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.

Blood-gas history and arterial oxygen target classification after acute brain injury: a temporal and multicentre external validation study

Shikun Li
Tongji Hospital
Xu Chen
Tongji Hospital
Rudong Chen
Tongji Hospital
Zhuo Yang
Tongji Hospital
Yuyang Hou
Tongji Hospital
Jiasheng Yu
yujiasheng2000@tjh.tjmu.edu.cn
Tongji Hospital
Hongkuan Yang
Tongji Hospital

Research Article

Keywords: acute brain injury, arterial blood gas, oxygen targets, pulse oximetry, electronic health records, learning health system, external validation, calibration

Posted Date: November 11th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11455413/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

Oxygen management after acute brain injury (ABI) is guided by arterial oxygen tension (PaO_2), whereas bedside monitoring relies on pulse oximetry (SpO_2). Every intensive care unit (ICU) stores a blood-gas history that is rarely reused. We tested whether this history improves PaO_2 target classification beyond SpO_2 across eras and hospitals and supports reliable in-target calls.

Methods

In this retrospective study of adults with ABI (MIMIC-IV 3.1, eICU 2.0), arterial blood gases from the first 72 ICU hours were paired with SpO_2 recorded ≤ 10 minutes earlier; only results available before each draw counted as history. PaO_2 categories were < 80, 80–120 and > 120 mmHg, the middle range reflecting consensus for mechanically ventilated ABI patients. Penalised multinomial models using current information (SpO_2, age, ICU time), blood-gas history or both were developed in MIMIC-IV and evaluated without refitting in patient-disjoint later-era MIMIC-IV and eICU data. The primary outcome was the external Brier score difference; a prespecified ≤ 5% error gate governed automated in-target calls.

Results

Cohorts comprised 2,070 MIMIC-IV patients (4,283 pairs) and 2,373 eICU patients (7,731 pairs). In 4,877 external measurements with history (1,493 patients, 103 hospitals), adding blood-gas history improved the Brier score from 0.5753 to 0.5289 (− 0.0464; 95% CI − 0.0536 to − 0.0390), an 8.1% relative reduction. Improvement was replicated in later-era MIMIC-IV, was directionally consistent in 21 of 22 larger hospitals and persisted in eICU for blood gases > 24 hours old. Post hoc adjustment for recorded inspired oxygen fraction, positive end-expiratory pressure and ventilation status yielded a similar difference (− 0.0486; −0.0556 to − 0.0413). In-target classification, however, failed: 46.6% of the combined model's external calls were wrong, two in five eICU measurements with SpO_2 94–98% had PaO_2 <80 mmHg, and no model passed the error gate.

Conclusions

Blood-gas history stored in ICU databases provides reproducible incremental information for PaO_2 target classification after ABI, including after adjustment for recorded respiratory support, but the evaluated models could not reliably confirm in-target PaO_2. These findings support displaying the latest blood gas and its age alongside SpO_2, retaining arterial confirmation, and requiring external validation and prespecified error gates before deploying database-derived tools.

Background

Oxygen management after acute brain injury (ABI) is anchored to arterial oxygen tension (PaO_2), yet continuous bedside monitoring provides peripheral oxygen saturation (SpO_2). The European Society of Intensive Care Medicine (ESICM) consensus on mechanical ventilation in ABI recommends a PaO_2 range of 80–120 mmHg [1]. Applying this range requires clinicians to judge, from the information at hand, whether PaO_2 lies below, within or above it—a judgement made many times each day in every neurocritical care unit.

Intensive care units (ICUs) already hold much of the information needed for this judgement, but rarely use it. Every arterial blood gas is stored, time-stamped, in the electronic health record (EHR) and remains available long after the decision it originally informed. The learning health system model holds that such routinely generated data should be systematically fed back to improve care [2, 3]. Open multicentre ICU databases make it possible to test, before any bedside deployment, whether a data element carries transferable information and whether algorithms built on it are reliable enough to act on.

Prior work has defined SpO_2 targets that keep PaO_2 within a physiological range in acutely ill patients [4], evaluated how well a pulse-oximetry-derived index detects hyperoxia in neurocritical care [5] and characterised the limitations of the ratio of SpO_2 to the fraction of inspired oxygen (FiO_2) for classifying and monitoring acute respiratory distress syndrome (ARDS) [6]. A machine-learning respiratory monitoring system (RMS) has incorporated prior PaO_2, paired SpO_2, elapsed time and previous prediction errors into PaO_2 estimation [7]. Two questions central to data-driven neurocritical care follow: how much does blood-gas history add to current SpO_2 for classifying PaO_2 against the ABI target range, and is the gain large enough to support low-error automated in-target calls?

We used MIMIC-IV and the multicentre eICU database as a learning-health-system test bed. Models were developed in MIMIC-IV, locked and evaluated in a patient-disjoint later-era MIMIC-IV cohort and in 103 eICU hospitals, with timestamp rules that admitted only blood-gas results available before each draw. By comparing current-information, prior-only and current-plus-prior models at identical measurement times, we isolated the information contributed by blood-gas history, mapped how long it remains informative, tested its stability across hospitals and applied a prespecified error gate to decide whether automated in-target calls could be released. A locked extension then tested whether the gain survives adjustment for recorded respiratory support. Together, the answers define what database-derived oxygenation tools can and cannot do at the bedside.

Methods

Study design and data sources

This retrospective study used MIMIC-IV version 3.1 for development and later-era evaluation and the eICU Collaborative Research Database version 2.0 for external evaluation [8–11]. MIMIC-IV contains records from Beth Israel Deaconess Medical Center in Boston, United States, during 2008–2022; eICU

contains records from participating United States hospitals during 2014–2015. The task was classification of the current PaO_2 category at observed blood-gas sampling times. The analysis plan was locked before model fitting, and models and thresholds were locked before later-era and external evaluation; both databases had been inspected for pairing feasibility beforehand, and the plan was not publicly registered. Reporting followed the TRIPOD + AI statement [12].

Participants and reference measurements

We included adults aged at least 18 years with diagnostic codes or structured diagnosis entries for traumatic brain injury, intracerebral haemorrhage, subarachnoid haemorrhage or ischaemic stroke; subtype flags could overlap. MIMIC flags were assigned at the hospital-admission level and eICU flags at the ICU-stay level. Admissions or ICU stays with detected cardiac-arrest, anoxic-injury or extracorporeal membrane oxygenation (ECMO) flags were excluded, and uncertain ABI labels were retained separately. The exact code and text rules are provided in Additional file 1. Eligibility did not require mechanical ventilation or complete Glasgow Coma Scale, FiO_2, respiratory-device or medication data, so the cohort reflects blood-gas-tested patients with coded ABI in routine ICU practice.

Eligible measurements occurred during the first 72 hours of each ICU stay. In MIMIC, blood-gas PaO_2 results (item label "pO2") were linked by specimen identifier to an explicit arterial specimen-type entry. In eICU, which lacks an individual specimen-source field, arterial measurements were identified by the arterial-named PaO_2 test ("paO2"), the arterial blood gas (ABG) laboratory category and compatible pressure units. We retained finite PaO_2 values greater than zero and no greater than 800 mmHg after unit conversion, excluding censored or cancelled entries and unresolved duplicate conflicts. No value was excluded because PaO_2 was low, because SpO_2 and PaO_2 disagreed or because a model made a large error. PaO_2 <80, 80–120 inclusive and > 120 mmHg defined the three reference categories. The 80–120 mmHg interval, recommended by consensus for mechanically ventilated patients with ABI [1], served as the analytical reference rather than an individualised treatment target.

Measurement pairing and timestamp-faithful history

Each accepted blood gas was paired with the most recent valid SpO_2 observation at or before the recorded sampling time and no more than 10 minutes earlier, without interpolation or use of future observations. SpO_2 values had to be finite and within (0,100]%. Exact duplicate observations were collapsed and unresolved same-time conflicts were excluded. MIMIC charttime denotes the observation time, whereas storetime records manual entry or validation [13]; the eICU vitalPeriodic saO2 field contains monitor-derived SpO_2 stored as five-minute medians [14]. The main pairing therefore reflects when measurements were made.

A measurement was history-eligible when a distinct earlier paired sample existed in the same ICU stay and its information was demonstrably available before the current draw. The earlier PaO_2 result and, in

MIMIC, its specimen metadata had to carry a valid storage or revision time strictly before the current draw, and its paired SpO_2 had to be recorded by the current draw. The most recent sample satisfying these conditions was selected; samples with missing or temporally impossible availability times did not count as history. The eICU revised-result offset was treated conservatively as the time the retained result became available [15]. No current-sample PaO_2, pH, arterial carbon dioxide tension ($PaCO_2$) or other contemporaneous blood-gas result was used as a predictor, which excludes information leakage from the measurement being classified. Measurements without eligible history remained in the cohort and in coverage denominators.

Patient partitioning and model comparisons

MIMIC admission-era intervals were reconstructed from anchor-year ranges and each ICU admission's offset from the patient's anchor year. Patients whose primary pairs all had interval upper bounds before 2020 formed the earlier group; those whose interval lower bounds were all 2020 or later formed the later-era test group. Patients with boundary-spanning intervals were described separately and excluded from fitting and later-era evaluation. A fixed patient-identifier hash allocated approximately 20% of earlier patients to threshold calibration and the remainder to training, with all encounters of a patient kept in the same partition. The later-era test set was therefore both temporally separated and patient-disjoint.

The fixed rule A mapped SpO_2 <94%, 94–98% and > 98% to the three PaO_2 categories; P0 carried forward the previous PaO_2 category. The prior-only model (P) used previous PaO_2 and elapsed time since that sample. The current-information model (B) used current SpO_2, age capped at 90 years and elapsed ICU time, information available at every bedside irrespective of ventilation status; recorded respiratory support was added in a post hoc extension (see below). The current-plus-prior model (C) included all predictors from B and P plus the SpO_2 paired with the prior blood gas. Transformations, including SpO_2 hinge terms and the prior-PaO_2/time interaction, are specified in Additional file 1. P, B and C used identical history-eligible training and evaluation samples, so that differences between them isolate the contribution of each information source. A separate model, B_all, used the B specification fitted to all training pairs, including those without history, for overall and no-history descriptions.

We deliberately used a transparent model class, L2-penalised multinomial logistic regression without class weighting, so that performance differences reflect information content rather than algorithmic capacity. Predictors were standardised within each training fold. The inverse regularisation strength was selected from 0.01, 0.1, 1 and 10 by pooled out-of-fold Brier score in three-fold patient-grouped cross-validation. Models were refitted on their training samples only and were applied to later-era and external data without refitting or recalibration. No oxygen measurement was imputed; the prespecified rule for missing age was imputation with the training median. Feature transformations, standardisation procedures and software versions are described in Additional files 1 and 2; access to fitted parameters is described under Availability of data and materials.

Performance measures and the in-target error gate

The primary measure was the pair-weighted external difference in multiclass Brier score between C and B. For each observation, the score was the sum of squared differences between the three predicted probabilities and the corresponding class indicators, without division by three (range 0–2). Negative C − B values favoured C; C − P was a secondary contrast. We also reported class-specific Brier scores, one-versus-rest areas under the receiver operating characteristic curve (AUROCs), classification metrics and calibration (fixed probability bins, calibration-in-the-large and logistic calibration slopes).

Ordinary in-target classification used the category with the highest predicted probability; the false in-target error was the proportion of such calls with measured PaO_2 <80 or > 120 mmHg. To emulate a deployment decision, a selective error gate was specified before fitting. In the MIMIC calibration set, candidate in-target probability thresholds from 0.50 to 0.99, in increments of 0.01, qualified if they yielded an empirical error ≤ 5% among at least 30 selected measurements from at least 20 patients. The qualifying threshold with the greatest coverage was to be carried forward, with ties broken in favour of the higher threshold; if none qualified, the policy made no in-target calls. Coverage was reported against both history-eligible and all paired measurements. External error–coverage curves were descriptive and were not used to choose thresholds.

Statistical analysis and prespecified supplementary evaluations

Confidence intervals (95%) were derived from 1,000 patient-cluster bootstrap replicates of the locked predictions, with paired resampling for model contrasts; all observations from a sampled patient, including those from different hospitals, were resampled together. The primary analysis weighted measurements equally; patient-equal and external hospital-equal comparisons were secondary. All eligible observations were used without a formal sample-size calculation; secondary and subgroup analyses were descriptive, and AUROCs and calibration coefficients are reported as point estimates.

We characterised the C − B loss difference continuously against log(1 + prior sampling interval) using restricted cubic splines with prespecified knots corresponding to 1, 6, 12 and 24 hours and patient-cluster bootstrap bands, and within prespecified strata of > 0–6, > 6–12, > 12–24 and > 24–72 hours. Hospital-specific estimates were presented for centres with at least 20 history-eligible patients, with smaller centres pooled; all centres remained in overall and leave-one-hospital-out analyses. Further analyses described overlapping ABI subtypes; applied the locked models to pairs formed within a 5-minute window and, in eICU, to SpO_2 observations 5–10 minutes before sampling; and described a reference group without a detected ABI code. Alternative PaO_2 ranges were used for descriptive fixed-rule tabulations.

To emulate real-time deployment, a record-availability analysis selected the most recent SpO_2 from observations already entered at the time of sampling, within the unchanged 10-minute window, with

duplicate conflicts assessed using only observations known at that time. We distinguished a correction that changed only the selection of current SpO_2, keeping the original history set, from a stricter analysis that also rebuilt history from record-available pairs. Both applied the locked models without retraining; in eICU, monitoring availability could be approximated only by observation time. Descriptive baseline and availability tables were completed after model evaluation and did not alter any analysis. Patient-level summaries used the earliest paired ICU admission within each MIMIC patient and a deterministic representative admission for each eICU patient. Age and included oxygen measurements had no missing values; recorded sex was other or unknown for two eICU patients.

Post hoc respiratory-support extension

After the frozen analysis and a descriptive audit of FiO_2 availability had been completed, a post hoc extension was specified and locked before any extension model was fitted (amendment v1.1; lock files and execution log in Additional file 2). Current FiO_2 and positive end-expiratory pressure (PEEP) were taken from the most recent bedside record both observed and entered before the draw within 4 hours, with explicitly documented room air assigned an FiO_2 of 0.21. Ventilation status was reconstructed from documented ventilator modes, delivery devices and ventilator on/off entries as positive-pressure ventilation, no positive-pressure ventilation or unknown; neither PEEP nor an airway device alone was taken as evidence of ventilation. Missing FiO_2 and PEEP were imputed with training-fold medians and accompanied by indicator variables. Model B+ added these variables to model B; model B + H added the five original history features to B+; and model C+ further added the prior PaO_2/FiO_2 (P/F) ratio and the change in FiO_2 since the prior sample. The primary extension contrast was the external B + H−B+ Brier difference, with C + − B + H as a secondary contrast. Fitting, threshold selection, the error gate and bootstrap procedures followed the original rules. Prespecified sensitivity analyses used a 6-hour window (S1), strict complete cases (S2) and same-draw blood-gas FiO_2 as a retrospective adjustment (S3) (Additional file 1: S6).

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

Results

Cohorts and availability of blood-gas history

The primary measurement cohorts contained 4,283 pairs from 2,070 MIMIC patients and 7,731 pairs from 2,373 eICU patients across 118 hospitals (Fig. 1). Of the MIMIC pairs, 2,611 were assigned to training, 605 to calibration, 830 to later-era testing and 237 to the boundary-era description (1,288, 316, 378 and 88 patients, respectively). Matched model training used 1,243 history-eligible pairs from 584 patients, and threshold calibration used 271 pairs from 135 patients.

Blood-gas history was available at the moment of sampling for 423 later-era MIMIC pairs (180 patients) and 4,877 eICU pairs (1,493 patients; 103 hospitals), which formed the matched comparison sets; nearly two-thirds (63.1%) of external measurements therefore had usable history. The remaining 407 and 2,854 pairs were retained in coverage denominators. Timestamp rules withheld history from 343 eICU and 39 MIMIC measurements whose chronologically earlier blood gas was not demonstrably available before the current draw (Additional file 1: Table S7). Median age was 64 years in both the later-era MIMIC and external cohorts, and recorded female sex accounted for 38.1% and 39.5%, respectively (Table 1; Additional file 1: Tables S1 and S2). The external coded case mix included more traumatic brain injury and less intracerebral haemorrhage than later-era MIMIC, providing a stringent test of transportability.

Blood-gas history improves probability classification

Blood-gas history improved the primary external probability score (Table 2; Fig. 2a). The Brier score fell from 0.5753 with the current-information model (B) to 0.5289 with the current-plus-prior model (C): paired difference −0.0464 (95% confidence interval −0.0536 to −0.0390), an 8.1% relative reduction in scoring loss. The current-plus-prior model also outperformed the prior-only model (P; 0.5920): C−P = −0.0631 (−0.0769 to −0.0483), showing that current SpO_2 and blood-gas history carry complementary information.

The increment was reproduced in later-era MIMIC, where Brier scores were 0.4819 for B and 0.4595 for C (difference −0.0224; −0.0383 to −0.0063); the corresponding C−P difference was −0.1024 (−0.1434 to −0.0572). External C−B remained clearly negative when patients were weighted equally, −0.0398 (−0.0483 to −0.0320), and when hospitals were weighted equally, −0.0313 (−0.0445 to −0.0177).

Class-specific scores favoured C in every PaO_2 category, with external Brier scores lower for C than for B below 80 mmHg (0.1278 versus 0.1395), within 80–120 mmHg (0.2230 versus 0.2372) and above 120 mmHg (0.1780 versus 0.1985) (Additional file 1: Table S3). Discrimination by the current-plus-prior model was high for out-of-range PaO_2, with external AUROCs of 0.8682 below 80 mmHg and 0.8361 above 120 mmHg, but markedly lower for the in-target category (0.6747).

In-target calls remain unreliable

External probabilities were systematically miscalibrated (Fig. 3). For the current-plus-prior model, calibration-in-the-large was +1.112 for PaO_2 below 80 mmHg and −0.673 for PaO_2 above 120 mmHg, indicating underprediction of low and overprediction of high PaO_2; the in-target calibration slope was 0.654. This pattern paralleled a shift in category prevalence: among history-eligible measurements, PaO_2 below 80 mmHg accounted for 175/1243 (14.1%) training and 1132/4877 (23.2%) external observations, and PaO_2 above 120 mmHg for 566/1243 (45.5%) and 1870/4877 (38.3%), respectively (Additional file 1: Table S2).

Under ordinary maximum-probability classification, the current-plus-prior model called 1,471 of 4,877 history-eligible eICU measurements in target, a history-eligible coverage of 30.2% and 19.0% of all 7,731 primary external pairs. Nearly half of these calls were wrong: 686 (46.6%; 95% confidence interval 43.9–49.4%) had PaO_2 outside the range, 529 (36.0%) below 80 mmHg and 157 (10.7%) above 120 mmHg. The current-information model called 24.8% of history-eligible measurements in target with 49.3% error, and the conventional SpO_2 94–98% rule called 33.3% with 50.7% error (Table 3). In later-era MIMIC, C covered 37.6% with 39.6% error (31.0–47.9%).

No probabilistic model came close to the prespecified error gate. Among calibration thresholds meeting the minimum measurement and patient counts, the lowest empirical false in-target error was 39.5% for both B and C and 32.4% for B_all, against the required ≤5%; P had no threshold meeting the count requirements. The gate therefore withheld all automated in-target calls in subsequent evaluation. The external error–coverage curves showed the same ceiling: in eICU, no threshold that selected at least 30 measurements reduced the false in-target fraction below 45% for any model (Fig. 2b).

Durability across information age and hospitals

Blood-gas history retained its value well beyond the first hours after sampling (Additional file 1: Fig. S1 and Table S4). In eICU, all four prespecified interval strata showed a negative C–B difference, including blood gases obtained more than 24 and up to 72 hours earlier: −0.0473 (−0.0606 to −0.0341). In later-era MIMIC, the increment was clear up to 12 hours (>0–6 hours, −0.0256; >6–12 hours, −0.0350), whereas estimates beyond 12 hours were imprecise. The data identified no common interval after which historical information ceased to contribute.

Twenty-two of the 103 history-eligible eICU hospitals had at least 20 eligible patients; the other 81 contributed 1,238 pairs from 405 patients to the pooled evaluation. Twenty-one of the 22 larger centres had a negative C–B point estimate, with confidence intervals excluding zero in 14 (Additional file 1: Fig. S2 and Table S6). Omitting each hospital in turn yielded overall C–B estimates from −0.0479 to −0.0433, so the average increment was not driven by any single hospital.

Sensitivity analyses

The external increment was insensitive to the pairing window: in eICU, C–B was −0.0469 (−0.0544 to −0.0396) with a five-minute window and −0.0458 (−0.0537 to −0.0387) with SpO_2 observations 5–10 minutes before sampling (Additional file 1: Table S5). The smaller later-era MIMIC five-minute analysis, with 197 pairs from 112 patients, was directionally consistent (−0.0170; −0.0398 to 0.0055).

Restricting current SpO_2 to already-entered observations preserved the increment in the later-era subset. This correction enlarged the record-available MIMIC subset from 551 pairs in 423 patients to 565 pairs in 433 patients without changing the main measurement pairs (Additional file 1: Table S7); in its later-era

portion (84 pairs, 69 patients), C–B was −0.0506 (−0.0856 to −0.0183) (Additional file 1: Table S5). Rebuilding both current and historical pairs under the strict record-availability definition left 257 pairs from 195 patients overall and 30 pairs from 25 patients in later-era testing, where the C–B estimate remained negative (−0.0092; −0.0415 to 0.0265).

Adjustment for recorded respiratory support

Respiratory-support records were available for most history-eligible measurements: current FiO_2 for 1,467 of 2,077 MIMIC (70.6%) and 3,477 of 4,877 eICU measurements (71.3%). Documented positive-pressure ventilation was identified for 1,375 MIMIC (66.2%) and 1,257 eICU measurements (25.8%), and ventilation status was unknown for 569 (27.4%) and 3,501 (71.8%), respectively.

Adjustment for these recorded variables left the increment intact (Additional file 1: Table S8). In eICU, the Brier score fell from 0.5787 with B+ to 0.5301 with B+H (difference −0.0486; 95% confidence interval −0.0556 to −0.0413), an 8.4% relative reduction that closely matched the unadjusted increment (−0.0464). The adjusted increment was reproduced in later-era MIMIC (−0.0229; −0.0366 to −0.0094), held under patient-equal (−0.0426; −0.0510 to −0.0345) and hospital-equal weighting (−0.0316; −0.0459 to −0.0180), was directionally consistent in 21 of 22 larger hospitals and persisted with a 6-hour window (−0.0449; −0.0516 to −0.0382). Recorded respiratory support itself did not improve the external score of the current-information model (B+ 0.5787 versus B 0.5753).

In the strict complete-case subset, composed predominantly of measurements with documented ventilation, point estimates again favoured blood-gas history in later-era MIMIC (−0.0330; −0.0596 to −0.0085) and eICU (−0.0183; −0.0399 to 0.0034; 732 measurements, 249 patients). With same-draw blood-gas FiO_2 as a retrospective adjustment, the external increment was −0.0212 (−0.0277 to −0.0153). Adding the prior P/F ratio and FiO_2 change provided no further gain (C+−B+H in eICU, −0.0001; −0.0019 to 0.0019) and worsened external scores in S1 (0.0115; 0.0088 to 0.0145) and S3 (0.0049; 0.0012 to 0.0085). No extension model passed the 5% error gate; the lowest calibration error for B+H was 40.0%.

Discussion

Blood-gas history already stored in ICU databases is a robust, transportable predictor of arterial oxygen target status after ABI. Adding prior blood-gas results to current SpO_2, age and ICU time cut the external probability loss by 8.1% across 103 eICU hospitals. The gain proved robust: it was reproduced in a later era, held under patient- and hospital-equal weighting and alternative pairing windows, was directionally consistent in 21 of 22 larger hospitals, persisted in eICU for blood gases more than a day old and was unchanged after adjustment for recorded FiO_2, PEEP and ventilation status. The same analysis exposes a hard ceiling: no rule or model achieved reliable in-target classification. About one in two external in-target calls was wrong whichever rule was used (46.6–51.9%), and the lowest calibration-set error of any model (32.4% for the all-pair baseline; 39.5% for the matched models) was more than six times the

prespecified 5% gate. Blood-gas history sharpens oxygenation assessment, but it cannot replace arterial measurement when PaO_2 must be confirmed within target.

These results extend saturation-based assessment of arterial oxygenation [4–6] to the absolute PaO_2 target used in ABI and complement machine-learning systems that incorporate blood-gas history into respiratory monitoring [7]. Three design features make the estimate credible and reusable. Models were compared at identical measurement times, isolating the information contributed by blood-gas history independently of any particular algorithm; history was admitted only when its availability before the draw could be audited; and locked models faced temporal and 103-hospital external validation followed by a prespecified deployment gate. Together, these features provide a template for evaluating database-derived oxygenation tools before they reach the bedside.

The value of blood-gas history was not a proxy for recorded respiratory support. Current FiO_2, PEEP and ventilation status did not improve the external score of the current-information model, whereas adding history reduced the loss by a similar margin with or without adjustment (− 0.0486 versus − 0.0464). This pattern is consistent with prior PaO_2 conveying patient-specific gas exchange that ventilator settings do not capture. Adding the prior P/F ratio and FiO_2 change together gave no further external gain and worsened external scores in two sensitivity analyses.

Better probability scores did not yield reliable in-target calls because the target band is the hardest category to identify. Brier scores reward accurate probabilities across all categories, whereas the false in-target fraction concerns only the measurements selected for that decision. In-target discrimination (AUROC 0.67) lagged far behind discrimination of below-range and above-range PaO_2 (0.87 and 0.84), consistent with the flat upper portion of the oxyhaemoglobin dissociation curve, on which a narrow SpO_2 interval spans a wide range of PaO_2. External calibration shifted with category prevalence—below-range PaO_2 was more common and above-range PaO_2 less common in eICU than in training—so transportable oxygenation tools require site-level calibration monitoring, not discrimination alone [16].

The data expose a clinically important blind spot of SpO_2-guided oxygen management. Among all primary pairs with SpO_2 94–98%, a widely used target band, PaO_2 was below 80 mmHg in one in five MIMIC measurements (328/1644; 19.95%) and two in five eICU measurements (970/2427; 39.97%). Because PaO_2 ranges associated with the lowest ICU mortality differ across ABI subtypes [17], a reassuring saturation can conceal a large share of patients below the consensus target.

In eICU, a single arterial sample stayed informative for more than a day: blood gases obtained more than 24 hours earlier still improved classification, whereas later-era MIMIC estimates beyond 12 hours were imprecise. These estimates describe information value at clinician-chosen testing times rather than a retesting schedule. For ICU learning health systems, the findings show both sides of data reuse: routinely stored data carry transportable signal, and the same data can demonstrate, before deployment, when that signal is insufficient to automate a decision.

These findings provide direct guidance for clinical practice and for ICU learning health systems:

(1) Oxygenation displays and decision-support tools should present the most recent blood gas and its age alongside SpO_2; across external hospitals, this history added information independently of recorded respiratory support and even when more than a day old.

(2) A reassuring SpO_2, alone or combined with earlier results, must not be taken as confirmation that PaO_2 lies within 80–120 mmHg; where precise PaO_2 targeting matters, a contemporaneous arterial blood gas is required.

(3) Database-derived oxygenation algorithms should use only inputs available at the time of use, be validated across eras and hospitals, be monitored for site-level calibration and be released only after passing prespecified error gates.

(4) Prospective studies should test whether displaying blood-gas history improves oxygen management, and next-generation models should integrate complete real-time ventilator data and SpO_2 trajectories and be evaluated across patient groups, including those with different skin pigmentation.

This study has limitations. ABI, exclusions and arterial specimen origin were identified from codes and structured records rather than chart review, and neurological severity was unavailable. Respiratory support was captured only as recorded: ventilation status was unknown for 71.8% of external measurements, the strict complete-case external estimate was imprecise, and the extension was specified after the primary results were known, although it was locked before fitting. The 80–120 mmHg reference, derived for ventilated patients, served as an analytical rather than individual treatment target. Current SpO_2 was paired by observation time, eICU five-minute monitoring medians and result-revision offsets approximated contemporaneous measurement and availability, and the strict real-time reconstruction (30 later-era pairs) was too small to confirm the increment under full real-time conditions. Results apply to clinician-chosen testing times rather than to time spent outside the target, the later-era sample was modest, bootstrap intervals were conditional on the locked models and observed hospitals, and fairness across skin pigmentation and demographic groups was not assessed. These factors bear on deployment rather than on the matched comparison, in which every model was evaluated at identical measurement times.

Conclusions

Blood-gas history stored in ICU databases is a robust, transportable predictor of arterial oxygen target status after ABI: its gain was reproduced across eras and 103 hospitals, persisted in eICU for blood gases more than a day old and was unchanged after adjustment for recorded respiratory support. SpO_2-based tools, with or without this history, nonetheless cannot confirm in-target PaO_2: about half of in-target calls were wrong, and no model passed a prespecified 5% error gate. Oxygenation displays should pair SpO_2 with the latest blood gas and its age while retaining contemporaneous arterial confirmation, and database-derived algorithms must clear external validation with time-of-use inputs and prespecified error gates before deployment.

Abbreviations

ABG, arterial blood gas; ABI, acute brain injury; ARDS, acute respiratory distress syndrome; AUROC, area under the receiver operating characteristic curve; CI, confidence interval; ECMO, extracorporeal membrane oxygenation; EHR, electronic health record; ESICM, European Society of Intensive Care Medicine; FiO_2, fraction of inspired oxygen; ICH, intracerebral haemorrhage; ICU, intensive care unit; IQR, interquartile range; P/F, ratio of PaO_2 to FiO_2; $PaCO_2$, arterial carbon dioxide tension; PaO_2, arterial oxygen tension; PEEP, positive end-expiratory pressure; RMS, respiratory monitoring system; SAH, subarachnoid haemorrhage; SpO_2, peripheral oxygen saturation measured by pulse oximetry; TBI, traumatic brain injury.

Declarations

Ethics approval and consent to participate

This study used existing deidentified records from MIMIC-IV and eICU without participant contact or attempted reidentification. The collection of patient information and creation and sharing of MIMIC-IV were reviewed by the Institutional Review Board of Beth Israel Deaconess Medical Center, which approved data sharing and waived informed consent. eICU was released after deidentification in accordance with the Safe Harbor standard of the Health Insurance Portability and Accountability Act [8–11]. No additional institutional ethics approval or individual informed consent was required for this secondary analysis of deidentified records. This secondary analysis was conducted in accordance with the ethical principles of the Declaration of Helsinki, including protection of patient privacy and confidentiality.

Consent for publication

Not applicable.

Availability of data and materials

MIMIC-IV version 3.1 and eICU version 2.0 are available through PhysioNet subject to credentialing, required training and data-use agreements. Data access for this study was authorised through PhysioNet after completion of these requirements. Patient-level source data and derived individual records are not redistributed. The analysis protocol, code and aggregate results are provided in Additional file 2. Fitted model objects and complete reconstruction parameters are excluded from this public supplement under PhysioNet's guidance for MIMIC-derived models. These materials have not been deposited; any future sharing would require a PhysioNet release under the source-data agreement.

Competing interests

The authors declare that they have no competing interests.

Funding

This research received no specific grant from any funding agency in the public, commercial or not-for-profit sectors.

Authors' contributions

SL designed the study, analysed the data and drafted the manuscript. XC reviewed the eligibility criteria, diagnostic codes and variable definitions and revised the corresponding Methods descriptions. RC contributed to the clinical interpretation of the findings and revised the Discussion. ZY checked the consistency of reported results, denominators and model comparisons across the main manuscript and supplementary tables and revised the relevant reporting. YH reviewed and revised the presentation of tables and figures, their legends and supplementary cross-references. JY supervised the study, interpreted the results and critically revised the manuscript. HY co-supervised the study, interpreted the results and critically revised the manuscript. All authors reviewed and approved the final manuscript.

Acknowledgements

Not applicable.

References

1. Robba C, Poole D, McNett M, Asehnoune K, Bösel J, Bruder N, et al. Mechanical ventilation in patients with acute brain injury: recommendations of the European Society of Intensive Care Medicine consensus. Intensive Care Med. 2020;46:2397–2410. https://doi.org/10.1007/s00134-020-06283-0.
2. Friedman C, Rubin J, Brown J, Buntin M, Corn M, Etheredge L, et al. Toward a science of learning systems: a research agenda for the high-functioning Learning Health System. J Am Med Inform Assoc. 2015;22:43–50. https://doi.org/10.1136/amiajnl-2014-002977.
3. Celi LA, Mark RG, Stone DJ, Montgomery RA. “Big data” in the intensive care unit: closing the data loop. Am J Respir Crit Care Med. 2013;187:1157–1160. https://doi.org/10.1164/rccm.201212-2311ED.
4. Röttgering JG, de Man AME, Schuurs TC, Wils EJ, Daniels JM, van den Aardweg JG, et al. Determining a target SpO_2 to maintain PaO_2 within a physiological range. PLoS One. 2021;16(5):e0250740. https://doi.org/10.1371/journal.pone.0250740.

5. de Courson H, Julien-Laferrière T, Georges D, Boyer P, Verchère E, Biais M. The ability of Oxygen Reserve Index® to detect hyperoxia in critically ill patients. Ann Intensive Care. 2022;12:40. https://doi.org/10.1186/s13613-022-01012-w.
6. Erlebach R, Pale U, Beck T, Markovic S, Seric M, David S, et al. Limitations of SpO_2/FiO_2-ratio for classification and monitoring of acute respiratory distress syndrome—an observational cohort study. Crit Care. 2025;29:82. https://doi.org/10.1186/s13054-025-05317-7.
7. Hüser M, Lyu X, Faltys M, Pace A, Berger D, Hoche M, et al. RMS: a ML-based system for ICU respiratory monitoring and resource planning. npj Digit Med. 2025;8:775. https://doi.org/10.1038/s41746-025-02081-4.
8. Johnson AEW, Bulgarelli L, Shen L, Gayles A, Shammout A, Horng S, et al. MIMIC-IV, a freely accessible electronic health record dataset. Sci Data. 2023;10:1. https://doi.org/10.1038/s41597-022-01899-x.
9. Johnson A, Bulgarelli L, Pollard T, Gow B, Moody B, Horng S, et al. MIMIC-IV (version 3.1). PhysioNet. 2024. https://doi.org/10.13026/kpb9-mt58.
10. Pollard TJ, Johnson AEW, Raffa JD, Celi LA, Mark RG, Badawi O. The eICU Collaborative Research Database, a freely available multi-center database for critical care research. Sci Data. 2018;5:180178. https://doi.org/10.1038/sdata.2018.178.
11. Pollard T, Johnson A, Raffa J, Celi LA, Badawi O, Mark R. eICU Collaborative Research Database (version 2.0). PhysioNet. 2019. https://doi.org/10.13026/C2WM1R.
12. 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.
13. MIT Laboratory for Computational Physiology. The chartevents table. MIMIC documentation. https://mimic.mit.edu/docs/iv/modules/icu/chartevents.html. Accessed 7 Oct 2026.
14. MIT Laboratory for Computational Physiology. vitalPeriodic. eICU Collaborative Research Database documentation. https://eicu.mit.edu/eicutables/vitalperiodic/. Accessed 7 Oct 2026.
15. MIT Laboratory for Computational Physiology. lab. eICU Collaborative Research Database documentation. https://eicu.mit.edu/eicutables/lab/. Accessed 7 Oct 2026.
16. Van Calster B, McLernon DJ, van Smeden M, Wynants L, Steyerberg EW. Calibration: the Achilles heel of predictive analytics. BMC Med. 2019;17:230. https://doi.org/10.1186/s12916-019-1466-7.
17. Premraj L, Camarda C, Romero-Garcia N, Seymour B, Subramaniam A, Badenes R, et al. Early arterial oxygen tension thresholds for mechanically ventilated patients with non-traumatic acute brain injury: a Bayesian analysis of three large neurocritical care datasets. Neurocrit Care. 2026. https://doi.org/10.1007/s12028-026-02659-0.

Tables

Table 1. Characteristics of the primary paired cohorts

Characteristic	MIMIC training	MIMIC calibration	MIMIC later era	eICU external
Patients, n	1288	316	378	2373
Age, years; median [IQR]	65 [52–76]	67 [55.8–78.2]	64 [52–74]	64 [49–76]
Female sex, n (%)	548 (42.5)	143 (45.3)	144 (38.1)	938 (39.5)
TBI, n (%)	325 (25.2)	73 (23.1)	88 (23.3)	1150 (48.5)
ICH, n (%)	364 (28.3)	105 (33.2)	121 (32.0)	157 (6.6)
SAH, n (%)	205 (15.9)	57 (18.0)	65 (17.2)	271 (11.4)
Ischaemic stroke, n (%)	595 (46.2)	134 (42.4)	181 (47.9)	877 (37.0)
Paired ICU stays, n	1347	326	398	2511
Paired measurements, n	2611	605	830	7731
Patients ever history-eligible, n (%)	584 (45.3)	135 (42.7)	180 (47.6)	1493 (62.9)

Patient characteristics use one representative admission per patient; subtype flags may overlap. Age was capped at 90 years. IQR, interquartile range; TBI, traumatic brain injury; ICH, intracerebral haemorrhage; SAH, subarachnoid haemorrhage. Age was complete; sex was other or unknown for two external patients. Boundary-era MIMIC patients are described separately in Table S1.

Table 2. Matched probability scores and paired model contrasts

Evaluation set	Model or contrast	Brier score or difference (95% CI)
MIMIC later era	P	0.5619 (0.5327 to 0.5912)
MIMIC later era	B	0.4819 (0.4342 to 0.5327)
MIMIC later era	C	0.4595 (0.4097 to 0.5062)
MIMIC later era	C−B	−0.0224 (−0.0383 to −0.0063)
MIMIC later era	C−P	−0.1024 (−0.1434 to −0.0572)
eICU external	P	0.5920 (0.5797 to 0.6030)
eICU external	B	0.5753 (0.5557 to 0.5964)
eICU external	C	0.5289 (0.5111 to 0.5492)
eICU external	C−B	−0.0464 (−0.0536 to −0.0390)
eICU external	C−P	−0.0631 (−0.0769 to −0.0483)

P, prior-only model; B, current-information model; C, current-plus-prior model. All three models use identical history-eligible evaluation times: 423 pairs/180 patients in MIMIC and 4,877 pairs/1,493

patients in eICU. Lower Brier scores and negative C–B or C–P differences favour C. The three class losses were summed, not divided by three. Intervals use patient-cluster bootstrap resampling of frozen predictions. The eICU C–B comparison is primary.

Table 3. Ordinary in-target calls and their false in-target fraction

Set	Rule	Calls, n	History coverage, %	All-pair coverage, %	False in-target, % (95% CI)
MIMIC later era	A	181	42.8	21.8	42.0 (34.0 to 50.0)
MIMIC later era	P	211	49.9	25.4	51.2 (44.0 to 58.9)
MIMIC later era	B	147	34.8	17.7	38.1 (29.8 to 46.7)
MIMIC later era	C	159	37.6	19.2	39.6 (31.0 to 47.9)
eICU external	A	1624	33.3	21.0	50.7 (47.8 to 53.4)
eICU external	P	2335	47.9	30.2	51.9 (49.5 to 54.2)
eICU external	B	1209	24.8	15.6	49.3 (46.2 to 52.5)
eICU external	C	1471	30.2	19.0	46.6 (43.9 to 49.4)

For P/B/C, an ordinary call means that the in-target category has the largest predicted probability. A calls SpO_2 94–98% in-target and is shown here on the same history-eligible times. History coverage divides calls by 423 MIMIC or 4,877 eICU pairs; all-pair coverage divides these same calls by 830 or 7,731 primary pairs. False in-target fractions divide measurements below 80 or above 120 mmHg by the number called in-target, not by all observations. Different rules select different measurements. The prespecified ≤5% error gate was not passed in calibration, so the locked selective policy released no in-target calls.

Figures

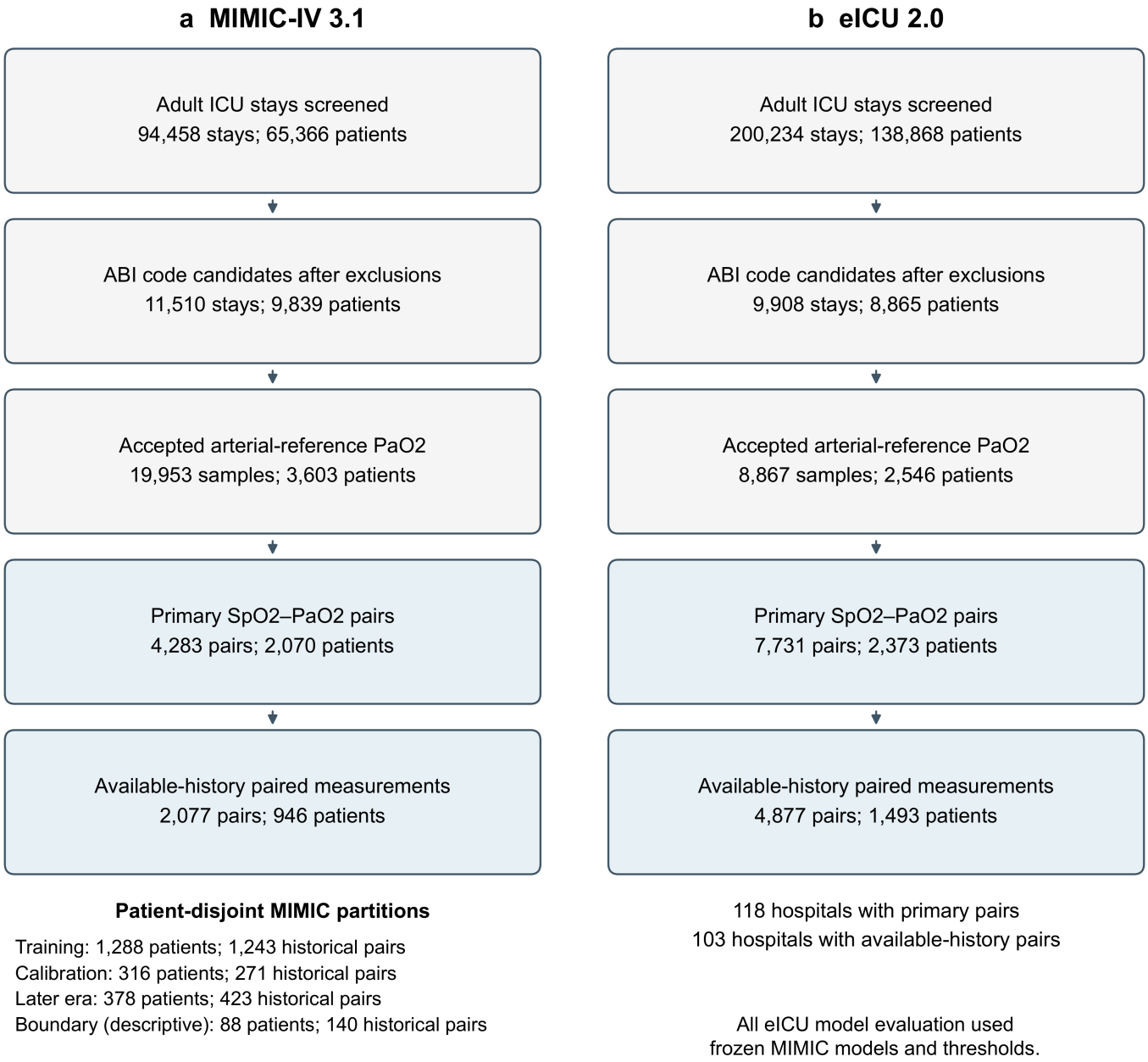
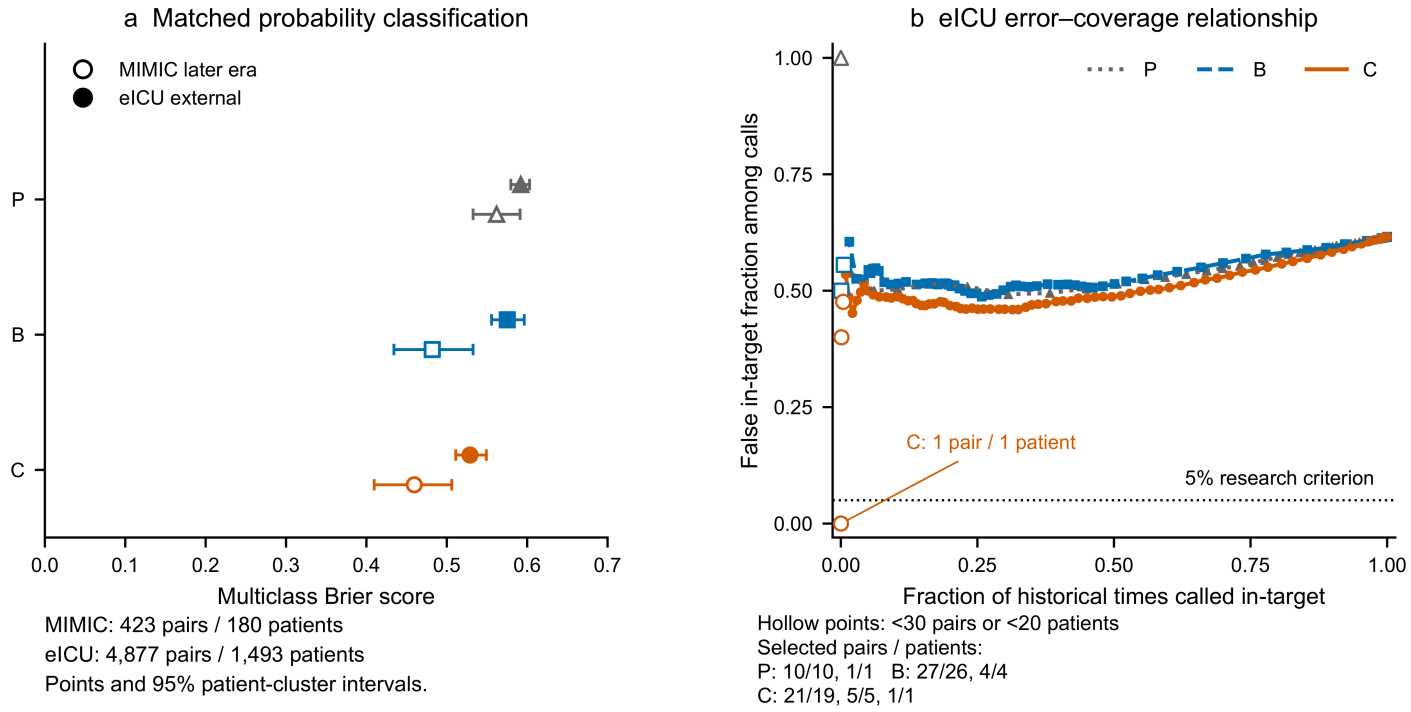


Figure 1

Cohort construction and evaluation partitions.

The flow shows adult ICU screening, acute brain injury (ABI) code candidates after documented special-background exclusions, accepted arterial-reference oxygen measurements, primary pulse oximetry–blood-gas pairs and available-history paired measurements. Current SpO_2 was the most recent valid

measurement at or before the blood-gas time proxy within 10 min, restricted to the first 72 h of each ICU stay. MIMIC used explicit arterial specimen documentation; eICU used arterial-named PaO_2 within the arterial blood-gas laboratory category with pressure units. Patient-level partitions kept all encounters together. MIMIC patients whose calendar-era intervals crossed 2020 contributed only to descriptive summaries. Partition entries show all paired patients and history-eligible measurements, whose patient denominators are 584, 135, 180 and 47 in training, calibration, later-era and boundary groups, respectively. Exclusion-branch patient counts can overlap and are not represented as sequential patient losses. Two eICU centre counts distinguish all primary pairs from the matched historical comparison.



P: prior information; B: current information; C: current plus prior information. Thresholds were not retuned externally.

Figure 2

Probability-classification increment and the error–coverage relationship.

(a) Pair-weighted three-class Brier scores for the prior-only model (P), current-information model (B) and current-plus-prior-information model (C), evaluated on identical history-eligible measurements: 423 measurements from 180 patients in the MIMIC later-era set and 4,877 measurements from 1,493 patients in eICU. The sum-of-three-class Brier score was not divided by 3; lower values indicate better probability predictions. Intervals are 95% percentile intervals from 1,000 patient-cluster bootstrap replicates. (b) Descriptive eICU false in-target fractions versus coverage over all 101 probability thresholds from 0 to 1. The false in-target fraction is the proportion below 80 or above 120 mmHg among selected measurements; coverage uses all 4,877 history-eligible measurements. Nonempty selections with fewer than 30 measurements or 20 patients are hollow and disconnected; their

denominators are printed. Zero-selection thresholds remain in the source table but cannot be plotted as an error fraction. The dotted line marks the prespecified 5% error criterion used as the deployment gate. This external curve was not used to select or revise thresholds. No model passed the gate in MIMIC calibration, so the locked selective policy released no in-target calls.

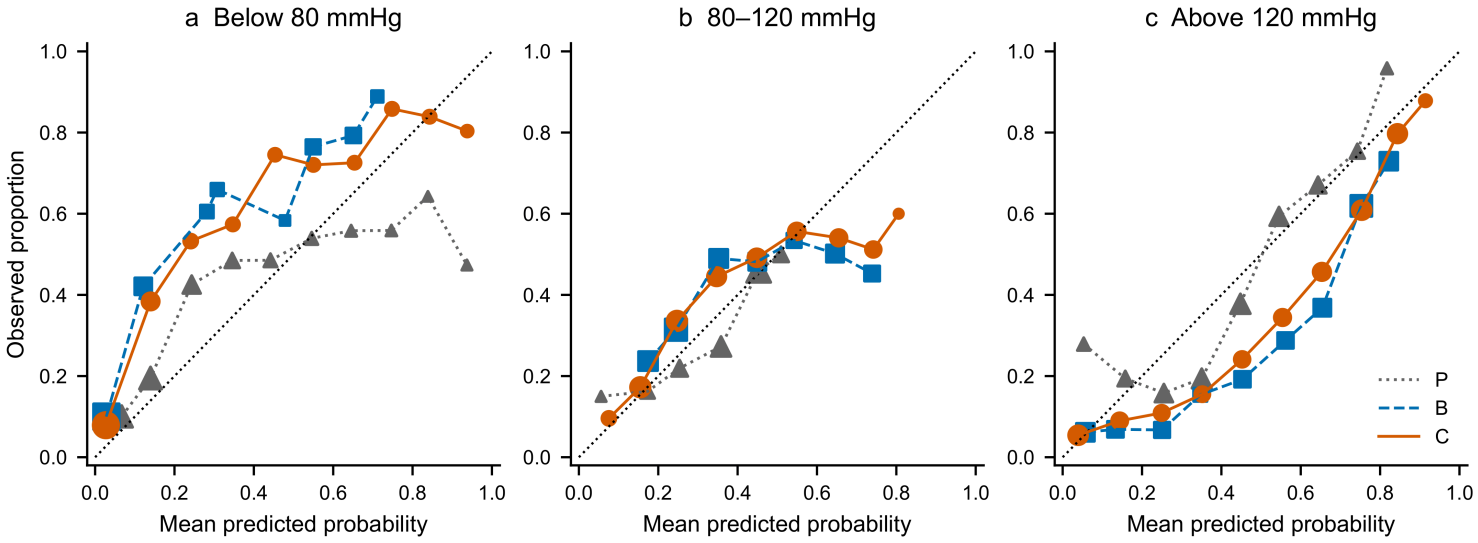


Figure 3

External calibration of the three oxygen-reference categories.

Calibration plots show observed versus mean predicted probabilities in fixed 0.1-wide probability bins for models P (prior information), B (current information) and C (current plus prior information) in 4,877 eICU history-eligible measurements from 1,493 patients. Categories are below 80, 80–120 inclusive and above 120 mmHg. The dotted diagonal denotes ideal calibration. Point size increases with the number of measurements in the bin; all occupied bins are retained. Lines connect ordered probability bins. Bin membership and patient counts are available in calibration_bins.csv. No external recalibration was applied.

Supplementary Files

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

- 09SupplementaryFigureS2.pdf
- 08SupplementaryFigureS1.pdf
- 03AdditionalFile1SupplementaryMethodsTables.docx
- 04AdditionalFile2Reproducibility.zip