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.

An EHR Data Passport for Renal-Risk AI: Cross-Database Transportability of Creatinine-Based AKI Prediction in Intensive Care

Xiaoxi Zhang
^1 Ansteel Group General Hospital
Shaonan Wang
Taizhou People's Hospital Affiliated to Nanjing Medical University
Giselle Chan
Cornell University
Lixin Yin
ylx3020129@gmail.com

Jiangyin People's Hospital

Research Article

Keywords: acute kidney injury, artificial intelligence, electronic health records, calibration, transportability, algorithmovigilance

Posted Date: July 1st, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10172948/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

Clinical deployment of acute kidney injury (AKI) AI models depends on whether labels, predictors, missingness, and calibration remain stable across electronic health record (EHR) systems.

Objective

To develop an AI-ready data passport and transportability stress-test framework for creatinine-based AKI prediction using MIMIC-IV v3.1 and eICU-CRD v2.0.

Methods

We performed a retrospective secondary analysis of adult ICU stays. The primary outcome was 7-day serum-creatinine-based operational AKI, requiring first-24h creatinine and post-24h follow-up creatinine through day 7. Predictors included age, sex, first-24h renal/electrolyte laboratories, missingness indicators, and first-24h creatinine count. We assessed endpoint computability and laboratory coverage, then trained L2-regularized logistic regression models in each database and externally validated them bidirectionally. Evaluation emphasized discrimination, Brier score, calibration, and measurement-density failure modes.

Results

The feasibility audit included 94,458 MIMIC-IV and 200,234 eICU adult ICU stays. First-24h creatinine coverage was 96.8% and 84.6%, respectively; 7-day creatinine-based AKI computability was 89.2% and 65.6%. Modeling cohorts included 84,237 MIMIC-IV and 131,429 eICU stays. Internal AUROC was 0.732 in MIMIC-IV and 0.723 in eICU. External AUROC was 0.703 for MIMIC-IV to eICU and 0.722 for eICU to MIMIC-IV. Calibration shifted directionally: MIMIC-IV to eICU overpredicted risk, whereas eICU to MIMIC-IV underpredicted higher-risk strata.

Conclusions

Creatinine-based AKI prediction is feasible across MIMIC-IV and eICU, but endpoint computability, laboratory coverage, measurement density, and calibration shift differ substantially. Renal-risk AI should include an EHR data passport and calibration-focused algorithmovigilance before transport or deployment claims.

1. Introduction

Acute kidney injury is a common and clinically consequential complication among critically ill patients. It is relevant to intensivists, anesthesiologists, hospitalists, and perioperative clinicians because renal dysfunction often reflects the combined effects of hemodynamic instability, systemic inflammation, nephrotoxic exposure, and baseline vulnerability. Predicting renal deterioration early enough to guide monitoring or intervention is therefore an attractive use case for clinical artificial intelligence. However, AKI prediction is not only a modeling problem. It is also a data construction problem: AKI labels depend on how often creatinine is measured, whether a baseline value is available, how urine output is recorded, and whether laboratory variables are comparable across EHR systems [1, 2].

Many clinical prediction studies emphasize discrimination, especially the area under the receiver operating characteristic curve (AUROC). AUROC is useful, but it does not establish whether a risk score can be transported across institutions. A model can preserve rank ordering while systematically overestimating or underestimating absolute risk in a new database. For clinical decision support, this distinction matters. Absolute risk estimates often determine alert thresholds, triage rules, resource allocation, and clinician trust. If a transported AKI model overpredicts risk, it may amplify alarm burden. If it underpredicts risk, it may miss patients who need closer surveillance. These risks are particularly relevant when models are developed on public ICU databases and then discussed as if they reflect general clinical reality [3, 4].

MIMIC-IV and the eICU Collaborative Research Database are two of the most widely used public ICU EHR datasets. Their availability has advanced reproducible clinical AI research, but they differ in data provenance, institutional structure, timestamp conventions, and laboratory measurement patterns [5–8]. These differences make them well suited for a transportability stress test. If an AKI model trained in one database performs differently in the other, the reasons may not lie only in patient physiology or algorithm choice. They may also reflect endpoint computability, laboratory coverage, measurement intensity, and missingness. EHR-based prediction studies face known challenges from incomplete data, variable documentation, and informatively collected measurements [19, 20]. AKI is a useful stress test for transportability because both the label and several predictors are coupled to observation practice.

This study treats renal-risk AI as a clinical systems problem. We developed an AI-ready data passport for creatinine-based AKI prediction and paired it with bidirectional cross-database validation between MIMIC-IV v3.1 and eICU-CRD v2.0. Our objective was not to produce the highest-performing AKI predictor. Instead, we aimed to show how endpoint construction, shared laboratory coverage, calibration, and failure-mode analysis can reveal whether a renal-risk model is ready to be transported across EHR systems. This framing aligns with algorithmovigilance: the systematic monitoring of where, why, and for whom an AI model may fail after moving beyond its development environment [9, 10]. It also connects to model cards and datasheets for datasets, which emphasize provenance, intended use, evaluation context, and limitations [21, 22].

2. Methods

2.1 Study Design and Data Sources

We performed a retrospective secondary analysis using locally stored PhysioNet releases of MIMIC-IV v3.1 and eICU-CRD v2.0. MIMIC-IV tables used in this analysis included $patients$, $admissions$, $diagnoses_icd$, $d_icd_diagnoses$, $d_labitems$, $labevents$, and $icustays$. eICU tables included $patient$, lab, $diagnosis$, and $hospital$. Both datasets are de-identified public research resources, and this study did not involve patient contact, intervention, or reidentification attempts. Dataset access and analysis followed the applicable data-use agreements [6, 8, 11].

The study was designed as a cross-database data-readiness and transportability analysis. We first quantified cohort eligibility, creatinine availability, operational AKI computability, and shared laboratory coverage. We then built a transparent baseline prediction model in each database and evaluated it internally and externally. The primary emphasis was calibration and transportability, not algorithmic novelty. The design was motivated by the broader dataset-shift problem in machine learning, where performance estimated in one source distribution may not hold in a different target distribution [23]. In clinical prediction modeling, external validation is specifically needed because healthcare settings differ in case mix, measurement practice, outcome definition, and treatment pathways [24, 25]. We therefore treated MIMIC-IV and eICU as two EHR systems with overlapping clinical content but different data-generating processes.

2.2 Cohort Construction

The source population comprised adult ICU stays. In MIMIC-IV, adult status was determined from $anch\ \text{or}\ _a \geq\ >\ =\ 18$, and ICU stay timing was derived from $icustays.\int ime$ and $icustays.\ ou\text{ime}$. In eICU, adult status was determined from $patient.\ a \geq$, with ages recorded as greater than 89 coded operationally as 90 for modeling. Stays were eligible for the feasibility audit if ICU-level identifiers and laboratory records were available. For modeling, stays also had to have a first-24h creatinine value and at least one post-24h creatinine value through day 7, allowing a leakage-controlled outcome definition.

Repeated stays were retained at the ICU-stay level in this working draft. This choice maximizes comparability with common ICU database analyses but should be revisited during S7 integrity review. If the final journal or reviewers require a patient-level analysis, a sensitivity analysis restricted to the first ICU stay per patient should be added.

2.3 Outcome Definition

The primary outcome was 7-day serum-creatinine-based operational AKI. For the modeling cohort, the baseline creatinine was defined as the lowest serum creatinine observed within the first 24 hours after ICU admission. AKI was defined as a post-24h creatinine value through day 7 that either reached at least 1.5 times the first-24h baseline or increased by at least 0.3 mg/dL from the first-24h baseline. This definition was chosen to separate the predictor window from the label window and to avoid using a value already present during feature construction as the outcome.

This endpoint is intentionally described as operational and creatinine-based. It is not full KDIGO AKI staging because the current local eICU table set did not include urine-output data. A 48-hour creatinine-rise definition was retained as a planned sensitivity endpoint, but the primary modeling results in this draft use the leakage-controlled 7-day operational endpoint.

2.4 Predictor Window and Shared Variables

The primary prediction window was the first 24 hours after ICU admission. The baseline model used a compact shared predictor set selected for clinical interpretability and cross-database availability: age, male sex indicator, latest first-24h creatinine, blood urea nitrogen (BUN), sodium, potassium, chloride, bicarbonate, anion gap, calcium, missingness indicators for these laboratory variables, and first-24h creatinine measurement count. Lower-coverage or workflow-sensitive variables, including lactate, magnesium, and phosphate, were retained for the data passport and sensitivity interpretation rather than the primary baseline model. MIMIC-IV laboratory mapping used item identifiers for the selected chemistry variables. eICU mapping used $labname$ values for the corresponding variables. Because database-specific laboratory names and measurement density can encode clinical workflow, harmonization was treated as part of the scientific object rather than a preprocessing detail.

2.5 AI-Ready Data Passport

We defined an AI-ready data passport as a structured set of pre-modeling and post-modeling checks that describe whether a database can support a transportable renal-risk model. The passport included six domains: cohort eligibility, endpoint computability, shared laboratory coverage, measurement density, missingness strata, and external calibration behavior. Each domain was summarized in manuscript-ready tables and figures. The passport was designed to make visible the data-generation context that would otherwise be hidden behind a single model-performance table. ### 2.6 Baseline Model and Validation Strategy

We trained L2-regularized logistic regression models using NumPy. Scikit-learn was not available in the active Python environment, and no package installation was performed for the modeling stage. Within each source database, the cohort was split into a 70% development set and 30% internal holdout set using a fixed random seed. One model was trained in MIMIC-IV and evaluated on the MIMIC-IV holdout set and all eligible eICU modeling stays. A second model was trained in eICU and evaluated on the eICU holdout set and all eligible MIMIC-IV modeling stays.

Continuous variables were median-imputed based on the development set and standardized using development-set means and standard deviations. Binary missingness indicators were included directly. The model was deliberately simple because the primary question concerned transportability and calibration, not optimization of predictive performance.

2.7 Evaluation Metrics and Failure-Mode Analysis

We evaluated discrimination with AUROC and AUPRC. We evaluated overall accuracy with Brier score and calibration with calibration intercept, calibration slope, decile-level observed versus predicted risk, and mean predicted risk. Failure-mode analyses stratified external predictions by first-24h creatinine measurement count and by the number of missing primary laboratory predictors. These strata were chosen because S3 feasibility results suggested that measurement density and missingness differed materially across databases.

The analysis followed the spirit of prediction-model reporting guidance by separating internal validation, external validation, calibration, and model-use limitations [3, 12]. No prospective validation, clinical deployment, alert simulation, or clinician-facing decision support evaluation was performed. We did not recalibrate the models in this stage because the purpose was to expose transportability failure modes before model updating [4, 24–26, 29].

3. Results

3.1 Cohort and Endpoint Computability

The feasibility audit included 94,458 adult ICU stays in MIMIC-IV and 200,234 in eICU (Table 1; Fig. 2). First-24h creatinine coverage was high in both datasets but higher in MIMIC-IV: 91,446 MIMIC-IV stays (96.8%) and 169,463 eICU stays (84.6%) had first-24h creatinine. A stricter 48-hour computability criterion requiring at least two creatinine measurements was met in 86,293 MIMIC-IV stays (91.4%) and 134,033 eICU stays (66.9%). The broader 7-day creatinine-based operational endpoint was computable in 84,237 MIMIC-IV stays (89.2%) and 131,429 eICU stays (65.6%).

Operational 7-day AKI rates differed among computable stays. In MIMIC-IV, 19,539 of 84,237 computable stays (23.2%) met the operational 7-day AKI definition in the feasibility analysis. In eICU, 20,825 of 131,429 computable stays (15.8%) met the corresponding definition. This difference should not be interpreted as a direct comparison of true AKI incidence. It reflects a mixture of population differences, creatinine measurement density, baseline handling, and database-specific documentation practices. For this reason, endpoint computability itself was treated as a data-readiness result.

3.2 Shared Laboratory Coverage

Shared early renal and electrolyte laboratory coverage supported the planned baseline model but revealed important cross-database gaps (Fig. 3; Table 3). In MIMIC-IV, first-24h coverage exceeded 96% for creatinine, BUN, sodium, potassium, chloride, bicarbonate, and anion gap. Calcium coverage was 90.0%, magnesium 93.8%, phosphate 90.7%, and lactate 53.7%. In eICU, first-24h coverage was 84.7% for creatinine, 84.4% for BUN, 84.7% for sodium, 85.4% for potassium, 84.2% for chloride, 79.7% for bicarbonate, and 67.6% for anion gap. Coverage was lower for magnesium (48.9%), phosphate (31.4%), and lactate (22.0%).

These findings shaped the primary predictor set. Core chemistry variables with stable coverage in both datasets were retained as primary predictors. Lactate, magnesium, and phosphate were interpreted as data-passport variables because their coverage differed enough to risk encoding measurement practice rather than patient physiology.

3.3 AI-Ready Data Passport

The data passport summarized each dataset's readiness for renal-risk AI (Table 2). Both datasets had large adult ICU cohorts and sufficient creatinine coverage for a creatinine-based operational AKI study. However, eICU had lower 7-day AKI computability and lower coverage for several clinically relevant laboratory variables. MIMIC-IV had denser early laboratory coverage and higher operational AKI computability. These differences provide a plausible mechanism for downstream transportability problems: a model may learn patterns that combine renal physiology with local measurement practices.

Figure 1 presents the study's conceptual framework. The EHR data source is not treated as a neutral input table. Instead, the model is embedded in a sequence that includes endpoint construction, data passport assessment, bidirectional external validation, and algorithmovigilance. This framework makes it possible to interpret calibration failure as a systems-level signal rather than only a modeling error.

3.4 Baseline Model Performance

The leakage-controlled modeling cohorts included 84,237 MIMIC-IV stays and 131,429 eICU stays. Event rates differed between cohorts: 22,623 MIMIC-IV modeling stays (26.9%) and 24,371 eICU modeling stays (18.5%) met the operational AKI label. Internal discrimination was moderate in both databases. The MIMIC-IV model achieved an internal holdout AUROC of 0.732 and AUPRC of 0.485. The eICU model achieved an internal holdout AUROC of 0.723 and AUPRC of 0.371 (Table 4; Fig. 4).

External discrimination showed only modest degradation. The MIMIC-IV-trained model achieved an AUROC of 0.703 and AUPRC of 0.343 when evaluated in eICU. The eICU-trained model achieved an AUROC of 0.722 and AUPRC of 0.482 when evaluated in MIMIC-IV. If AUROC alone were used, one might conclude that cross-database transportability was acceptable. The calibration results told a more cautious story. The two transport directions implied different implementation risks: potential over-alerting in eICU for the MIMIC-IV model and missed high-risk surveillance in MIMIC-IV for the eICU model. Direction-specific calibration error was therefore more actionable than a pooled external validation score.

3.5 Bidirectional Calibration Shift

Calibration shifted in opposite directions depending on the training and validation database. The MIMIC-IV-trained model overpredicted risk in eICU. Its mean predicted risk in eICU was 0.239 compared with an observed event rate of 0.185, with a calibration intercept of -0.538 and calibration slope of 0.830. Decile-

level analysis showed overprediction across all ten external risk deciles, with the largest error in the highest-risk decile, where mean predicted risk exceeded observed risk by 0.133 (Fig. 5; Table 5a).

The eICU-trained model underpredicted risk in MIMIC-IV. Its mean predicted risk in MIMIC-IV was 0.211 compared with an observed event rate of 0.269, with a calibration intercept of 0.684 and calibration slope of 1.261. Underprediction was most prominent in mid- and high-risk deciles. In decile 9, observed risk exceeded mean predicted risk by 0.180. These results support the view that rank discrimination can persist while clinically relevant absolute risk calibration fails.

3.6 Failure-Mode Strata

Measurement-density and missingness strata identified where transported risk estimates became less reliable (Table 5b; Fig. 5). For MIMIC-IV to eICU, the model overpredicted risk most strongly among eICU stays with at least five first-24h creatinine measurements. In this stratum, observed event rate was 0.293, mean predicted risk was 0.491, and prediction-observation error was + 0.198. For eICU to MIMIC-IV, the model underpredicted risk among MIMIC-IV stays with two or three to four first-24h creatinine measurements, with risk errors of -0.088 and − 0.104, respectively.

Missingness strata provided a second failure-mode signal. For MIMIC-IV to eICU, stays with one to two missing primary laboratory predictors had a risk error of + 0.090, compared with + 0.041 among stays with no missing primary labs. For eICU to MIMIC-IV, stays with one to two missing primary labs had a risk error of -0.217. Some high-missingness strata were small and should be interpreted cautiously, but the pattern supports including missingness profiles as part of the renal-risk AI data passport.

4. Discussion

4.1 Principal Findings

This cross-database study found that creatinine-based AKI prediction is feasible in both MIMIC-IV and eICU, but the conditions supporting model transportability differ substantially. MIMIC-IV had higher first-24h creatinine coverage, denser creatinine follow-up, and higher 7-day operational AKI computability. eICU provided a larger cohort but had lower endpoint computability and lower coverage for several laboratory variables. A transparent logistic baseline achieved moderate internal and external AUROC, yet external calibration shifted in clinically meaningful and direction-specific ways.

The key finding is not that one database is better than the other. The key finding is that renal-risk AI can be sensitive to the EHR measurement process. The MIMIC-IV-trained model tended to overestimate eICU risk, whereas the eICU-trained model tended to underestimate MIMIC-IV risk in higher-risk strata. Measurement density and missingness modified these errors. These patterns support an algorithmovigilance approach in which endpoint computability, laboratory coverage, calibration, and failure modes are audited before any clinical transport or deployment claim is made. The point is not that calibration failure is surprising; it is that calibration failure should be expected, measured, and explained. In routine care, creatinine is ordered in response to clinical concern, fluid shifts, drug exposure, sepsis, operative stress, and institutional monitoring pathways. Because these drivers are partly unmeasured in public EHR data, a transported model can inherit differences in observation intensity even when the visible predictor names appear harmonized.

4.2 Why Data Passporting Matters for AKI Prediction

AKI is a useful case study for clinical AI data readiness because its label is partly dependent on measurement. A patient cannot meet a creatinine-based endpoint unless creatinine is measured, and the probability of detection increases when measurement is more frequent. Therefore, AKI modeling is vulnerable to observation-intensity bias. A model may learn that frequent creatinine testing predicts AKI, but frequent testing may itself reflect clinician concern, ICU workflow, or institutional practice. When transported across databases, such learned patterns can produce calibration errors.

The AI-ready data passport proposed here makes this problem explicit. Instead of starting with model performance, the passport starts with cohort eligibility, endpoint computability, laboratory coverage, measurement density, and missingness. These elements are not administrative details. They are part of the evidence chain needed to judge whether an EHR model can move across systems. In this study, the data passport helped explain why external AUROC remained moderate while absolute risk calibration shifted.

4.3 Implications for Clinical AI Systems

Clinical AI systems are often evaluated in development settings where labels and predictors are defined under local data conditions. Deployment requires a different standard. A renal-risk model used for alerting, triage, or surveillance must estimate risk well enough in the target setting, not merely rank patients acceptably in a source dataset. The present results suggest that external calibration tables, calibration deciles, and measurement-density failure-mode analyses should accompany cross-database validation. Similar concerns have motivated work on calibration drift in AKI models, national EHR-derived AKI risk stratification, and prospective evaluation of AKI prediction tools [29–31]. Recent multi-center AKI machine-learning studies also show that strong external discrimination can be achieved under carefully engineered pipelines, but these studies still require attention to endpoint definition, validation source, and target-site calibration before clinical use [32].

For hospital AI governance, this implies that implementation teams should ask a different sequence of questions. Before asking whether an AKI model has a high AUROC, they should ask whether AKI labels are computable in the target EHR, whether key renal/electrolyte predictors are measured with comparable frequency, whether missingness profiles differ from the development setting, and whether calibration remains acceptable after transport. ### 4.4 Relation to Existing Reporting Guidance

Prediction-model reporting guidance emphasizes transparent development, validation, and reporting of model performance [3]. Observational reporting guidance emphasizes cohort construction and measurement transparency [12]. Clinical AI guidance increasingly stresses evaluation beyond model

development [13]. The present study complements these approaches by operationalizing a data-passport layer for renal-risk AI. It does not replace formal reporting checklists. It provides an analysis structure that helps populate them with cross-database evidence.

4.5 Strengths

The study has several strengths. First, it uses two large public ICU databases, enabling reproducibility and cross-database stress testing. Second, it emphasizes a clinically interpretable endpoint and predictor set rather than algorithmic complexity. Third, it evaluates bidirectional transportability, avoiding the assumption that one database is always the development source and the other always the validation target. Fourth, it includes calibration deciles and failure-mode strata, which provide more actionable information than AUROC alone. Fifth, it preserves a transparent modeling strategy. By avoiding an opaque high-capacity model in the primary analysis, the study keeps attention on measurement practice, missingness, and calibration. This choice should make the work more accessible to clinicians, informaticians, and implementation teams who need to audit model behavior rather than simply rank algorithms.

4.6 Limitations

The main limitation is outcome definition. The endpoint is serum-creatinine-based operational AKI, not complete KDIGO AKI staging, because urine-output data were not available in the current local eICU table set. Baseline creatinine was derived operationally from the first 24 hours rather than from pre-illness outpatient values. This choice improves leakage control but may misclassify patients who already had renal dysfunction at ICU admission.

Second, this study is retrospective and uses public ICU databases. It cannot establish prospective clinical utility, alert safety, or clinician response. Third, the baseline model is intentionally simple. This supports interpretability but does not test whether more flexible algorithms could improve calibration after transport. Fourth, race/ethnicity, diagnosis groups, medication exposure, urine output, and renal replacement therapy were not included in the primary model. These variables may be clinically relevant but require careful harmonization before use. Fifth, some missingness strata were small, so subgroup failure-mode results should be interpreted as exploratory. Sixth, this draft does not yet include a formal first-stay-only sensitivity analysis, site-level eICU heterogeneity analysis, or recalibration experiment. These additions would strengthen the paper but would also shift the scope from data-passport and transportability stress testing toward model updating. Seventh, although the study cites clinical AI governance guidance, it does not test clinician interaction, alert fatigue, human factors, or workflow consequences. Those questions require prospective or simulation-based designs [16–18, 35, 36].

4.7 Future Work

Future work should extend the data passport to include urine output, renal replacement therapy, medication exposures, and site-level clinical workflow variables where available. Recalibration strategies should be evaluated after transport, including intercept adjustment, slope adjustment, and target-site recalibration. Additional modeling approaches, such as gradient boosting, may be useful as benchmarks, but algorithmic complexity should not replace data-readiness assessment. Ultimately, prospective evaluation is needed before renal-risk AI can be considered clinically deployable. Future versions of the passport could also include equity and subgroup monitoring domains, especially because measurement patterns may differ across demographic groups, hospitals, and care pathways [37, 38]. Finally, the framework could be tested beyond AKI, including perioperative hypotension, postoperative pulmonary complications, sepsis-associated organ dysfunction, and medication-related adverse events. These conditions share a common feature: the model target is shaped by both patient physiology and the way the hospital chooses to observe that physiology.

5. Conclusion

Creatinine-based AKI prediction across MIMIC-IV and eICU is feasible, but transportability depends on more than discrimination. Endpoint computability, laboratory coverage, measurement density, missingness, and calibration shift all shaped cross-database performance. A renal-risk AI system should therefore be accompanied by an EHR data passport and calibration-focused algorithmovigilance before clinical deployment claims are made. The practical message is deliberately modest: before a hospital asks which AKI model is most accurate, it should ask whether its EHR can reproduce the label, predictor window, missingness structure, and calibration assumptions required by that model. For renal-risk AI, that question is not peripheral to modeling; it is part of model validity.

Declarations

Ethics approval and consent to participate: This retrospective secondary analysis used de-identified public research databases from PhysioNet under credentialed data-use agreements. MIMIC-IV and eICU-CRD were de-identified before public research distribution and were made available by their respective data custodians under approved data-use procedures. All analyses in the present study used de-identified data only and involved no patient contact, intervention assignment, prospective recruitment, or attempt at reidentification. Therefore, additional informed consent and new institutional ethics approval were not required for this secondary analysis of de-identified public data. The authors will provide institutional confirmation if requested by the journal.

Consent for publication: Not applicable.

Availability of data and materials: MIMIC-IV v3.1 and eICU-CRD v2.0 are available from PhysioNet to credentialed users under the applicable data-use agreements. Derived analysis outputs, figure source data, and scripts are stored in the local workflow directory and should be deposited or shared according to journal and data-use requirements before submission.

Competing interests: The authors declare that they have no competing interests.

Funding: No specific funding was received for this work.

Authors' contributions: XZ and SW contributed equally to this work. XZ contributed to data curation, reproducible analysis design, statistical verification, manuscript drafting, and critical revision. SW contributed to study conception, clinical interpretation, manuscript drafting, and critical revision. GC contributed to background review, clinical and biological interpretation, manuscript editing, and critical revision. LY supervised the project, contributed to study design, medical informatics framing, manuscript revision, and correspondence. All authors read and approved the final manuscript.

Acknowledgements: Not applicable.

AI-assisted workflow disclosure: The manuscript preparation workflow used AI-assisted drafting, organization, citation-checking support, and formatting support under author direction. No large-model API was used for data computation or statistical analysis. All analyses were performed locally using de-identified public datasets. The author(s) are responsible for the accuracy, integrity, and final content of the manuscript.

References

1. Kidney Disease: Improving Global Outcomes (KDIGO) Acute Kidney Injury Work Group (2012) KDIGO clinical practice guideline for acute kidney injury. Kidney Int Suppl 2:1–138. https://kdigo.org/wp-content/uploads/2016/10/KDIGO-2012-AKI-Guideline-English.pdf
2. Hoste EAJ, Bagshaw SM, Bellomo R et al (2015) Epidemiology of acute kidney injury in critically ill patients: the multinational AKI-EPI study. Intensive Care Med 41:1411–1423. https://doi.org/10.1007/s00134-015-3934-7
3. Logullo P, Collins GS, Moons KGM et al (2024) TRIPOD + AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ 385:e078378. https://doi.org/10.1136/bmj-2023-078378
4. Van Calster B, McLernon DJ, van Smeden M et al (2019) Calibration: the Achilles heel of predictive analytics. BMC Med 17:230. https://doi.org/10.1186/s12916-019-1466-7
5. Johnson AEW, Bulgarelli L, Shen L et al (2023) MIMIC-IV, a freely accessible electronic health record dataset. Sci Data 10:1. https://doi.org/10.1038/s41597-022-01899-x
6. Johnson A, Bulgarelli L, Pollard T, Gow B, Moody B, Horng S, Celi LA, Mark R (2024) MIMIC-IV (version 3.1). PhysioNet. RRID:SCR_007345. https://doi.org/10.13026/kpb9-mt58
7. Pollard TJ, Johnson AEW, Raffa JD, Celi LA, Mark RG, Badawi O (2018) The eICU Collaborative Research Database, a freely available multi-center database for critical care research. Sci Data 5:180178. https://doi.org/10.1038/sdata.2018.178
8. Pollard T, Johnson A, Raffa J, Celi LA, Badawi O, Mark R (2019) eICU Collaborative Research Database (version 2.0). PhysioNet. RRID:SCR_007345. https://doi.org/10.13026/C2WM1R
9. Embi PJ (2021) Algorithmovigilance: advancing methods to analyze and monitor artificial intelligence-driven health care for effectiveness and equity. JAMA Netw Open 4:e214622. https://doi.org/10.1001/jamanetworkopen.2021.4622
10. Bedoya AD, Economou-Zavlanos NJ, Goldstein BA et al (2022) A framework for the oversight and local deployment of safe and high-quality prediction models. J Am Med Inform Assoc 29:1631–1636. https://doi.org/10.1093/jamia/ocac078
11. Goldberger AL, Amaral LAN, Glass L et al (2000) PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. Circulation 101:e215-e220. https://doi.org/10.1161/01.CIR.101.23.e215
12. von Elm E, Altman DG, Egger M et al (2007) The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. PLoS Med 4:e296. https://doi.org/10.1371/journal.pmed.0040296
13. Vasey B, Nagendran M, Campbell B et al (2022) Reporting guideline for the early-stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. Nat Med 28:924–933. https://doi.org/10.1038/s41591-022-01772-9
14. Feng Y, Wang AY, Jun M et al (2023) Characterization of risk prediction models for acute kidney injury: a systematic review and meta-analysis. JAMA Netw Open 6:e2313359. https://doi.org/10.1001/jamanetworkopen.2023.13359
15. Vagliano I, Chesnaye NC, Leopold JH, Jager KJ, Abu-Hanna A, Schut MC (2022) Machine learning models for predicting acute kidney injury: a systematic review and critical appraisal. Clin Kidney J 15:2266–2280. https://doi.org/10.1093/ckj/sfac181
16. Wilson FP, Martin M, Yamamoto Y et al (2021) Electronic health record alerts for acute kidney injury: multicenter, randomized clinical trial. BMJ 372:m4786. https://doi.org/10.1136/bmj.m4786
17. Wilson FP, Shashaty M, Testani J et al (2015) Automated, electronic alerts for acute kidney injury: a single-blind, parallel-group, randomised controlled trial. Lancet 385:1966–1974. https://doi.org/10.1016/S0140-6736(15)60266-5
18. Fu Z, Hao X, Lv Y et al (2024) Effect of electronic alerts on the care and outcomes in patients with acute kidney injury: a meta-analysis and trial sequential analysis. BMC Med 22:408. https://doi.org/10.1186/s12916-024-03639-x
19. Sun M, Engelhard MM, Bedoya AD et al (2024) Incorporating informatively collected laboratory data from EHR in clinical prediction models. BMC Med Inform Decis Mak 24:206. https://doi.org/10.1186/s12911-024-02612-1
20. Goldstein BA, Navar AM, Pencina MJ, Ioannidis JPA (2017) Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic review. J Am Med Inform Assoc 24:198–208. https://doi.org/10.1093/jamia/ocw042
21. Mitchell M, Wu S, Zaldivar A et al (2019) Model cards for model reporting. Proceedings of the Conference on Fairness, Accountability, and Transparency:220–229. https://doi.org/10.1145/3287560.3287596
22. Gebru T, Morgenstern J, Vecchione B et al (2021) Datasheets for datasets. Commun ACM 64:86–92. https://doi.org/10.1145/3458723
23. Quiñonero-Candela J, Sugiyama M, Schwaighofer A, Lawrence ND, eds (2009) Dataset Shift in Machine Learning. MIT Press, Cambridge

24. Riley RD, Archer L, Snell KIE et al (2024) Evaluation of clinical prediction models (part 1): from development to external validation. BMJ 384:e074819. https://doi.org/10.1136/bmj-2023-074819
25. Riley RD, Archer L, Snell KIE et al (2024) Evaluation of clinical prediction models (part 2): how to undertake an external validation study. BMJ 384:e074820. https://doi.org/10.1136/bmj-2023-074820
26. Fehr J, Piccininni M, Kurth T et al (2023) Assessing the transportability of clinical prediction models for cognitive impairment using causal models. BMC Med Res Methodol 23:187. https://doi.org/10.1186/s12874-023-02003-6
27. Wolff RF, Moons KGM, Riley RD et al (2019) PROBAST: a tool to assess the risk of bias and applicability of prediction model studies. Ann Intern Med 170:51–58. https://doi.org/10.7326/M18-1376
28. Moons KGM, de Groot JAH, Bouwmeester W et al (2014) Critical appraisal and data extraction for systematic reviews of prediction modelling studies: the CHARMS checklist. PLoS Med 11:e1001744. https://doi.org/10.1371/journal.pmed.1001744
29. Davis SE, Lasko TA, Chen G, Siew ED, Matheny ME (2017) Calibration drift in regression and machine learning models for acute kidney injury. J Am Med Inform Assoc 24:1052–1061. https://doi.org/10.1093/jamia/ocx030
30. Cronin RM, VanHouten JP, Siew ED et al (2015) National Veterans Health Administration inpatient risk stratification models for hospital-acquired acute kidney injury. J Am Med Inform Assoc 22:1054–1071. https://doi.org/10.1093/jamia/ocv051
31. Flechet M, Falini S, Bonetti C et al (2019) Machine learning versus physicians' prediction of acute kidney injury in critically ill adults: a prospective evaluation of the AKIpredictor. Crit Care 23:282. https://doi.org/10.1186/s13054-019-2563-x
32. Zappalà S, Alfieri F, Ancona A et al (2024) Development and external validation of a machine learning model for the prediction of persistent acute kidney injury stage 3 in multi-centric, multi-national intensive care cohorts. Crit Care 28:189. https://doi.org/10.1186/s13054-024-04954-8
33. Davis SE, Embí PJ, Matheny ME (2024) Sustainable deployment of clinical prediction tools: a 360 degree approach to model maintenance. J Am Med Inform Assoc 31:1195–1198. https://doi.org/10.1093/jamia/ocae036
34. Salwei ME, Davis SE, Reale C et al (2025) Human-centered design of an artificial intelligence monitoring system: the Vanderbilt Algorithmovigilance Monitoring and Operations System. JAMIA Open 8:ooaf136. https://doi.org/10.1093/jamiaopen/ooaf136
35. Liu X, Rivera SC, Moher D et al (2020) Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. Nat Med 26:1364–1374. https://doi.org/10.1038/s41591-020-1034-x
36. Rivera SC, Liu X, Chan AW et al (2020) Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension. Nat Med 26:1351–1363. https://doi.org/10.1038/s41591-020-1037-7
37. World Health Organization (2021) Ethics and governance of artificial intelligence for health: WHO guidance. World Health Organization, Geneva. https://www.who.int/publications/i/item/9789240029200
38. Obermeyer Z, Powers B, Vogeli C, Mullainathan S (2019) Dissecting racial bias in an algorithm used to manage the health of populations. Science 366:447–453. https://doi.org/10.1126/science.aax2342
39. Pearl J, Bareinboim E (2011) Transportability of causal and statistical relations: a formal approach. Proceedings of the AAAI Conference on Artificial Intelligence 25:247–254. https://doi.org/10.1609/aaai.v25i1.7861
40. Collins GS, Reitsma JB, Altman DG, Moons KGM (2015) Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. Ann Intern Med 162:55–63. https://doi.org/10.7326/M14-0697

Tables

Table 1. Cohort and endpoint computability in MIMIC-IV and eICU-CRD.

dataset	adult_icu_stays	mortality_events	mortality_rate	stays_with_24h_creatinine	stays_with_2plus_48h_creatinine	stays_with_7d_aki_c
MIMIC-IV	94,458	11,350	0.120	91,446	86,293	84,237
eICU-CRD	200,234	18,204	0.091	169,463	134,033	131,429

Table 2. AI-ready data passport for cross-database renal-risk AI transportability.

domain	MIMIC-IV	eICU-CRD	readiness	MIMIC-IV->eICU	eICU->MIMIC-IV
Cohort eligibility	94,458 adult ICU stays	200,234 adult ICU stays	green		
24h creatinine coverage	96.8%	84.6%	yellow		
7d AKI computability	89.2%	65.6%	yellow		
Core renal/electrolyte labs	>90% for core labs	mostly 79.7-85.4%	yellow		
Lactate/magnesium/phosphate	moderate-high	lower coverage	red/yellow		
External calibration			red/yellow	overprediction	underprediction

Table 3. Primary predictor dictionary and cross-database laboratory mapping.

lab_group	mimic_itemid	eicu_labname	included_for_s3
creatinine	50,912	creatinine	True
bun	51,006	BUN	True
sodium	50,983	sodium	True
potassium	50,971	potassium	True
chloride	50,902	chloride	True
bicarbonate	50,882	bicarbonate	True
anion_gap	50,868	anion gap	True
calcium	50,893	calcium	True
magnesium	50,960	magnesium	True
phosphate	50,970	phosphate	True
lactate	50,813	lactate	True

Table 4. Internal and external model performance and calibration.

run	train_dataset	eval_dataset	n	events	event_rate	mean_predicted_risk	auroc	auprc	brier	calibration_inte
internal_holdout	MIMIC-IV	MIMIC-IV	25,272	6,725	0.266	0.268	0.732	0.485	0.171	0.060
external	MIMIC-IV	eICU-CRD	131,429	24,371	0.185	0.239	0.703	0.343	0.144	-0.538
internal_holdout	eICU-CRD	eICU-CRD	39,429	7,361	0.187	0.185	0.723	0.371	0.139	0.232
external	eICU-CRD	MIMIC-IV	84,237	22,623	0.269	0.211	0.722	0.482	0.181	0.684

Table 5. External calibration and failure-mode sensitivity analyses. Table 5a reports external calibration deciles, and Table 5b reports sensitivity strata by creatinine measurement density and missingness.

Table 5a. External calibration deciles for bidirectional validation.

run	train_dataset	eval_dataset	decile	n	events	observed_rate	mean_predicted_risk	calibration_error
external	MIMIC-IV	eICU-CRD	1	13,143	975	0.074	0.100	0.026
external	MIMIC-IV	eICU-CRD	2	13,143	1,080	0.082	0.130	0.048
external	MIMIC-IV	eICU-CRD	3	13,143	1,214	0.092	0.149	0.057
external	MIMIC-IV	eICU-CRD	4	13,143	1,456	0.111	0.167	0.056
external	MIMIC-IV	eICU-CRD	5	13,143	1,720	0.131	0.186	0.055
external	MIMIC-IV	eICU-CRD	6	13,142	2,090	0.159	0.208	0.049
external	MIMIC-IV	eICU-CRD	7	13,143	2,498	0.190	0.237	0.047
external	MIMIC-IV	eICU-CRD	8	13,143	3,151	0.240	0.279	0.039
external	MIMIC-IV	eICU-CRD	9	13,143	4,299	0.327	0.354	0.027
external	MIMIC-IV	eICU-CRD	10	13,143	5,888	0.448	0.581	0.133
external	eICU-CRD	MIMIC-IV	1	8,424	724	0.086	0.110	0.024
external	eICU-CRD	MIMIC-IV	2	8,424	938	0.111	0.132	0.021
external	eICU-CRD	MIMIC-IV	3	8,423	1,196	0.142	0.146	0.004
external	eICU-CRD	MIMIC-IV	4	8,424	1,560	0.185	0.158	-0.027
external	eICU-CRD	MIMIC-IV	5	8,424	1,711	0.203	0.171	-0.032
external	eICU-CRD	MIMIC-IV	6	8,423	2,022	0.240	0.186	-0.054
external	eICU-CRD	MIMIC-IV	7	8,424	2,383	0.283	0.205	-0.078
external	eICU-CRD	MIMIC-IV	8	8,423	2,959	0.351	0.233	-0.118
external	eICU-CRD	MIMIC-IV	9	8,424	3,940	0.468	0.288	-0.180
external	eICU-CRD	MIMIC-IV	10	8,424	5,190	0.616	0.485	-0.131

Table 5b. Failure-mode sensitivity by measurement density and missingness strata.

run	train_dataset	eval_dataset	stratifier	stratum	n	events	event_rate	mean_predicted_risk	risk_error
external	MIMIC-IV	eICU-CRD	creatinine_count_stratum	1	82,463	11,891	0.144	0.199	0.055
external	MIMIC-IV	eICU-CRD	creatinine_count_stratum	2	31,441	7,220	0.230	0.261	0.032
external	MIMIC-IV	eICU-CRD	creatinine_count_stratum	3-4	13,316	4,027	0.302	0.355	0.052
external	MIMIC-IV	eICU-CRD	creatinine_count_stratum	>=5	4,209	1,233	0.293	0.491	0.198
external	MIMIC-IV	eICU-CRD	missing_count_stratum	0.000	96,178	17,857	0.186	0.227	0.041
external	MIMIC-IV	eICU-CRD	missing_count_stratum	1-2	34,660	6,425	0.185	0.276	0.090
external	MIMIC-IV	eICU-CRD	missing_count_stratum	3-4	26	6	0.231	0.305	0.074
external	MIMIC-IV	eICU-CRD	missing_count_stratum	>=5	565	83	0.147	0.089	-0.058
external	eICU-CRD	MIMIC-IV	creatinine_count_stratum	1	27,016	3,893	0.144	0.161	0.017
external	eICU-CRD	MIMIC-IV	creatinine_count_stratum	2	35,178	10,093	0.287	0.199	-0.088
external	eICU-CRD	MIMIC-IV	creatinine_count_stratum	3-4	18,810	7,137	0.379	0.275	-0.104
external	eICU-CRD	MIMIC-IV	creatinine_count_stratum	>=5	3,233	1,500	0.464	0.398	-0.066
external	eICU-CRD	MIMIC-IV	missing_count_stratum	0.000	77,937	20,208	0.259	0.215	-0.044
external	eICU-CRD	MIMIC-IV	missing_count_stratum	1-2	6,227	2,393	0.384	0.168	-0.217
external	eICU-CRD	MIMIC-IV	missing_count_stratum	3-4	58	20	0.345	0.145	-0.200
external	eICU-CRD	MIMIC-IV	missing_count_stratum	>=5	15	2	0.133	0.152	0.019

Figures

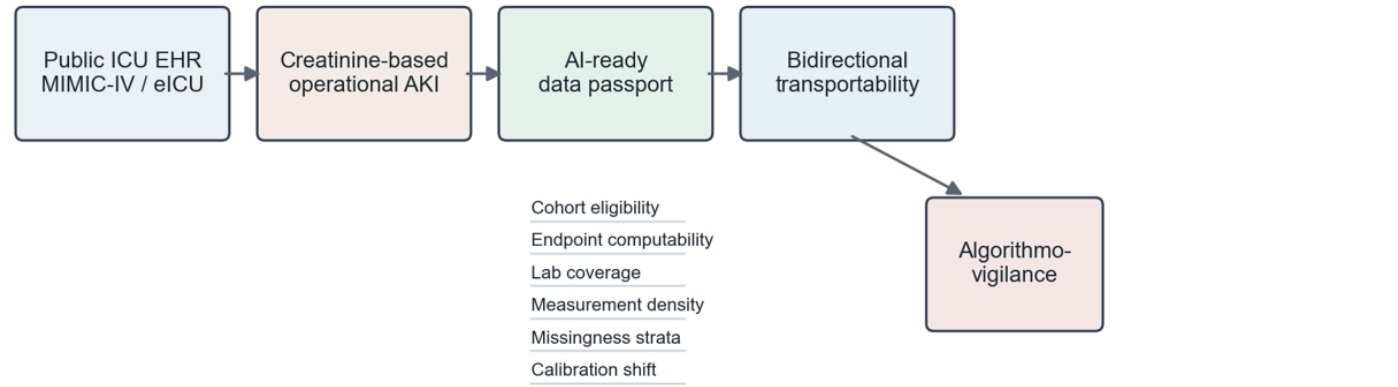


Figure 1

Data-passport and algorithmovigilance framework for renal-risk AI. The figure shows the study logic from public ICU EHR data through creatinine-based operational AKI construction, AI-ready data passport assessment, bidirectional transportability testing, and algorithmovigilance. Passport domains include cohort eligibility, endpoint computability, laboratory coverage, measurement density, missingness strata, and calibration shift.

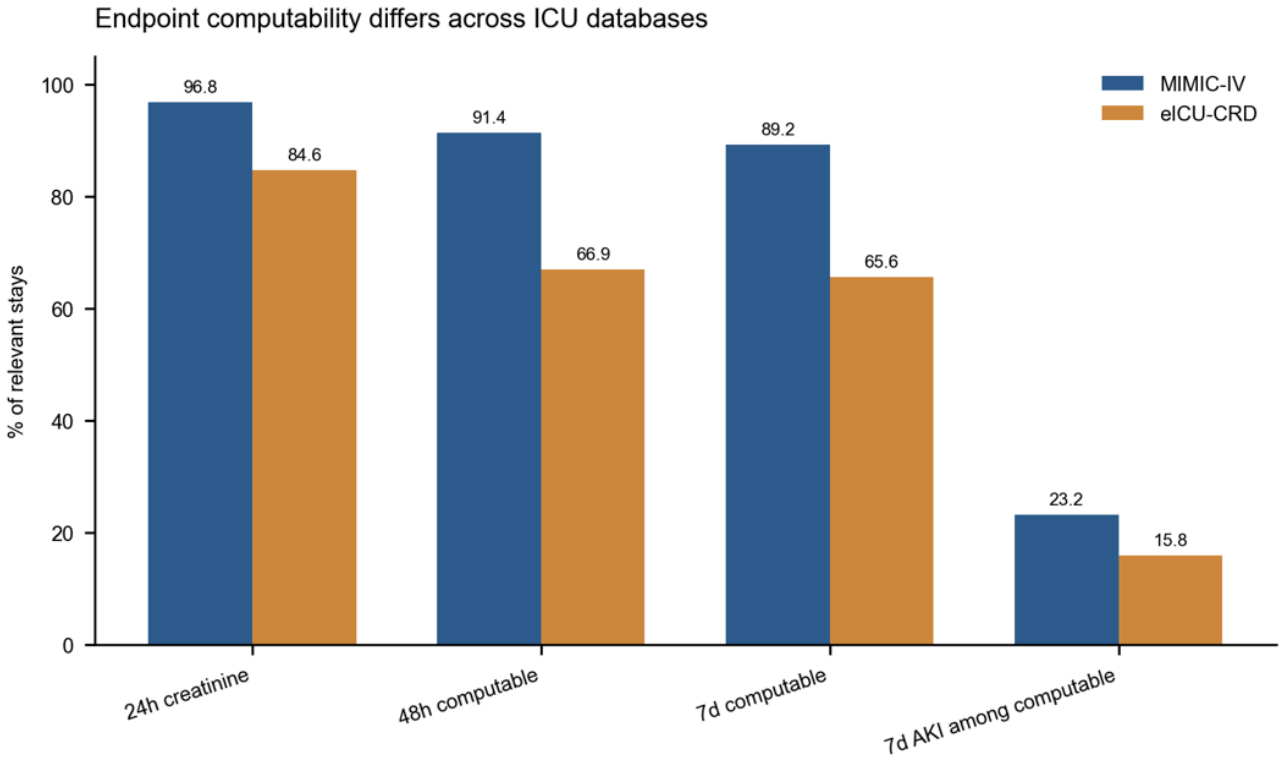


Figure 2

Endpoint computability differs across MIMIC-IV and eICU. Bars show first-24h creatinine coverage, 48-hour computability, 7-day computability, and operational 7-day AKI among computable stays. MIMIC-IV had higher creatinine coverage and endpoint computability than eICU.

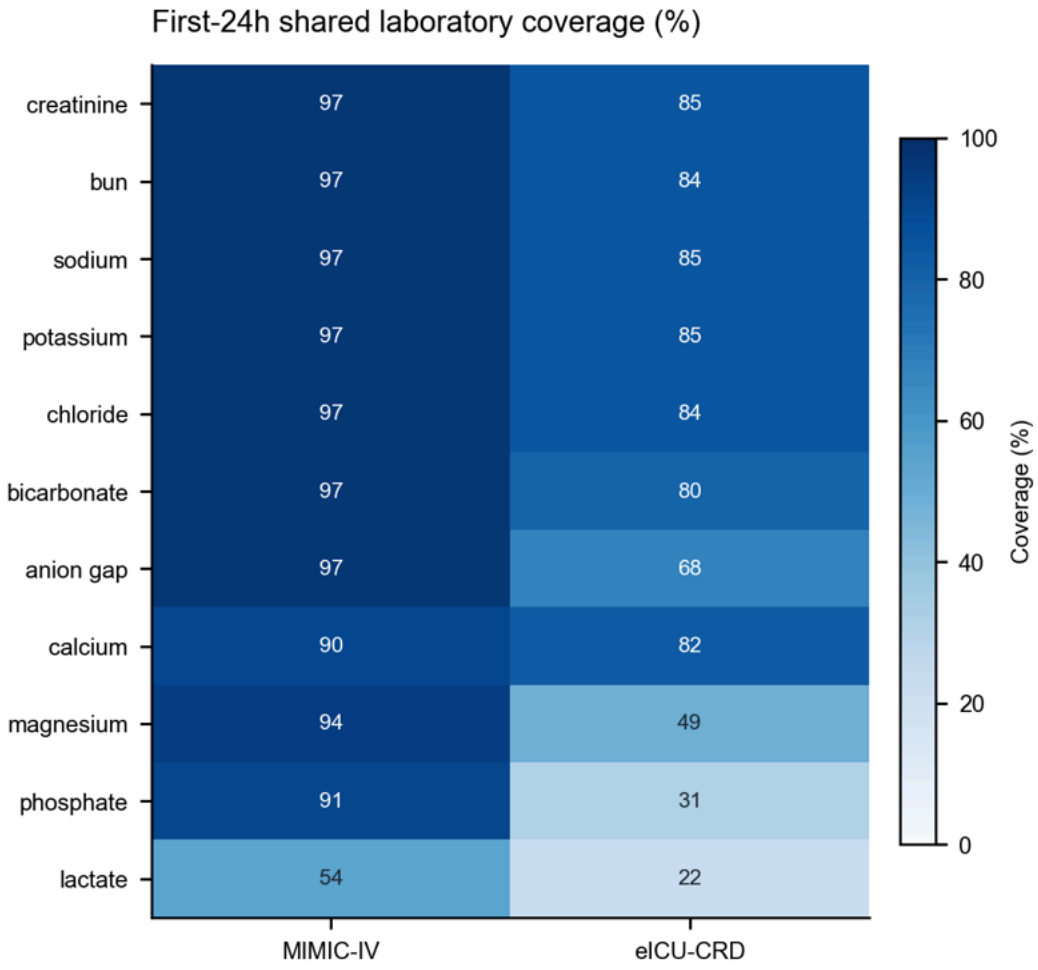


Figure 3

First-24h shared laboratory coverage. Heatmap shows the percentage of adult ICU stays with each shared renal or electrolyte laboratory variable measured in the first 24 hours. Core chemistry variables were broadly available, whereas lactate, magnesium, and phosphate showed larger cross-database coverage gaps.

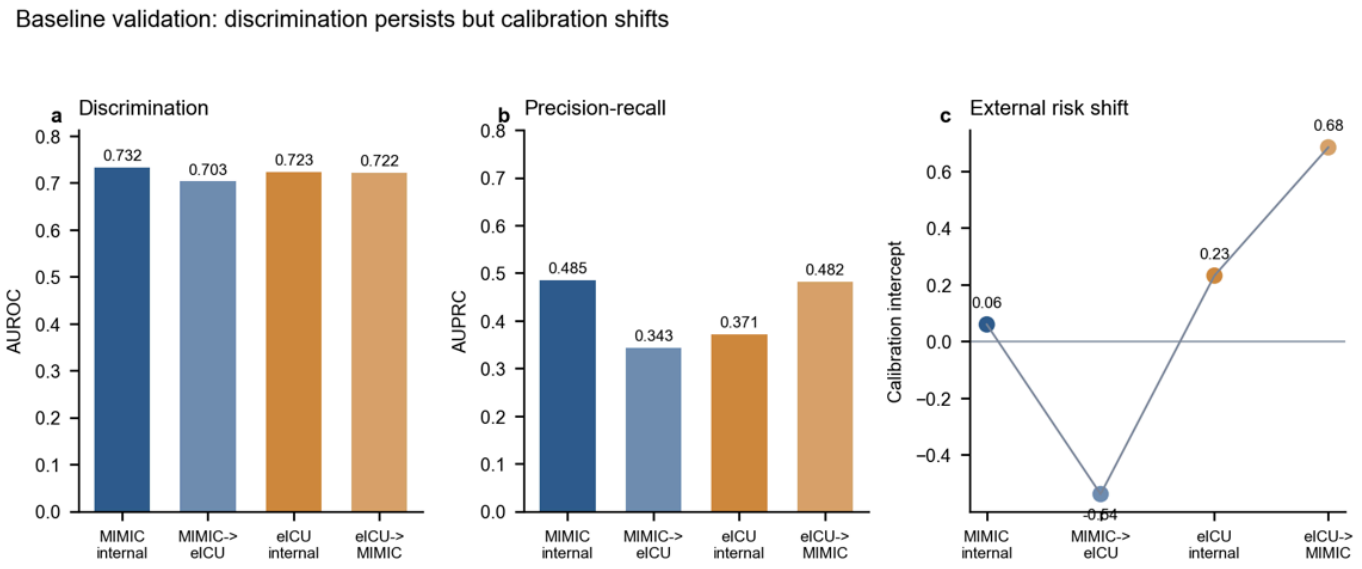


Figure 4

Baseline bidirectional validation. Panels show AUROC, AUPRC, and calibration intercept for internal and external validation. External AUROC remained moderate, but calibration shifted when models were transported across databases.

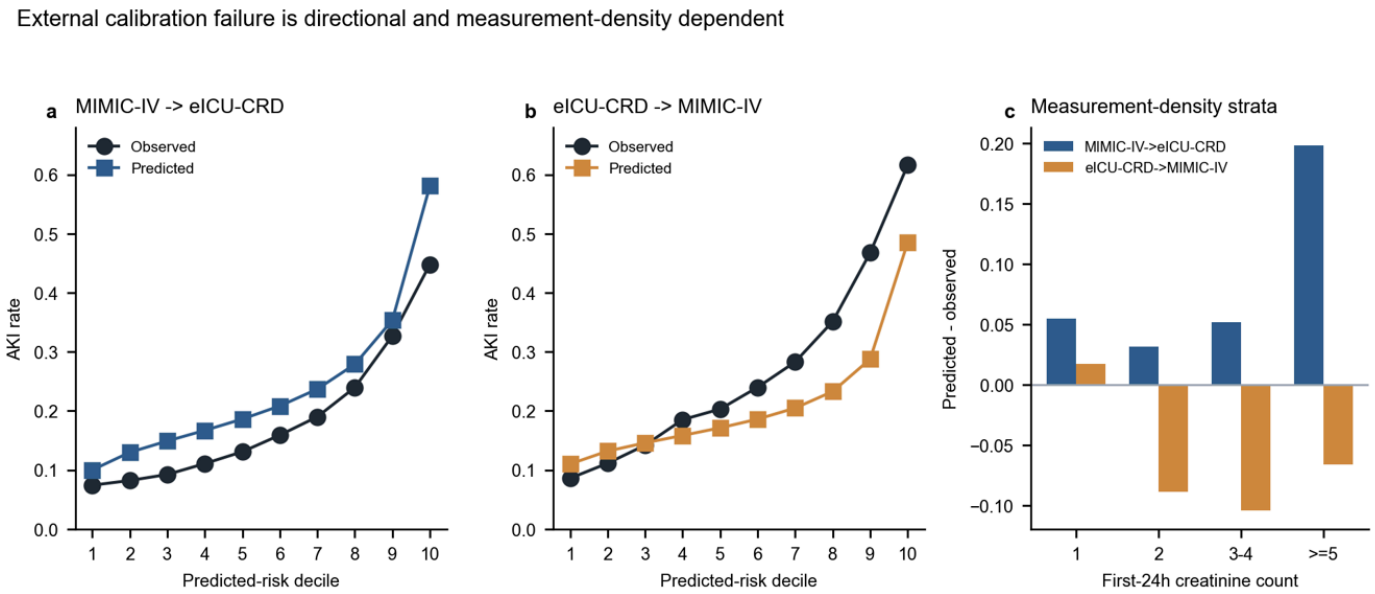


Figure 5

External calibration failure and measurement-density strata. Panels show external calibration deciles for MIMIC-IV to eICU and eICU to MIMIC-IV, plus prediction-observation error stratified by first-24h creatinine measurement count. The transported models showed direction-specific calibration failure and measurement-density-dependent risk error.