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.

Calibration and Missingness in External Confirmation of a Transparent Next-Day Extubation Model

Alexis M. Collier

alexis.collier@ung.edu

University of North Georgia

Research Article

Keywords: clinical prediction model, external validation, calibration, missing data, mechanical ventilation, extubation, clinical artificial intelligence, reproducibility

Posted Date: November 4th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11051386/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

Discrimination alone does not show whether a clinical prediction model transfers adequately. Calibration, predictive value, threshold behavior, and missingness can change the interpretation of favorable external discrimination.

Methods

We conducted a locked secondary analysis of the credentialed SCRIPT X2B8 version 1.0.0 dataset. A 10-feature respiratory comparator and a 37-feature logistic model were trained only on the supplied internal training cohort. Preprocessing and model specifications were frozen before confirmation. Thresholds selected on the internal validation cohort by maximizing Youden's J were applied unchanged to the internal test and institutional external cohorts. We assessed discrimination, calibration, prediction error, threshold metrics, patient-cluster uncertainty, and missingness-stratified performance.

Results

The internal test cohort contained 831 patient-day observations and 65 events. The full-model area under the receiver operating characteristic curve was 0.8553 versus 0.7916 for the comparator, a difference of 0.0636 (95% confidence interval 0.0236 to 0.1046). The external cohort contained 2,835 observations and 192 events. External area under the curve was 0.8490 versus 0.8183, a difference of 0.0307 (95% confidence interval 0.0030 to 0.0584). The full model had a worse external Brier score than the comparator, 0.0611 versus 0.0589, and a calibration slope of 0.675. At the frozen threshold, sensitivity was 0.885, specificity 0.683, positive predictive value 0.168, and negative predictive value 0.988. Full-model area under the curve declined from 0.8839 in the lowest external missingness quartile to 0.7681 in the highest.

Conclusions

The full model retained external discrimination and exceeded the respiratory comparator, but calibration, positive predictive value, and missingness findings limit readiness claims. These results support continued evaluation, not clinical use, deployment readiness, or patient benefit.

Background

Clinical prediction models may retain rank discrimination across settings while producing probabilities or threshold behavior that do not transfer adequately [1]. For health-system evaluation, an external area under the receiver operating characteristic curve is therefore only one part of the evidence. Calibration,

predictive value, missingness, cohort construction, and prespecified thresholds affect whether a result can support a local decision. TRIPOD + AI provides reporting guidance for prediction models developed using regression or machine learning [2].

The SCRIPT X2B8 dataset contains deidentified patient-day data for patients receiving mechanical ventilation. The dataset creators assembled 37 clinical features measured from midnight through 8:00 AM and next-day extubation labels. The released data include internal training, validation, and test partitions plus an institutional external cohort [3]. The creators previously reported multiple machine-learning models, including logistic regression and recurrent models, using the same clinical task [4]. The present study is not a first analysis of the dataset and is not a historically untouched external validation. Its distinct purpose is a transparent confirmation audit that compares a prespecified respiratory feature set with all 37 released features while retaining calibration, fixed-threshold, missingness, and blocked-analysis findings beside discrimination.

The primary question was whether a fixed logistic model trained only on the supplied internal training cohort could predict successful next-day extubation in the internal test and institutional external cohorts. The primary incremental contrast was the full model minus the respiratory comparator. Clinical utility, treatment benefit, workflow effect, and deployment readiness were not estimands.

Methods

Study design and data source

This was a retrospective secondary analysis of SCRIPT X2B8 Dataset version 1.0.0, distributed through PhysioNet under credentialed access [3, 5, 6]. The local source archive had SHA-256 hash 89A2A646D88B4A1B1D7C1DF07AB23D56CD62757DB9653C93C1317B360DF8156A.

The released internal cohort originated from the Successful Clinical Response in Pneumonia Therapy program at a quaternary-care hospital. The institutional external cohort was collected from a community-hospital intensive care unit. The dataset custodians report that the original cohorts were governed by Northwestern University Institutional Review Board protocols STU00204868 and STU00216678 [3, 4]. The current secondary analysis used only deidentified data, did not access direct identifiers, did not attempt reidentification, and did not redistribute row-level source data.

The supplied partitions contained 2,264 training rows, 594 validation rows, 831 internal test rows, and 2,835 external rows. Patient identifiers were derived from the prefix of the released Patient_stay field. The internal training, validation, and test partitions had no patient or intensive-care-stay identifier overlap. Twelve deidentified patient-number prefixes occurred in both the internal and external files. The public documentation does not establish that these numbers are globally unique across the two institutions. We therefore retain this as unresolved identifier collision and do not treat the external cohort as historically or identifier-proven untouched.

Outcome

Outcome

The positive class was successful next-day extubation, encoded as Labels = = 0 in the released main cohort files. Patients with tracheostomy or failed extubation were excluded by the dataset creators from the primary task and supplied in separate files [3, 4]. We preserved that primary-task construction. A secondary analysis compared unchanged model scores in successful-extubation event rows with scores in the supplied failed-extubation rows. It did not retrain the model or redefine the primary outcome. For this secondary comparison only, internal successful-event rows were pooled across training, validation, and test partitions; the external comparison used the external cohort.

Models and predictors

The respiratory comparator used positive end-expiratory pressure, fraction of inspired oxygen, plateau pressure, lung compliance, minute ventilation, arterial blood gas pH, arterial carbon dioxide pressure, arterial oxygen pressure, respiratory rate, and oxygen saturation. The full model used all 37 clinical features listed in the released data dictionary [3].

Both models used median imputation with missingness indicators, standard scaling, and logistic regression with L2 regularization and C = 1.0. Training used only the internal training file. There was no hyperparameter search and no model modification after review of the external results. This single-imputation approach did not propagate uncertainty in imputed values; missing-data assumptions and imputation uncertainty remain limitations [7].

Threshold and evaluation

One threshold for each model was selected on the internal validation cohort by maximizing Youden's J. The selected thresholds, 0.0601 for the respiratory comparator and 0.0549 for the full model, were applied unchanged to the internal test and external cohorts.

Reported measures were area under the receiver operating characteristic curve, average precision, Brier score, 10-bin expected calibration error, calibration intercept, calibration slope [1], sensitivity, specificity, positive predictive value, and negative predictive value. Confidence intervals used 1,000 patient-cluster bootstrap resamples with random seed 20260829. Patients were sampled with replacement, retaining all observations for each sampled patient. Resamples containing only one outcome class were omitted. The 2.5th and 97.5th percentiles of finite bootstrap estimates defined the 95% confidence intervals. The full-minus-comparator area-under-the-curve difference used paired patient-cluster bootstrap samples.

Missingness sensitivity analyses divided observations within each evaluation cohort into quartiles of the fraction of missing predictor values. No demographic subgroup, hospital-within-cohort, or monitor-vendor analysis was performed because the released files did not contain the required fields.

Historical-access and reproducibility controls

Earlier project records contained performance results from the external cohort. Before the current confirmation run, the protocol, cohort use, predictors, preprocessing, threshold selection, metrics, bootstrap approach, secondary analyses, and blocked analyses were frozen. The analysis is therefore a controlled confirmation audit, not a prospective preregistration or first-look external validation. Positive, null, adverse, and non-estimable findings were retained.

Use of generative artificial intelligence

OpenAI ChatGPT and Codex assisted with code drafting, manuscript organization, language editing, formatting, and submission-package quality assurance. The author defined the research question and evidence rules, reviewed or executed the analysis, verified all reported values against the frozen outputs, interpreted the findings and limitations, and accepts full responsibility for the work. These tools did not create outcome labels, alter source data, independently determine scientific conclusions, or make the submission decision.

Results

Cohort and split audit

The validation cohort contained 594 observations and 48 successful next-day extubations. The internal test cohort contained 831 observations and 65 events. The institutional external cohort contained 2,835 observations and 192 events. Internal partitions had no patient or intensive-care-stay identifier overlap. The unresolved 12-prefix internal-external collision is reported as a limitation.

Model performance

Table 1
Discrimination and calibration by cohort and model

Cohort	Model	AUROC	Average precision	Brier	Calibration intercept	Calibration slope
Internal test	Respiratory comparator	0.7916	0.2167	0.0677	-0.413	0.984
Internal test	Full	0.8553	0.3183	0.0653	-0.578	0.774
External	Respiratory comparator	0.8183	0.2286	0.0589	-0.146	1.248
External	Full	0.8490	0.2512	0.0611	-0.873	0.675

The internal full-minus-comparator area-under-the-curve difference was 0.0636 (95% confidence interval 0.0236 to 0.1046). The external difference was 0.0307 (95% confidence interval 0.0030 to 0.0584). The

full model therefore retained external discrimination and exceeded the comparator on discrimination, but its external Brier score was worse than the comparator and its calibration slope was farther from 1. Model performance is summarized in Table 1.

Frozen-threshold performance

Table 2
Performance at frozen validation-selected thresholds

Cohort	Model	Sensitivity	Specificity	PPV	NPV
Internal test	Respiratory comparator	0.908	0.492	0.132	0.984
Internal test	Full	0.908	0.645	0.178	0.988
External	Respiratory comparator	0.953	0.479	0.117	0.993
External	Full	0.885	0.683	0.168	0.988

At the frozen threshold, the external full model produced 170 true positives, 839 false positives, 1,804 true negatives, and 22 false negatives. Its positive predictive value was 0.168. These values are properties of the historical cohort and threshold procedure, not recommended clinical operating points. Threshold metrics are summarized in Table 2.

Missingness sensitivity analysis

Table 3
Performance in the lowest and highest missingness quartiles

Cohort	Missingness quartile	Mean missing fraction	Full-model AUROC	Full-model Brier
Internal test	Q1, lowest	0.136	0.9456	0.0270
Internal test	Q4, highest	0.350	0.7570	0.1009
External	Q1, lowest	0.102	0.8839	0.0332
External	Q4, highest	0.359	0.7681	0.1007

Discrimination and Brier performance deteriorated in the highest-missingness quartile in both evaluation cohorts. This analysis is descriptive and does not establish a causal effect of missingness. Table 3 presents the lowest and highest missingness quartiles.

Secondary failed-extubation characterization

Without retraining, full-model scores distinguished successful-extubation event rows from the supplied failed-extubation rows with an area under the receiver operating characteristic curve of 0.7186 internally and 0.7033 externally. This secondary comparison did not use the failed-extubation subset as part of the

primary training or evaluation task and should not be interpreted as validated prediction of extubation failure. The internal secondary comparison pooled successful-event rows from the training, validation, and test partitions. It is therefore partly in-sample descriptive characterization, not held-out internal validation.

Discussion

Principal findings

A fixed, transparent 37-feature logistic model retained discrimination in the institutional external cohort and modestly exceeded a 10-feature respiratory comparator. The positive result was not uniform across evaluation domains. External probability accuracy by Brier score was worse for the full model than for the comparator, the full-model calibration slope was 0.675, positive predictive value at the frozen threshold was 0.168, and performance degraded as missingness increased.

These findings illustrate why a favorable external area under the curve should not stand alone. Ranking performance can persist while probability calibration and threshold consequences remain inadequate for an intended use [1]. The appropriate conclusion is continued evaluation with local calibration, missingness, subgroup, workflow, and utility assessment, not clinical deployment.

Relation to prior work

The dataset creators developed and evaluated several model classes on this task and reported strong discrimination, including logistic regression and recurrent models [4]. The present analysis is narrower. It does not claim a new state-of-the-art extubation model. Its contribution is a transparent comparator design, prespecified threshold transfer, paired patient-cluster uncertainty, explicit calibration reporting, missingness-stratified performance, and preservation of adverse and blocked findings in a reproducible confirmation audit.

Limitations

First, the study is retrospective and relies on a released derivative dataset rather than independent extraction from source records. Second, prior project records contained external results, so this is not a first-look external validation. Third, 12 deidentified patient-number prefixes occurred in both institutional files, and the documentation does not establish global uniqueness across institutions. Fourth, patient-day observations may be correlated within stays and patients; uncertainty was clustered by the released patient prefix, but the data structure still limits stronger longitudinal modeling claims. Fifth, the outcome reflects next-day extubation status in the dataset creators' task and does not establish appropriateness of an extubation decision. Sixth, failed-extubation and tracheostomy records were excluded from the primary task by the dataset creators. Seventh, demographic subgroup, hospital-within-cohort, and monitor-vendor fields were absent. Eighth, the respiratory comparator is clinically interpretable but is not every possible clinical benchmark. Ninth, the missingness analysis is descriptive. Finally, there were no

prospective decisions, clinician actions, workflow outcomes, treatment effects, safety outcomes, or patient-benefit endpoints.

Conclusions

In a controlled historical confirmation analysis of SCRIPT X2B8, a fixed 37-feature logistic model retained external discrimination and exceeded a respiratory comparator. Worse external Brier performance, calibration slope 0.675, positive predictive value 0.168, missingness-related degradation, historical access to the external results, and unavailable subgroup and vendor fields limit the inference. The study supports a complete-evidence approach to model evaluation. It does not establish clinical utility, treatment benefit, deployment readiness, or validation of a complete product.

Declarations

Ethics approval and consent to participate

The dataset custodians report original study approval under Northwestern University Institutional Review Board protocols STU00204868 and STU00216678, with consent or waiver procedures described in the dataset documentation [3, 4]. This secondary analysis used only deidentified data accessed under the PhysioNet credentialing, training, license, and data-use requirements. No direct identifiers were accessed, no reidentification was attempted, and no row-level source data are redistributed. The author had no access to patient-identification keys and did not link datasets in a way that could identify individuals. The University of North Georgia Institutional Review Board reviewed this research under protocol 2026-027 and determined it exempt under 45 CFR 46.104(d)(4) on March 24, 2026. The author confirms that this secondary analysis is covered by that determination.

Consent for publication

Not applicable.

Competing interests

The author is the founder of a company developing clinical AI evaluation and governance tools and has intellectual property interests that could be perceived as related to the subject matter of this work. This manuscript does not evaluate a marketed product.

Funding

The research reported in this paper was supported by AIM-AHEAD Coordinating Center at the University of North Texas Health Science Center at Fort Worth. This research was, in part, funded by the National Institutes of Health (NIH) Agreement No. 1OT2OD032581. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the NIH.

Author Contribution

A.M.C was responsible for conceptualization, protocol specification, methodology, investigation, validation, analysis governance, interpretation, writing, and final approval.

Acknowledgements

Not applicable.

Data Availability

SCRIPT X2B8 version 1.0.0 is available to qualified credentialed users through PhysioNet under its access requirements and data use agreement: https://doi.org/10.13026/235w-zn26. Restricted row-level data are not redistributed. A private, row-free reviewer package containing the frozen protocol, analysis code, aggregate results, audit summaries, and integrity manifest has been prepared and can be supplied through an approved anonymous-review route. No unrestricted public source-data release is claimed.

List of abbreviations

AI, artificial intelligence; AUROC, area under the receiver operating characteristic curve; ICU, intensive care unit; IRB, institutional review board; NPV, negative predictive value; PPV, positive predictive value; SCRIPT, Successful Clinical Response in Pneumonia Therapy; TRIPOD + AI, Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis plus artificial intelligence.

References

1. Van Calster B, McLernon DJ, van Smeden M, Wynants L, Steyerberg EW. Calibration: the Achilles heel of predictive analytics. BMC Med. 2019;17:230. 10.1186/s12916-019-1466-7.
2. Collins GS, Moons KGM, Dhiman P, et al. TRIPOD + AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ. 2024;385:e078378. 10.1136/bmj-2023-078378.
3. Fenske S, Peltekian A, Kang M et al. SCRIPT X2B8 Dataset: per-day clinical features to model successful next-day extubation. Version 1.0.0. PhysioNet. 2025. 10.13026/235w-zn26

4. Fenske SW, Peltekian A, Kang M, et al. Developing and validating machine learning models to predict next-day extubation. Sci Rep. 2025;15:27552. 10.1038/s41598-025-12264-4.
5. Pollard T, Moody BE, Lehman L, Gow B, Fernandes C, Xie C, et al. PhysioNet as a global platform for biomedical research. Nat Health. 2026;1:792–5. 10.1038/s44360-026-00096-z.
6. Goldberger AL, Amaral LAN, Glass L, Hausdorff JM, Ivanov PC, Mark RG, et al. PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. Circulation. 2000;101(23):e215–20. 10.1161/01.CIR.101.23.e215.
7. Sterne JAC, White IR, Carlin JB, et al. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. BMJ. 2009;338:b2393. 10.1136/bmj.b2393.