Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Categorical Versus Continuous Encoding of Routine CSF Variables for Diagnosing Tuberculous Meningitis: Development in Iran and External Validation in Peru

Hamid Reza Naderi
Mashhad University of Medical Sciences
Vahideh Zolfaghari
vahidehzolfagharii@gmail.com

Mashhad University of Medical Sciences
Majid Khadem Rezayian
Mashhad University of Medical Sciences
Hamid Reza Mirzaei
Mashhad University of Medical Sciences
Fereshte Sheybani
Mashhad University of Medical Sciences
Mahboubeh Haddad
Mashhad University of Medical Sciences

Research Article

Keywords: tuberculous meningitis, cerebrospinal fluid, clinical prediction model, external validation, transportability, point score

Posted Date: November 5th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11272234/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

Background

Diagnostic scores for tuberculous meningitis (TBM) based on routine cerebrospinal fluid (CSF) measurements often perform inconsistently across settings. We assessed whether representing routine CSF measurements as fixed, clinically motivated categories, rather than continuous values, may improve the transportability of a TBM diagnostic score across countries.

Methods

From four routine CSF variables (white cell count, protein, protein to glucose ratio, lymphocyte percentage), we developed a continuous logistic model and an integer score with fixed cut points in an Iranian cohort (69 TBM, 70 controls). Both were locked and evaluated without refitting in a Peruvian cohort (40 TBM; 62 bacterial, cryptococcal, or viral meningitis; mononuclear cell percentage replacing lymphocyte percentage), using harmonized multiple imputation; the study was reported according to TRIPOD + AI.

Results

In Iranian cross validation, discrimination was similar (AUC 0.822, 95% CI 0.744 to 0.886, continuous; 0.821, categorical). In Peru, the AUC was 0.583 (95% CI 0.473 to 0.691) for the continuous and 0.699 (95% CI 0.599 to 0.794) for the categorical score (paired difference 0.116, 95% CI 0.048 to 0.182); a three-predictor continuous model did not perform better (0.553). The advantage was evident against bacterial but not cryptococcal or viral meningitis. Calibration slopes were 0.471 (95% CI 0.163 to 0.779) and 0.150, both well below 1. Neither model discriminated well between TBM and neurobrucellosis (apparent AUC 0.71 and 0.66). Both cohorts were smaller than required (385 and 581).

Conclusions

In one external cohort, fixed categories showed higher discrimination than continuous values, but absolute discrimination was modest, results varied across mimics, and encoding was not isolated from other differences between the models. Fixed categories may improve transportability, but these findings are preliminary and warrant confirmation in larger, prospective, multicountry cohorts.

Introduction

Tuberculous meningitis (TBM) is the most severe and life-threatening manifestation of tuberculosis. Modelling estimates indicate that approximately 164,000 adults developed TBM in 2019 (95% uncertainty interval, 129,000 to 199,000), with an estimated mortality of approximately 48%; among

treated patients, mortality was still estimated at 27% (1). Tuberculosis remains an important public health concern in Iran, where recent national estimates indicate an incidence of approximately 10 to 13 per 100,000 population, corresponding to an intermediate burden according to the WHO classification (2). Population based spatial analyses have also demonstrated substantial geographic variation in tuberculosis incidence across Iranian provinces (3). Although central nervous system involvement, including TBM, represents a minority of extrapulmonary tuberculosis cases, it is associated with a disproportionate burden of morbidity and mortality (4). Timely diagnosis and treatment are important in the management of TBM, as diagnostic delay has been associated with increased mortality and persistent neurological impairment (4, 5).

Diagnostic delay remains an important challenge in TBM, in part because microbiological confirmation can be difficult to obtain. CSF smear microscopy has limited sensitivity, whereas mycobacterial culture, although more sensitive, generally requires several weeks to yield a result (4). Xpert MTB/RIF Ultra has greater sensitivity than earlier molecular assays, but a negative result does not reliably exclude TBM (5). As a result, clinicians may need to decide whether to initiate antituberculous treatment before definitive microbiological confirmation is available. The consensus case definition (6) was developed primarily as a framework for research classification rather than as a tool for estimating individual patient risk at the bedside.

One of the most established clinical scoring systems, the Thwaites diagnostic rule, combines age, symptom duration, peripheral white cell count, CSF white cell count, and CSF neutrophil percentage to distinguish TBM from bacterial meningitis (7). Additional diagnostic rules have subsequently been developed for other comparator groups and clinical settings, including a rule for differentiating TBM from viral meningitis (8) and, in Peru, a rule incorporating cough duration, CSF cell count, and CSF adenosine deaminase (9). However, the performance of these rules has varied across populations and clinical settings. In an Indian validation study, a modified version of the Thwaites rule showed higher sensitivity and specificity after the age criterion was removed, indicating that the original age threshold may have had limited transportability to that population (10). Similarly, in a multicenter study of subacute and chronic meningitis, both the Thwaites rule and the consensus case definition score showed poor to fair discrimination when evaluated against non-tuberculous causes of meningitis (11).

Differences in case mix may account for some of the variation in score performance across settings, but are unlikely to explain all observed differences. CSF measurements are not necessarily obtained or reported in the same way across laboratories, as analyzers, differential cell counting practices, and reporting conventions may vary. Consequently, a score that represents each measurement as a precise continuous value may partly reflect characteristics of the laboratory in which it was developed. Efforts to improve model transportability have focused primarily on post development recalibration (12), with comparatively less attention given to the way model inputs are encoded during development. One possible reason is the established statistical guidance against categorizing continuous predictors, because categorization can result in information loss, residual confounding, and bias when cut points are selected in a data driven manner (13). However, these concerns primarily relate to statistical

efficiency and bias associated with categorization and do not directly address whether a clinically motivated categorical representation might be more robust when a model is transported to a different laboratory. In clinical practice, CSF findings are often interpreted using familiar ranges, such as lymphocytic predominance, marked hypoglycorrhachia, and a protein to glucose ratio above a conventional threshold. This raises the possibility that a coarser, clinically motivated encoding could retain clinically relevant diagnostic information while being less sensitive to differences in measurement practices across laboratories.

To examine this question, we developed two models using the same four routine CSF measurements: white cell count, protein, protein to glucose ratio, and lymphocyte percentage, all of which have biological relevance to TBM. In TBM, basal meningeal inflammation increases CSF protein, impairs glucose transport, and, following an early neutrophilic phase, is associated with lymphocytic predominance (4). The two models differed only in how these measurements were encoded. One represented each measurement as a continuous variable, whereas the other used fixed bands that were summed into an integer score. Both models were developed in an Iranian cohort, locked, and applied without refitting to an independent Peruvian cohort with a different case mix and laboratory panel. In the Peruvian cohort, mononuclear cell percentage was used instead of lymphocyte percentage, and HIV status was unavailable, which is particularly relevant to the interpretation of cryptococcal meningitis. Because the two models used the same candidate predictors, development data, and outcome definition, their external performance could be compared while reducing differences attributable to predictor selection or development data. As detailed in the Methods under “Protocol deviations”, the Peruvian external validation cohort and the comparison of the two encodings were added beyond the originally approved single country protocol. We therefore treat the underlying hypothesis and its evaluation in Peru as exploratory rather than confirmatory throughout. The aim of this study was to assess whether categorical encoding improves the external transportability of a CSF based TBM score.

Methods

Study design and participants

We conducted a prediction model development and external validation study and reported the study in accordance with TRIPOD + AI (14). The development cohort comprised 139 patients identified from the Community-acquired Central Nervous System Infections Clinical Data Registry of Mashhad University of Medical Sciences, which enrolls patients hospitalized with community-acquired CNS infections at Imam Reza Hospital, Mashhad, Iran, including 69 patients with TBM and 70 controls. Registry data are entered by trained staff from patients' medical records. Registry records were extracted on 21 September 2025 using the information available at the time of extraction. Eligibility criteria were age ≥ 16 years (observed age range of included patients, 16 to 104 years), clinical features suggestive of meningitis, and at least one CSF abnormality consistent with meningeal inflammation, defined as protein above the standard upper limit of normal (> 45 mg/dL), glucose below the standard lower limit of normal (< 40 mg/dL), or lymphocytosis (> 50%). These thresholds represent conventional CSF reference ranges rather than a

single published TBM specific diagnostic criterion (15). Controls were patients without TBM who had suspected central nervous system infection and an established alternative diagnosis: viral infection (n = 29), neurobrucellosis (n = 21), bacterial meningitis (n = 16), rickettsial infection (n = 2), cryptococcal meningitis (n = 1), and non-infectious malignancy (n = 1) (Table 1, panel C). No patient was excluded because of missing predictor data: all 139 patients in the registry extract had complete CSF white cell count, protein, and glucose values, and the 4 patients whose CSF differential count was unavailable were retained and handled by imputation (see “Missing data”).

TBM diagnostic certainty was classified as definite, probable, or possible using the uniform case definition proposed by Marais et al. (6). The definition assigns points across clinical, CSF, neuroimaging, and evidence of tuberculosis elsewhere criteria, with points deducted when an alternative diagnosis is identified. Definite TBM requires microbiological confirmation from CSF by a positive smear, culture, or nucleic acid amplification test. Probable TBM is defined by a total score of ≥ 10 when neuroimaging is available or ≥ 12 when neuroimaging is unavailable, whereas possible TBM is defined by a score of 6 to 9 with neuroimaging or 6 to 11 without neuroimaging. Of the 69 patients classified as having TBM in the development cohort, 66 met the criteria for definite TBM based on microbiological confirmation, 2 were classified as probable TBM, and 1 as possible TBM (Table 1, panel B). TBM certainty was classified by an independent infectious disease specialist who was blinded to the predictions of the models.

All data were de-identified before analysis. The present analysis of registry data was approved by the Research Ethics Committee of Mashhad University of Medical Sciences (IR.MUMS.MEDICAL.REC.1404.410).

The external cohort was derived from a publicly available, de identified dataset comprising 450 CSF samples analyzed at a private ISO certified laboratory in Lima, Peru, between January 2011 and December 2022. The original authors accessed the dataset on 5 July 2023 (16). The study received ethical approval from the Institutional Ethics Committee of Universidad Privada de Tacna (FACSA CEI/008 04 2023), with a waiver of informed consent. For the present analysis, we restricted the dataset to patients with TBM or one of three alternative diagnoses, resulting in 102 patients: 40 with TBM, 27 with bacterial meningitis, 27 with cryptococcal meningitis, and 8 with viral meningitis. In the source study, TBM and the comparison diagnoses were assigned from laboratory results, and 97.5% of TBM cases were positive on Xpert MTB/RIF or Xpert MTB/RIF Ultra (16). The Marais case definition was not applied to this cohort in the present analysis, because the clinical data required to do so were not available. TBM status in both cohorts therefore rests predominantly on microbiological confirmation, although the reference standards were not identical. The Peruvian dataset reported the percentage of mononuclear cells rather than lymphocyte percentage specifically (16). Because mononuclear cells comprise lymphocytes and monocytes, mononuclear cell percentage was used as the closest available proxy for lymphocyte percentage, although the two measures are not identical. Systematic differences between the two measures could affect a model that weights this predictor heavily differently from a score that uses it only in three bands; this was not directly evaluated. HIV status was unavailable in the Peruvian dataset (16) and was not routinely recorded in the Iranian registry.

Data cleaning

Before analysis, the Iranian registry extract was cleaned, and every change was recorded in a cleaning log. CSF lymphocyte and neutrophil (PMN) percentages, which had been recorded as fractions (0 to 1) in 134 patients and as percentages in 1 patient, were harmonized to a 0 to 100 scale. In four patients (3 with TBM and 1 control, all with a CSF white cell count of 0), lymphocyte and PMN values were both recorded as 0; these were set to missing, because a 0%/0% differential indicates that no differential count was available rather than a measured value. Aetiology labels of controls were standardized, and TBM diagnostic certainty (definite, probable, or possible) was reconstructed from the registry diagnostic codes.

Protocol deviations

The present analysis extends the originally approved study protocol in three respects. First, the protocol specified a single country cohort in which machine learning algorithms would be compared with the Marais case definition. During the analysis, an independent external validation cohort from Peru was added to assess model transportability, although this cohort was not included in the original protocol. Second, rather than developing a single best performing machine learning model, we developed and compared two models that differed only in predictor encoding, continuous versus categorical. This comparison was introduced after initial within cohort analyses and was intended to allow a more direct assessment of the potential role of predictor encoding. Third, the planned comparison with the “Marais scoring system” was not implemented as originally specified because the Marais case definition was used to define TBM status in the study and therefore could not be independently evaluated against the same outcome without circularity. Instead, we assessed the discrimination of the independent Thwaites clinical diagnostic rule (7) (Supplementary Methods S6). These deviations are reported to ensure transparency regarding the analyses undertaken.

Candidate predictors

Four CSF predictors were selected a priori based on clinical knowledge: white cell count, protein, protein to glucose ratio, and lymphocyte percentage, with mononuclear cell percentage used in the Peruvian cohort. Right skewed variables were transformed using log(1 + x) and standardized using the Iranian cohort. The linearity of the association between each predictor and the logit of TBM status was assessed post hoc using two complementary approaches: the Box Tidwell test and a likelihood ratio test comparing a linear specification with a 3 degree of freedom B spline specification (Supplementary Methods S4). All four predictors showed evidence of nonlinearity in the spline analysis (log white cell count: $\chi^2=13.17$, df = 2, $p = 0.0014$; log protein: $\chi^2=8.86$, $p = 0.0119$; lymphocyte percentage: $\chi^2=10.40$, $p = 0.0055$; log protein to glucose ratio: $\chi^2=32.50$, $p < 0.0001$), and the protein to glucose ratio was also significant in the Box Tidwell test ($p < 0.0001$). These findings provide a rationale for considering categorical representations of the predictors, but do not establish that the selected cut points optimally

capture the observed nonlinearity or exclude the possibility that a flexible continuous model could achieve comparable performance.

Missing data

Lymphocyte percentage was missing for 4 of 139 Iranian patients (2.9%) (3 with TBM and 1 control, all with a CSF white cell count of 0 and therefore no differential count), whereas mononuclear cell percentage was missing for 26 of 102 Peruvian patients (25.5%); one additional Peruvian patient had an undefined protein to glucose ratio. Missing values were imputed using chained equations with scikit learn's IterativeImputer, posterior sampling, and the default BayesianRidge estimator, generating 40 imputed datasets with up to 15 iterations per imputation. Log transformed variables were constrained to values ≥ 0, and percentage variables were constrained to the range of 0 to 100. For the Peruvian cohort, the imputation model included the meningitis mimic category as an auxiliary variable. During cross validation, the imputation procedure was refitted independently within each training fold to avoid information leakage.

For the categorical score, each patient's final category for a predictor was defined as the modal category across the 40 imputed datasets. Ties were resolved in favor of the category that was more common for that predictor in the Iranian cohort. Among the 27 Peruvian patients with at least one missing predictor, 26 (96%) had at least one predictor for which the assigned category was not consistent across imputations. In these patients, the resulting total score varied by up to 10 points, and in several cases this variation was sufficient to cross the rule out, indeterminate, or rule in boundaries described in Supplementary Methods S6. Although multiple imputation was used to estimate model performance under missing data, this variability highlights the need for a prespecified and deterministic approach to handling missing predictor values before the score could be considered for bedside implementation.

Development of the continuous (Core) model

A multivariable logistic regression model was fitted to the full Iranian cohort using the four standardized predictors, with no variable selection. Because the cross validated C statistic used to anticipate model performance was derived from the same dataset and analysis, sample size adequacy was assessed post hoc rather than as a prospective planning step. The assessment followed the development sample size criteria of Riley et al. (17), as implemented in the R package pmsampsize. Using the cross validated C statistic of 0.822 and an outcome prevalence of 0.4964, the corresponding Cox Snell R^2 was 0.297 and the Nagelkerke R^2 was 0.397. The required sample size was 100 according to the global shrinkage criterion of ≥ 0.9, 88 according to the criterion targeting a small difference between apparent and adjusted R^2, and 385 according to the criterion for precise estimation of overall outcome risk, which was the binding criterion. The resulting overall minimum sample size was therefore 385 participants, compared with 139 participants available in the development cohort.

Development of the categorical score

The categorical score was developed using the integer point score method described by Sullivan et al. (18). Each predictor was divided into three bands: white cell count (≤ 200, > 200 to 1000, > 1000 cells/μL), lymphocyte or mononuclear percentage (≤ 50%, > 50 to 80%, > 80%), protein to glucose ratio (≤ 1, > 1 to 3, > 3), and protein (≤ 100, > 100 to 200, > 200 mg/dL), with intervals defined as left open and right closed. These bands were specified by the study team on the basis of conventional CSF reference ranges and clinical convention (15). Because the Iranian data were available to the study team and the Peruvian dataset is publicly accessible, and because no time-stamped record documenting the bands before the data were examined is available, we cannot exclude that knowledge of the data influenced band selection. We did not identify a single published source that had validated this specific three level banding scheme for TBM diagnosis. Accordingly, these bands are described as fixed, clinically motivated cut points rather than as externally validated or formally prespecified diagnostic thresholds. Each categorized predictor was entered as a single ordinal term coded 0, 1, or 2.

Protein was retained only if its coefficient exceeded one half of the smallest absolute coefficient among the other three predictors. Because this retention rule uses outcome data, its application was considered exploratory. The final model was then refitted using the retained predictors, and the resulting coefficients were divided by the smallest absolute coefficient and rounded to the nearest integer to derive the final point weights (Supplementary Methods S1). Scaling by the smallest absolute coefficient is a convention whose result becomes unstable when a retained coefficient is close to zero (Supplementary Methods S1). The scaling constant was not fixed in advance, and neither a fixed-constant version nor a reduced-predictor version of the score was evaluated.

Internal validation

Internal validation was performed using repeated stratified 5-fold cross validation with 20 repetitions, yielding 100 folds in total. All data dependent procedures, including imputation, the protein retention decision, and derivation of integer weights, were repeated independently within each training fold. For the Core model, we report both the mean fold wise AUC with its SD and the pooled out of fold AUC. Using the pooled out of fold predictions, we identified the highest probability threshold that achieved ≥ 95% sensitivity. Because this threshold was selected and evaluated using the same out of fold predictions, the resulting threshold specific sensitivity should be regarded as indicative rather than as an unbiased estimate of performance at a prespecified threshold. Given the substantial variation in integer weights across folds, we also recorded, for each fold, the smallest absolute coefficient used as the scaling denominator and the predictor corresponding to that coefficient.

External validation

Both locked models were applied to the Peruvian cohort without refitting. Discrimination was assessed using the AUC overall and within each meningitis mimic group, with confidence intervals estimated by bootstrap resampling stratified by mimic group. The same bootstrap samples were applied across the

40 imputed datasets, and estimates were pooled by combining between imputation variance using Rubin's rules (19) with within imputation bootstrap variance, following the general approach described by Schomaker and Heumann (20). Differences in AUC between the models were estimated using the same paired bootstrap resamples to preserve the paired structure of the comparison.

Calibration was assessed using the Brier score, calibration slope, and calibration in the large. These measures were pooled using the same approach and presented graphically in Fig. 1. Both intercept only and intercept and slope recalibration were applied to the Peruvian cohort using the harmonized 40 imputed datasets (Fig. 1). The pooled Brier score, calibration slope, and calibration in the large reported in Table 3 and the Results refer to the apparent models before recalibration.

Given the relatively small validation cohort, we also performed a post hoc assessment of whether the available sample size met the minimum requirements for precise external validation, using target standard errors of 0.0255 for the C statistic, 0.051 for the calibration slope, and 0.085 for calibration in the large, following the criteria described by Riley et al. (21).

Prespecified versus exploratory analyses

Following TRIPOD + AI guidance, an analysis was considered prespecified only when its methods and comparisons had been defined in a written protocol before outcome linked results were examined. No analysis involving the Peruvian cohort meets this strict definition because the external validation cohort was not included in the original protocol (see "Protocol deviations"). We designated the comparison of the categorical and continuous models with respect to external AUC and calibration as the primary comparative analysis of this manuscript. However, because this comparison was not prespecified according to the stricter definition, it was treated as exploratory and hypothesis generating throughout.

All other analyses were considered exploratory, including the protein exclusion rule, rule out and rule in thresholds, Firth re derivation, the three-predictor continuous model, the extended demographic model, LASSO stability selection, algorithm comparisons, comparison with the Thwaites rule, neurobrucellosis and age and sex subgroup analyses, the two-step reflex algorithm, and sensitivity analyses of boundary conventions.

Potential contributors to differences in external performance were explored by comparing results from multiple imputation with complete case analysis and cohort specific median imputation, examining Spearman correlations between lymphocyte percentage in Iran and mononuclear cell percentage in Peru and the other predictors, fitting a continuous model restricted to the three predictors retained in the categorical score, and repeating external validation after excluding eight patients with predictor values exceeding 3 SD from the Iranian mean. These analyses were intended to assess the sensitivity of the observed findings to alternative analytic assumptions and cohort characteristics rather than to establish a specific causal explanation for differences in external performance.

Additional robustness analyses compared logistic regression with random forest (300 trees, maximum depth of 2, minimum 3 samples per leaf, max features=√p, with hyperparameters tuned using an inner 3-fold grid search), an RBF kernel support vector machine (C = 1, gamma="scale", fixed and not tuned), and a multilayer perceptron (one hidden layer with 8 units, alpha = 0.05, early stopping, fixed and not tuned). These comparisons were performed using the Nadeau Bengio corrected t test (22). The Core model was additionally extended to include age and sex and evaluated using paired nested cross validation. LASSO stability selection was conducted across 200 bootstrap resamples using penalties from {0.05, 0.1, 0.3, 0.5, 1, 2}, with the penalty selected by 3-fold cross validation within each resample. Selection frequency was recorded for each candidate predictor, without applying a prespecified inclusion threshold; these frequencies are therefore presented descriptively rather than as a formal variable selection procedure. Model weights were also re derived using Firth's bias reduced regression (23). Model performance was additionally compared with the Thwaites clinical diagnostic rule (7), although two of its five variables, peripheral blood white cell count and symptom duration, were unavailable in the present dataset.

Subgroup analyses included comparison of TBM with culture or serology confirmed neurobrucellosis (n = 90: 69 TBM and 21 neurobrucellosis) and descriptive assessment of model performance across sex and median age strata; these subgroup analyses were exploratory and were not cross validated. In addition, a two-step reflex approach combining the locked score with a secondary log protein/PMN% based model for cases with indeterminate scores was examined as a supplementary illustrative exercise only (Supplementary Methods S6). The rule out and rule in thresholds used in this illustration were fixed for demonstration and were not derived using a formal threshold selection method. Accordingly, this analysis is presented only to illustrate a possible two step approach and is not proposed as a candidate clinical algorithm.

Clinical utility

Decision curve analysis (DCA) was used to estimate net benefit relative to treat all and treat none strategies across threshold probabilities from 0.01 to 0.99 (24). The full threshold range was retained because no independent rationale was available for restricting the analysis to a narrower clinically relevant range. The categorical score was converted to a predicted probability using univariate logistic recalibration fitted to the apparent Iranian score (intercept − 4.263, slope 0.289), with the recalibration parameters fixed before application to the Peruvian cohort. Net benefit was calculated using two approaches: averaging net benefit calculated separately within each imputed dataset (primary approach), and averaging predicted probabilities across imputed datasets before calculating net benefit (secondary approach, included as a sensitivity analysis). LOESS smoothing was performed using a span of 0.15.

Statistical software and reproducibility

Analyses were conducted in Python 3.13.1 using pandas 3.0.2, NumPy 2.4.4, SciPy 1.17.1, scikit learn 1.8.0 (25), and statsmodels 0.15.0. Development sample size calculations were performed using the R package pmsampsize (17). Bootstrap confidence intervals were based on 2000 resamples with seed =

42. A two-sided $p < 0.05$ threshold was used only for the primary comparison, which was considered exploratory rather than confirmatory.

The analysis code is not publicly available at present because of institutional data governance requirements. The analysis pipeline may be made available by the corresponding author upon reasonable request, subject to institutional approval. The underlying patient level data are not publicly available.

Results

Participants

The Iranian cohort comprised 139 patients, including 69 with TBM and 70 controls (Table 1). The Peruvian cohort comprised 102 patients, including 40 with TBM, 27 with bacterial meningitis, 27 with cryptococcal meningitis, and 8 with viral meningitis, selected from 450 available samples (Fig. 2). In the Iranian cohort, median age was similar between the TBM and control groups (43 vs 46 years, $p = 0.571$), as was sex distribution ($p = 1.000$). CSF protein, glucose, lymphocyte percentage, and protein to glucose ratio were statistically different between the groups (all $p \leq 0.002$), whereas CSF white cell count showed no evidence of a difference ($p = 0.779$; AUC = 0.514). Controls comprised viral (n = 29), neurobrucellosis (n = 21), bacterial (n = 16), rickettsial (n = 2), cryptococcal (n = 1), and non-infectious malignancy (n = 1) diagnoses (Table 1, panel C), a spectrum that differed from the bacterial, cryptococcal, and viral comparators of the Peruvian cohort. The CSF differential count was unavailable for 4 Iranian patients (3 with TBM and 1 control).

Table 1
Baseline characteristics, TBM diagnostic certainty, and final diagnoses of controls, Iranian development cohort

Characteristic	TBM (n = 69)	Control (n = 70)	p value
Panel A. Demographics and data availability			
Age, years, median [IQR]	43.0 [30.0–62.0]	46.0 [30.0–66.0]	0.571
Female sex, n/N (%)	31/69 (44.9%)	31/70 (44.3%)	1.000
CSF differential count unavailable, n (%)	3 (4.3)	1 (1.4)	–
Panel B. TBM diagnostic certainty, n (% of TBM)			
Definite	66 (95.7)	–	–
Probable	2 (2.9)	–	–
Possible	1 (1.4)	–	–
Panel C. Final diagnosis in controls, n (% of controls)			
Viral	–	29 (41.4)	–
Herpes simplex virus	–	23	–
Varicella-zoster virus	–	3	–
SARS-CoV-2	–	3	–
Neurobrucellosis	–	21 (30.0)	–
Bacterial	–	16 (22.9)	–
Pneumococcal	–	9	–
Streptococcus (other)	–	2	–
Meningococcal	–	2	–
Escherichia coli	–	1	–
Listeria	–	1	–
Haemophilus	–	1	–
Rickettsial	–	2 (2.9)	–
Cryptococcal	–	1 (1.4)	–
Non-infectious malignancy	–	1 (1.4)	–

Note: p values are from the Mann–Whitney U test (age) and Fisher's exact test (sex). TBM diagnostic certainty was assigned with the uniform case definition (6). The CSF differential count was unavailable in patients with a CSF white cell count of 0.

Continuous Core model: internal performance

In cross validation, the Core model had a mean fold wise AUC of 0.823 (SD 0.068), with a pooled out of fold AUC of 0.822 (95% CI 0.744 to 0.886). The standardized coefficients were − 0.565 for log white cell count, 1.009 for log protein, 1.516 for lymphocyte percentage, and 0.687 for log protein to glucose ratio. A probability threshold of 0.244 yielded an observed sensitivity of 95.7%, classifying 66 of 69 TBM cases as positive (Table 2). Because the threshold was selected and evaluated using the same out of fold predictions, this sensitivity estimate may be optimistic.

Table 2
Univariate CSF cytochemistry, Iranian development cohort

CSF variable	TBM, median [IQR]	Control, median [IQR]	p value	AUC
White cell count, cells/μL	118.0 [59.0–225.0]	120.0 [52.0–200.0]	0.779	0.514
Protein, mg/dL	177.0 [93.0–280.0]	95.0 [56.5–167.5]	< 0.001	0.681
Glucose, mg/dL	32.0 [21.0–43.0]	54.5 [23.2–77.8]	< 0.001	0.664
Lymphocyte, %	85.0 [77.8–90.0]	80.0 [25.0–90.0]	0.002	0.652
Protein/glucose ratio	5.6 [2.5–11.0]	1.5 [0.7–7.5]	< 0.001	0.701

Continuous Core model: external validation and calibration

When applied to the Peruvian cohort without refitting and using multiple imputation harmonized with the categorical score, the Core model had an AUC of 0.583 (95% CI 0.473 to 0.691), representing a decrease of 0.239 compared with the pooled out of fold cross validation AUC (Table 3). The calibration slope was 0.150 (95% CI − 0.080 to 0.379), and calibration in the large was − 0.817 (95% CI − 1.176 to − 0.457). The mean predicted probability was 0.521, compared with an observed TBM prevalence of 0.392. The Brier score was 0.307 (95% CI 0.249 to 0.365). The validation cohort included 102 participants, compared with the 581 participants estimated to be required for a more precise external evaluation (21).

Table 3
Core continuous model: internal versus external performance

Metric	Setting	Value (95% CI)
AUC	Internal (Iran, nested CV)	0.822 (0.744–0.886)
AUC	External (Peru, locked, harmonized imputation)	0.583 (0.473–0.691)
Brier score	External (Peru, harmonized imputation)	0.307 (0.249–0.365)
Calibration slope	External (Peru)	0.150 (− 0.080 to 0.379)
Calibration-in-the-large	External (Peru)	−0.817 (− 1.176 to − 0.457)
Minimum n, external validation	Required / actual (21)	581 / 102
Minimum n, model development	Required (lower bound) / actual (17)	385 / 139

Note: Brier score, calibration slope, and calibration in the large refer to the apparent models before recalibration (harmonized imputation, m = 40); post recalibration Brier scores were not computed.

Sources of the external performance gap

The Core model had an AUC of 0.543 against bacterial meningitis, 0.606 against cryptococcal meningitis, and 0.642 against viral meningitis. The same general pattern was observed across the five imputation strategies, including outcome aware and outcome blind multiple imputation, complete case analysis restricted to 75 patients, and cohort specific median imputation, with Core model AUCs ranging from 0.511 to 0.608. A continuous model restricted to the same three predictors retained in the categorical score, excluding protein, had an AUC of 0.553 (95% CI 0.435 to 0.669), compared with 0.583 for the four-predictor continuous model. Thus, the three-predictor model had a numerically lower AUC. These results indicate that the observed external performance difference was not attributable solely to the exclusion of protein. They do not test whether the use of mononuclear cell percentage in place of lymphocyte percentage contributed, because lymphocyte percentage, which had the largest standardized coefficient in the continuous model, was retained in both continuous models; this possibility therefore remains untested.

Categorical score: development, internal performance, and weight instability

The protein band had a coefficient of 0.039, which was less than half the smallest coefficient among the other three predictors, and was therefore excluded according to the exploratory rule. The resulting integer weights were − 1 for white cell count, + 5 for lymphocyte percentage, and + 6 for protein to glucose ratio (Table 4). The apparent AUC in Iran was 0.837, while nested cross validation yielded a mean AUC of 0.821 (SD 0.069; median 0.819). The decision to exclude protein varied across folds, with

protein excluded in 52 of 100 folds, and the resulting integer weights showed substantial variation, ranging from 2 to 107 for lymphocyte percentage and from 3 to 121 for the protein to glucose ratio. This variation coincided with differences in the scaling denominator. The smallest absolute coefficient within each fold had a median of 0.246 (IQR 0.150 to 0.390; minimum 0.017), and corresponded to white cell count in 81 folds and protein in 19 folds, but never to lymphocyte percentage or the protein to glucose ratio. Of the 10 folds with the smallest denominator coefficient, 9 were also among the 10 folds with the most extreme integer weights. This pattern suggests that variation in the scaling denominator was closely associated with the observed variation in the integer weights, although it does not by itself establish the source of coefficient instability.

Table 4
Locked categorical score: weights and fold-to-fold stability (100-fold nested CV)

Category	Locked weight	Nested-CV median [min, max]
WBC category	−1	−1 [− 2, 2]
Lymphocyte/mononuclear-% category	+ 5	5.5 [2, 107]
Protein/glucose-ratio category	+ 6	7 [3, 121]
Protein category	dropped (exploratory rule)	dropped in 52/100 folds

Across the 100 folds, the smallest absolute coefficient used as the denominator for deriving the integer weights had a median of 0.246 (IQR 0.150 to 0.390; minimum 0.017). This coefficient corresponded to white cell count in 81 folds and to protein in 19 folds, and did not correspond to lymphocyte percentage or the protein to glucose ratio in any fold. Of the 10 folds with the smallest denominator coefficient, 9 were also among the 10 folds with the most extreme integer weight.

Categorical score: external validation and calibration

Using the same harmonized imputation scheme as the Core model, the categorical score had an overall AUC of 0.699 (95% CI 0.599 to 0.794) in Peru. This was 0.116 higher than the four-predictor continuous model (paired difference 95% CI 0.048 to 0.182, bootstrap $p < 0.001$) and 0.146 higher than the three-predictor continuous model (AUC 0.553, 95% CI 0.435 to 0.669). Because the three-predictor continuous model did not show higher discrimination than the four-predictor model, these comparisons are consistent with a contribution from predictor encoding to the observed difference in external performance, although they do not establish that encoding alone accounted for the difference. By mimic, the AUC was 0.744 for bacterial meningitis (95% CI 0.614 to 0.856), 0.638 for cryptococcal meningitis (95% CI 0.481 to 0.776), and 0.749 for viral meningitis (95% CI 0.527 to 0.920). The difference between the categorical score and the continuous model reached statistical significance only for bacterial meningitis (paired difference 0.202, 95% CI 0.116 to 0.294). For cryptococcal and viral meningitis, the corresponding confidence intervals included zero (0.032 [95% CI − 0.068 to 0.130] and 0.107 [95% CI − 0.008 to 0.234], respectively); the viral comparison included only eight cases (Table 5).

The categorical score had a calibration slope of 0.471 (95% CI 0.163 to 0.779), compared with the corresponding estimates for the continuous models, and a Brier score of 0.237 (95% CI 0.187 to 0.286), compared with 0.307 to 0.328 for the continuous models. The calibration in the large was − 0.319 (95% CI − 0.638 to 0.001). However, the calibration slope of the categorical score remained well below the ideal value of 1, indicating predicted risks that were too extreme in the Peruvian cohort, and its Brier score was close to that of a non-informative model assigning the observed prevalence to every patient (0.392 × 0.608 ≈ 0.238). In the eight patients with viral meningitis, all three models had calibration in the large values ranging from + 2.0 to + 2.5. Given the small size of this subgroup, these estimates should be interpreted cautiously. The observed TBM prevalence in this subgroup was 83%, which was higher than that in the Iranian development cohort.

Table 5
Categorical score performance, internal and external

Setting	AUC	95% CI
Apparent (non-CV), Iran, n = 139	0.837	n/a
Nested CV, Iran (100 folds)	0.821 (SD 0.069)	n/a
External, Peru overall, n = 102	0.699	0.599–0.794
External, Peru vs Bacterial (n = 67)	0.744	0.614–0.856
External, Peru vs Cryptococcal (n = 67)	0.638	0.481–0.776
External, Peru vs Viral (n = 48)	0.749	0.527–0.920
Paired difference vs Core (4-predictor continuous)	+ 0.116	0.048–0.182
Paired difference vs Core3 (3-predictor continuous)	+ 0.146	0.063–0.227

Comparison with a traditional clinical rule

As specified in the study protocol, we attempted to compare the models with a traditional bedside scoring system. The Marais consensus definition (6) was used to define TBM status in this cohort and therefore could not be independently evaluated against the same outcome. We instead assessed the Thwaites clinical diagnostic rule (7), which incorporates five variables: age, peripheral blood white cell count, duration of illness, CSF white cell count, and CSF neutrophil percentage, with a lower total score indicating TBM. Peripheral blood white cell count and duration of illness were not available in either the cleaned or raw Iranian dataset, so the complete rule could not be reconstructed. A partial version using the three available variables, age ≥ 36 years, CSF white cell count ≥ 900, and CSF neutrophil percentage ≥ 75%, with the original published thresholds and without refitting, had an apparent AUC of 0.648 (95% CI 0.564 to 0.731; n = 135/139 complete cases). This apparent estimate was lower than the nested cross validation AUCs of the two study models (0.821 to 0.822), but the estimates are not directly comparable because the partial Thwaites rule was evaluated using apparent performance and did not represent the original five variable rule (Table 6).

Table 6
Head-to-head comparison: algorithms and a traditional clinical rule, Iranian cohort

Method	AUC (95% CI or SD)	vs LR	p value
Partial Thwaites rule (age, CSF WBC, CSF neutrophil% only; not the validated 5-variable rule), apparent (not cross-validated) (7)	0.648 (0.564–0.731)	—	—
Logistic regression / Core model (reference), nested CV	0.823 (SD 0.068)	—	—
Categorical score, nested CV	0.821 (SD 0.069)	−0.002	n.s.
Random Forest, nested CV	0.849 (SD 0.068)	+ 0.026	0.438
Support vector machine (RBF), nested CV (fixed, untuned)	0.835 (SD 0.065)	+ 0.012	0.704
Multilayer perceptron, nested CV (fixed, untuned)	0.657 (SD 0.098)	−0.166	< 0.001

Note: the partial Thwaites-rule AUC is apparent (non-cross-validated) and was not evaluated in the Peruvian cohort, whereas all other rows are nested cross-validated estimates from the Iranian cohort; the first row is therefore not a directly comparable head-to-head figure. The lower multilayer perceptron AUC may reflect its fixed, untuned settings and the small sample rather than a limitation of the method.

Robustness, neurobrucellosis, and decision curve analysis

The random forest (AUC 0.849) and SVM (AUC 0.835) did not show statistically significant differences from logistic regression (AUC 0.823; p = 0.438 and 0.704, respectively), whereas the multilayer perceptron had a lower AUC (0.657; p < 0.001), which may reflect its fixed, untuned settings and the small sample size rather than a limitation of the method. Firth re derivation did not materially change the observed variation in integer weights, with lymphocyte weights ranging from 2 to 112 and ratio weights from 3 to 142; the mean AUC remained 0.821 (SD 0.070). Adding age and sex resulted in an AUC of 0.823 compared with 0.822 for the Core model (p = 0.975). Age and male sex were selected in 76% and 65% of LASSO bootstrap resamples, respectively. These frequencies were reported descriptively, with no selection frequency threshold applied, and the individual coefficients were not statistically significant (age p = 0.251; male sex p = 0.152).

In the neurobrucellosis subgroup (n = 90: 69 TBM and 21 neurobrucellosis), the apparent AUC was 0.7095 (95% CI 0.567 to 0.827) for the Core model and 0.6556 (95% CI 0.512 to 0.796) for the categorical score. No cross validated estimate was obtained for this subgroup.

In the exploratory decision curve analysis using the harmonized 40 dataset imputation scheme in the Peruvian cohort, the categorical score had a positive net benefit relative to the treat all strategy across most of the evaluated threshold range from 0.01 to 0.99, apart from a small interval below 0.02 (Fig. 3).

The four-predictor continuous Core model and the three-predictor continuous model had net benefit exceeding that of the treat all strategy above threshold probabilities of 0.33 and 0.36, respectively. Relative to the treat none strategy, all three models had positive net benefit up to approximately 0.42 to 0.51, specifically from 0.01 to 0.51 for the categorical score, 0.01 to 0.43 for the Core model, and 0.01 to 0.42 for the three-predictor model (Table 7).

Using the secondary approach, in which predicted probabilities were first averaged across imputations before calculating net benefit, the corresponding ranges were 0.08 to 0.99 for the categorical score versus treat all and 0.01 to 0.55 versus treat none; 0.31 to 0.99 and 0.01 to 0.45 for the Core model; and 0.37 to 0.99 and 0.01 to 0.44 for the three-predictor model, respectively. Given the exploratory nature of these analyses and the broad threshold range evaluated, these decision curve results are presented as descriptive estimates of net benefit rather than evidence of clinical superiority. The illustrative two step reflex algorithm is reported only in Supplementary Methods S6 because its zone thresholds were not derived using a documented formal method.

Table 7
Summary of sensitivity and robustness analyses

Analysis	Purpose	Key finding
Random Forest / SVM / MLP (nested CV)	Flexible model class vs logistic regression	No significant difference from LR (RF p = 0.438, SVM p = 0.704); MLP lower AUC (0.657 vs 0.823; p < 0.001); fixed, untuned settings, small sample
Firth bias-reduced regression	Test whether weight instability reflects separation	Weight range widened, not narrowed; AUC unchanged; ordinary logistic retained
Extended model (+ age, +sex)	Test demographic contribution	No improvement (0.823 vs 0.822, p = 0.975)
Three-predictor continuous model (Core3), harmonized imputation	Isolate encoding from predictor selection	AUC 0.553 (0.435–0.669) — not better than 4-predictor model (0.583)
Exclusion of 8 extreme Peru patients	Outlier influence	Score AUC materially unchanged (0.678–0.699); continuous model 0.608–0.655
TBM vs neurobrucellosis (n = 90: 69 TBM and 21 neurobrucellosis)	Hardest single pairwise comparison	Core apparent AUC 0.7095 (0.567–0.827); score apparent AUC 0.6556 (0.512–0.796)
Two-step reflex algorithm (Supplementary S6 only)	Illustrative combination of score with secondary PMN%-based model	Overall AUC 0.654 (SD 0.007); ad hoc thresholds, not a proposed clinical rule
Linearity of continuous predictors	Test whether categorization matches predictors' functional form	All four predictors non-linear by spline LRT (p = 0.0014–<0.0001)
LASSO stability selection (200 bootstrap)	Descriptive predictor selection frequency	No fixed threshold applied; age 75.5%, male 64.5%, log_wbc 90%, log_protein 100%, lym_pct 100%, log_ratio 80%
Development sample size (pmsampsize) (17)	Assess adequacy of Iranian cohort for model development	Minimum n = 385 vs 139 available — underpowered

Discussion

Principal findings

In this two-cohort study, a three-band integer score built from routine CSF variables retained modest discrimination in an external Peruvian cohort (AUC, 0.699), whereas a continuous logistic model developed from the same four candidate variables fell close to chance (AUC, 0.583; paired difference, 0.116; 95% CI, 0.048 to 0.182), despite virtually identical cross-validated performance in the Iranian

development cohort (0.821 vs 0.822). The gap persisted across five strategies for handling the missing mononuclear cell percentage and was not explained by the exclusion of protein, because a continuous model restricted to the three retained predictors performed no better (AUC, 0.553 vs 0.583). The advantage of the categorical score was evident against bacterial meningitis but could not be distinguished from chance against cryptococcal or viral meningitis. Importantly, absolute performance was limited for both models: the better-performing score reached an external AUC below 0.70, and its calibration slope of 0.471 remained far from the ideal value of 1 (12).

A decline in discrimination between internal and external validation is expected and well documented. It reflects overfitting in the development sample, differences in case mix between settings, differences in predictor–outcome associations, and differences in measurement procedures (12, 26, 27). The drop of 0.239 AUC points for the continuous model nevertheless illustrates how fragile a model estimated from 139 patients can be, and it reinforces the view that favorable performance in a single external dataset cannot be taken as validation of a model (27).

The partially reconstructed Thwaites rule (apparent AUC, 0.648 in the Iranian cohort) should not be regarded as a benchmark that either model surpassed. The rule was developed to distinguish TBM from bacterial meningitis in Vietnamese adults and depends on peripheral blood leukocyte count and symptom duration, which were unavailable here (7). Its performance has been inconsistent across settings (10, 11), and in our data it was evaluated only in the Iranian cohort and with apparent rather than cross-validated performance, so it cannot be compared with the external estimates obtained in Peru. Other published rules were built for different comparators or used different inputs: a rule for differentiating TBM from viral meningitis (8), and a Peruvian rule that combined cough duration, CSF cell count, and CSF adenosine deaminase (9). The latter suggests that routine CSF cytochemistry alone may offer limited discrimination in this setting.

Interpretation: is predictor encoding the explanation?

Our results are compatible with, but do not demonstrate, the hypothesis that fixed, clinically motivated bands are less sensitive than continuous values to between-laboratory differences. The statistical literature has long cautioned against categorizing continuous predictors because of loss of information, and empirical work shows that categorized models discriminate and calibrate worse than models that retain continuous predictors when both are evaluated in data from the same population (13, 28, 29). Our findings do not contradict that literature, because those studies did not examine transport to a setting in which measurement procedures and case mix differ. Coarse, near-unit-weight representations are known to be remarkably robust when development samples are small (30), and an encoding that depends less on exact numerical values could plausibly be advantageous when laboratory practices differ.

Several features of the design prevent attributing the difference to encoding alone. First, the categorical score differs from the continuous model not only in encoding but also in its integer weights, which are of similar magnitude for the two dominant predictors (+ 5 and + 6), and in the omission of protein; the three-

predictor continuous model addresses only the latter. Second, the comparator is the simplest continuous specification, a log-linear logistic model, although all four predictors showed non-linear associations with TBM status. Categorization performed worse than approaches that retain continuous predictors, whether modelled linearly or non-linearly with fractional polynomials or splines (29), so a spline-based or penalized continuous comparator, and a model using the same bands with unrounded logistic coefficients, are needed to separate the contribution of encoding from that of weighting and shrinkage. Third, the largest standardized coefficient in the continuous model belonged to lymphocyte percentage (1.516), which was replaced in Peru by mononuclear cell percentage, a related but not identical measure. Any systematic difference between these two variables would affect a model that relies heavily on one continuous value more than a score that distinguishes only three bands. The three-predictor comparison retained the lymphocyte variable and therefore could not test this possibility; a model excluding it would be informative.

Differences in case mix and case ascertainment are equally plausible contributors. The Peruvian cohort was restricted to patients with an identified causative pathogen, and spectrum effects arising from restricted sampling of cases or controls are a recognized source of variation in diagnostic accuracy estimates (31–33). The composition of the mimic group (27 bacterial, 27 cryptococcal, and 8 viral cases) is also unlikely to mirror the diagnoses encountered in routine practice in either country.

The mimics that were hardest to separate from TBM with the categorical score, cryptococcal meningitis in Peru and neurobrucellosis in Iran, are plausible sources of residual biological overlap. In adults with HIV infection, clinical and laboratory findings have been reported to be similar in tuberculous and cryptococcal meningitis (34), and cryptococcal meningitis also occurs in other states of impaired cell-mediated immunity (35). HIV status, which was unavailable in both of our cohorts, is therefore relevant to case mix and to prognosis, because HIV co-infection substantially reduces survival in TBM (36). Neurobrucellosis can present as a lymphocytic meningitis with raised protein and low glucose and may be mistaken for tuberculosis (37, 38); consistent with this, neither model discriminated well between TBM and neurobrucellosis (apparent AUC, 0.71 and 0.66, respectively). Routine CSF cytochemistry therefore appears to have an inherent ceiling for these comparisons, and pathogen-specific testing is likely to be needed to resolve them.

Calibration and clinical utility

The calibration of the categorical score was more favorable only in relative terms. Its calibration slope of 0.471 (95% CI, 0.163 to 0.779) indicates predicted risks that remain too extreme in the new setting, and the interval excludes the ideal value of 1; exclusion of zero is not a meaningful benchmark. The slope of the continuous model (0.150) indicates almost no relationship between predicted and observed risk (12). Moreover, the Brier score of the categorical score (0.237) was almost identical to the value of approximately 0.238 expected from a non-informative model that assigns the observed prevalence (0.392) to every patient (0.392 × 0.608), so the gain in overall accuracy over a null model was minimal (39); the continuous models (0.307 to 0.328) performed worse than the null model. Predicted probabilities also depend on the outcome prevalence of the development sample: TBM accounted for

49.6% of the Iranian cohort and 39.2% of the Peruvian cohort, and the mean predicted risk in Peru exceeded the observed risk (0.521 vs 0.392). Any clinical use would therefore require recalibration to the local prevalence (12).

The decision curve analysis should be read with the same caution. Net benefit is informative only across threshold probabilities that reflect plausible clinical trade-offs (24, 40, 41), whereas we evaluated the full range from 0.01 to 0.99. Because the consequences of delayed antituberculous treatment are severe (4), clinically relevant thresholds for starting empirical therapy are likely to be low, although the appropriate range should be defined by clinicians in advance. In that region, the categorical score showed no advantage over treating all patients below a threshold of 0.02 under the primary approach and below 0.08 under the secondary approach. Because the validation sample had a TBM prevalence of 39%, the net benefit estimates are specific to that prevalence and will differ where TBM is less common among patients with meningitis.

Robustness and weight instability

Flexible algorithms did not consistently improve on logistic regression in the Iranian data (random forest AUC, 0.849; support vector machine, 0.835; logistic regression, 0.823), in line with systematic evidence that machine learning methods offer no consistent advantage over logistic regression for clinical prediction models (42). The lower performance of the multilayer perceptron (AUC, 0.657) may reflect an unfavorable ratio of parameters to events at this sample size and its fixed, untuned settings rather than a property of the method (17).

The integer weights of the categorical score varied substantially across folds (2 to 107 for lymphocyte percentage and 3 to 121 for the protein to glucose ratio), and this variation was closely associated with the scaling denominator. Dividing all coefficients by the smallest absolute coefficient is a convenient convention, but it makes the weights explode whenever that coefficient approaches zero, which is likely when a weakly informative predictor such as white cell count (univariate AUC, 0.514) is retained. This is a property of the scaling rule rather than of the data. Fixing the scaling constant in advance, as in the original Framingham point system (18), would remove this source of instability. The negative weight retained for white cell count may also reflect noise, and a simpler two-predictor score might perform equally well; this should be tested. More generally, instability of individual predictions is a recognized consequence of model building in small samples and manifests as miscalibration in new data (43, 44), so the locked weights (− 1, + 5, and + 6) should be viewed as one realization among many plausible ones. Firth's penalized estimation, which addresses small-sample bias and separation (23), did not reduce the variation, which is consistent with instability arising from the scaling step rather than from the coefficient estimates themselves.

A practical limitation for deployment

Multiple imputation is an appropriate way to estimate model performance when predictors are missing in a validation dataset, but it does not transfer to practice, where a missing value for a single patient must be handled at the bedside without access to the outcome or to a reference dataset (45–47).

Approaches that can be applied to one new patient, such as submodels based on observed predictors only, marginalization, or regression imputation, should be specified and evaluated before deployment (45, 47). Our imputation model for the Peruvian cohort also used the mimic category as an auxiliary variable, which is not available at the point of care; reassuringly, outcome-aware and outcome-blind imputation gave similar results for the continuous model (AUC range across five strategies, 0.511 to 0.608). Among the 27 Peruvian patients with at least one missing predictor, 26 (96%) had at least one band that varied across the imputed datasets, and the total score varied by up to 10 points, in several cases crossing the rule-out, indeterminate, and rule-in boundaries. Because a small change in one value can move a patient into a different band, a banded score may be particularly vulnerable to this problem, and a prespecified deterministic missing-data rule remains an unresolved prerequisite for bedside use.

Strengths and limitations

Strengths of this study include the application of locked models, without refitting, to an independent cohort from a different continent; a paired bootstrap comparison under harmonized imputation; the reporting of calibration and net benefit in addition to discrimination; and transparent documentation of protocol deviations in line with TRIPOD + AI (14).

The study also has important limitations. The development cohort (139 patients) and the external validation cohort (102 patients) were smaller than the 385 and 581 participants estimated to be required by the corresponding sample size criteria (17, 21), so estimates are imprecise and subgroup results, particularly for the eight viral cases, are highly uncertain. The Peruvian cohort was restricted to patients with an identified pathogen, which may have introduced spectrum bias of unknown direction (31, 32). In the Iranian cohort, eligibility also required at least one CSF abnormality defined by protein, glucose, or lymphocyte percentage, which overlaps with the candidate predictors and may have altered the spectrum of controls and the apparent performance of these predictors; the control spectrum in Iran (mainly viral, neurobrucellosis, and bacterial) also differed from the Peruvian comparators (bacterial, cryptococcal, and viral). The reference standards were not identical between cohorts: in Iran, 66 of 69 TBM cases were microbiologically confirmed and classified according to the uniform case definition (6), whereas in Peru TBM status was taken from the laboratory-based classification of the source dataset, in which 97.5% of TBM cases were positive on Xpert MTB/RIF or Xpert MTB/RIF Ultra (16), and was not reclassified using the uniform case definition. Because both cohorts consist predominantly of microbiologically confirmed TBM, the score has not been evaluated in culture- and Xpert-negative or clinically diagnosed disease, whose CSF profile may differ. The Iranian cohort had a TBM proportion set by the sampling framework that is higher than would be expected in most clinical populations. Mononuclear cell percentage was used as a proxy for lymphocyte percentage and was missing in one quarter of Peruvian patients, and HIV status was unavailable in both cohorts. Socioeconomic markers and prior antituberculous or antibiotic treatment were not recorded in either dataset, so health inequalities between sociodemographic groups could not be examined and any effect of prior treatment on CSF values could not be assessed. The retrospective design limits control over measurement and ascertainment. Because CSF findings contribute to the uniform case definition, some incorporation bias

is possible in the Iranian cohort, although 66 of 69 cases were microbiologically confirmed. The cut points of the categorical score were investigator-defined rather than formally prespecified, and because the Peruvian dataset is public and no time-stamped record of the bands is available, some influence of knowledge of the data on their choice cannot be excluded. Many analyses were exploratory and were not adjusted for multiplicity; this is consistent with their hypothesis-generating purpose (48), but nominal p-values should not be read as confirmatory. Fairness was examined only through exploratory sex and age subgroup analyses, and no formal fairness metrics were evaluated. A formal appraisal with PROBAST or PROBAST + AI would be advisable, and the small samples would be expected to place the analysis domain at high risk of bias (49, 50). Finally, the study extended beyond the originally registered protocol, as described under "Protocol deviations", and the comparison of categorical and continuous encoding should therefore be interpreted as hypothesis generating.

Clinical implications and future directions

The score is not suitable for clinical deployment in its current form. Overall discrimination was modest, its advantage over the continuous model was not consistent across mimics, no evaluated threshold cleanly separated TBM from all mimics, and a deterministic approach to missing predictors has not been established. Future work should evaluate the score in larger, prospective, multicountry cohorts with directly measured lymphocyte percentage and systematically documented HIV status, following current guidance for external validation studies (51) and, where several datasets are available, internal–external cross-validation (52). Protocol-registered comparisons of predictor encodings should include spline-based continuous models and banded models with unrounded coefficients, together with a prespecified missing-data strategy (45, 47). Because routine CSF cytochemistry has an apparent ceiling, the score should also be compared with the complete five-variable Thwaites rule (7) and examined as a triage step before pathogen-specific testing, such as Xpert MTB/RIF Ultra, which detected more TBM than Xpert or culture in HIV-infected adults but cannot exclude the disease when negative (5, 53), and CSF Brucella agglutination testing for patients with indeterminate scores (37).

Conclusions

In a single external cohort from a different country and laboratory, a CSF score based on fixed, clinically motivated bands showed higher discrimination than a continuous model, although calibration remained poor for both models. The comparison with continuous models using the same predictors suggests that predictor encoding may have contributed to this difference, although the study does not establish that encoding alone accounted for it. The difference was not consistent across all meningitis mimics, absolute discrimination remained modest, and both the development and external validation cohorts were smaller than the sample sizes estimated by the Riley criteria. These findings support further investigation of predictor encoding as a potential determinant of model transportability, using larger, prospective, multicountry, protocol-registered cohorts and incorporating a prespecified, deployable strategy for handling missing predictor values.

Declarations

Consent for Publication

Not applicable.

Competing Interests

None declared.

Funding

This work was supported by a grant from the Vice Chancellor for Research at Mashhad University of Medical Sciences (Grant No. 4041316) and was part of the M.D. thesis of Hamid Reza Mirzaei. The funder had no role in the design, analysis, interpretation, or reporting of the study.

Ethics Approval and Consent to Participate

This study was conducted in accordance with the Declaration of Helsinki. The analysis of registry data was approved by the Research Ethics Committee of Mashhad University of Medical Sciences (IR.MUMS.MEDICAL.REC.1404.410). All participants, or their legal guardians where applicable, provided informed consent for inclusion in the registry, and all data were de identified before analysis. The Peruvian external validation cohort is a publicly available, de-identified dataset; its original study was approved by the Institutional Ethics Committee of Universidad Privada de Tacna (FACSA CEI/008 04 2023), with a waiver of informed consent.

Clinical trial number: not applicable.

Study Registration and Protocol

This study was not registered in a public registry. The study protocol was approved by the Research Ethics Committee of Mashhad University of Medical Sciences (IR.MUMS.MEDICAL.REC.1404.410) and has not been deposited in a public repository.

Patient and Public Involvement

Patients and the public were not involved in the design, conduct, or reporting of this study.

Authors' Contributions

Hamid Reza Naderi: Conceptualization, project administration, supervision, writing – review and editing.

Vahideh Zolfaghari: Data curation, investigation, methodology, formal analysis, validation, writing – original draft preparation, writing – review and editing.

Hamid Reza Mirzaei: Data curation, Validation, Review and editing.

Majid Khadem Rezayian: Review and editing.

Fereshte Sheybani: Investigation, Data curation, Validation.

Mahboubeh Haddad: Investigation, Data curation, Validation.

All authors read and approved the final manuscript.

Acknowledgments

AI-assisted tools were used for language editing and translation. All substantive content was produced and reviewed by the authors.

Availability of Data and Materials

The analysis code is not publicly available at present because of institutional data governance requirements. The analysis pipeline may be made available by the corresponding author upon reasonable request, subject to institutional approval. The underlying patient level data are not publicly available.

References

1. Dodd PJ, Osman M, Cresswell FV, Stadelman AM, Lan NH, Thuong NT, Muzyamba M, Glaser L, Dlamini SS, Seddon JA. The global burden of tuberculous meningitis in adults: a modelling study. PLOS Global Public Health. 2021;1(12):e0000069.
2. World Health Organization. Global tuberculosis report 2023. Geneva: World Health Organization; 2023.
3. Kiani B, Raouf Rahmati A, Bergquist R, Hashtarkhani S, Firouraghi N, Bagheri N, Moghaddas E, Mohammadi A. Spatio-temporal epidemiology of the tuberculosis incidence rate in Iran 2008 to 2018. BMC Public Health. 2021;21(1):1093.
4. Wilkinson RJ, Rohlwink U, Misra UK, Van Crevel R, Mai NT, Dooley KE, Caws M, Figaji A, Savic R, Solomons R, Thwaites GE. Tuberculous meningitis. Nat reviews Neurol. 2017;13(10):581–98.
5. Donovan J, Cresswell FV, Thuong NT, Boulware DR, Thwaites GE, Bahr NC. Xpert MTB/RIF ultra for the diagnosis of tuberculous meningitis: a small step forward. Clin Infect Dis. 2020;71(8):2002–5.
6. Marais S, Thwaites G, Schoeman JF, Török ME, Misra UK, Prasad K, Donald PR, Wilkinson RJ, Marais BJ. Tuberculous meningitis: a uniform case definition for use in clinical research. Lancet Infect Dis. 2010;10(11):803–12.
7. Thwaites GE, Chau TT, Stepniewska K, Phu NH, Chuong LV, Sinh DX, White NJ, Parry CM, Farrar JJ. Diagnosis of adult tuberculous meningitis by use of clinical and laboratory features. lancet. 2002;360(9342):1287–92.

8. Hristea A, Olaru ID, Baicus C, Moroti R, Arama V, Ion M. Clinical prediction rule for differentiating tuberculous from viral meningitis. Int J Tuberculosis Lung Disease: Official J Int Union Against Tuberculosis Lung Disease. 2012;16(6):793–8.
9. Solari L, Van der Soto A. Development of a clinical prediction rule for tuberculous meningitis in adults in Lima, Peru. Tropical Med Int Health. 2018;23(4):367–74.
10. Vibha D, Bhatia R, Prasad K, Srivastava MP, Tripathi M, Kumar G, Singh MB. Validation of diagnostic algorithm to differentiate between tuberculous meningitis and acute bacterial meningitis. Clin Neurol Neurosurg. 2012;114(6):639–44.
11. Sulaiman T, Medi S, Erdem H, Senbayrak S, Ozturk-Engin D, Inan A, Civljak R, Nechifor M, Akbulut A, Crisan A, Ozguler M. The diagnostic utility of the Thwaites' system and lancet consensus scoring system in tuberculous vs. non-tuberculous subacute and chronic meningitis: multicenter analysis of 395 adult patients. BMC Infect Dis. 2020;20(1):788.
12. Van Calster B, McLernon DJ, Van Smeden M, Wynants L, Steyerberg EW. Topic Group 'Evaluating diagnostic tests and prediction models' of the STRATOS initiative Bossuyt Patrick Collins Gary S. Macaskill Petra McLernon David J. Moons Karel GM Steyerberg Ewout W. Van Calster Ben van Smeden Maarten Vickers Andrew J. Calibration: the Achilles heel of predictive analytics. BMC Med. 2019;17(1):230.
13. Royston P, Altman DG, Sauerbrei W. Dichotomizing continuous predictors in multiple regression: a bad idea. Stat Med. 2006;25(1):127–41.
14. Collins GS, Moons KG, Dhiman P, Riley RD, Beam AL, Van Calster B, Ghassemi M, Liu X, Reitsma JB, Van Smeden M, Boulesteix AL. TRIPOD + AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ. 2024;385.
15. Seehusen DA, Reeves MM, Fomin DA. Cerebrospinal fluid analysis. Am Family Phys. 2003;68(6):1103–9.
16. Hueda-Zavaleta M, de la Torre JC, Barletta-Carrillo C, Tapia-Sequeiros G, Flores C, Piscoche C, Miranda C, Mendoza A, Sánchez-Tito M, Benites-Zapata VA. Cytochemical analysis of cerebrospinal fluid in tuberculous meningitis versus other etiologies. PLoS ONE. 2025;20(3):e0318398.
17. Riley RD, Snell KI, Ensor J, Burke DL, Harrell FE Jr, Moons KG, Collins GS. Minimum sample size for developing a multivariable prediction model: PART II-binary and time-to-event outcomes. Stat Med. 2019;38(7):1276–96.
18. Sullivan LM, Massaro JM, D'Agostino Sr RB. Presentation of multivariate data for clinical use: the Framingham Study risk score functions. Stat Med. 2004;23(10):1631–60.
19. Rubin DB. Multiple imputations in sample surveys-a phenomenological Bayesian approach to nonresponse. InProceedings of the survey research methods section of the American Statistical Association 1978 (Vol. 1, pp. 20–34). Alexandria, VA: American Statistical Association.
20. Schomaker M, Heumann C. Bootstrap inference when using multiple imputation. Stat Med. 2018;37(14):2252–66.

21. Riley RD, Debray TP, Collins GS, Archer L, Ensor J, van Smeden M, Snell KI. Minimum sample size for external validation of a clinical prediction model with a binary outcome. Stat Med. 2021;40(19):4230–51.
22. Nadeau C, Bengio Y. Inference for the generalization error. Adv Neural Inf Process Syst. 1999;12.
23. Firth D. Bias reduction of maximum likelihood estimates. Biometrika. 1993 Mar;1:27–38.
24. Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. Med Decis Making. 2006;26(6):565–74.
25. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J. Scikit-learn: Machine learning in Python. J Mach Learn Res. 2011;12:2825–30.
26. Debray TP, Vergouwe Y, Koffijberg H, Nieboer D, Steyerberg EW, Moons KG. A new framework to enhance the interpretation of external validation studies of clinical prediction models. J Clin Epidemiol. 2015;68(3):279–89.
27. Van Calster B, Wynants L, Timmerman D, Steyerberg EW, Collins GS. Predictive analytics in health care: how can we know it works? J Am Med Inform Assoc. 2019;26(12):1651–4.
28. Bennette C, Vickers A. Against quantiles: categorization of continuous variables in epidemiologic research, and its discontents. BMC Med Res Methodol. 2012;12(1):21.
29. Collins GS, Ogundimu EO, Cook JA, Manach YL, Altman DG. Quantifying the impact of different approaches for handling continuous predictors on the performance of a prognostic model. Stat Med. 2016;35(23):4124–35.
30. Dawes RM. The robust beauty of improper linear models in decision making. Am Psychol. 1979;34(7):571.
31. Ransohoff DF, Feinstein AR. Problems of spectrum and bias in evaluating the efficacy of diagnostic tests. N Engl J Med. 1978;299(17):926–30.
32. Mulherin SA, Miller WC. Spectrum bias or spectrum effect? Subgroup variation in diagnostic test evaluation. Ann Intern Med. 2002;137(7):598–602.
33. Rutjes AW, Reitsma JB, Vandenbroucke JP, Glas AS, Bossuyt PM. Case–control and two-gate designs in diagnostic accuracy studies. Clin Chem. 2005;51(8):1335–41.
34. Cohen DB, Zijlstra EE, Mukaka M, Reiss M, Kamphambale S, Scholing M, Waitt PI, Neuhann F. Diagnosis of cryptococcal and tuberculous meningitis in a resource-limited African setting. Tropical Med Int Health. 2010;15(8):910–7.
35. Williamson PR, Jarvis JN, Panackal AA, Fisher MC, Molloy SF, Loyse A, Harrison TS. Cryptococcal meningitis: epidemiology, immunology, diagnosis and therapy. Nat Reviews Neurol. 2017;13(1):13–24.
36. Thwaites GE, Duc Bang N, Huy Dung N, Thi Quy H, Thi Tuong Oanh D, Thi Cam Thoa N, Quang Hien N, Tri Thuc N, Ngoc Hai N. Thi Ngoc Lan N, Ngoc Lan N. The influence of HIV infection on clinical

presentation, response to treatment, and outcome in adults with tuberculous meningitis. J Infect Dis. 2005;192(12):2134–41.

37. Guven T, Ugurlu K, Ergonul O, Celikbas AK, Gok SE, Comoglu S, Baykam N, Dokuzoguz B. Neurobrucellosis: clinical and diagnostic features. Clin Infect Dis. 2013;56(10):1407–12.
38. Ceran N, Turkoglu R, Erdem I, Inan A, Engin D, Tireli H, Goktas P. Neurobrucellosis: clinical, diagnostic, therapeutic features and outcome. Unusual clinical presentations in an endemic region. Brazilian J Infect Dis. 2011;15(1):52–9.
39. Steyerberg EW, Vickers AJ, Cook NR, Gerds T, Gonen M, Obuchowski N, Pencina MJ, Kattan MW. Assessing the performance of prediction models: a framework for some traditional and novel measures. Epidemiol (Cambridge Mass). 2010;21(1):128.
40. Vickers AJ, Van Calster B, Steyerberg EW. Net benefit approaches to the evaluation of prediction models, molecular markers, and diagnostic tests. BMJ. 2016;352.
41. Vickers AJ, Van Calster B, Steyerberg EW. A simple, step-by-step guide to interpreting decision curve analysis. Diagn prognostic Res. 2019;3(1):18.
42. Christodoulou E, Ma J, Collins GS, Steyerberg EW, Verbakel JY, Van Calster B. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. J Clin Epidemiol. 2019;110:12–22.
43. Riley RD, Collins GS. Stability of clinical prediction models developed using statistical or machine learning methods. Biom J. 2023;65(8):2200302.
44. Riley RD, Ensor J, Snell KI, Harrell FE, Martin GP, Reitsma JB, Moons KG, Collins G, Van Smeden M. Calculating the sample size required for developing a clinical prediction model. BMJ. 2020;368.
45. Hoogland J, van Barreveld M, Debray TP, Reitsma JB, Verstraelen TE, Dijkgraaf MG, Zwinderman AH. Handling missing predictor values when validating and applying a prediction model to new patients. Stat Med. 2020;39(25):3591–607.
46. Sperrin M, Martin GP, Sisk R, Peek N. Missing data should be handled differently for prediction than for description or causal explanation. J Clin Epidemiol. 2020;125:183–7.
47. Sisk R, Sperrin M, Peek N, van Smeden M, Martin GP. Imputation and missing indicators for handling missing data in the development and deployment of clinical prediction models: a simulation study. Stat Methods Med Res. 2023;32(8):1461–77.
48. Bender R, Lange S. Adjusting for multiple testing—when and how? J Clin Epidemiol. 2001;54(4):343–9.
49. Wolff RF, Moons KG, Riley RD, Whiting PF, Westwood M, Collins GS, Reitsma JB, Kleijnen J, Mallett S, PROBAST Group†. PROBAST: a tool to assess the risk of bias and applicability of prediction model studies. Ann Intern Med. 2019;170(1):51–8.
50. Moons KG, Damen JA, Kaul T, Hooft L, Navarro CA, Dhiman P, Beam AL, Van Calster B, Celi LA, Denaxas S, Denniston AK. PROBAST + AI: an updated quality, risk of bias, and applicability assessment tool for prediction models using regression or artificial intelligence methods. BMJ. 2025;388.

51. Collins GS, Dhiman P, Ma J, Schlussel MM, Archer L, Van Calster B, Harrell FE, Martin GP, Moons KG, Van Smeden M, Sperrin M. Evaluation of clinical prediction models (part 1): from development to external validation. BMJ. 2024;384.
52. Steyerberg EW, Harrell FE Jr. Prediction models need appropriate internal, internal-external, and external validation. J Clin Epidemiol. 2015;69:245.
53. Bahr NC, Nuwagira E, Evans EE, Cresswell FV, Bystrom PV, Byamukama A, Bridge SC, Bangdiwala AS, Meya DB, Denkinger CM, Muzoora C. Diagnostic accuracy of Xpert MTB/RIF Ultra for tuberculous meningitis in HIV-infected adults: a prospective cohort study. Lancet Infect Dis. 2018;18(1):68–75.

Figures

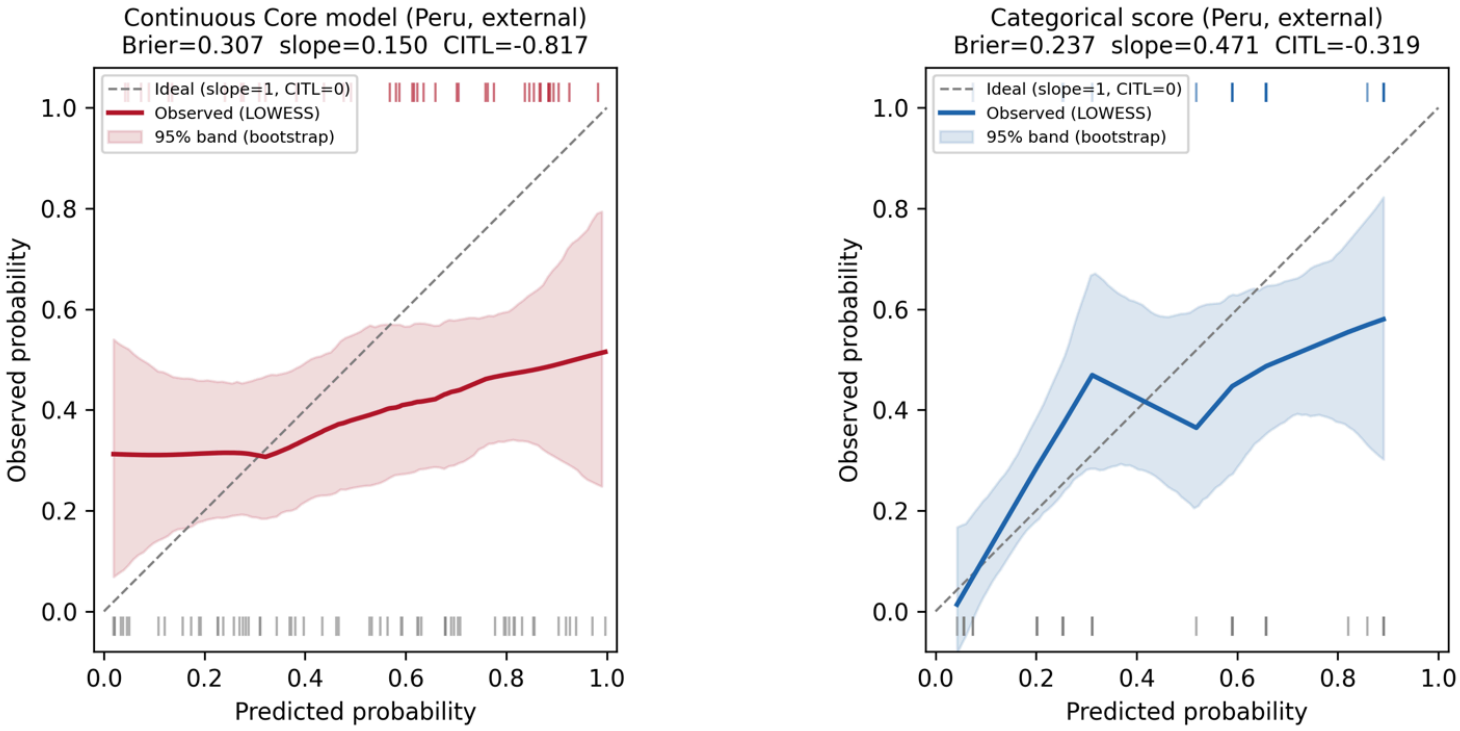


Figure 1

Calibration plots for the Core model and categorical score in the Peruvian external validation cohort.

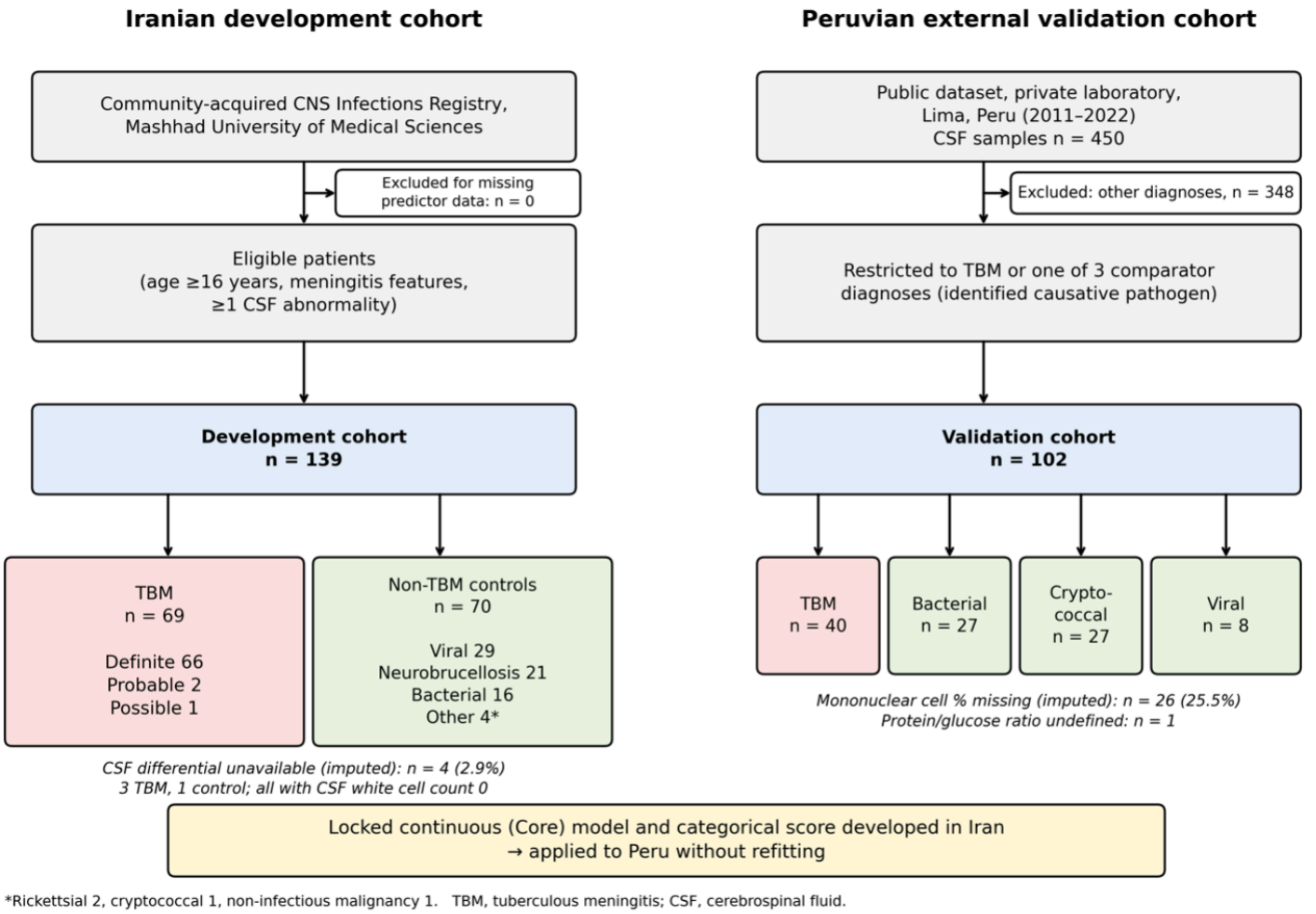


Figure 2

Participant flow for the Iranian development cohort and Peruvian external validation cohort. Control diagnoses and TBM diagnostic certainty are shown for the Iranian cohort, and the numbers of patients with missing predictor values (imputed) are shown for each cohort. In the Peruvian dataset, 348 of the 450 CSF samples had diagnoses other than TBM or the three comparator diagnoses and were not included. No patient was excluded because of missing predictor data.

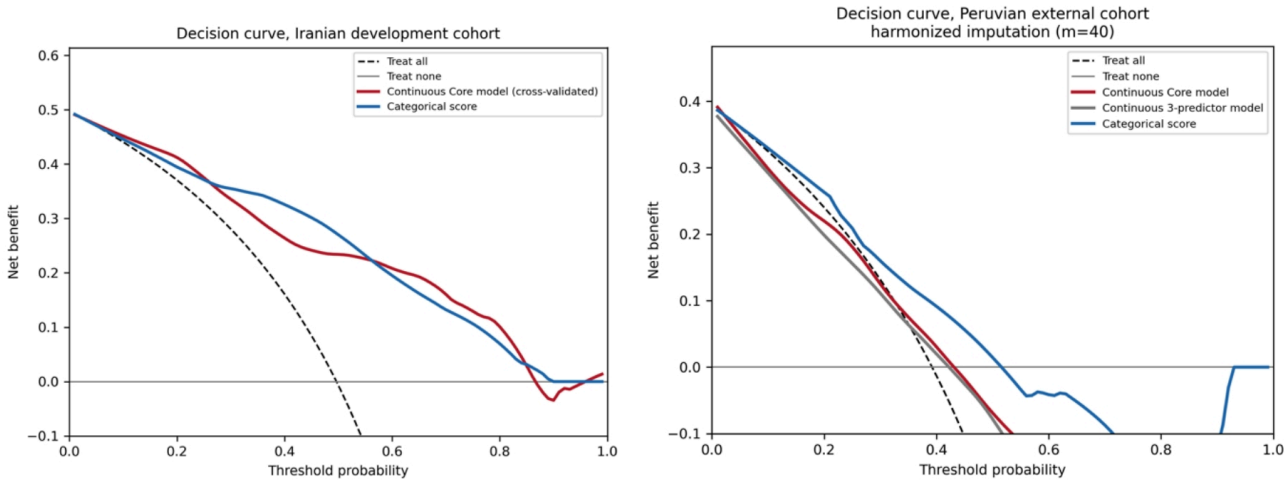


Figure 3

Decision curve analysis in the Iranian development cohort (internal) and Peruvian external validation cohort (external, harmonized imputation).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryMethods.docx