delay indicator P2. The comparison sought to determine whether the score provides prognostic information beyond initial severity.

No sample size calculation was done in advance: the sample is that of the available cohort. According to the criterion of Riley et al. [33], with three predictors, a target shrinkage factor of 0.90, and an assumed Cox–Snell R^2 of 0.05, about 525 patients would be needed. With 150 patients and 16 deaths, the sample is insufficient for the mortality model, whose results are exploratory.

We evaluated the models by bootstrap internal validation: 2,000 samples of the size of the cohort, drawn with replacement. In each, the Firth model is refitted and then evaluated in the bootstrap sample and in the original cohort. The mean optimism, the difference between the two performances, is subtracted from the apparent estimates. Splitting into training and validation samples would have been inefficient with sixteen deaths [34,35]. No variable selection took place in the analysis, but the initial choice of GCS and age, made on the same data, is not resampled: the reported optimism is therefore probably underestimated.

Discrimination is measured by the C statistic, calibration by the calibration intercept and slope, and overall performance by the Brier score [36], reported with that of the null model and in normalized form. We did not use the Hosmer–Lemeshow test, whose power depends too much on sample size [37]. The optimism-corrected slope measures overfitting; it is not applied to the coefficients of Table V. Clinical utility was explored by decision curve analysis [38,39], whose net benefit is written $NB = TP/n - (FP/n) \cdot [p_t/(1 - p_t)]$, TP and FP being the true and false positives and p_t the probability threshold that triggers action. Such an analysis makes sense only if a concrete action follows the crossing of the threshold. For mortality, we retained enhanced postoperative surveillance or admission to critical care, for which a false positive has few consequences, hence a threshold range of 5 to 30%. For functional outcome, the action is close follow-up and early rehabilitation, with an inexpensive false positive, with a range of 20 to 70%.

Risk-adjusted mortality is expressed as the observed/expected (O:E) ratio. Expected deaths come from the severity-only model fitted on the whole cohort, with confidence intervals from the Poisson approximation. We compared the 2016–2020 and 2021–2026 periods on crude rates and O:E ratios. The significance threshold is 0.05, without correction for multiple testing: all analyses are exploratory and hypothesis-generating. Analyses were performed with Python 3.12 (pandas, NumPy, SciPy, statsmodels, Matplotlib) and an implementation of Firth regression written for the study; the code is available from the corresponding author. No analysis protocol was registered.

III. RESULTS

A. Population, Descriptive Data, and Missing Data

We analyzed 150 observations, corresponding to 148 patients (the flow of inclusions appears in the main report). Median age was 26 years (IQR: 14–50) and the sex ratio was 2.3. Another health facility had referred 77 patients (51.3%). Median GCS at admission was 13 (IQR: 10–15); 12.7% of patients had a GCS below 9. Brain abscess accounted for 49% of cases, and 125 patients (83.3%) received combined medical and surgical management. Table II describes the 142 patients whose GOS was evaluable, according to their outcome.

The framework produced 987 evaluable care opportunities, i.e., a median of 7 applicable indicators per patient (range: 3 to 8). A non-operated patient generates no opportunity for P2 and P3, and a deceased patient generates none for P6 or P8. The database was 97.3% complete for the expected values (4,602/4,732 across 34 variables), and 100% complete for the eight variables of the analytic core. Table III lists the variables that were not completely recorded.

TABLE II. PATIENT CHARACTERISTICS BY FUNCTIONAL OUTCOME (N = 142 EVALUABLE)

Variable	Favorable Outcome (n = 85)	Unfavorable Outcome (n = 57)	p
Age (years), median [IQR]	30 [14–50]	25 [15–51]	0.71
Age ≥ 65 years	5 (5.9%)	8 (14.0%)	0.14
Male sex	57 (67.1%)	42 (73.7%)	0.46
Referred from another facility	43 (50.6%)	29 (50.9%)	1.00
Symptom-to-admission delay (days), median [IQR]	18 [15–24]	19 [16–25]	0.18
Facilities passed through, median [IQR]	3 [2–5]	3 [1–4]	0.13
GCS at admission, median [IQR]	14 [13–15]	10 [8–11]	< 0.001
GCS < 9	0 (0.0%)	17 (29.8%)	< 0.001

Measuring the Quality of Neurosurgical Care Through Intracranial Suppurations: A Twenty-Indicator Framework Tested on 150 Cases at Bouaké University Hospital

Variable	Favorable Outcome (n = 85)	Unfavorable Outcome (n = 57)	P
Mass effect	59 (69.4%)	47 (82.5%)	0.12
Maximum lesion size (mm), median [IQR]	32 [29–38]	43 [38–50]	< 0.001
Hydrocephalus	10 (11.8%)	4 (7.0%)	0.40
Brain abscess (isolated form)	50 (58.8%)	19 (33.3%)	0.004
Immunosuppression	7 (8.2%)	6 (10.5%)	0.77
Combined medical and surgical management	66 (77.6%)	54 (94.7%)	0.008
Admission-to-surgery delay (days), median [IQR]	2 [2–3]	2 [0–2]	< 0.001
Pus sample sent	65 (76.5%)	48 (84.2%)	0.30
Positive culture	29 (34.1%)	21 (36.8%)	0.86
ISI-QI score (%), median [IQR]	86 [80–100]	100 [86–100]	0.013
Applicable indicators (n), median [IQR]	7 [6–8]	7 [6–7]	0.78
Postoperative complication	13 (15.3%)	16 (28.1%)	0.09
Length of stay (days), median [IQR]	10 [7–13]	15 [10–17]	< 0.001

Fisher's exact test for qualitative variables, Mann–Whitney test for quantitative variables. GCS: Glasgow Coma Scale; IQR: interquartile range. Postoperative complications are reported relative to the whole of each group (and not to the operated patients only); 29 of the 31 complications concern patients whose GOS is evaluable.

TABLE III. DATA COMPLETENESS: VARIABLES NOT COMPLETELY RECORDED

Variable	Recorded	Missing	Denominator
C-reactive protein	93/150	38.0%	Entire cohort
Facilities passed through before admission	132/150	12.0%	Entire cohort
Admission-to-surgery delay	115/125	8.0%	Operated patients
Leukocytes	138/150	8.0%	Entire cohort
Mode of transport	140/150	6.7%	Entire cohort
Follow-up duration	125/134	6.7%	Survivors at discharge
GOS	142/150	5.3%	Entire cohort
Maximum lesion size	144/150	4.0%	Entire cohort

The 26 other variables analyzed, including the eight of the analytic core, are complete relative to their theoretical denominator. All 34 variables: 4,602/4,732 values recorded (97.3%). A non-applicable variable (the antibiogram without a positive culture, for example) is not counted as missing.

B. Structure Domain

Imaging capacity did not limit management: all 150 patients had a CT scan or an MRI before the therapeutic decision (a partly tautological finding, since the imaging that documents the diagnosis is an inclusion criterion). MRI access was restricted: only 47 patients (31.3%) had one, mostly as a second-line examination. Access to the operating theater was complete: all 125 patients for whom a surgical indication had been established were operated on, with no documented cancellation or postponement for unavailability of the room or team (an unrecorded postponement cannot be excluded in paper records).

The structural limits were concentrated elsewhere. The laboratory had only aerobic cultures, with no anaerobic medium or molecular biology throughout the period, whereas ICS are often polymicrobial with anaerobes. Blood cultures were performed in only 39 of the 132 patients for whom the information existed (29.5%). Only 40 of 140 patients (28.6%) were transported by a medicalized means. The median number of health facilities passed through before admission was 3 (IQR: 2–5), and a quarter of patients had attended five or more (n = 132).

C. Process Domain

Overall compliance with the eight process indicators was 89.0% (878 opportunities met out of 987; 95% CI: 86.9–90.8). This value hides large differences between indicators (Fig. 2).

Four indicators reached their target: cross-sectional imaging before any decision (150/150, 100%), antibiotic duration of at least six weeks in survivors (127/134, 94.8%), treatment of the portal of entry when it was identified (122/129, 94.6%), and documented

Measuring the Quality of Neurosurgical Care Through Intracranial Suppurations: A Twenty-Indicator Framework Tested on 150 Cases at Bouaké University Hospital

follow-up at thirty days in survivors (114/134, 85.1%). Two others came close: empirical anaerobic coverage (139/150, 92.7%; target 95%) and the antibiogram reported if the culture was positive (44/50, 88.0%; target 90%). Intraoperative sampling, performed in 105 of 125 operated patients (84.0%), remained eleven points below the 95% target.

The two timeliness indicators stood out clearly. Among patients with a surgical indication and a documented delay, 77 of 115 (67.0%; 95% CI: 58.0–75.0) were operated on within 48 hours, and 35.7% within 24 hours. Above all, pre-hospital access collapsed: 3 of 150 patients (2.0%) were admitted within seven days of symptom onset, and 22.7% within fourteen days. This indicator is not part of the composite score, because it does not depend on the care delivered by the facility.

At the individual level, the median ISI-QI score was 87.5% (IQR: 83.3–100; range: 50 and 100%). Eight patients (5.3%) had a score below 75%. The distribution showed a clear ceiling effect with a low variance, which limits the ability of the score to distinguish patients.

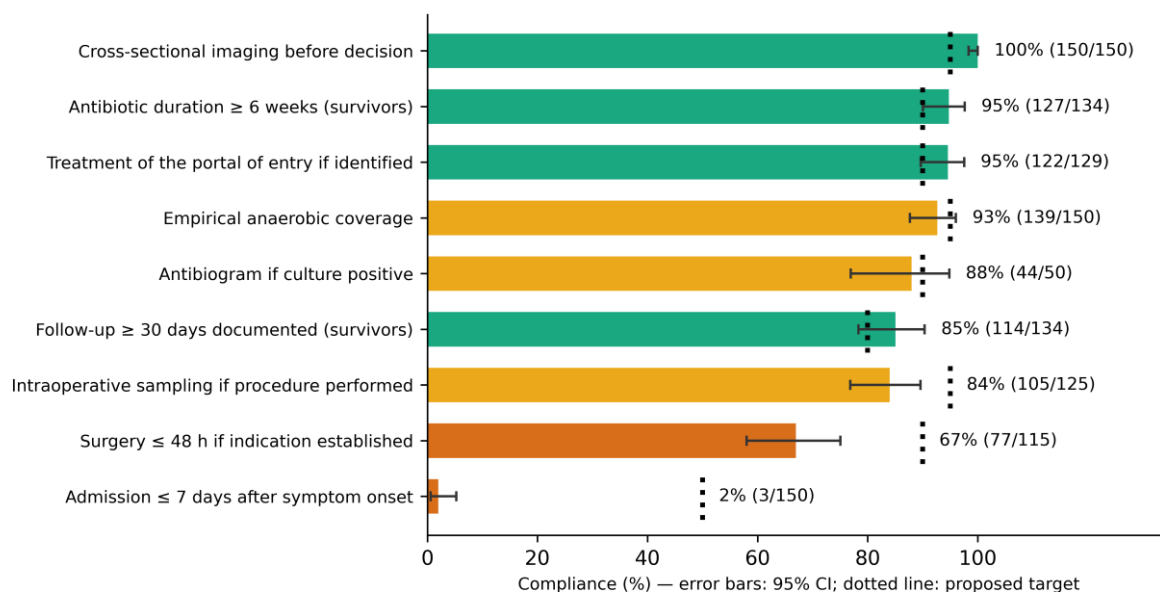


Fig. 2. Compliance with process indicators (variable denominators). Error bars represent the 95% confidence intervals (Jeffreys method); dotted lines indicate the proposed targets. Green: target met; orange: gap below 20 points; red: gap above 20 points.

D. The Microbiological Cascade

Of 150 patients admitted, 116 (77.3%) had a sample sent to the laboratory: 105 of 125 operated patients (84.0%) and 11 of 25 medically treated patients (diagnostic puncture or wound sample). Fifty (33.3%) had a positive culture, 44 (29.3%) a reported antibiogram, and only 34 (22.7%) an antibiotic therapy actually adapted to that antibiogram (Fig. 3).

Thirty-four observations had no sample at all: 14 in exclusive medical treatment and 20 in operated patients (16% of procedures), a loss that stems from operative practice rather than from the laboratory. The main break comes next: only 43.1% of the samples sent yielded a positive culture. This single step accounts for 66 of the 116 cumulative losses, more than all the others combined. The last loss, 10 of 44 patients whose available antibiogram did not lead to a documented adaptation of treatment, reflects a broken loop between laboratory and clinic, not a lack of technical means.

Measuring the Quality of Neurosurgical Care Through Intracranial Suppurations: A Twenty-Indicator Framework Tested on 150 Cases at Bouaké University Hospital

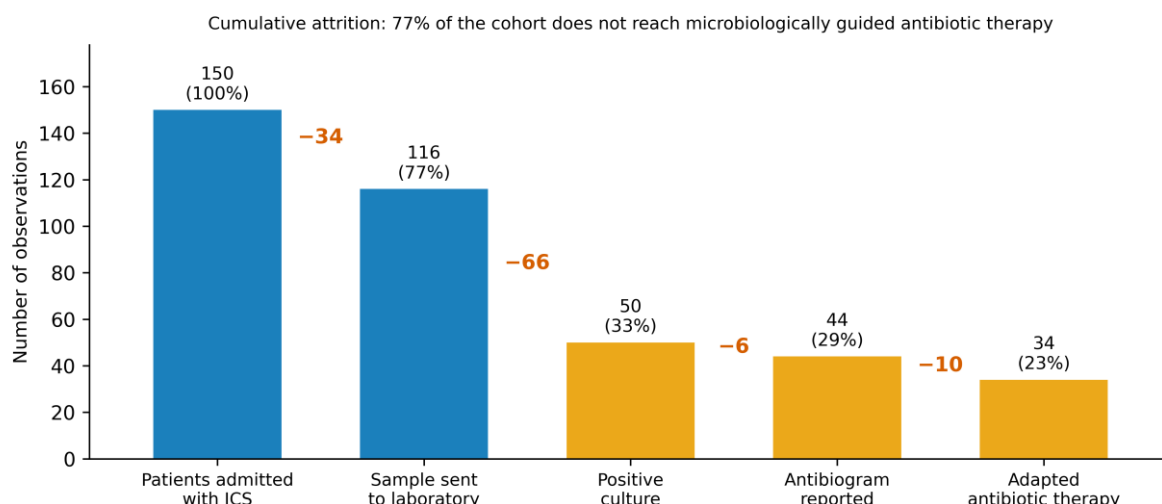


Fig. 3. Neurosurgical quality cascade applied to microbiological documentation (N = 150). Values in orange indicate the number of observations lost at each step (34 without a sample: 14 in exclusive medical treatment, 20 operated patients). The terminal yield, an antibiogram-guided antibiotic therapy, concerns 22.7% of the cohort.

E. Breakdown of Delays

The median delay between symptom onset and admission to the hospital was 19 days (IQR: 15–25), and that between admission and surgery was 2 days (IQR: 1–3). Relative to a median total delay of 20 days, the pre-hospital segment accounts for about 95% of the elapsed time (a ratio of medians, which do not add up; the individual proportion was not calculated) (Fig. 4).

This imbalance dictates the hierarchy of actions. Halving the in-hospital delay, which would already be demanding, would save one day in twenty. Halving the pre-hospital delay would save ten. A strategy focused on the hospital alone would act on the 5% least decisive portion of the pathway.

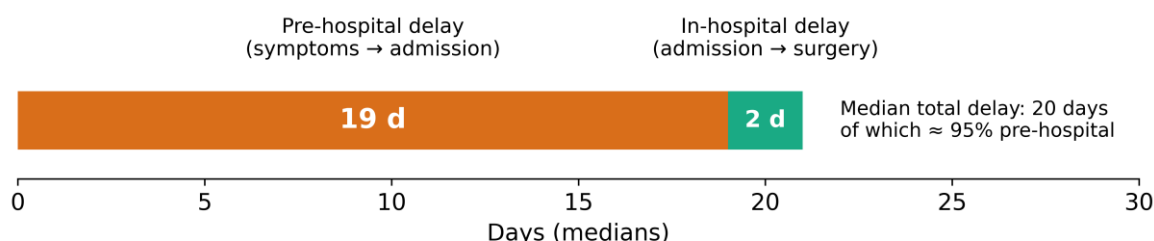


Fig. 4. Breakdown of the total delay to surgery (median values). The pre-hospital segment accounts for about 95% of the total delay.

F. Outcomes Domain

In-hospital mortality was 10.7% (16/150; 95% CI: 6.5–16.4). A favorable outcome (GOS 4 or 5) concerned 59.9% of evaluable patients (85/142; 95% CI: 51.7–67.7). Among the 125 operated patients, 31 had a postoperative complication (24.8%; 95% CI: 17.9–32.9) and 13 a reoperation (10.4%; 6.0–16.6). Recurrence occurred in 6 patients (4.0%; 1.7–8.1). Median length of stay was 11 days (IQR: 8–15).

G. Process and Outcome: Confounding by Indication

The ISI-QI score was associated neither with crude mortality (OR 1.31 per 10-point increment; 95% CI: 0.80–2.28) nor with mortality adjusted for GCS and age (aOR 1.18; 0.72–2.06; $p = 0.54$). In this model, a GCS below 9 (aOR 4.05; 1.16–16.8) and an age of at least 65 years (aOR 4.29; 1.11–19.7) remained the only two determinants of death. All intervals in this paragraph are penalized profile-likelihood intervals.

For functional outcome, the result is baffling: higher compliance went with more unfavorable outcomes, in crude analysis (OR 1.55; 1.11–2.25) as after adjustment (aOR 1.52; 1.04–2.29; $p = 0.031$). Taken literally, this would mean that better care does harm.

Table II sheds light on this association. None of the 85 patients with a favorable outcome had a GCS below 9, against 17 of the 57 patients with an unfavorable outcome (29.8%; $p < 0.001$): this quasi-complete separation justifies the Firth penalization and

Measuring the Quality of Neurosurgical Care Through Intracranial Suppurations: A Twenty-Indicator Framework Tested on 150 Cases at Bouaké University Hospital

explains the width of the odds ratio interval. Lesion size differed markedly between the groups (median 32 versus 43 mm; $p < 0.001$) without being in the model, which leaves residual confounding. Finally, patients with an unfavorable outcome were more often operated on (94.7% versus 77.6%; $p = 0.008$) and faster (median delay 2 days [IQR 0–2] versus 2 days [2–3]; $p < 0.001$), and their ISI-QI was mechanically higher (median 100% versus 86%; $p = 0.013$). The pre-hospital delay, the number of facilities passed through, the sampling rate, and the culture yield did not differ between the groups.

Indicator-by-indicator analysis resolves the ambiguity. None of the indicators P3 to P8 was associated with outcome. All the signal came from P2: surgery within 48 hours went with an unfavorable outcome in crude analysis (OR 4.88; 1.84–14.05) and after adjustment (aOR 3.72; 1.31–11.41; $p = 0.013$). Among patients operated on within 48 hours, 18.2% had a GCS below 9, against 2.6% of those operated on later (14/77 versus 1/38; $p = 0.019$, Fisher's exact test; a single patient with a GCS < 9 in the second group, hence a very imprecise estimate). The earliness of the intervention is not independent of severity: it is its expression. The path linking severity to the operative decision and then to outcome remains open, and its effect masks the specific effect of surgical earliness, whose direction remains undetermined and could be protective.

Without indicator P2, the association between the score and unfavorable outcome disappears (aOR 1.15; 0.80–1.69; $p = 0.44$), as does that with mortality (aOR 1.07; 0.64–1.86; $p = 0.80$). Table IV summarizes these estimates.

TABLE IV. ASSOCIATIONS BETWEEN PROCESS COMPLIANCE AND CLINICAL OUTCOME

Exposure	Outcome	Crude OR [Profile 95% CI]	aOR [Profile 95% CI]*
ISI-QI score (per 10 points)	Death	1.31 [0.80–2.28]	1.18 [0.72–2.06]
ISI-QI score (per 10 points)	GOS 1–3	1.55 [1.11–2.25]	1.52 [1.04–2.29]
ISI-QI score excluding P2 (per 10 points)	GOS 1–3	1.13 [0.82–1.61]	1.15 [0.80–1.69]
ISI-QI score excluding P2 (per 10 points)	Death	—	1.07 [0.64–1.86]
Surgery ≤ 48 h (P2)	GOS 1–3	4.88 [1.84–14.05]	3.72 [1.31–11.41]
Surgery ≤ 48 h (P2)	Death	—	1.49 [0.29–8.52]
Intraoperative sampling (P3)	GOS 1–3	—	1.05 [0.31–3.81]
Antibiogram reported (P4)	GOS 1–3	—	0.29 [0.03–1.84]
Anaerobic coverage (P5)	GOS 1–3	—	1.15 [0.22–6.54]
Antibiotic therapy ≥ 6 weeks (P6)	GOS 1–3	—	0.93 [0.18–4.66]
Treatment of the portal of entry (P7)	GOS 1–3	—	1.10 [0.18–6.68]
Follow-up ≥ 30 days (P8)	GOS 1–3	—	2.28 [0.54–10.66]

*Adjusted for GCS < 9 at admission and age ≥ 65 years (Firth penalized logistic regression). OR: odds ratio; aOR: adjusted odds ratio; GOS: Glasgow Outcome Scale. Confidence intervals are obtained by penalized profile likelihood.

H. Prognostic Models and Internal Validation

Table V gives the complete coefficients of the two models, intercepts included, which allows any prediction to be recalculated. These are the unshrunk Firth coefficients. The equations are written:

$$\hat{p}_{\text{death}} = 1 / \{ 1 + \exp[-(-4.132 + 0.166 \cdot (\text{ISI-QI}/10) + 1.399 \cdot \text{I}(\text{GCS} < 9) + 1.456 \cdot \text{I}(\text{age} \geq 65))] \}$$

$$\hat{p}_{\text{GOS 1-3}} = 1 / \{ 1 + \exp[-(-4.552 + 0.419 \cdot (\text{ISI-QI}/10) + 4.174 \cdot \text{I}(\text{GCS} < 9) + 0.708 \cdot \text{I}(\text{age} \geq 65))] \}$$

TABLE V. COEFFICIENTS OF THE PROGNOSTIC MODELS (FIRTH PENALIZED LOGISTIC REGRESSION)

Parameter	Coefficient β (Firth, Unshrunk)	Profile 95% CI of β	OR
Model 1 — in-hospital mortality (n = 150; 16 events)			
Intercept β_0	−4.132	[−8.36; −1.01]	—
ISI-QI (per 10 points) β_1	+0.166	[−0.329; +0.724]	1.18
GCS < 9 β_2	+1.399	[+0.15; +2.82]	4.05
Age ≥ 65 years β_3	+1.456	[+0.10; +2.98]	4.29
Model 2 — unfavorable outcome, GOS 1–3 (n = 142; 57 events)			
Intercept γ_0	−4.552	[−7.73; −1.83]	—
ISI-QI (per 10 points) γ_1	+0.419	[+0.039; +0.829]	1.52
GCS < 9 γ_2	+4.174	[+1.84; +8.10]	65.0
Age ≥ 65 years γ_3	+0.708	[−0.52; +2.06]	2.03

Measuring the Quality of Neurosurgical Care Through Intracranial Suppurations: A Twenty-Indicator Framework Tested on 150 Cases at Bouaké University Hospital

Unshrunk Firth coefficients; 95% confidence intervals of β by penalized profile likelihood; $OR = \exp(\beta)$. The optimism-corrected calibration slopes (0.842 for model 1 and 1.006 for model 2) measure calibration; they are not applied to these coefficients. The odds ratio of 65.0 for $GCS < 9$ in model 2 reflects a quasi-complete separation and its interval is very wide: it is an order of magnitude, not a reliable point estimate.

These equations must not be used to decide for a patient before external validation, for the reasons set out below and in Section IV-E.

Table VI and Figs. 5 to 7 report the internal validation. It can be summarized as follows: the ISI-QI improves no corrected performance of the mortality model; for functional outcome, it improves it in appearance, but this gain is due to the surgical delay indicator.

1) **Mortality:** The severity model (GCS and age) had an apparent C statistic of 0.668, corrected to 0.651 after bootstrap. Adding the ISI-QI raised the apparent statistic to 0.700, but the corrected statistic fell to 0.646: all the apparent gain was optimism. The corrected Brier score went from 0.0948 to 0.0954, for a null model at 0.0953. The full model therefore does no better than the mean frequency of the event, with a corrected normalized Brier score of -0.001 . The ISI-QI alone had a corrected C statistic of 0.523, i.e., almost no discrimination.

2) **Functional Outcome:** Adding the ISI-QI improved the corrected measures in appearance: the C statistic went from 0.676 to 0.739 and the normalized Brier score from 0.191 to 0.211. But in the model without indicator P2, the corrected C statistic fell back to 0.685 and the normalized Brier score to 0.189, at the level of the severity-only model. The apparent contribution of the score therefore tells nothing about the quality of care: it is an artifact of the confounding by indication described in Section III-G.

3) **Calibration:** The mortality model had a corrected calibration slope of 0.842: moderate overfitting, predictions too dispersed. The functional outcome model was well calibrated, with a corrected slope of 1.006 (Fig. 6). The corrected calibration intercepts were close to zero: -0.02 for mortality and $+0.01$ for functional outcome.

4) **Clinical Utility:** On optimism-corrected predictions, the mortality model provided a net benefit superior to the “intervene in all” and “intervene in none” strategies for thresholds of about 9% to 23%, and the functional outcome model for thresholds of about 25% to 61% (Fig. 7). These results are exploratory. Because the ISI-QI contributes nothing discriminating, this net benefit is that of the severity model, not of the quality score.

5) **Sensitivity to Missing GOS:** We reran the unfavorable-outcome model under three scenarios. In complete cases (142 patients, 57 events), the aOR of the ISI-QI per 10-point increment was 1.52 (95% profile CI: 1.04–2.29; $p = 0.031$; corrected C 0.739), and 1.15 (0.80–1.69) without P2. If the eight patients had all had a favorable outcome (150 patients, 57 events), the aOR was 1.49 (1.03–2.21; $p = 0.035$; corrected C 0.731), and 1.13 (0.79–1.64) without P2. If they had all had an unfavorable outcome (150 patients, 65 events), the aOR was 1.39 (0.98–2.01; $p = 0.066$; corrected C 0.716), and 1.10 (0.78–1.58) without P2. The apparent association of the ISI-QI with unfavorable outcome is therefore not robust to missing data, whereas its disappearance without P2 is found in all three scenarios.

TABLE VI. INTERNAL VALIDATION BY BOOTSTRAP (2,000 RESAMPLES)

Model	Apparent C	Corrected C	Corrected Brier	Normalized Brier	Calibration Slope
Outcome: in-hospital mortality (16 events; null-model Brier = 0.0953)					
Severity alone (GCS + age)	0.668	0.651	0.0948	+0.005	0.920
ISI-QI alone	0.579	0.523	0.0986	-0.035	0.354
ISI-QI + severity (model 1)	0.700	0.646	0.0954	-0.001	0.842
ISI-QI excluding P2 + severity	0.691	0.640	0.0956	-0.003	0.849
Outcome: unfavorable outcome, GOS 1–3 (57 events; null-model Brier = 0.2403)					
Severity alone (GCS + age)	0.681	0.676	0.1944	+0.191	1.000
ISI-QI alone	0.633	0.625	0.2330	+0.030	1.170
ISI-QI + severity (model 2)	0.751	0.739	0.1897	+0.211	1.006
ISI-QI excluding P2 + severity	0.710	0.685	0.1949	+0.189	0.956

C: C statistic (area under the ROC curve). The normalized Brier score equals $1 - (\text{corrected Brier} / \text{null-model Brier})$; a negative value means that the model does worse than the mean frequency of the event. The calibration slope is corrected for optimism. Corrected calibration intercepts of the two full models: -0.02 (model 1, mortality) and $+0.01$ (model 2, unfavorable outcome).

Measuring the Quality of Neurosurgical Care Through Intracranial Suppurations: A Twenty-Indicator Framework Tested on 150 Cases at Bouaké University Hospital

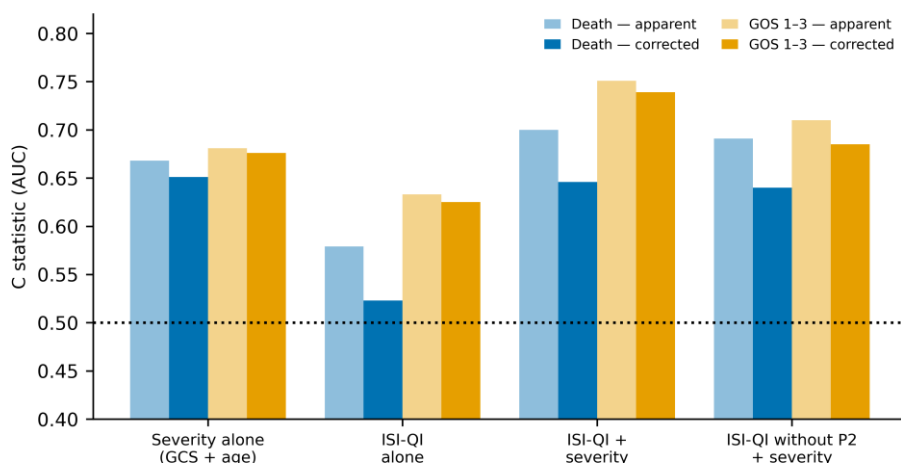


Fig. 5. Apparent and optimism-corrected C statistic for the four specifications of each model. The gap between the light bars and the solid bars represents the optimism.

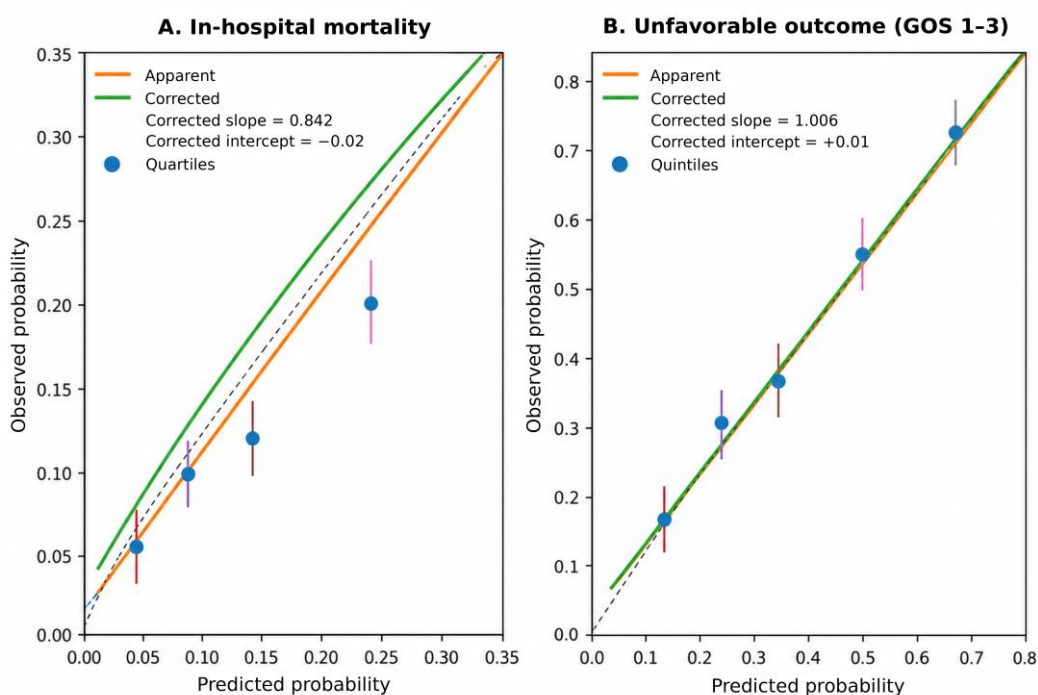


Fig. 6. Calibration curves of the two prognostic models. The corrected curve is reconstructed from the calibration slope and intercept corrected by bootstrap. Points represent the risk groups (quartiles for mortality, quintiles for functional outcome).

Measuring the Quality of Neurosurgical Care Through Intracranial Suppurations: A Twenty-Indicator Framework Tested on 150 Cases at Bouaké University Hospital

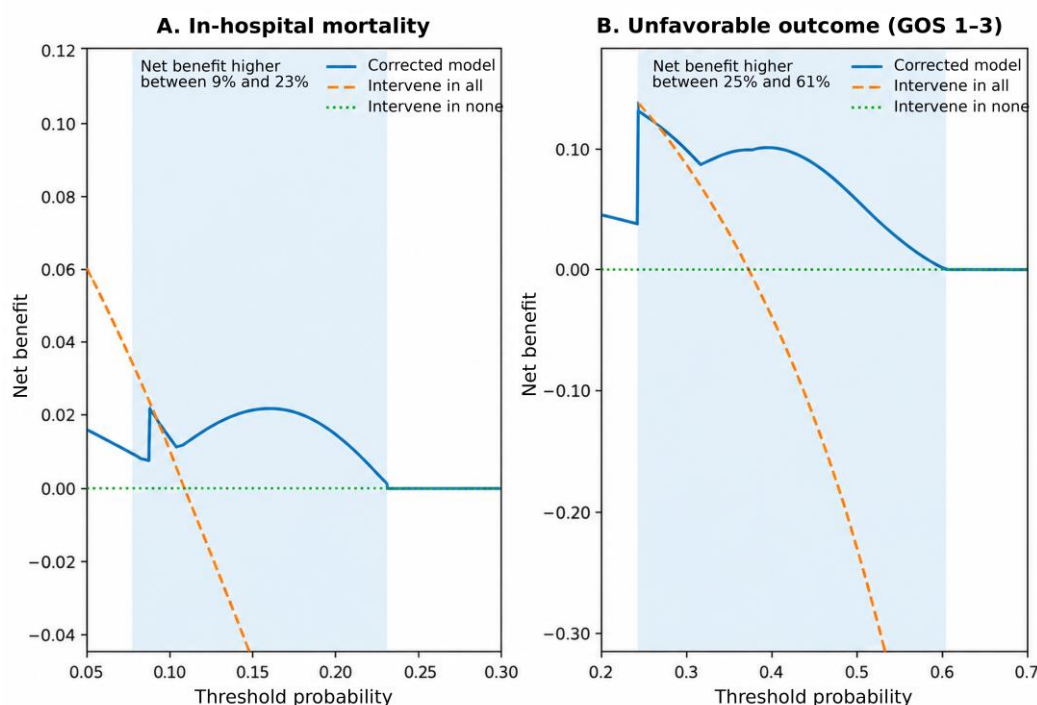


Fig. 7. Decision curve analyses. The net benefit of the model is compared with the strategies of intervening in all patients or in none.

I. Risk Adjustment and Temporal Comparison

Crude mortality went from 12.7% (10/79) in 2016–2020 to 8.5% (6/71) in 2021–2026, with no significant difference ($p = 0.44$). The median ISI-QI score was 87.5% in both periods. After adjustment for GCS and age, the observed/expected ratio went from 1.17 (95% CI: 0.56–2.16) to 0.80 (0.29–1.74), i.e., from a slightly unfavorable position to a slightly favorable position relative to the expected risk. The intervals are wide and almost entirely overlap: with sixteen deaths, no firm conclusion is allowed.

On the process side, two trends diverged. Culture yield improved, from 38.1% to 49.1% of the samples sent ($p = 0.26$). Empirical anaerobic coverage, on the contrary, declined, from 97.5% to 87.3% ($p = 0.026$). This signal of prescribing drift, which permanent monitoring would have detected in real time, illustrates the value of a dashboard over a one-off retrospective analysis. Delays did not change: the median symptom-to-admission delay was 19 days in both periods ($p = 0.44$) and the admission-to-surgery delay 2 days ($p = 0.24$).

IV. DISCUSSION

A. Main Findings

The framework is feasible: twenty indicators were filled in from existing records, without dedicated data collection, i.e., 987 care opportunities with a completeness of 97.3%. It also locates failures instead of delivering an overall verdict. Clinical decision-making reaches or approaches its targets (treatment of the portal of entry 94.6%, anaerobic coverage 92.7%, antibiogram 88.0%), and imaging and the operating theater do not limit management, even though 33% of procedures exceed 48 hours. The faults lie elsewhere: access to the hospital, with 2% of admissions within seven days and a median delay of 19 days, and the microbiological chain, at the end of which fewer than one patient in four receives a guided antibiotic therapy.

The relation between process and outcome postulated by Donabedian is not found in our data, and looking for it naively leads to a false conclusion: the early-surgery indicator seems to harm outcome because it reflects severity. Finally, after optimism correction, the composite score contributes nothing beyond initial severity. This result is thankless to state, but it protects the instrument from misuse: the ISI-QI measures the compliance of an organization with standards; it does not predict the outcome of a patient.

B. Intracranial Suppurations as a Tracer Condition

Kessner's criteria are met in an unusual way [8]. The functional impact is heavy and measurable by the GOS (GOS 1–3 in 40.1% of evaluable patients). The incidence, about fifteen cases a year in our series, allows an annual follow-up. Management is codified by recent guidelines [11]. Drainage combined with an adapted antibiotic therapy changes the prognosis. And the pathway weighs heavily on the result, since most of the delay is pre-hospital.

Measuring the Quality of Neurosurgical Care Through Intracranial Suppurations: A Twenty-Indicator Framework Tested on 150 Cases at Bouaké University Hospital

We propose adding three requirements for urgent surgical tracer conditions in resource-limited settings. The condition must successively draw on heterogeneous subsystems. Its indicators must be measurable on ordinary records. And the evaluation must reveal a failing link on which action is possible. Without the first, the condition illuminates only one segment of the system; without the second, the measurement does not hold in routine; without the third, one obtains a diagnosis without a lever. Hernia [9] draws on a shorter, non-urgent pathway. ICS add a microbiological dimension and a prolonged follow-up. Performance there is bounded by the weakest link, not by the average: a sample sent in 84.0% of operated patients is cancelled out by a culture yield of 43.1%. Equipping the laboratory with anaerobic media without shortening the 19-day delay would amount to better documenting infections that are already advanced.

C. The Pre-Hospital Bottleneck

About 95% of the time elapsed between the first symptoms and the operation lies before the hospital. This reverses the spontaneous order of priorities, which looks at what one sees from the hospital. Our data can be read within the framework of the three delays of Thaddeus and Maine: deciding to seek care, reaching an appropriate facility, being cared for once there [40]. The first two are clear. Three facilities passed through at the median, and five or more for a quarter of patients, signal diagnostic wandering; 28.6% of medicalized transport speaks to the weakness of transfer. The third delay is under control, with two days median to surgery. These findings echo the barriers described in Cameroon for nontraumatic neurosurgical diseases [41]. For ICS, the most useful global surgery indicator would therefore be access within two hours to a facility capable of cross-sectional imaging and referral, more than operative volume or the number of neurosurgeons.

D. The Microbiological Cascade

The cascade, borrowed from HIV programs [42], turns an uninformative indicator (“a third of cultures are positive”) into a breakdown that points to the failing step. We propose to call it the neurosurgical quality cascade. Here, the main loss is neither the decision nor the procedure, but the passage from sample to culture result. Three mechanisms probably contribute: antibiotic therapy started before sampling, given the admission delay; the absence of anaerobic culture, whereas these infections are often polymicrobial with anaerobes [43,44]; and the conditions under which samples are transported. Series that use 16S rRNA sequencing or metagenomics identify an etiology in the majority of culture-negative abscesses [45–47], which gives an order of magnitude of the expected gain.

The last loss, ten patients whose available antibiogram did not lead to a documented adaptation, is of a different nature. Neither equipment nor competence is at issue: the loop between the laboratory and the clinical team is broken. A well-constructed indicator makes it visible, and an inexpensive measure, such as feeding back the results at the weekly staff meeting, can correct it.

E. Confounding by Indication

The most baffling result is the link between early surgery and unfavorable outcome. It is a textbook case of confounding by indication [48,49]. Patients operated on within 48 hours do not resemble those operated on later: they are more severe, and it is their severity that made the intervention urgent. In our cohort, 18.2% had a GCS below 9, against 2.6% of those operated on later, a gap that rests on a single patient in this second group. Severity is a common cause of operative earliness and of outcome. Adjustment for GCS and age does not close this path, especially as lesion size and mass effect, which also differ between the groups (Table II), remain outside the model.

The implications go beyond ICS. A facility that receives many severe cases will operate quickly and display worse crude results; one that transfers its severe cases will operate less quickly and appear better. Without adjustment, the ranking would be inverted, with the risk of perverse incentives if it were used to fund or accredit. Three rules follow. A surgical timeliness indicator is read together with the severity distribution of the patients. Facilities and periods are compared by observed/expected ratios, not by crude rates. And before validating a composite score, one must verify that none of its items carries an inverse signal of confounding origin: our analysis without item P2 gives the minimal protocol for this verification.

The absence of a link between the score and mortality has three explanations. The score suffers from a ceiling effect (median 87.5%, low variance). Power is limited, with sixteen deaths. And the variability of in-hospital processes is probably too low to weigh on outcomes already fixed upstream by a 19-day delay. This does not mean that quality of care has no effect: once the door of the hospital is crossed, in-hospital quality is no longer the limiting factor.

F. Intended Use: What the ISI-QI Is and Is Not, and the ISI-QI-10 Dashboard

An instrument is defined as much by its uses as by its misuses. The ISI-QI makes it possible to describe the compliance of a department with standards, indicator by indicator; to identify failing steps and prioritize actions; to follow an organization over time, provided the indicators are read separately; and to structure a prospective registry. It does not make it possible to predict the outcome of a patient, since internal validation shows no contribution beyond severity. Nor does it make it possible to rank facilities on the score alone, to evaluate a practitioner (the framework measures an organization), or to allocate resources or accreditation without external validation. The equations of Table V must not be used to decide for a patient before external validation.

Measuring the Quality of Neurosurgical Care Through Intracranial Suppurations: A Twenty-Indicator Framework Tested on 150 Cases at Bouaké University Hospital

From this experience we derive a minimal dashboard of ten indicators, the ISI-QI-10, measurable without additional resources and covering the three Donabedian domains. A core of six indicators is feasible in any department that has paper records and a bacteriology laboratory: median delay between symptom onset and admission (target ≤ 7 days; observed 19 days); cross-sectional imaging before decision ($\geq 95\%$; 100%); median delay between admission and surgery (≤ 48 h; 2 days); pus sampling in operated patients ($\geq 95\%$; 84.0%); risk-adjusted in-hospital mortality (O:E < 1 ; 1.17 then 0.80 depending on the period); favorable outcome ($\geq 70\%$; 59.9%). Four other indicators form an extension that presupposes an organized follow-up: facilities passed through before admission (≤ 2 ; 3), culture yield ($\geq 60\%$; 43.1%), antibiotic therapy adapted to the available antibiogram ($\geq 90\%$; 77.3%), documented follow-up at 30 days in survivors ($\geq 80\%$; 85.1%). The targets combine guidelines and internal consensus; they are to be discussed before any comparative use. Culture yield is the only indicator whose improvement requires a material investment, which makes it a quantified advocacy argument. Adaptation of antibiotic therapy is conversely the least costly to improve.

G. Comparisons, Equity, and Registries

Crude rates of mortality or unfavorable outcome depend on the patient profile, and a center that receives more severe or later cases may display less bad results despite a better quality [6,7]. O:E ratios are a first answer, but their validity depends on the risk model. Ours, limited to GCS and age, has a corrected C statistic of 0.651 for mortality: it will need to be enriched before any comparative use between facilities. Table II points to the candidates: lesion size, mass effect, and type of suppurations all differed according to outcome without being in the model.

Quality cannot be separated from equity. The pre-hospital delay is also an indirect marker of territorial and social inequalities. Our study evaluated neither equity nor algorithmic bias: no equity variable was collected. A registry should include a few simple ones: distance to the center, urban or rural setting, mode of transport, insurance coverage, direct payment, interruption of treatment for financial reasons. They would make it possible to verify that the average performance does not mask systematically disadvantaged subgroups [50,51]. Because the chain of care crosses several levels of the system, it would be unfair to attribute the results to the neurosurgery department alone: the framework can support broadened morbidity and mortality meetings, from families and primary care providers to the laboratory.

The move from the framework to a prospective registry is the next step. A registry should serve to learn (measure, spot gaps, intervene, reassess) and not to produce publications from time to time [52]. Sobriety matters: a registry overloaded with variables sees its completeness erode, as shown by the Ethiopian perioperative registry designed around sixty-seven variables, which approached completeness only after reorganization [53].

H. Strengths, Limitations, and Generalizability

1) **Strengths:** The framework is explicit: indicators, numerators, denominators, and targets were set before the analysis, and the applicability of each indicator is handled with care, which composite scores often neglect. Confounding by indication is formalized and tested by a sensitivity analysis. Internal validation is complete (bootstrap, calibration, Brier score, decision curves) and makes it possible to state a solid negative conclusion rather than a flattering apparent performance.

2) **Limitations:** The study is single-center: it cannot assess the variability between facilities, which is nonetheless the intended use of these indicators. The absence of external validation is the main limitation. The reported performances, even corrected for optimism, are development performances, and these almost always degrade when a model changes population [19,54]. The risk of overfitting is real: sixteen deaths for three predictors make 5.3 events per variable, well below the rule of ten [33,55], and the corrected calibration slope of the mortality model is 0.842. The criterion of Riley et al. would require about 525 patients for an assumed Cox–Snell R^2 of 0.05 (Section II-E). The choice of GCS and age was made on the same data, which probably underestimates the optimism. No analysis protocol was registered.

The cohort comes from a referral center. Patients who died before transfer, were directed elsewhere, or never reached a neurosurgical center are absent from it: the observed mortality does not represent that of the source population, and our rates have no epidemiological reading. Nor can it be directly compared with that of series with different criteria [24,25], and it cannot serve alone as a performance indicator. Recruitment, largely retrospective, exposes the study to information bias: a care that was not recorded cannot be distinguished from a care that was not delivered, and 8% of undocumented surgical delays give its measure. A single investigator collected the data, without double entry or measure of agreement, so that the reliability of the rating (portal of entry, applicability, GOS) is not established. The GOS itself was assigned from sequelae and autonomy, without a standardized time point, and is missing in 8 survivors: it is not an independent measure of outcome.

Two limitations concern the framework. Its content validity rests on the literature and an internal consensus, without a Delphi procedure. Patients and the public did not take part in it: the cost of the pathway, the distance, the burden of transport, the clarity of the information received, and the social impact of sequelae are represented by no indicator. Risk adjustment, limited to GCS and age, is moreover rudimentary, and Table II shows a possible residual confounding. The logical next step is a validation program: content validity by a multinational Delphi with patient representatives, reliability by double rating and a kappa coefficient, sensitivity to change through improvement cycles, then multicenter application.

Measuring the Quality of Neurosurgical Care Through Intracranial Suppurations: A Twenty-Indicator Framework Tested on 150 Cases at Bouaké University Hospital

3) **Generalizability:** It decreases across three levels. The method (domains, applicability, denominators, cascade, risk adjustment) transposes widely. The indicators are probably transposable to neurosurgical centers of sub-Saharan Africa with a comparable technical platform, with adaptation of the thresholds. The observed values, for their part, hold only for this department, this period, and this recruitment: a median delay of 19 days or a culture yield of 43% are local findings, not regional references. The targets are benchmarks for discussion, not norms. As it stands, the ISI-QI must be used neither to compare, nor to decide, nor to prognosticate.

V. CONCLUSIONS

Applied to 150 observations, a structure–process–outcome framework built from ordinary records shows a generally solid clinical decision-making and access to the operating theater (compliance of 89.0%, complete access to the theater, median operative delay of two days), and two breaking points: pre-hospital access, which concentrates about 95% of the total delay, and the microbiological chain, which loses 77% of the cohort between admission and a guided antibiotic therapy.

The study also highlights two pitfalls of quality measurement in surgery. The first is confounding by indication: compliance with the operative delay is so tied to severity that it reverses the direction of the association with outcome. The second is the optimism of development performance: after correction, the score contributes nothing beyond initial severity. The ISI-QI must therefore serve as a descriptive instrument of the quality of an organization, and any comparison between facilities or periods must go through observed/expected ratios. The ten-indicator dashboard we propose requires no additional resources; it can serve as the foundation of a regional registry.

DECLARATIONS

Conflicts of interest. The authors declare no conflicts of interest.

Funding. No specific funding.

Ethical approval and consent. Bouaké University Hospital did not have a constituted ethics committee during the study period; the study was conducted after authorization from the head of the neurosurgery department and from the hospital administration, in accordance with the principles of the Declaration of Helsinki. It has since received the approval of the Institutional Ethics Committee of Bouaké University Hospital (approval no. CEI/CHU-BKE/133/2026).

Author contributions. Serge Konan Yao: conception, methodology, writing. Landry Stéphane Faozo Teti: conception, analysis, writing and revision. Djoko Traoré: data collection and validation. Adama Kouadio: data collection and resources. Koffi Jean-Jacques Kouassi: methodology, validation, and supervision. Aderehime Haidara: supervision, validation, and revision. All authors approved the final version.

Patient and public involvement. None (see Section II-F).

Data and code availability. The anonymized analytic database, the operational definitions (Table S1), and the code for calculating the indicators and the models are available on reasoned request from the corresponding author. The complete coefficients, intercepts included, appear in Table V and allow any prediction to be recalculated.

Reproducibility. All analyses were performed with Python 3.12 (pandas, NumPy, SciPy, statsmodels, Matplotlib) and an implementation of Firth regression written for the study. The seed of the random number generator for the 2,000 resamples is fixed and documented in the code.

Use of machine learning. No machine learning method was employed; the modeling relies exclusively on penalized logistic regression. The TRIPOD+AI recommendations are followed because they constitute the current version of the TRIPOD guideline.

Note on the article. This work is a secondary analysis of the cohort described in the main epidemiological and prognostic report, with which it shares the population and the data collection.

REFERENCES

- 1) Kruk ME, Gage AD, Arsenault C, Jordan K, Leslie HH, Roder-DeWan S, et al. High-quality health systems in the Sustainable Development Goals era: time for a revolution. *Lancet Glob Health*. 2018;6(11):e1196-e1252.
- 2) Meara JG, Leather AJM, Hagander L, Alkire BC, Alonso N, Ameh EA, et al. Global Surgery 2030: evidence and solutions for achieving health, welfare, and economic development. *Lancet*. 2015;386(9993):569-624.
- 3) Dewan MC, Rattani A, Fieggen G, Arráez MA, Servadei F, Boop FA, et al. Global neurosurgery: the current capacity and deficit in the provision of essential neurosurgical care. *J Neurosurg*. 2018;130(4):1055-1064.
- 4) Donabedian A. Evaluating the quality of medical care. *Milbank Mem Fund Q*. 1966;44(3 Suppl):166-206.
- 5) Donabedian A. The quality of care: how can it be assessed? *JAMA*. 1988;260(12):1743-1748.
- 6) Lilford R, Mohammed MA, Spiegelhalter D, Thomson R. Use and misuse of process and outcome data in managing performance of acute medical care: avoiding institutional stigma. *Lancet*. 2004;363(9415):1147-1154.
- 7) Iezzoni LI. The risks of risk adjustment. *JAMA*. 1997;278(19):1600-1607.
- 8) Kessner DM, Kalk CE, Singer J. Assessing health quality: the case for tracers. *N Engl J Med*. 1973;288(4):189-194.

Measuring the Quality of Neurosurgical Care Through Intracranial Suppurations: A Twenty-Indicator Framework Tested on 150 Cases at Bouaké University Hospital

- 9) NIHR Global Health Research Unit on Global Surgery. Access to and quality of elective care: a prospective cohort study using hernia surgery as a tracer condition in 83 countries. *Lancet Glob Health*. 2024;12(7):e1094-e1103.
- 10) Brouwer MC, Tunkel AR, McKhann GM II, van de Beek D. Brain abscess. *N Engl J Med*. 2014;371(5):447-456.
- 11) Bodilsen J, D'Alessandris QG, Humphreys H, Iro MA, Klein M, Last K, et al. European Society of Clinical Microbiology and Infectious Diseases guidelines on diagnosis and treatment of brain abscess in children and adults. *Clin Microbiol Infect*. 2024;30(1):66-89.
- 12) Nathoo N, Nadvi SS, van Dellen JR. Taming an old enemy: a profile of intracranial suppuration. *World Neurosurg*. 2012;77(3-4):484-490.
- 13) Djientcheu VP, Mouafo Tambo FF, Esiene A, Nguefack S, Mewa Kinoo S, Bello F, et al. Intracranial suppurations in the African child: a severe but preventable condition. *Childs Nerv Syst*. 2013;29(1):119-123.
- 14) Chikani MC, Mezue WC, Ohaegbulam SC, Ndubuisi CA, Chikani UN. Subdural empyema: clinical presentations and management options for an uncommon neurosurgical emergency in a developing country. *Niger J Clin Pract*. 2017;20(10):1221-1225.
- 15) Jolayemi EO, Olaniyi EO, Shokunbi MT, Malomo AO. Contemporary management of intracranial subdural empyema: an institutional experience. *J West Afr Coll Surg*. 2022;12(4):31-38.
- 16) Boukassa L, Ngackosso OB, Kinata SB, Ekouele Mbaki HB. Intracranial suppurations in Africa: systematic review. *Open J Mod Neurosurg*. 2020;10(2):244-253.
- 17) Ooi SZY, Sichimba D, Dalle DU, Higginbotham G, Takoutsing BD, Bankole NDA, et al. Management and outcomes of paediatric intracranial suppurations in low- and middle-income countries: a scoping review. *Front Surg*. 2021;8:690895.
- 18) Nathoo N, Nadvi SS, Narotam PK, van Dellen JR. Brain abscess: management and outcome analysis of a computed tomography era experience with 973 patients. *World Neurosurg*. 2011;75(5-6):716-726.
- 19) Moons KGM, Wolff RF, Riley RD, Whiting PF, Westwood M, Collins GS, et al. PROBAST: a tool to assess risk of bias and applicability of prediction model studies: explanation and elaboration. *Ann Intern Med*. 2019;170(1):W1-W33.
- 20) Collins GS, Moons KGM, Dhiman P, Riley RD, Beam AL, Van Calster B, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ*. 2024;385:e078378.
- 21) von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP. The Strengthening the Reporting of Observational Studies in Epidemiology statement: guidelines for reporting observational studies. *Lancet*. 2007;370(9596):1453-1457.
- 22) Ogrinc G, Davies L, Goodman D, Batalden P, Davidoff F, Stevens D. SQUIRE 2.0: revised publication guidelines from a detailed consensus process. *BMJ Qual Saf*. 2016;25(12):986-992.
- 23) Brouwer MC, Coutinho JM, van de Beek D. Clinical characteristics and outcome of brain abscess: systematic review and meta-analysis. *Neurology*. 2014;82(9):806-813.
- 24) Bodilsen J, Duerlund LS, Mariager T, Brandt CT, Petersen PT, Larsen L, et al. Clinical features and prognostic factors in adults with brain abscess. *Brain*. 2023;146(4):1637-1647.
- 25) Tseng JH, Tseng MY. Brain abscess in 142 patients: factors influencing outcome and mortality. *Surg Neurol*. 2006;65(6):557-562.
- 26) Mainz J. Defining and classifying clinical indicators for quality improvement. *Int J Qual Health Care*. 2003;15(6):523-530.
- 27) Campbell SM, Braspenning J, Hutchinson A, Marshall MN. Research methods used in developing and applying quality indicators in primary care. *BMJ*. 2003;326(7393):816-819.
- 28) Nolan T, Berwick DM. All-or-none measurement raises the bar on performance. *JAMA*. 2006;295(10):1168-1170.
- 29) Little RJA, Rubin DB. Statistical analysis with missing data. 3rd ed. Hoboken: Wiley; 2019.
- 30) Sterne JAC, White IR, Carlin JB, Spratt M, Royston P, Kenward MG, et al. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *BMJ*. 2009;338:b2393.
- 31) Firth D. Bias reduction of maximum likelihood estimates. *Biometrika*. 1993;80(1):27-38.
- 32) Heinze G, Schemper M. A solution to the problem of separation in logistic regression. *Stat Med*. 2002;21(16):2409-2419.
- 33) Riley RD, Snell KI, Ensor J, Burke DL, Harrell FE Jr, Moons KG, et al. Minimum sample size for developing a multivariable prediction model: Part II — binary and time-to-event outcomes. *Stat Med*. 2019;38(7):1276-1296.
- 34) Steyerberg EW, Harrell FE Jr, Borsboom GJ, Eijkemans MJ, Vergouwe Y, Habbema JD. Internal validation of predictive models: efficiency of some procedures for logistic regression analysis. *J Clin Epidemiol*. 2001;54(8):774-781.
- 35) Harrell FE Jr, Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Stat Med*. 1996;15(4):361-387.
- 36) Steyerberg EW, Vickers AJ, Cook NR, Gerds T, Gonen M, Obuchowski N, et al. Assessing the performance of prediction models: a framework for traditional and novel measures. *Epidemiology*. 2010;21(1):128-138.

Measuring the Quality of Neurosurgical Care Through Intracranial Suppurations: A Twenty-Indicator Framework Tested on 150 Cases at Bouaké University Hospital

- 37) Van Calster B, McLernon DJ, van Smeden M, Wynants L, Steyerberg EW. Calibration: the Achilles heel of predictive analytics. *BMC Med.* 2019;17(1):230.
- 38) Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. *Med Decis Making.* 2006;26(6):565-574.
- 39) Vickers AJ, van Calster B, Steyerberg EW. Net benefit approaches to the evaluation of prediction models, molecular markers, and diagnostic tests. *BMJ.* 2016;352:i6.
- 40) Thaddeus S, Maine D. Too far to walk: maternal mortality in context. *Soc Sci Med.* 1994;38(8):1091-1110.
- 41) Nguembu S, Kanmounye US, Tétinou F, Djiofack D, Takoukam R. Barriers to management of nontraumatic neurosurgical diseases at 2 Cameroonian neurosurgical centers: cross-sectional study. *World Neurosurg.* 2020;139:e774-e783.
- 42) Gardner EM, McLees MP, Steiner JF, Del Rio C, Burman WJ. The spectrum of engagement in HIV care and its relevance to test-and-treat strategies for prevention of HIV infection. *Clin Infect Dis.* 2011;52(6):793-800.
- 43) Al Masalma M, Lonjon M, Richet H, Dufour H, Roche PH, Drancourt M, et al. Metagenomic analysis of brain abscesses identifies specific bacterial associations. *Clin Infect Dis.* 2012;54(2):202-210.
- 44) Cobo F, Rodríguez-Granger J, Sampedro A, Aliaga-Martínez L, Navarro-Marí JM. Brain abscesses caused by anaerobic microorganisms: a systematic review. *Anaerobe.* 2023;79:102668.
- 45) Kommedal Ø, Wilhelmsen MT, Skrede S, Meisal R, Jakovljević A, Gaustad P, et al. Massive parallel sequencing provides new perspectives on bacterial brain abscesses. *J Clin Microbiol.* 2014;52(6):1990-1997.
- 46) John DV, Ramakrishnan R, Thomas B, Nair S, Abraham OC. Identification of microbial agents in culture-negative brain abscess samples by 16S/18S ribosomal RNA gene PCR and sequencing. *J Clin Microbiol.* 2023;61(1):e01254-22.
- 47) Wang X, Guo X, Liu H, Wang B, Wu J, Chen S, et al. Augmented pathogen detection in brain abscess using metagenomic next-generation sequencing: a retrospective cohort study. *Microbiol Spectr.* 2024;12(10):e0032524.
- 48) Walker AM. Confounding by indication. *Epidemiology.* 1996;7(4):335-336.
- 49) Kyriacou DN, Lewis RJ. Confounding by indication in clinical research. *JAMA.* 2016;316(17):1818-1819.
- 50) Alkire BC, Raykar NP, Shrimel MG, Weiser TG, Bickler SW, Rose JA, et al. Global access to surgical care: a modelling study. *Lancet Glob Health.* 2015;3(6):e316-e323.
- 51) Shrimel MG, Dare AJ, Alkire BC, O'Neill K, Meara JG. Catastrophic expenditure to pay for surgery worldwide: a modelling study. *Lancet Glob Health.* 2015;3(Suppl 2):S38-S44.
- 52) Institute of Medicine. Best care at lower cost: the path to continuously learning health care in America. Washington DC: National Academies Press; 2013.
- 53) Jaraczewski TJ, Abebe BM, Diehl T, Esayas T, Melaku W, Nigussie Y, et al. Implementation of a perioperative registry in Ethiopia to enhance surgical quality improvement. *World J Surg.* 2024;48(8):1829-1839.
- 54) Riley RD, Ensor J, Snell KIE, Debray TPA, Altman DG, Moons KGM, et al. External validation of clinical prediction models using big datasets from e-health records or IPD meta-analysis: opportunities and challenges. *BMJ.* 2016;353:i3140.
- 55) Peduzzi P, Concato J, Kemper E, Holford TR, Feinstein AR. A simulation study of the number of events per variable in logistic regression analysis. *J Clin Epidemiol.* 1996;49(12):1373-1379.