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.

Six-Hour Prediction of Documented Hypophosphatemia Across Different ICU Repeat-Testing Practices: Development and Multicenter External Validation

Tianyi Yu

tianyi.1.yu@kcl.ac.uk

King's College London

Research Article

Keywords: clinical prediction model, hypophosphatemia, intensive care, electronic health records, external validation, informative observation, laboratory testing, calibration

Posted Date: November 12th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11172913/v1

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

Additional Declarations: No competing interests reported.

Six-Hour Prediction of Documented Hypophosphatemia Across Different ICU Repeat-Testing Practices: Development and Multicenter External Validation

Article type: Research article

Author: Tianyi Yu

Affiliation: Department of Biostatistics and Health Informatics, Denmark Hill Campus, King's College London, London, United Kingdom

Corresponding author: Tianyi Yu; yutianyi16@gmail.com; tianyi.1.yu@kcl.ac.uk

Abstract

Background: Prediction of a future laboratory abnormality depends on whether it is measured. We examined whether an early intensive care model predicts documented low phosphate across different repeat-testing practices, and what selective testing leaves unresolved.

Methods: Models were developed in MIMIC-IV 3.1 and evaluated after freezing in eICU-CRD 2.0. Eligible adults remained in intensive care at six hours, had phosphate available by then, and had no baseline specimen below 2.0 mg/dL (approximately 0.65 mmol/L); some baseline specimens were reported later. The endpoint was a recorded low result from a specimen collected after six hours and by discharge or 48 hours; reporting could occur later. Primary evaluation was conditional on repeat testing. Penalized logistic regression using available laboratory results, age and sex was compared with phosphate alone. Internal patient and external hospital-cluster bootstraps quantified uncertainty. Observation analyses included testing patterns, inverse-probability weighting and extreme assignments of unknown labels.

Results: Of 31118 eligible MIMIC patients, 26469 (85.1%) were retested, with 2934 events (11.1%). Of 15568 eligible eICU patients from 186 hospitals, 10133 (65.1%) were retested, with 1646 events (16.2%). Internal area under the receiver operating characteristic curve was 0.716. Externally, it was 0.717 (95% confidence interval 0.704–0.735), area under the precision-recall curve was 0.330 (0.304–0.368), Brier score was 0.125 (0.117–0.135), and calibration slope was

0.827 (0.754–0.914). The paired external discrimination gain over phosphate alone was 0.067 (0.054–0.083). Weighting under conditional missing-at-random assumptions yielded an external discrimination estimate of 0.712. Hypothetically assigning all untested eICU patients a binary assay label of zero or one gave a completion range of 10.6%–45.5%; this does not bound biological incidence.

Conclusions: The model retained moderate discrimination for documented low phosphate among retested patients. Imperfect external calibration, selective ascertainment and retrospectively determined eligibility limit clinical use. Prospective evaluation requires systematic outcome measurement and a defined clinical action; these data do not establish that tests can safely be omitted.

Keywords: clinical prediction model; hypophosphatemia; intensive care; electronic health records; external validation; informative observation; laboratory testing; calibration

Background

Clinical and methodological context

Observational cohorts and reviews have examined associations between phosphate and outcomes in critical illness; their findings differ by population and adjustment strategy and do not establish that correcting a low laboratory value improves outcomes [1-8]. Nutrition initiation and refeeding are important settings in which phosphate may fall, and treatment or monitoring decisions depend on broader clinical context [9-14]. Existing prediction work has addressed refeeding hypophosphatemia, neurocritical refeeding syndrome, and phosphate abnormalities in oncology; these are useful comparators but define different populations, endpoints, or decision moments [15-17].

Laboratory values in electronic records are collected in response to care. Methodological analyses show that informative presence and outcome observability can bias clinical prediction [18-22]. Studies of laboratory density, cross-database measurement features, and testing workflow add practical reasons why model meaning may change when transported [23-25]. This matters especially when a prediction target is a future laboratory result. A patient without a second

phosphate test has an unknown observed label, not a documented normal value. Standardized prediction reporting and risk-of-bias guidance call for precise outcome definitions, realistic predictor availability, external validation, calibration assessment, and transparency about missing information and clustered data [26-35].

Research question and contribution

At 6 hours after ICU admission, among adults retrospectively confirmed to have no low phosphate in specimens collected during the first 6 hours, how accurately can information already available predict an eventually recorded phosphate result below 2.0 mg/dL from a specimen collected after six hours and by discharge or 48 hours, and how does selective repeat testing constrain external interpretation? Figure 1 depicts the landmark, distinct development and validation sources, and the observation process. The empirical contribution is the joint evaluation of a frozen early ICU prediction model and its measurement process across two data systems. We do not claim to have introduced a new learning algorithm, established biological incidence, or demonstrated a benefit from withholding tests.

Methods

Study design, data, and eligibility

This retrospective development and external-validation study used MIMIC-IV version 3.1 for model development and internal testing and eICU-CRD version 2.0 for external evaluation [36-40]. MIMIC represents one hospital system and includes source records from 2008-2022; eICU provides a multicenter sample from 2014-2015. These are source-database accrual periods, not a claim that each year contributes equally to this analytic cohort. We selected one source stay per adult before applying the remaining eligibility criteria, required ICU presence at 6 hours, and required at least one phosphate result collected during the first 6 hours. All collected baseline phosphate values had to be at least 2.0 mg/dL; physiologically implausible values below 0.2 or above 20 mg/dL were excluded. At least one baseline phosphate result had to be available by the landmark, and model predictors used only available results. MIMIC age was derived from its anchor age and year and capped at 90; recorded eICU ages above 89 were mapped to 90. The selected MIMIC stay was

earliest by admission time. In eICU, the lowest stay identifier was chosen per patient before requiring unit visit number 1; the identifier does not establish cross-encounter chronology.

The prediction moment was ICU admission plus 6 hours. Source-unit thresholds were retained exactly: phosphate 2.0 mg/dL is approximately 0.65 mmol/L; multiplication by 0.3229 converts phosphate from mg/dL to mmol/L. Rounded SI equivalents are descriptive and were not used to reclassify patients. The target was any recorded serum or plasma phosphate below 2.0 mg/dL from a specimen collected after 6 hours and through the earlier of ICU discharge or 48 hours. Follow-up eligibility used specimen time, without requiring that the result be reported by the end of that window; the endpoint is therefore an eventually recorded result from that sampling window. Patients with at least one repeat phosphate and no low measurement in that window were observed negatives. Patients without a repeat measurement were unlabeled, and the primary performance estimand was conditional on being retested. Results collected before the landmark but reported afterward did not enter predictors, although they could exclude a patient from the retrospectively defined baseline cohort. MIMIC used laboratory collection and store times; eICU used result and revision offsets, conservatively requiring the revised result to be available by 6 hours. The operational harmonization and its limitations are detailed in Additional file 1.

Predictors and model development

The prespecified candidate predictors were age, sex, first and minimum baseline phosphate, and the latest available potassium, magnesium, calcium, creatinine, glucose, and sodium. The principal model was penalized logistic regression, with imputation and missingness indicators learned in MIMIC development data. A phosphate-only penalized logistic model was the simple comparator. We separately added the number of phosphate tests before 6 hours to examine dependence on measurement intensity, and fitted gradient boosting as an exploratory complexity comparator. Treatment and phosphate replacement were not modeled; the target is the laboratory event under historical care, which could be altered by interventions after the landmark. Penalization was chosen by five-fold stratified cross-validation within a deterministic patient-level 70% MIMIC development split; preprocessing was fitted afresh in each fold. The originally evaluated gradient-boosting settings were fixed rather than tuned; its iteration setting was a maximum, with automatic early stopping. Because the extracted cohort contains one row per

person, validation folds did not mix records from a person. Internal testing used the remaining 30%. Code, source-extract hashes, model hashes, the chosen hyperparameters, and two development-defined alert thresholds were frozen before the single external evaluation. The frozen protocol and deviations appear in Additional file 1. Extraction and analysis code, the locked protocol and aggregate outputs are supplied in Additional file 2.

Post hoc hyperparameter sensitivity

After the original external evaluation, we assessed whether broader tuning offered a material improvement. Six gradient-boosting configurations were compared using the original five-fold MIMIC development cross-validation and mean Brier score. The selected configuration was compared with the frozen booster on the previously used internal test set, with 600 paired patient bootstraps. A separate development-only analysis extended the logistic penalty grid to C=100 and C=1000. This post hoc exercise used no external patient records, did not alter the principal model, and does not constitute new independent validation. Full configurations and results are in Additional files 1 and 2.

Performance and uncertainty

We report AUROC, average precision (reported as area under the precision-recall curve, AUPRC), Brier score, jointly fitted calibration intercept and slope, and a flexible calibration plot. Individual metric intervals used 400 bootstrap replicates. In a post hoc calibration analysis, an intercept-only logistic model with the frozen logit prediction as an offset estimated calibration-in-the-large with slope fixed at 1. We also report mean predicted probability, observed event proportion and their observed-to-expected ratio, with 600 patient or hospital-cluster bootstrap replicates. This analysis describes the frozen probabilities and does not recalibrate them. AUPRC is prevalence-sensitive and supplements AUROC when outcomes are relatively uncommon [41]. The paired difference between the multivariable and phosphate-only models was estimated by 600 patient bootstraps internally and 600 hospital-cluster bootstraps externally; external intervals preserve within-hospital dependence. We also examined discrimination among patients staying in the ICU through 48 hours and, externally, hospitals with at least 50 eligible patients and 70% repeat testing. Hospital displays are descriptive; site-level observed minus predicted risk was shown for hospitals with at

least 50 eligible patients, 50 tested patients, and 10 observed events. No site identifiers are reported. Although decision-curve analysis can assess the net benefit of a defined intervention at specified threshold preferences, the relevant consequences and threshold trade-offs for changing phosphate testing were not established, and target labels were selectively ascertained; we therefore did not present a clinical net-benefit estimate [42,43].

The two alert thresholds were set at the development score cutoffs corresponding to its top 10% and 20% of predicted risk. We carried their numeric values unchanged into the internal and external test sets and report resulting alert fraction, sensitivity, and positive predictive value. These are hypothetical triage characteristics among tested patients. They are not measured test savings, a recommendation to omit tests, or evidence of benefit from acting on an alert.

Observation-process analyses

We reported the fraction retested, time to first repeat, selected tested-versus-untested characteristics, hospital variation, discharge before 48 hours, and early ICU deaths. Discharge and death shorten opportunities to observe the endpoint. As a sensitivity analysis, we fitted a repeat-testing propensity model using only information available by 6 hours and applied inverse-probability weights to retested patients, truncated at the 99th percentile. The propensity model used the principal baseline covariates; it did not include hospital indicators or postlandmark care and does not provide a hospital-specific selection correction. It addresses repeat testing only within the baseline-selected cohort, not selection into that cohort. This analysis relies on conditional exchangeability of labels between tested and untested patients and positivity; these assumptions cannot be established from the data. It does not reconstruct missed transient biological events. To illustrate missing assay information, we completed a hypothetical binary label vector by assigning every untested patient 0 or 1 while holding tested labels fixed. These completions are not bounds on the actually recorded full-cohort event fraction, which is known, nor on a standardized-testing counterfactual or biological incidence. No qualifying assay result was recorded for these patients in the extracted tables, and even a negative recorded sample can miss a transient low value. The exercise displays the arithmetic consequences of missing labels without identifying a physiological target.

Reporting, ethics, and data access

The analysis was organized against TRIPOD+AI and TRIPOD-Cluster, with a PROBAST+AI self-assessment [27,28,30]. A completed reporting map with explicitly identified limitations is provided in Additional file 3. The two source databases are deidentified research resources with their own access and governance processes [36-40]. This secondary analysis used credentialed-access data; no new patient contact or intervention occurred. Source-database governance is described in the Declarations; this analysis does not claim a newly issued institutional waiver. Patients and the public were not involved in this retrospective study. The protocol was documented before outcome extraction but was not registered in an external registry. Source row data, individual predictions, and fitted model objects are excluded from the shareable supplement because database access and redistribution are controlled by the data providers. Reproducible extraction and analysis code and aggregate results are provided, subject to the source data agreements.

Use of generative artificial intelligence

OpenAI Codex (GPT-6) assisted with code debugging and manuscript grammar checking. Analyses were run in a credentialed local research workspace; manuscript text and figures report aggregate numerical results. Bibliographic metadata and numerical claims were checked against DOI registries and saved result files. The sole author approved release of the manuscript and retains responsibility for all content, analytic choices, references, and conclusions. Original AI interaction records are retained in the local research environment; restricted patient-level material must not be redistributed in a public appendix.

Results

Study population and repeat testing

In MIMIC-IV, 64767 selected adult ICU stays met the age and six-hour presence screen and 31118 had a qualifying baseline phosphate profile. Of these, 26469 (85.1%) had a repeat measurement; 2934 of the retested patients (11.1%) had a documented low result. The MIMIC split produced 18644 retested development patients with 2090 events and 7825 internal test patients with 844 events. In eICU, 116271 selected adult stays met the initial screen and 15568

from 186 hospitals met baseline eligibility. There were 10133 retested patients (65.1%) across 173 hospitals, with 1646 events (16.2%). Baseline eligibility was met by 48.0% of screened MIMIC stays and 13.4% of screened eICU stays, showing that cohort selection differed before repeat testing was considered. Thus, the external observed outcome proportion is calculated among a smaller and differently selected fraction of eligible patients; it should not be interpreted as a higher physiological incidence. Median time from ICU admission to the first repeat phosphate was 11.67 hours in MIMIC and 13.87 hours in eICU.

Repeat testing was associated with differences in measured case mix and varied substantially by institution (Table 1). The first phosphate median was 3.6 versus 3.5 mg/dL among tested and untested MIMIC patients and 3.5 versus 3.5 mg/dL in eICU; eICU baseline creatinine medians were 1.12 versus 1.00 mg/dL. Among eICU hospitals with at least 50 eligible patients, the median repeat-testing fraction was 59.6% (interquartile range 51.4%-68.0%); 16 hospitals satisfied the prespecified high-testing definition. There were 15028 MIMIC and 7545 eICU eligible patients discharged before 48 hours; early ICU deaths numbered 1253 and 737, respectively. These events shorten observation windows and are a reason not to treat the 48-hour target as uniformly observed.

Table 1. Baseline-eligible patients by repeat phosphate ascertainment.

Characteristic	MIMIC retested	MIMIC no repeat	eICU retested	eICU no repeat
Patients	26469	4649	10133	5435
Age, years	66	66	63	62
First available phosphate, mg/dL	3.6	3.5	3.5	3.5
Baseline creatinine, mg/dL	1.00	0.90	1.12	1.00
Baseline potassium, mmol/L or mEq/L	4.1	4.1	4.1	4.0
ICU death within 48 h, n	809	444	367	370
ICU discharge before 48 h, n	10546	4482	3796	3749

Continuous values are medians. ICU death and early discharge categories can overlap. Absence of a recorded repeat does not prove no blood specimen was obtained. Phosphate medians of 3.6 and 3.5 mg/dL correspond to approximately 1.16 and 1.13 mmol/L; creatinine medians of 1.00, 0.90 and 1.12 mg/dL correspond to 88.4, 79.6 and 99.0 µmol/L. Laboratory source units are retained for comparability with extraction code.

Internal and external prediction performance

In the MIMIC internal test set, multivariable AUROC was 0.716, AUPRC was 0.259, and Brier score was 0.089, compared with 0.656, 0.195, and 0.093 for phosphate alone. The paired internal AUROC difference was 0.060 (95% patient-bootstrap CI 0.044-0.075). External multivariable AUROC was 0.717 (95% hospital-bootstrap CI 0.704-0.735), AUPRC 0.330 (0.304-0.368), and Brier score 0.125 (0.117-0.135). The external phosphate-only model had AUROC 0.651, AUPRC 0.258, and Brier score 0.133. Paired external improvements were 0.067 in AUROC (95% CI 0.054-0.083), 0.072 in AUPRC (0.049-0.103), and -0.0083 in Brier score (-0.0118 to -0.0062). Figure 2 shows the model estimates and uncertainty. The near-equal internal and external AUROC should not be read as identical population performance: the measured outcome fraction, calibration, and testing patterns differed.

The jointly fitted external calibration intercept was -0.053 and slope was 0.827 (95% hospital-bootstrap CI 0.754-0.914), compared with an internal slope of 0.932. In the post hoc intercept-only analysis, external calibration-in-the-large was +0.235 (95% hospital-bootstrap CI 0.138–0.322). Mean predicted risk was 13.58%, compared with an observed event proportion of 16.24%; the

observed-to-expected ratio was 1.196 (1.112–1.273). Internally, calibration-in-the-large was −0.051 (95% patient-bootstrap CI −0.129 to 0.022). These results identify external overall underprediction in addition to excessive dispersion. Figure 3 shows the calibration curves and hospital-level observed-minus-predicted event fractions. In 37 eICU hospitals with adequate size and event counts for this descriptive display, the median absolute site difference between observed and mean predicted risk was 0.035. The Spearman association between site repeat-testing fraction and observed event risk was -0.237 (descriptive P=.157); this ecological association is not evidence that testing causes lower risk. Adding the prelandmark phosphate test count changed performance little. Exploratory gradient boosting improved external AUROC by 0.013 over the main model (95% paired CI 0.002-0.020), but its Brier difference was -0.0009 (95% CI -0.0022 to 0.0006). This does not justify replacing the prespecified principal model.

At the development 10% alert cutoff, 17.3% of external tested patients were flagged, sensitivity was 38.1%, and positive predictive value was 35.9%. At the development 20% cutoff, 29.6% were flagged, sensitivity was 55.7%, and positive predictive value was 30.6%. Transported cutoffs produced different alert fractions as the score and outcome distributions changed. These figures describe only a retrospectively observed subset and do not quantify a safer or cheaper testing policy.

Post hoc tuning results

The shallow booster had the lowest development cross-validation Brier score (0.090486 versus 0.090521 for the fixed booster). On the previously used internal test set, however, its Brier difference was +0.000063 (95% paired bootstrap interval −0.000213 to 0.000359), and its AUROC difference was −0.000225 (−0.003464 to 0.002851). Extending logistic C beyond 10 did not improve development cross-validation Brier score. These small differences provide no convincing evidence that the additional search improved performance; the frozen principal model and external results remain the basis of inference.

Observation and sensitivity analyses

The eICU propensity model's AUROC for repeat testing was 0.595. After 99th-percentile weight truncation (maximum weight 2.26; effective tested sample size approximately 9996), the

multivariable external AUROC was 0.712 rather than 0.717 and AUPRC was 0.322 rather than 0.330. This small conditional adjustment cannot rule out unmeasured determinants of testing. Among 4654 tested patients in the 16 high-testing eICU hospitals, AUROC was 0.691. Among 6337 external tested patients still in the ICU at 48 hours, AUROC was 0.703; in the analogous internal subset of 4633, AUROC was 0.706. In the post-freeze subset whose every baseline phosphate result was available by six hours, 7335 internally tested and 9904 externally tested patients remained; main-model AUROCs were 0.718 and 0.717, respectively. This restriction cannot restore patients excluded because a later-reported baseline specimen was low. These conditional subsets have different case mix and time at risk; differences from the main results are not intervention effects.

The actual proportion with a recorded qualifying event in the full eICU cohort was 1646/15568=10.6%. As a separate hypothetical exercise, fixing the observed labels and assigning all 5435 untested patients zero or one gave completion proportions of 10.6% and 45.5% (7081/15568). The corresponding MIMIC completion range was 9.4%–24.4%. Figure 4 distinguishes these hypothetical completions from observed event proportions among retested patients. Neither the completions nor the recorded full-cohort proportion establish the occurrence of biological hypophosphatemia or risk under a standardized testing schedule.

Discussion

Principal findings and interpretation

An early, interpretable model improved prediction of a subsequent *documented* low-phosphate result over phosphate alone and retained moderate discrimination in a distinct multicenter database. Its external calibration slope below 1 indicates overly dispersed predictions after transport, while calibration-in-the-large and the observed-to-expected ratio show overall underprediction. These are distinct aspects of calibration failure; a new setting would require local calibration and evaluation before use. Repeat phosphate measurement was substantially less frequent in the external cohort and heterogeneous across hospitals. The central scientific finding is therefore conditional: the model transports useful ranking information among patients who were

retested under historical care, while biological risk and prediction performance for all eligible patients remain unidentified without assumptions or additional measurements.

Prior phosphate, nutrition, and related prediction studies describe clinical associations and mechanisms, but they do not make every low result actionable or establish the benefit of a prediction-triggered test [1-17]. Informative-observation analyses explain why clinician-driven measurement complicates interpretation [18-22]; laboratory-workflow studies illustrate how measurement density and site practice may differ [23-25]. Our paired external comparison supports incremental predictive information beyond initial phosphate, while the tested-versus-untested contrast and hypothetical label-completion analysis prevent the stronger claim that a hospital can safely withhold tests in patients given low scores. Reviews of ICU and other external validations document variable methodological conduct [44,45]. Simulation and methodological work address sample-size planning and heterogeneity in external validation [46,47]; reporting guidance and a journal editorial emphasize transparent model comparisons, calibration, and limitations [48,49].

Clinical implications and next study

The clinical use case is prioritizing attention to a future laboratory event, not diagnosing one at the landmark. A prospective study would need to predefine what action a risk estimate triggers, maintain scheduled or randomized ascertainment in at least a validation subset, track treatments and patient-important outcomes, and evaluate safety and resource use. Recalibration should be prespecified within each deployment setting before the alert threshold is used. Decision-analytic benefit should be evaluated only against a concrete policy that compares the consequences of missed lows, extra tests, and any downstream treatment [42,43]. Uncertainty in any estimated net benefit should inform the need for further research rather than be treated as a simple significance gate [50].

Limitations

First, entry into the baseline cohort and subsequent repeat testing were both selective. The much lower baseline eligibility fraction in eICU limits extension even to its wider screened ICU population. The primary endpoint was observable only after repeat testing. Weighting depends on

measured covariates and cannot remove unmeasured selection or missed transient events. Second, discharge and death before 48 hours make the possible observation period unequal; restriction to 48-hour stays changes the target population and does not solve competing risks. Third, all baseline specimens had to be nonlow even if a result was reported after the landmark; consequently, the retrospectively selected population cannot be identified exactly in real time. Restriction to cases with complete baseline result availability was reassuring for discrimination but could not recover excluded late-reported low values. Follow-up specimens could also be reported after the observation window, so the endpoint does not require a result to have been known by 48 hours. Fourth, the two databases span different institutions and periods and differ in result-timing fields. Requiring revised eICU results by six hours was a conservative availability rule but could omit initially available results whose final revision occurred later. Fifth, the reproducible first eICU stay by identifier need not be chronologically first, and age top-coding and imperfect cross-database laboratory harmonization may affect transport. Sixth, this was retrospective evaluation; neither alert threshold was prospectively enacted, and no effect on testing, correction of phosphate, or patient outcomes was measured. Seventh, no independent contemporary non-US validation cohort was available. The supplementary tuning exercise was initiated after the external results were known and reused the original internal test set; its results are exploratory. Finally, the site-level analysis is descriptive with varying numbers of tested patients and events; it is not a causal hospital comparison.

Conclusions

In two large ICU research databases, a six-hour multivariable model predicted a later *documented* phosphate result below 2.0 mg/dL better than baseline phosphate alone among patients with a repeat test. External calibration and differences in repeat-testing practice limit direct use. The next evidentiary step is prospective evaluation with systematic outcome ascertainment and a defined clinical action, rather than deployment based on retrospective discrimination alone.

List of abbreviations

AUPRC: area under the precision-recall curve; AUROC: area under the receiver operating characteristic curve; CI: confidence interval; eICU-CRD: eICU Collaborative Research Database; ICU: intensive care unit; MIMIC-IV: Medical Information Mart for Intensive Care IV.

Declarations

Ethics approval and consent to participate

This study used deidentified MIMIC-IV and eICU-CRD data under credentialed-access data-use agreements, with no new recruitment, patient contact or intervention. The MIMIC-IV 3.1 documentation reports that the Beth Israel Deaconess Medical Center Institutional Review Board approved creation and sharing of the resource and waived informed consent [36]. The eICU source report describes deidentification to HIPAA safe-harbor standards and credentialed access [39]. No participant identity key was available to this analysis. The author classified the work as secondary analysis of existing anonymized research data, for which additional ethical clearance was not required under the existing-data provision in section 4.1.4 of the King's College London Research Ethics Policy [51]. No new committee approval, exemption certificate or reference number is claimed. No additional participant consent was sought for this secondary analysis.

Consent for publication

Not applicable. No identifiable individual information is reported.

Availability of data and materials

MIMIC-IV 3.1 and eICU-CRD 2.0 are available through PhysioNet to appropriately credentialed researchers under their respective data-use agreements [36,38,40]. Patient-level extracts, individual predictions and fitted model objects are not redistributed. Source access must be obtained from the data providers. Additional file 2 contains extraction and analysis code, the frozen protocol and machine-readable aggregate results supporting the article. Additional file 1 describes the reproducibility requirements and post hoc analyses; the reporting map is supplied in Additional file 3.

Competing interests

TY declares no financial or non-financial competing interests.

Funding

No funding was received for this study.

Authors' contributions

TY is the sole author and accepts responsibility for the study conception, design, analysis, interpretation, manuscript and final approval. The use of generative artificial intelligence is disclosed in the Methods.

Acknowledgements

Not applicable.

Additional files

Additional file 1 (.docx). Additional methods and aggregate results. Cohort assembly, source mapping, model specifications, missingness, full performance results, observation analyses, post hoc calibration-in-the-large and hyperparameter sensitivity.

Additional file 2 (.zip). Reproducible code and aggregate outputs. Analysis and figure code, locked protocol, source mapping, aggregate JSON results and post hoc tuning configurations. Credentialed source data, individual predictions and fitted models are excluded.

Additional file 3 (.docx). TRIPOD+AI reporting map. Item-by-item self-assessment identifying manuscript locations and remaining applicability or reporting limitations.

Figure legends

Figure 1. Study design and selective ascertainment of a future low-phosphate result. The prediction target is an eventually recorded phosphate result below 2.0 mg/dL (approximately 0.65 mmol/L) from a specimen collected after six hours and by ICU discharge or 48 hours among patients retrospectively confirmed to have no low baseline specimen. Model inputs were available

by six hours, but some baseline specimens used in eligibility were reported later. An untested patient has an unknown measured label; the diagram does not imply biological normality.

Figure 2. Internal and external discrimination of the phosphate-only, multivariable, and exploratory gradient-boosting models. Circles denote MIMIC-IV internal testing and squares eICU-CRD external testing; horizontal lines are 95% patient or hospital-cluster bootstrap intervals. Dotted lines on the precision-recall panel indicate the corresponding observed event fractions among retested patients.

Figure 3. Calibration of the prespecified multivariable model. Panel A displays spline-smoothed observed event fractions with decile summaries and the ideal line. Panel B shows hospital-level observed minus predicted risk in 37 eICU hospitals with at least 50 eligible patients, 50 retested patients, and 10 events; point area reflects retested sample size. Hospital comparisons are descriptive.

Figure 4. Repeat phosphate measurement and hypothetical label completion. Panel A shows fractions retested in each cohort; panel B shows repeat-testing fractions in eICU hospitals with at least 50 eligible patients and marks the 70% high-testing criterion; panel C assigns all untested patients hypothetical binary assay labels of 0 or 1 while fixing observed labels. Diamonds are observed event fractions among retested patients. The ranges are hypothetical completions, not bounds on recorded-event risk, standardized-testing risk or biological incidence.

References

1. Attokaran AG, White KC, Doola R, McIlroy P, Senthuran S, Luke S, et al. Epidemiology of hypophosphatemia in critical illness: A multicentre, retrospective cohort study. Anaesthesia Critical Care & Pain Medicine. 2024;43(5):101410. https://doi.org/10.1016/j.accpm.2024.101410

2. Sin JCK, King L, Ballard E, Llewellyn S, Laupland KB, Tabah A. Hypophosphatemia and Outcomes in ICU: A Systematic Review and Meta-Analysis. Journal of Intensive Care Medicine. 2021;36(9):1025-1035. https://doi.org/10.1177/0885066620940274

3. Reintam Blaser A, Gunst J, Ichai C, Casaer MP, Benstoem C, Besch G, et al. Hypophosphatemia in critically ill adults and children – A systematic review. Clinical Nutrition. 2021;40(4):1744-1754. https://doi.org/10.1016/j.clnu.2020.09.045

4. Irandoost P, Ansari MM, Lahiji MN, Talebi S, Torabinasab K, Pourmoradian S, et al. Association between blood phosphorus level and clinical outcomes in critically ill patients: A systematic review and meta-analysis. Journal of Parenteral and Enteral Nutrition. 2026;50(6):767-780. https://doi.org/10.1002/jpen.70115

5. Wei S, Li Y, Zhang C, Guo X, Liang X, Huang Y, et al. Prognostic value of serum phosphate levels in sepsis: a systematic review and meta-analysis. PeerJ. 2023;11:e16241. https://doi.org/10.7717/peerj.16241

6. Federspiel CK, Itenov TS, Thormar K, Liu KD, Bestle MH. Hypophosphatemia and duration of respiratory failure and mortality in critically ill patients. Acta Anaesthesiologica Scandinavica. 2018;62(8):1098-1104. https://doi.org/10.1111/aas.13136

7. Statlender L, Raphaeli O, Bendavid I, Hellerman M, Kagan I, Fishman G, et al. Correlations between First 72 h Hypophosphatemia, Energy Deficit, Length of Ventilation, and Mortality—A Retrospective Cohort Study. Nutrients. 2022;14(7):1332. https://doi.org/10.3390/nu14071332

8. Statlender L, Raphaeli O, Shochat T, Robinson E, Hellerman Itzhaki M, Bendavid I, et al. Contributing factors to hypophosphatemia development in critically Ill ventilated patients: a retrospective cohort study. Scientific Reports. 2024;14(1):19771. https://doi.org/10.1038/s41598-024-68688-x

9. Marik PE, Bedigian MK. Refeeding hypophosphatemia in critically ill patients in an intensive care unit: A prospective study. Archives of Surgery. 1996;131(10):1043-1047. https://doi.org/10.1001/archsurg.1996.01430220037007

10. Reintam Blaser A, van Zanten ARH. Electrolyte disorders during the initiation of nutrition therapy in the ICU. Current Opinion in Clinical Nutrition & Metabolic Care. 2021;24(2):151-158. https://doi.org/10.1097/MCO.0000000000000730

11. Naik NM, Li J, Seres D, Freedberg DE. Assessment of refeeding syndrome definitions and 30-day mortality in critically ill adults: A comparison study. Journal of Parenteral and Enteral Nutrition. 2023;47(8):993-1002. https://doi.org/10.1002/jpen.2560

12. Olthof LE, Koekkoek WACK, van Setten C, Kars JCN, van Blokland D, van Zanten ARH. Impact of caloric intake in critically ill patients with, and without, refeeding syndrome: A retrospective study. Clinical Nutrition. 2018;37(5):1609-1617. https://doi.org/10.1016/j.clnu.2017.08.001

13. Lemon SJ, Zack SD, Voils SA. No difference in mechanical ventilation–free hours in critically ill patients who received intravenous, oral, or enteral phosphate replacement. Journal of Critical Care. 2017;39:31-35. https://doi.org/10.1016/j.jcrc.2017.01.002

14. Lauwers C, Langouche L, Wouters PJ, Wilmer A, Van den Berghe G, Casaer MP, et al. Early phosphate changes as potential indicator of unreadiness for artificial feeding: a secondary analysis of the EPaNIC RCT. Critical Care. 2025;29(1):48. https://doi.org/10.1186/s13054-025-05273-2

15. Choi TY, Chang M, Heo S, Jang JY. Explainable machine learning model to predict refeeding hypophosphatemia. Clinical Nutrition ESPEN. 2021;45:213-219. https://doi.org/10.1016/j.clnesp.2021.08.022

16. Zhang W, Zhang S, Chen S, Yu T, Tang Y. Development and validation of risk prediction model for refeeding syndrome in neurocritical patients. Frontiers in Nutrition. 2023;10:1083483. https://doi.org/10.3389/fnut.2023.1083483

17. Scanlon LA, Monaghan PJ, Adam S. Development of a Machine Learning Algorithm to Predict Abnormalities in Serum Phosphate in a Large Oncology Cohort. JCO Clinical Cancer Informatics. 2025;9:e2400312. https://doi.org/10.1200/CCI-24-00312

18. Sisk R, Lin L, Sperrin M, Barrett JK, Tom B, Diaz-Ordaz K, et al. Informative presence and observation in routine health data: A review of methodology for clinical risk prediction. Journal of the American Medical Informatics Association. 2021;28(1):155-166. https://doi.org/10.1093/jamia/ocaa242

19. Sun M, Engelhard MM, Bedoya AD, Goldstein BA. Incorporating informatively collected laboratory data from EHR in clinical prediction models. BMC Medical Informatics and Decision Making. 2024;24(1):206. https://doi.org/10.1186/s12911-024-02612-1

20. McGee G, Haneuse S, Coull BA, Weisskopf MG, Rotem RS. On the Nature of Informative Presence Bias in Analyses of Electronic Health Records. Epidemiology. 2022;33(1):105-113. https://doi.org/10.1097/EDE.0000000000001432

21. Yan M, Pencina MJ, Boulware LE, Goldstein BA. Observability and its impact on differential bias for clinical prediction models. Journal of the American Medical Informatics Association. 2022;29(5):937-943. https://doi.org/10.1093/jamia/ocac019

22. Sharafoddini A, Dubin JA, Maslove DM, Lee J. A New Insight Into Missing Data in Intensive Care Unit Patient Profiles: Observational Study. JMIR Medical Informatics. 2019;7(1):e11605. https://doi.org/10.2196/11605

23. Abedi V, Li J, Shivakumar MK, Avula V, Chaudhary DP, Shellenberger MJ, et al. Increasing the Density of Laboratory Measures for Machine Learning Applications. Journal of Clinical Medicine. 2020;10(1):103. https://doi.org/10.3390/jcm10010103

24. Zou J, Wang H, Xin Z, Zhou Q, Chen J, Huang J. From predictive signal to deployable value: EHR measurement-process features in ICU AKI prediction. International Journal of Medical Informatics. 2026;221:106668. https://doi.org/10.1016/j.ijmedinf.2026.106668

25. Rabbani N, Ma SP, Li RC, Winget M, Weber S, Boosi S, et al. Targeting repetitive laboratory testing with electronic health records-embedded predictive decision support: A pre-implementation study. Clinical Biochemistry. 2023;113:70-77. https://doi.org/10.1016/j.clinbiochem.2023.01.002

26. Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. BMJ. 2015;350:g7594. https://doi.org/10.1136/bmj.g7594

27. Collins GS, Moons KGM, Dhiman P, Riley RD, Beam AL, Van Calster B, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ. 2024;385:e078378. https://doi.org/10.1136/bmj-2023-078378

493 28. Moons KGM, Damen JAA, Kaul T, Hooft L, Andaur Navarro C, Dhiman P, et al. PROBAST+AI:
494 an updated quality, risk of bias, and applicability assessment tool for prediction models using
495 regression or artificial intelligence methods. BMJ. 2025;388:e082505. https://doi.org/10.1136/bmj-
496 2024-082505

497 29. Wolff RF, Moons KGM, Riley RD, Whiting PF, Westwood M, Collins GS, et al. PROBAST: A
498 Tool to Assess the Risk of Bias and Applicability of Prediction Model Studies. Annals of Internal
499 Medicine. 2019;170(1):51-58. https://doi.org/10.7326/M18-1376

500 30. Debray TPA, Collins GS, Riley RD, Snell KIE, Van Calster B, Reitsma JB, et al. Transparent
501 reporting of multivariable prediction models developed or validated using clustered data (TRIPOD-
502 Cluster): explanation and elaboration. BMJ. 2023;380:e071058. https://doi.org/10.1136/bmj-2022-
503 071058

504 31. Efthimiou O, Seo M, Chalkou K, Debray T, Egger M, Salanti G. Developing clinical prediction
505 models: a step-by-step guide. BMJ. 2024;386:e078276. https://doi.org/10.1136/bmj-2023-078276

506 32. Riley RD, Archer L, Snell KIE, Ensor J, Dhiman P, Martin GP, et al. Evaluation of clinical
507 prediction models (part 2): how to undertake an external validation study. BMJ.
508 2024;384:e074820. https://doi.org/10.1136/bmj-2023-074820

509 33. Riley RD, Snell KIE, Archer L, Ensor J, Debray TPA, van Calster B, et al. Evaluation of clinical
510 prediction models (part 3): calculating the sample size required for an external validation study.
511 BMJ. 2024;384:e074821. https://doi.org/10.1136/bmj-2023-074821

512 34. Riley RD, Debray TPA, Collins GS, Archer L, Ensor J, van Smeden M, et al. Minimum sample
513 size for external validation of a clinical prediction model with a binary outcome. Statistics in
514 Medicine. 2021;40(19):4230-4251. https://doi.org/10.1002/sim.9025

515 35. Van Calster B, McLernon DJ, van Smeden M, Wynants L, Steyerberg EW. Calibration: the
516 Achilles heel of predictive analytics. BMC Medicine. 2019;17(1):230.
517 https://doi.org/10.1186/s12916-019-1466-7

518 36. Johnson A, Bulgarelli L, Pollard T, Gow B, Moody B, Horng S, et al. MIMIC-IV (version 3.1).
519 PhysioNet. 2024. https://doi.org/10.13026/kpb9-mt58

520 37. Johnson AEW, Bulgarelli L, Shen L, Gayles A, Shammout A, Horng S, et al. MIMIC-IV, a freely
521 accessible electronic health record dataset. Scientific Data. 2023;10(1):1.
522 https://doi.org/10.1038/s41597-022-01899-x

523 38. Pollard T, Johnson A, Raffa J, Celi LA, Badawi O, Mark R. eICU Collaborative Research
524 Database (version 2.0). PhysioNet. 2019. https://doi.org/10.13026/C2WM1R

525 39. Pollard TJ, Johnson AEW, Raffa JD, Celi LA, Mark RG, Badawi O. The eICU Collaborative
526 Research Database, a freely available multi-center database for critical care research. Scientific
527 Data. 2018;5(1):180178. https://doi.org/10.1038/sdata.2018.178

528 40. Pollard T, Moody BE, Lehman LH, Gow BJ, Fernandes C, Xie C, et al. PhysioNet as a global
529 platform for biomedical research. Nature Health. 2026;1(8):792-795.
530 https://doi.org/10.1038/s44360-026-00096-z

531 41. Saito T, Rehmsmeier M. The Precision-Recall Plot Is More Informative than the ROC Plot
532 When Evaluating Binary Classifiers on Imbalanced Datasets. PLOS ONE. 2015;10(3):e0118432.
533 https://doi.org/10.1371/journal.pone.0118432

534 42. Van Calster B, Wynants L, Verbeek JFM, Verbakel JY, Christodoulou E, Vickers AJ, et al.
535 Reporting and Interpreting Decision Curve Analysis: A Guide for Investigators. European Urology.
536 2018;74(6):796-804. https://doi.org/10.1016/j.eururo.2018.08.038

537 43. Vickers AJ, Elkin EB. Decision Curve Analysis: A Novel Method for Evaluating Prediction
538 Models. Medical Decision Making. 2006;26(6):565-574.
539 https://doi.org/10.1177/0272989X06295361

540 44. Rockenschaub P, Akay EM, Carlisle BG, Hilbert A, Wendland J, Meyer-Eschenbach F, et al.
541 External validation of AI-based scoring systems in the ICU: a systematic review and meta-
542 analysis. BMC Medical Informatics and Decision Making. 2025;25(1):5.
543 https://doi.org/10.1186/s12911-024-02830-7

544 45. Collins GS, de Groot JA, Dutton S, Omar O, Shanyinde M, Tajar A, et al. External validation of
545 multivariable prediction models: a systematic review of methodological conduct and reporting.
546 BMC Medical Research Methodology. 2014;14(1):40. https://doi.org/10.1186/1471-2288-14-40

46. Snell KIE, Archer L, Ensor J, Bonnett LJ, Debray TPA, Phillips B, et al. External validation of clinical prediction models: simulation-based sample size calculations were more reliable than rules-of-thumb. Journal of Clinical Epidemiology. 2021;135:79-89. https://doi.org/10.1016/j.jclinepi.2021.02.011

47. Riley RD, Ensor J, Snell KIE, Debray TPA, Altman DG, Moons KGM, et al. External validation of clinical prediction models using big datasets from e-health records or IPD meta-analysis: opportunities and challenges. BMJ. 2016;353:i3140. https://doi.org/10.1136/bmj.i3140

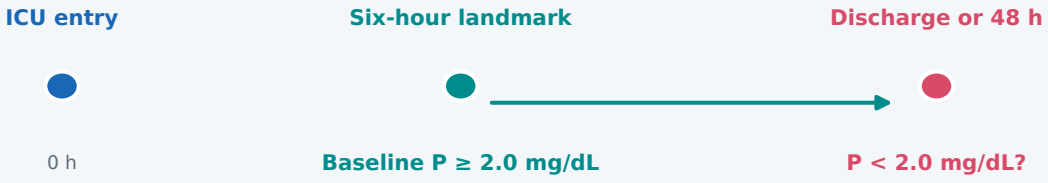
48. El Emam K, Klement W, Malin B. Reporting and Methodological Observations on Prognostic and Diagnostic Machine Learning Studies. JMIR AI. 2023;2:e47995. https://doi.org/10.2196/47995

49. Luo W, Phung D, Tran T, Gupta S, Rana S, Karmakar C, et al. Guidelines for Developing and Reporting Machine Learning Predictive Models in Biomedical Research: A Multidisciplinary View. Journal of Medical Internet Research. 2016;18(12):e323. https://doi.org/10.2196/jmir.5870

50. Vickers AJ, Van Calster B, Wynants L, Steyerberg EW. Decision curve analysis: confidence intervals and hypothesis testing for net benefit. Diagnostic and Prognostic Research. 2023;7(1):11. https://doi.org/10.1186/s41512-023-00148-y

51. King's College London. Research Ethics Policy. Effective 6 March 2024, section 4.1.4. https://www.kcl.ac.uk/assets/policyzone/research/research-ethics-policy.pdf. Accessed 26 September 2026.

A | THE PATIENT-LEVEL QUESTION



Model inputs available by 6 h; baseline cohort confirmed later

B | DEVELOPMENT → EXTERNAL TEST

MIMIC-IV
31,118 eligible
70% develop · 30% test

→

eICU
15,568 eligible
186 hospitals

Model selection stays within MIMIC development data.

eICU receives a frozen model and fixed alert thresholds.

C | THE OBSERVATION LENS

RETESTED
10,133 in eICU

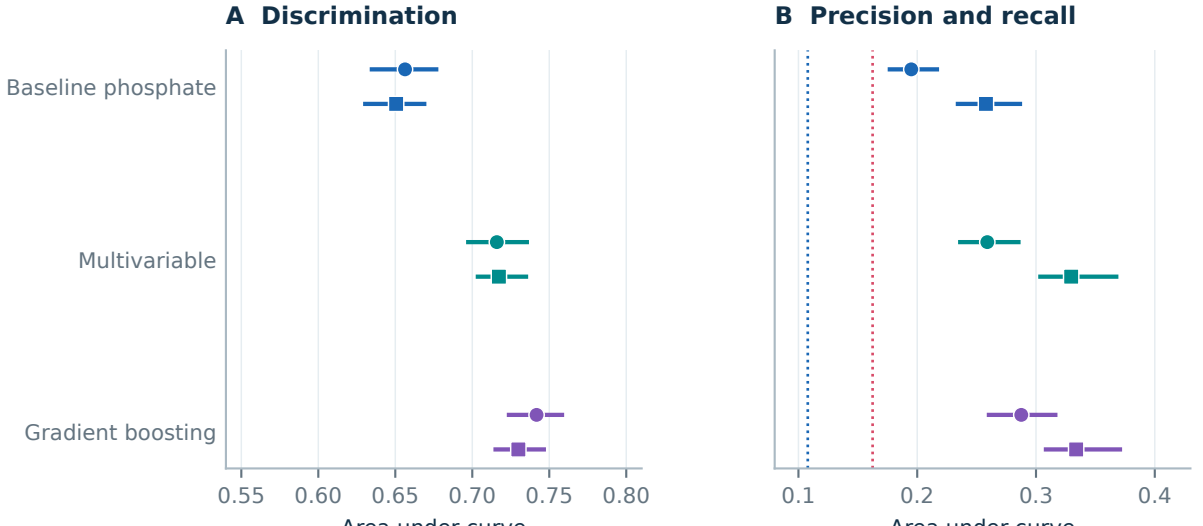
NOT RETESTED
5,435 unknown

Documented event ≠ biological incidence

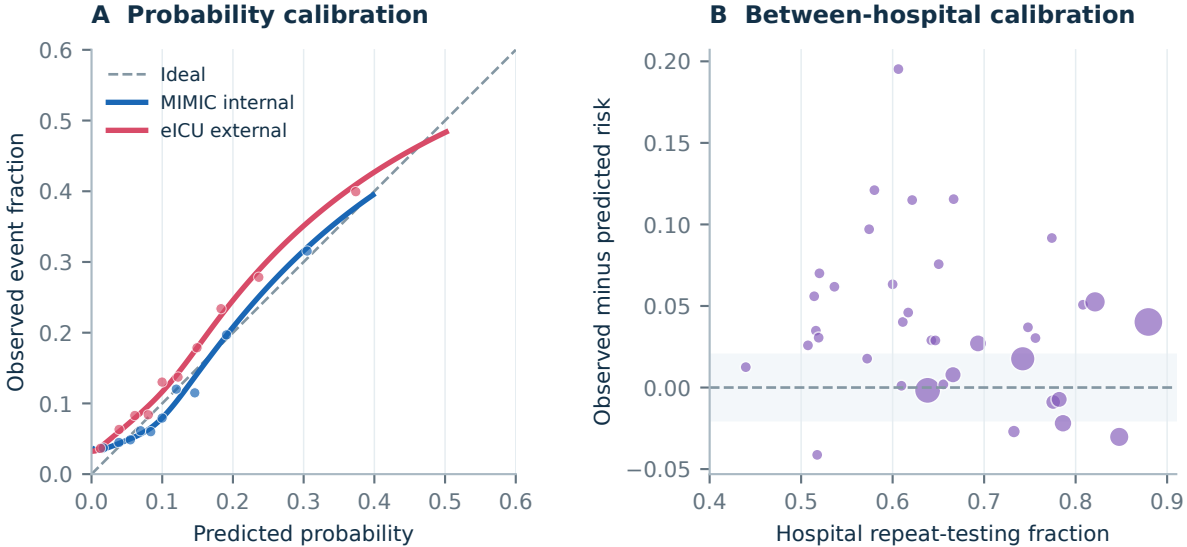
Report testing, label completions, weighting assumptions and between-hospital calibration before interpretation.

RESEARCH QUESTION

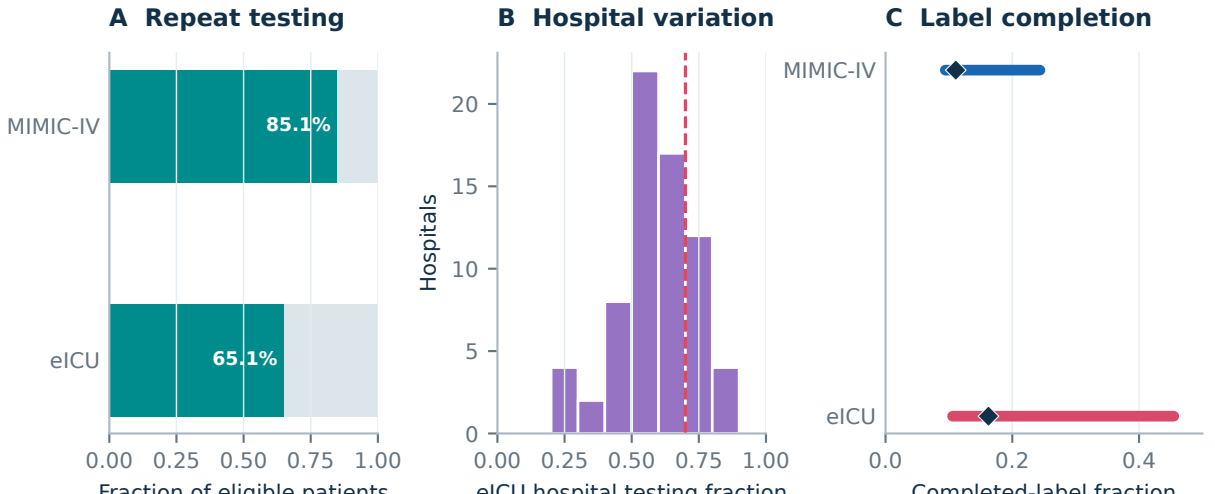
Does prediction of a recorded low result transport across different repeat-testing practices?



Circles: MIMIC internal; squares: eICU external. Lines: 95% bootstrap intervals.



Curves: logistic splines; points: decile means. Hospitals: ≥50 eligible, ≥50 tested, ≥10 events.



Ranges: hypothetical label completions. Diamonds: event fractions among retested patients.

Supplementary Files

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

- Additionalfile1.docx
- Additionalfile3.docx