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.

Claim-specific evidence for reusing data-derived patient labels: a reporting survey and cross-database evaluation in sepsis

Tianyi Wang
First Affiliated Hospital of Soochow University
Yukun Zhang
First Affiliated Hospital of Soochow University
Shouhua Zang
First Affiliated Hospital of Soochow University
Xiaomin Wu
First Affiliated Hospital of Soochow University
Chunlei Fan
First Affiliated Hospital of Soochow University
Jin Xie
First Affiliated Hospital of Soochow University
Jun Chen
chenjun@suda.edu.cn

First Affiliated Hospital of Soochow University

Research Article

Keywords: Electronic Health Records, Cluster Analysis, Sepsis, Phenotype, Reproducibility of Results

Posted Date: November 22nd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11438034/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

Patient labels derived from electronic health record data, including sepsis partitions, are stored and reused across studies and sites for risk stratification, assignment of new patients and treatment comparisons. Which stored objects and comparisons each use requires is seldom specified. We aimed to define the objects and comparisons needed to evaluate claims about reused patient labels and to assess how frequently they are reported.

Methods

We specified requirements for composition, discreteness, incremental information, external application and treatment comparison. We examined 54 sampled sepsis-partition reports and a K = 2 partition (K2) of 20,049 patients in the Medical Information Mart for Intensive Care (MIMIC)-IV database, assigned to 17,465 stays across 199 hospitals in the eICU Collaborative Research Database.

Results

K2 mainly recoded renal and acid-base physiology. It increased the mortality area under the curve by 0.050 over a limited clinical model but by < 0.001 over a model containing its source features. In post hoc external analyses, a fixed source-median imputation rule rendered the recovered assignment rule executable for new patients and reproduced 98.9% of the original MIMIC-IV labels; 9 of 199 hospitals lay outside 99.8% prevalence limits, and the mortality model with K2 had calibration slope 0.780. Intercept-and-slope updating in separate hospitals lowered its Brier score from 0.1219 to 0.1182 (difference −0.0037, 95% CI -0.0055 to -0.0022); the model without K2 improved as well. Of the 54 reports surveyed, six evaluated incremental prediction, none against a comparator containing the clustering inputs, and none of 50 reports claiming distinct classes compared them with a continuous reference. Although allocation was reproducible, held-out shape evidence did not establish discrete classes. Renal replacement therapy after hour 24 failed prespecified support criteria in both groups; no effect was estimated.

Conclusions

Reuse of a patient label requires distinct evidence for assignment, predictive increment and risk-model transport. Archiving the label with its generating objects and reporting its version alongside the comparator makes these conclusions interpretable and identifies when adaptation or an alternative study design is required.

Background

Data-driven patient partitions are widely used to characterize clinical heterogeneity. In sepsis, published groups have been associated with prognosis and potential treatment-response differences; however, their composition varies according to the selected features, algorithms, timing, and setting.[1–6] A clustering analysis can produce a patient-level label that is subsequently reused in outcome models, assigned to patients in another database, or applied to treatment comparisons. These applications do not necessarily evaluate the same object: the original assignments may be retained, the partition regenerated, or the missing-input rule adapted. Evidence obtained for one use therefore cannot be assumed to support another.

Common evaluations include refitting stability, outcome separation, external assignment, and subgroup treatment analyses.[7, 8] Each addresses a more specific question than 'validated phenotype' suggests. Refitting tests label reproducibility, not the distributional shape. For renal measurement partitions, the predictive advantage over a sparse clinical model may disappear when these measurements are included directly in the comparator. Successful code execution on new records demonstrates technical feasibility, whereas determining whether labels retain their meaning requires an independent reference. A scoping review of sepsis phenotyping found validity assessment in approximately half of the included studies,[9] although reporting of comparisons that distinguish these interpretations remains unknown.

PhenoFit evaluates the fitness of computable phenotypes for purpose and reuse in hypothetical settings.[10] Lisik et al. organized the clinical utility of machine-learning clusters around clinical relevance, stability and generalizability, and ease of identification,[11] and TRoCA is developing cluster-analysis reporting guidance.[12] Other phenotyping guidance details features, temporal definitions and executable rules.[13–19] For fixed labels, however, the stored objects, comparators, and evidentiary boundaries required to support downstream claims remain less clearly defined.

Reused data-derived labels should not be treated as broadly “validated.” Rather, the evidence required to justify reuse depends on the specific claim, the version of the label under evaluation, and the comparator against which the claim is assessed. We operationalize this as a specification that links each claim to the objects and comparisons it requires (Table 1). Label lineage identifies which version a comparison evaluates, even when successive implementations share a partition name, and an object inventory states what teams should retain before treating an existing label as ready for reuse. A survey examines how published sepsis partitions report these objects and comparisons; a worked K = 2 example shows how they change interpretation.

Our central question was which apparently adequate validation results change their meaning when an existing patient label is reused under a different version or judged against a different comparator. The specification defines the requirements for each claim, the survey assesses how frequently published sepsis partitions report them, and the cross-database analysis of one electronic health record (EHR)-derived K = 2 sepsis partition illustrates the extent to which the required comparisons alter interpretation.

This partition demonstrates the specification in practice; it neither validates the specification across phenotype types nor proposes a new sepsis subtype.

Methods

Specification

Table 1 translates the requirements from cluster validation, prediction model reporting, phenotype portability, and causal inference into obligations conditional on the claim made. Step 0 characterizes the encoded information; subsequent questions address latent discreteness (D1), incremental information (D2), external application (D3), and treatment-comparison support (D4). Additional file 1: Table S25 maps the required objects to storage locations; thus, each evaluation starts from an explicit record of what is available.

Table 1
Claim-specific specification for a fixed data-derived label

Claim	Required objects	Key comparison or design	Judgments the evidence can support	If an object is unavailable
Composition (Step 0)[7]	Fixed labels; source features in the original coordinate system	Reconstruct the label by prespecified feature blocks and examine group profiles	Identifies the measured information recorded by the label. Same-input analyses do not establish mechanisms or biological subtypes.	Without source features, only group profiles and outcomes were described.
Discreteness (D1)[20, 21]	Generation procedure, preprocessing, and source features sufficient to refit	Compare with a plausible continuous reference and held-out shape; use known-truth controls to define the detection range	Discrete structure, no detection within the stated range, or unresolved result. A stable assignment does not establish discreteness.	Without the full procedure, a pipeline-matched refit is unavailable, and retained source features may still permit limited shape analyses.
Incremental information (D2)[22]	The same label object and folds across models; source features	Compare a recognizable clinical model and a source-complete model, with and without the label in the same folds	Report gains against both comparators. A Base-only gain may reflect omitted inputs; no detected gain is not proof of none.	Without source features, report only the limited-comparator result and identify the missing comparison.
External application (D3)[15–17, 23]	Verified new-patient rule and harmonized external features; outcomes for risk evaluation; independent references for individual meaning	Evaluate technical assignment, risk performance, and individual meaning separately; examine missingness and site variation	Assignment does not establish risk calibration or individual label meaning; each requires its own set of evidence.	No rule: name regenerated or reconstructed labels. No outcomes: do not claim risk of transport. No reference: Individual meaning remains unassessed.

Note. The specification serves as an extraction form and not as a study-quality instrument. Each row links its cited requirements to a fixed label. Thresholds for the intended setting were specified before the analysis; worked-example thresholds are presented in Additional file 1: Table S12C.

Claim	Required objects	Key comparison or design	Judgments the evidence can support	If an object is unavailable
Treatment comparison (D4)[24–26]	Fixed labels; verified time-stamped treatment; eligibility, event, outcome, and covariate histories	Set time zero after label observation; assess eligible patients, arm counts, weights, overlap, and balance before a separately specified effect design	Support diagnostics identify information limits, not treatment effect modification. Report the original criteria as specified and label the later checks post-hoc.	Unverified treatment is an exposure measurement limit; invalid timing or strategy blocks the effect comparison. Name the limitation.

Note. The specification serves as an extraction form and not as a study-quality instrument. Each row links its cited requirements to a fixed label. Thresholds for the intended setting were specified before the analysis; worked-example thresholds are presented in Additional file 1: Table S12C.

Literature survey of published sepsis partitions

We searched PubMed and Europe PMC for English-language reports published from 2016 to 9 October 2026 that partitioned adults with sepsis using routinely collected clinical data. Eligible records were screened in publication-period-stratified random order toward 60 reports; sampling closed at 54 retrieved reports. The form recorded each claim and, only among reports making it, whether its corresponding object or comparison appeared in the article, obtainable supplements, or identified code. Two authors (S.Z. and X.W.) independently extracted core items for 53 reports; one additional report was extracted after those entries were visible and was treated as non-independent. Once extraction was complete, a third author (T.W.) compared the two sets of entries, resolved discrepancies, and verified every count-changing entry. Additional file 1: Section S13 details the search, sampling, extraction, and review processes.

Worked example: design and data sources

Each original worked-example module specified its label, comparator and decision thresholds before analysis. Additional file 1: Table S23 lists the post hoc additions and their triggers. Figure 1 links modules to label versions and comparisons; D4 examined recorded renal replacement therapy (RRT).

MIMIC-IV v3.1 supplied the development cohort and eICU-CRD v2.0 cross-database context.[27, 28] MIMIC included 20,049 adults' first ICU stays meeting the operational Sepsis-3 definition.[29] eICU included 17,465 adult first-unit stays from 16,212 patients in 199 eICU hospitals with a diagnosis string containing "sepsis" or "septic", without a diagnosis-time restriction (proxy sepsis), and observed first-day creatinine. Both excluded recorded chronic kidney disease or end-stage renal disease (cohort flow: Additional file 1: Table S31). Clustering included thirty-three first-day physiological variables, excluding

outcomes and RRT. Values outside prespecified physiological ranges were set missing, continuous variables were winsorized at the 1st and 99th percentiles, and multiple imputation by chained equations (MICE) used five completed datasets, 10 iterations, and seed 2024.[30] D2 and D4 used death within 30 days of hospital admission in MIMIC-IV; D3 used in-hospital mortality in both databases. Additional file 1: Section S1.4 defines the exploratory MIMIC-IV renal composite and distinct eICU death-or-last-ICU-creatinine composite (eICU composite). Reporting followed RECORD and TRIPOD + AI where applicable; [22, 31] simulations followed ADEMP.[32]

Partition construction and label lineage

The label lineage identifies the object to which each comparison applies. Consensus K-means screened K = 2–6 on a 5,000-patient median-imputed matrix with 1,000 resamples,[33] and final locked labels came from plain Lloyd K-means on MICE completed dataset 1 (K = 2, 25 starts, 100 iterations). Exact reuse retained this vector; refitting re-estimated parameters for the same target definition; regeneration reclustered after resampling or imputation. Reconstructed and deployed rules assigned new patients (D3), whereas risk-set restriction retained locked labels among time-eligible patients. We tracked these relationships separately, even when the versions were closely aligned. This distinction also separates changes in assignments from changes in the eligible population, which can occur under the same partition name.

Step 0: clinical identity and semantic composition

Step 0 characterized the locked partition's physiological information using standardized feature differences, Fisher loading geometry, out-of-fold reconstruction from pre-specified feature blocks, ablation analyses, Kidney Disease: Improving Global Outcomes (KDIGO) criterion decomposition,[34] and outcome associations.

D1: latent discreteness

Cluster-significance testing is one approach to discreteness assessment.[20] Per split, we refitted Lloyd K = 2 in training data and compared pooled-within Mahalanobis separation with a continuous Gaussian-copula reference under the same allocation. A training-derived Fisher direction and the first five principal components (PC1-PC5) underwent held-out Benjamini-Hochberg-adjusted dip tests.[35] Only the adjusted Fisher-axis P value could satisfy the shape criterion; PC1-PC5 entered the correction, not the verdict. Discrete-class evidence required both separation and held-out shape; discordance remained inconclusive, and full cohort separation did not determine split-allocation conclusions. Continuous-severity and outcome-independent discrete controls characterized the joint rule; independent Gaussian-copula realizations checked separation calibration (Additional file 1).

D2: incremental information

We compared out-of-sample area under the receiver-operating-characteristic curve (AUC) changes from adding K2 to two models: a limited clinical Base model (age, sex, infection-anchored rolling Sequential Organ Failure Assessment [SOFA] score, and acute kidney injury [AKI] stage during ICU hours 0–24) and Raw-EN, an elastic-net model containing all 33 source features with age, sex, total SOFA, and KDIGO stage.[36] Outcomes were mortality within 30 days of hospital admission and the exploratory renal composite. Analyses separately reused locked labels or examined imputation-specific and strictly fold-wise regenerated labels.[37] Each regenerated-label analysis used the same label object for Base and Raw-EN contrasts. Because K2 thresholds an affine score of Raw-EN features, a near-null increment against Raw-EN was expected, apart from possible threshold nonlinearity.

D3: external application

Post hoc D3 analyses recovered the original centroids and scaling from the saved MICE-completed matrix and fixed labels, verifying the exact reapplication (Additional file 1: Table S23). For new patients, we retained the recovered label-rule parameters but replaced cohort-wide MICE with fixed source-median imputation, and refitted only label-dependent Base+K2 source models. After a source audit identified omitted eICU nurse-charted temperatures, we corrected these first-day inputs and repeated external evaluation, keeping MIMIC-IV labels and source coefficients fixed. We also transferred external missingness patterns to complete source records. Hospital C1 counts were assessed against exact binomial funnel limits and with a random-intercept model. For mortality recalibration, hospitals sharing patients stayed together; intercept-only and intercept-and-slope updates fitted in 139 hospitals were evaluated in 60 others under a split fixed before the analysis. Brier score was primary; 1,000 paired hospital-component bootstrap attempts refitted update parameters.

D4: treatment comparison

The ICU-hour-24 landmark was placed for treatment assessment after K2's first-day observation window. It included patients who were alive, still in the ICU, and without an RRT-positive record (Additional file 1: Figure S1). Exposure was the first positive record before death, ICU exit, or 72 h.[24] The original stopping rule required at least 1% exposed and at least 10 admission-anchored 30-day deaths in each record-status cell within each K2 group, with prespecified propensity-score support checks before effect modeling (Additional file 1: Table S11). Post-hoc support diagnostics and loop diuretic timing checks were performed after the original analysis. The diagnostics assessed effective sample size (ESS) by exposure arm, weight concentration and covariate balance after whole-cohort and overlap weighting.[25, 38, 39]

Implementation checks and reproducibility

Implementation checks of selected calculations used simulated data with known answers to examine their diagnostic behavior.[32] Continuous and discrete generators examined the D1 joint rule; a noisy-oracle label assessed D2 detection beyond source features. Other generators examined D3 assignment

under distribution shift and D4 support under different exposure patterns. Additional file 1 reports scenario-specific operating characteristics and failure accounting results. Shared code supports reproducible EHR research,[40] with restricted repeat-level outputs retained in the authorized environment. Their operating characteristics are conditional on the selected generating scenarios and thresholds, which define the detection range of the implemented comparisons.

Use of large language models

OpenAI Codex was used to assist with code development and statistical figure visualization, and Anthropic Claude assisted with language and logic review. The final literature survey was independently redone by the authors without using earlier AI-assisted outputs. Two authors (S.Z. and X.W.) independently extracted 53 reports; the additional report and the adjudication process are described in Additional file 1: Section S13. T.W. resolved discrepancies. The authors verified the analyses, numerical results, references, and interpretations and retain full responsibility for the submitted work.

Results

Literature survey - the comparisons each claim requires were rarely reported

Of 54 reports (Table 2), 50 described distinct classes, but none reported comparison with a continuous reference; five compared a one-class model. All reported prognostic associations; six reported incremental performance, none against a comparator containing the clustering inputs. Seven of 20 external-application or reuse reports applied a fixed rule to new patients. Of these, two reported assignment parameters (one partially), two preprocessing parameters, and two a new-patient missing-input rule; one provided public assignment code. Twelve reports compared a treatment across phenotypes; three placed treatment after the label window, all comparing randomized assignments from two trials, and none reported within-phenotype support diagnostics.

Table 2
Reporting of the specified objects and comparisons in 54 published sepsis partitions

Claim	Reports making the claim or applicable subset	Object or comparison reported	n/N	% (exact 95% CI)
Composition	27 multivariable partitions	Contribution of clustering inputs quantified	8/27	29.6 (13.8–50.2)
Discreteness	50 distinct-class claims	Comparison with a continuous reference	0/50	0.0 (0.0-7.1)
Incremental information	54 prognostic claims	Incremental performance reported	6/54	11.1 (4.2–22.6)
	54 prognostic claims; 6 incremental comparisons	Comparator contained the clustering inputs	0/54; 0/6	0.0 (0.0-6.6); 0.0 (0.0-45.9)
External application	20 external-application or reuse reports	Fixed rule applied to new patients	7/20	35.0 (15.4–59.2)
	7 fixed-rule applications	Parameters; preprocessing; missing-input rule; public code	2/7; 2/7; 2/7; 1/7	28.6 (3.7–71.0) each; 14.3 (0.4–57.9)
Treatment comparison	12 treatment comparisons	Treatment began after the label window	3/12	25.0 (5.5–57.2)
		Within-phenotype support diagnostics	0/12	0.0 (0.0-26.5)

Note. Frequencies record reporting in the article, obtainable supplements, and identified code, not whether an analysis was performed or valid. The denominators included only reports making the corresponding claims or applicable subsets. Five of the 50 distinct-class reports compared a one-class model, which does not test against continuous structure; within-phenotype treatment/control counts appeared in 4 of 8 binary comparisons. Intervals are Clopper-Pearson limits conditional on this retrieved sample. Additional file 1: Tables S26-S28 give frequencies by partition type, the full assignment inventory, and a material-access sensitivity analysis.

Worked example: Step 0 - K2 mainly recoded renal and acid-base physiology

Renal and acid-base variables accounted for 61% of the squared Fisher loadings and reconstructed C1 with an out-of-fold AUC of 0.972, compared with 0.589 for non-renal SOFA proxies. The fixed K2 partition assigned 3,992/20,049 MIMIC-IV patients to the higher-mortality group (C1) and 16,057 to the lower-mortality group (C2). Mortality was 42.1% in C1 versus 12.3% in C2; the exploratory renal composite occurred in 67.8% and 29.0%, respectively. Removing renal and acid-base variables reduced the assignment agreement to an adjusted Rand index (ARI) of 0.284, whereas removing non-renal SOFA proxies retained an ARI of 0.927. C1 prevalence was 11.0% when neither KDIGO creatinine nor urine-

output criterion was met, 13.0% with urine output only, 41.8% with creatinine only, and 59.3% with both (Fig. 2).

D1 - Reproducible allocation accompanied unresolved evidence of discreteness

Across five imputations, pairwise ARI averaged 0.931 (range 0.927–0.936). Training-refitted K2 partitions agreed with fixed labels at ARI 0.977 (empirical 95% interval 0.943–0.991) in the 10,024/10,025 allocation. Separation exceeded its continuous reference in all 300 split-allocation refits, while every empirical allocation failed the held-out Fisher-axis shape criterion, leaving the joint result inconclusive (Fig. 3A-B). In a separate 200-repeat implementation check, a non-Gaussian continuous-severity process produced no joint alerts, although 49.5% of repeats triggered separation alone; a discrete process triggered both criteria in all 200 (Additional file 1: Section S4.2; Table S13B). The inconclusive state therefore includes some non-Gaussian continuous structures as well as unresolved empirical partitions. Stability and a separation alert therefore supplied different evidence from the held-out shape test.

D2 - K2 added information to a limited model but no detectable increment beyond its source variables

Adding K2 to the source-complete Raw-EN model changed AUC by 0.0000 for 30-day mortality and −0.0002 for the exploratory renal composite in the primary locked-label contrast; Raw-EN AUCs were 0.806 and 0.739. Base comparisons used regenerated labels: adding K2 raised AUC by 0.050 and 0.038 in the imputation-specific analysis and by 0.051 and 0.039 in the strictly fold-wise analysis, while the corresponding Raw-EN increments were less than 0.001. Base gains reflected K2 source physiology absent from that limited comparator (Fig. 3C; Additional file 1: Table S9). In the selected simulation, the noisy-oracle implementation check detected information beyond source inputs reliably once label accuracy reached 0.90 (Fig. 3D; Additional file 1: Figure S2).

D3 - The recovered rule assigned every eICU stay, but the Base+K2 risk model needed local updating

Applying the recovered centroids and standardization reproduced all 20,049 fixed MIMIC-IV labels without reclustering. Source-median imputation changed 224 labels, giving 98.9% agreement (ARI 0.949); the earlier reconstructed rule agreed with 98.6% (Fig. 4A). In MIMIC-IV, changes reached 56/2,049 among records missing at least six inputs. Correcting the eICU temperature source changed 116 of 17,465 external assignments and reduced the fraction missing at least six inputs from 73.9% to 54.9%. Applying eICU missingness patterns to complete MIMIC-IV records changed 6.9% of assignments, mostly from C1 to C2 (Additional file 1: Section S6.1 and Table S10F).

C1 prevalence under the deployed rule was 19.46% in MIMIC-IV and 19.00% with corrected eICU inputs. Missing inputs shifted assignments toward C2 in the preceding experiment, linking pooled prevalence to input availability. Hospital-size-adjusted comparison identified 9 of 199 hospitals outside exact binomial 99.8% limits around MIMIC-IV prevalence (Fig. 4B). In 17,284 stays with an in-hospital mortality outcome and required Base-model predictors, Base+K2 yielded AUC 0.761 and calibration slope 0.780 (95% CI

0.749–0.812). The interval excluded 1, indicating overly extreme predictions; it crossed the prespecified 0.80 threshold despite a lower point estimate (Fig. 4C; Additional file 1: Table S10G). The label-free Base model also showed miscalibration (AUC 0.747; slope 0.854), and the prespecified earlier rule gave a similar Base+K2 slope (0.775).

In 6,478 stays from 60 hospitals withheld from updating, intercept-only updating lowered the Base Brier score (paired difference − 0.0029, 95% CI -0.0048 to -0.0010) but not clearly the Base+K2 score (-0.0009, -0.0027 to 0.0008). Intercept-and-slope updating lowered both: Base+K2 from 0.1219 to 0.1182 (-0.0037, -0.0055 to -0.0022) and Base from 0.1247 to 0.1205. The updated Base+K2 slope was 1.038 (95% CI 0.949–1.107), but residual calibration still differed between C1 and C2 (slopes 1.020 and 1.489; Fig. 4D; Additional file 1: Tables S10H-I). Before updating, Base+K2 overpredicted risk among held-out C1 stays (calibration-in-the-large − 0.647) but not C2 stays (-0.035).

D4 - Sparse RRT records and weighting limited treatment comparison

RRT-positive records occurred in 265 of 3,151 C1 patients (8.41%) and 49 of 14,374 C2 patients (0.34%) in the hour-24 risk set of 17,525 patients. C2 fell below the prespecified 1% record-frequency minimum. In C1, 10.76% of patients lay outside the empirical score-overlap range and 2.44% had weights above 10, exceeding the original 10% and 1% limits. C2 also exceeded the outside-overlap limit (45.81%). Pooled ESS/N was 0.0739 in C1 and 0.0139 in C2, below 0.25 in both. Within RRT-positive arms, ESS was 54.9 of 265 in C1 and 23.9 of 49 in C2 (Fig. 5A-B).

After overlap weighting, 8 of 15 interpretable C2 covariates remained above an absolute standardized difference of 0.1, led by creatinine (0.927); imbalance persisted when SOFA was omitted and when K2 was removed from the allocation model (Fig. 5C; Additional file 1: Section S7.5.1 and Table S24). These balance diagnostics describe the fitted pooled model. Both groups failed the original support criteria, so no treatment-effect or interaction model was fitted. Table 3 connects these findings to the survey.

Table 3

Common checks, related survey findings, and the specified evidence in the worked example

Claim	Common check in this example; related reporting in the survey	Evidence the specification required	What this example established	What remains open
Composition (Step 0)	Outcome separation (mortality 42.1% vs 12.3%) suggests a high-risk clinical subtype. Survey: input contributions quantified in 8/27 multivariable reports.	Source features in the original coordinates; block reconstruction and ablation.	K2 mainly recodes renal and acid-base physiology (61% of squared loadings; block AUC 0.972; ablation ARI 0.284).	A distinct biological mechanism or subtype.
Discreteness (D1)	Stability across imputations (ARI 0.931) suggests discrete classes. Survey: continuous reference in 0/50 distinct-class reports; one-class comparison in 5/50.	Procedure-matched refits, a continuous reference, held-out shape evidence, and known-truth controls.	A reproducible operational grouping. A continuous-reference comparison alone would have alerted (separation above the reference in all 300 refits); held-out shape did not.	Discrete classes or their absence beyond the specified test.
Incremental information (D2)	An AUC gain of 0.050 over Base suggests value unique to the label. Survey: incremental metric in 6/54 reports; input-containing comparator in 0/6 of these.	The same label object and folds against limited and source-complete comparators.	Gain over Base, but less than 0.001 over Raw-EN in regenerated-label analyses; a noisy-oracle control detected added signal at accuracy 0.90.	Whether binary K2 is useful compression relative to its continuous score or a named source-feature block.

Note. Common checks are isolated indicators applied to this example; the stronger reading in column 2 is not attributed to any published study. Survey frequencies (Table 2) describe reporting in the retrieved sample, not what investigators did or concluded. Abbreviations: ARI, adjusted Rand index; AUC, area under the receiver-operating-characteristic curve; eICU, eICU Collaborative Research Database; K2, fixed K = 2 partition; RRT, renal replacement therapy.

Claim	Common check in this example; related reporting in the survey	Evidence the specification required	What this example established	What remains open
External application (D3)	Complete assignment and similar pooled C1 prevalence (19.46% vs 19.00%) suggest successful transport. Survey: fixed rule in 7/20 external or reuse reports; parameters in 2/7 and code in 1/7.	Verified new-patient rule; separate assessment of assignment, risk, and meaning; site variation beyond sampling error.	Every eICU stay was assigned after imputation adaptation; 9/199 hospitals were outside 99.8% limits (post hoc); unmodified slope 0.780; intercept-and-slope updating lowered held-out error for Base and Base+K2.	Individual label meaning without reference labels; uniform calibration across hospitals and K2 groups.
Treatment comparison (D4)	RRT-positive records in both groups suggest a treatment-by-label model can be fitted directly. Survey: treatment after the label window in 3/12 comparisons; within-phenotype support diagnostics in 0/12.	A decision time after label observation; within-group arm counts, weights, overlap, and measured balance.	Both groups failed original support criteria: C1 had concentrated weights and limited overlap; C2 had scarce records (0.34%). No effect model was fitted.	Any RRT effect or interaction; verified treatment starts and a target-trial design.

Note. Common checks are isolated indicators applied to this example; the stronger reading in column 2 is not attributed to any published study. Survey frequencies (Table 2) describe reporting in the retrieved sample, not what investigators did or concluded. Abbreviations: ARI, adjusted Rand index; AUC, area under the receiver-operating-characteristic curve; eICU, eICU Collaborative Research Database; K2, fixed K = 2 partition; RRT, renal replacement therapy.

Discussion

Principal findings

The survey of 54 sepsis-partition reports identified recurring gaps in the objects and comparisons required to support reuse claims (Table 2). The worked example demonstrated the implications of these requirements: comparator selection changed the interpretation of predictive gain, whereas external assignment required separate evaluation of risk performance (Table 3). For informatics practice, the relevant unit of evaluation is an identified label version together with its generating objects and the comparison appropriate to its intended use.

The specification organizes established requirements as claim-conditional obligations for a stored label (Table 1), which explains why survey denominators differ by row. The survey identifies reporting gaps,

whereas the worked example illustrates how the specified comparisons alter interpretation. Reporting reviews of latent-class and trajectory analyses beyond sepsis have identified similar deficiencies, including omitted one-class comparisons and missing-data handling.[41, 42]

PhenoFit assesses fitness for intended use and reuse through hypothetical settings.[10] Lisik et al.'s checklist spans planning, modelling and post-modelling,[11] while TRoCA is developing cluster-analysis reporting guidance.[12] Our specification explicitly defines the requirements for fixed-label reuse. Lisik et al. suggest label-based outcome regressions and classifiers using influential variables for new-patient assignment, alongside treatment-response comparisons. For claims of incremental prediction, outcome-model comparisons should include the label-generating inputs. A new assignment classifier creates a label object whose agreement with the original must be measured; treatment-response comparison requires decision timing after label observation and adequate within-group support.

Implications for informatics practice

A reusable label archive should connect the assigned vector to its generation procedure and source-feature coordinates (Additional file 1: Table S25). Application to new patients also requires the fitted assignment parameters and a missing-input rule; risk evaluation requires outcomes, while interpretation of individual labels requires independent reference labels. Store patient-linked objects in the authorized environment and share code or aggregate parameters after the applicable data-use and disclosure review. Reports should specify the exact label version and its relationship to the original, including retained assignments, refitted or regenerated partitions, adapted deployment rules, and restrictions to a decision-time risk set, as these address distinct questions. State the comparator alongside that version so later users can recover the basis for the claim.

Identity, latent structure, and comparator completeness

K2 provides a reproducible operational summary of measured renal and acid-base physiology. Establishing a biological subtype requires evidence independent of K2's source inputs. Cluster validation distinguishes stable assignment from evidence of distinct classes,[7, 8] and sepsis partitions differ across algorithms, time points and settings.[1–6]

Source-complete comparison locates the apparent value of a reused label: the gain over Base in this analysis reflected inputs absent from that model, and none of the six surveyed incremental comparisons included the clustering inputs. Evaluations should pair a clinically recognizable comparator with one containing the label-generating inputs. The complementary noisy-oracle implementation check indicated that D2 could detect information beyond those inputs.

Transportability requires a new-patient assignment rule

Fixed source-median imputation rendered the recovered rule executable for new patients, with agreement measured against the original labels. Only two of seven surveyed fixed-rule applications reported a missing-input rule for new patients. Every eICU stay received an assignment, while unmodified

mortality predictions remained miscalibrated. C1 prevalence in 9 hospitals varied beyond sampling error; case mix and measurement may contribute, since eICU missingness patterns shifted complete MIMIC-IV records toward C2. Updating in separate hospitals lowered both Base and Base+K2 Brier scores, locating the improvement in the risk models; residual calibration differed between C1 and C2.[23]

Support diagnostics and the timing of partition-guided treatment

The original D4 support screen failed in both K2 groups, albeit for different reasons. In C1, pooled ESS/N (0.0739) was about one quarter of its constant-score reference (0.3081), 2.44% of weights exceeded 10, and 10.76% of patients lay outside the empirical score-overlap range. C2 had only 49 record-positive patients, and pooled ESS/N (0.0139) was near its reference (0.0136). With a well-calibrated score, 4p(1-p) approximates an upper bound on pooled ESS/N; the prespecified 0.25 threshold therefore could not be met below about 6.7% record frequency, whatever the overlap; that criterion was ill-suited to rare exposures. Arm-specific ESS and empirical overlap separate rarity from additional information loss.

The landmark revealed a second constraint: K2 uses first-day data and can guide treatment decisions only after hour 24. RRT met this ordering but followed the renal physiology defining the label. The loop-diuretic check was halted at the same timing screen because most first records preceded hour 24 (Additional file 1: Section S7.5). In the survey, three of 12 treatment comparisons placed treatment after the label window, all in randomized trials. Treatment-effect modification claims require decisions after the phenotype observation window, adequate within-group treatment support, and a design that addresses time-varying confounding; treatment may legitimately depend on phenotype-defining measurements.[24, 26]

Limitations

Table 1 emerged from one iteratively developed K = 2 partition. D1 conditioned on fixed K, preprocessing and selected projections. Diagnosis-code exclusions may leave chronic disease; C1 prevalence varied with sepsis onset relative to ICU admission (24.2% vs 14.7%). Trajectory or latent-class labels may need different objects and comparisons.

Nonprobability sampling closed at 54 retrieved reports after preliminary frequencies were available; intervals ignore selection and shared trial cohorts. Shared definitions and adjudication by one author could carry a common misreading into both extractions. Unreported does not mean unperformed.

External evaluation used proxy sepsis and repeated eICU stays, with incomplete feature harmonization and no semantic reference. MIMIC-IV's derived Glasgow Coma Scale (GCS) assigned 15 to assessments flagged unable; eICU used charted totals. Labels remained fixed, original D3 implementation checks evaluated the earlier reconstructed rule, and composites differed. Updating was evaluated in 60 held-out hospitals of 199, with residual calibration differences across K2 groups.

The D4 exposure used RRT-positive records without independent verification of treatment initiation; ICU exits before a record counted as record-negative. Real-time covariate availability remained uncertain, including 151 cohort-defining SOFA assessments after hour 24. Admission-anchored thirty-day mortality windows had elapsed for 54 patients by hour 24. Four negative urine totals were excluded from interpretable balance summaries. Allocation-only simulations examined weights rather than the complete longitudinal stopping rule.

Conclusions

Reuse a patient label with an explicit version and an archive of its generating objects, then report the comparison required by the intended claim. The survey and K2 example show why routine checks need this context: they support specific interpretations rather than a general claim that a phenotype is validated.

Abbreviations

ADEMP
aims, data-generating mechanisms, estimands, methods and performance measures
AI
artificial intelligence
AKI
acute kidney injury
ARI
adjusted Rand index
ATE
average treatment effect
AUC
area under the receiver-operating-characteristic curve
BP
blood pressure
BUN
blood urea nitrogen
C1/C2
higher-risk and lower-risk K2 groups
CI
confidence interval
D1-D4
discreteness, incremental-information, external-application and treatment-comparison questions
EHR
electronic health record
eICU-CRD

eICU Collaborative Research Database
ESS
effective sample size
GCS
Glasgow Coma Scale
ICU
intensive care unit
INR
international normalized ratio
K2
fixed K = 2 partition
KDIGO
Kidney Disease:Improving Global Outcomes
L0/L1/L2
original completed matrix, deployed rule with source-median imputation, and earlier reconstructed rule
MICE
multiple imputation by chained equations
MIMIC-IV
Medical Information Mart for Intensive Care IV
PC
principal component
Raw-EN
elastic-net model containing the source features
Raw-RF
random-forest model containing the source features
RECORD
REporting of studies Conducted using Observational Routinely-collected health Data
RRT
renal replacement therapy
SAPS II
Simplified Acute Physiology Score II
SOFA
Sequential Organ Failure Assessment
TRIPOD + AI
Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis, updated for artificial intelligence
TRoCA
Transparent Reporting of Cluster Analyses

Declarations

Ethics approval and consent to participate

MIMIC-IV and eICU-CRD contain de-identified data released under their source studies' ethics approvals and informed-consent waivers. The authors accessed both databases under PhysioNet credentialed-use agreements. No additional local ethics approval or informed consent was required for this secondary analysis. The study was conducted in accordance with the Declaration of Helsinki.

Clinical trial number

Not applicable.

Consent for publication

Not applicable.

Availability of data and materials

MIMIC-IV v3.1 and eICU-CRD v2.0 are available to credentialed users through PhysioNet under their data-use agreements.[43,44] PhysioNet is described in its original resource paper and current platform citation.[45,46] Patient-level extracts and derived patient-level data cannot be redistributed. Analysis and figure code are available at https://github.com/wty1599/claim-specific-ehr-label-audit (release v1.0.1; https://doi.org/10.5281/zenodo.23396216).[47] The repository contains code and documentation, including the post hoc eICU temperature-source correction and final figure scripts, but no patient-level data, aggregate source datasets, or literature-survey materials. Numerical results and literature-survey methods and summaries are reported in this article and Additional file 1; full texts of included reports are not redistributed.

Competing interests

The authors declare that they have no competing interests.

Funding

This study received no funding.

Authors' contributions

TW: Methodology, Formal analysis, Software, Validation, Visualization, Writing – original draft. YZ: Conceptualization, Data curation, Methodology, Validation, Writing – review & editing. SZ: Investigation, Validation. XW: Investigation. CF: Formal analysis, Software. JX: Methodology. JC: Resources, Supervision, Writing – review & editing. TW and YZ contributed equally. All authors read and approved the final manuscript.

Acknowledgements

Not applicable.

References

1. Smit JM, Jonkman AH, Krijthe JH. Sub-phenotyping in critical care: a valuable strategy or methodologically fragile path? Intensive Care Med Exp. 2025;13:59. 10.1186/s40635-025-00769-1.
2. Seymour CW, Kennedy JN, Wang S, Chang CCH, Elliott CF, Xu Z, et al. Derivation, validation, and potential treatment implications of novel clinical phenotypes for sepsis. JAMA. 2019;321:2003–17. 10.1001/jama.2019.5791.
3. van Amstel RBE, Rademaker E, Kennedy JN, Bos LDJ, Peters-Sengers H, Butler JM, et al. Clinical subtypes in critically ill patients with sepsis: validation and parsimonious classifier model development. Crit Care. 2025;29:58. 10.1186/s13054-025-05256-3.
4. Yoon CH, Sjöholm D, Wickstrøm KE, Danielsen AS, Prebensen C, Valik JK, et al. Multicenter validation of clinical sepsis phenotypes. JAMA Netw Open. 2026;9(6):e2616134. 10.1001/jamanetworkopen.2026.16134.
5. Scherger SJ, Kalil AC. Sepsis phenotypes, subphenotypes, and endotypes: are they ready for bedside care? Curr Opin Crit Care. 2024;30:406–13. 10.1097/MCC.0000000000001178.
6. Antcliffe DB, Burrell A, Boyle AJ, Gordon AC, McAuley DF, Silversides J. Sepsis subphenotypes, theragnostics and personalized sepsis care. Intensive Care Med. 2025;51:756–68. 10.1007/s00134-025-07873-6.
7. Hennig C. What are the true clusters? Pattern Recognit Lett. 2015;64:53–62. 10.1016/j.patrec.2015.04.009.
8. Ullmann T, Hennig C, Boulesteix AL. Validation of cluster analysis results on validation data: a systematic framework. WIREs Data Min Knowl Discov. 2022;12:e1444. 10.1002/widm.1444.
9. Li H, Markal A, Balch JA, Loftus TJ, Efron PA, Ozrazgat-Baslanti T, et al. Methods for phenotyping adult patients in sepsis and septic shock: a scoping review. Crit Care Explor. 2022;4(4):e0672. 10.1097/CCE.0000000000000672.
10. Wiley LK, Rasmussen LV, Levinson RT, Malinowski J, Manemann SM, Wilson MP, et al. PhenoFit: a framework for determining computable phenotyping algorithm fitness for purpose and reuse. J Am Med Inf Assoc. 2026;33(2):536–42. 10.1093/jamia/ocaf195.
11. Lisik D, De Kok JWTM, Bermúdez Barón N, Vanfleteren LEGW, Nwaru BI, Basna R. Good things come in threes: evaluating clinical utility of machine learning-derived clusters. JAMIA Open. 2026;9(3):ooag098. 10.1093/jamiaopen/ooag098.
12. Lisik D, Shah SA, Basna R, Dinh T, Browne RP, Andrews JL, et al. Protocol for development of a checklist and guideline for transparent reporting of cluster analyses (TRoCA). BMJ Open. 2025;15(8):e099609. 10.1136/bmjopen-2025-099609.
13. Carrell DS, Floyd JS, Gruber S, Hazlehurst BL, Heagerty PJ, Nelson JC, et al. A general framework for developing computable clinical phenotype algorithms. J Am Med Inf Assoc. 2024;31:1785–96.

10.1093/jamia/ocae121.

14. Wei WQ, Rowley R, Wood A, MacArthur J, Embi PJ, Denaxas S. Improving reporting standards for phenotyping algorithm in biomedical research: 5 fundamental dimensions. J Am Med Inf Assoc. 2024;31:1036–41. 10.1093/jamia/ocae005.
15. Mo H, Thompson WK, Rasmussen LV, Pacheco JA, Jiang G, Kiefer R, et al. Desiderata for computable representations of electronic health records-driven phenotype algorithms. J Am Med Inf Assoc. 2015;22:1220–30. 10.1093/jamia/ocv112.
16. Kirby JC, Speltz P, Rasmussen LV, Basford M, Gottesman O, Peissig PL, et al. PheKB: a catalog and workflow for creating electronic phenotype algorithms for transportability. J Am Med Inf Assoc. 2016;23:1046–52. 10.1093/jamia/ocv202.
17. Pacheco JA, Rasmussen LV, Kiefer RC, Campion TR, Speltz P, Carroll RJ, et al. A case study evaluating the portability of an executable computable phenotype algorithm across multiple institutions and electronic health record environments. J Am Med Inf Assoc. 2018;25:1540–6. 10.1093/jamia/ocy101.
18. Pfaff ER, Girvin AT, Crosskey M, Gangireddy S, Master H, Wei WQ, et al. De-black-boxing health AI: demonstrating reproducible machine learning computable phenotypes using the N3C-RECOVER Long COVID model in the All of Us data repository. J Am Med Inf Assoc. 2023;30:1305–12. 10.1093/jamia/ocad077.
19. Denaxas S, Gonzalez-Izquierdo A, Direk K, Fitzpatrick NK, Fatemifar G, Banerjee A, et al. UK phenomics platform for developing and validating electronic health record phenotypes: CALIBER. J Am Med Inf Assoc. 2019;26:1545–59. 10.1093/jamia/ocz105.
20. Liu Y, Hayes DN, Nobel A, Marron JS. Statistical significance of clustering for high-dimension, low-sample size data. J Am Stat Assoc. 2008;103:1281–93. 10.1198/016214508000000454.
21. Bauer DJ, Curran PJ. Distributional assumptions of growth mixture models: implications for overextraction of latent trajectory classes. Psychol Methods. 2003;8(3):338–63. 10.1037/1082-989X.8.3.338.
22. 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. 10.1136/bmj-2023-078378.
23. Van Calster B, McLernon DJ, van Smeden M, Wynants L, Steyerberg EW. Calibration: the Achilles heel of predictive analytics. BMC Med. 2019;17:230. 10.1186/s12916-019-1466-7.
24. Hernán MA, Sauer BC, Hernández-Díaz S, Platt R, Shrier I. Specifying a target trial prevents immortal time bias and other self-inflicted injuries in observational analyses. J Clin Epidemiol. 2016;79:70–5. 10.1016/j.jclinepi.2016.04.014.
25. Petersen ML, Porter KE, Gruber S, Wang Y, van der Laan MJ. Diagnosing and responding to violations in the positivity assumption. Stat Methods Med Res. 2012;21:31–54. 10.1177/0962280210386207.

26. Kent DM, Paulus JK, van Klaveren D, D'Agostino R, Goodman S, Hayward R, et al. The Predictive Approaches to Treatment effect Heterogeneity (PATH) statement. Ann Intern Med. 2020;172(1):35–45. 10.7326/M18-3667.
27. 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. 10.1038/s41597-022-01899-x.
28. Pollard TJ, Johnson AEW, Raffa JD, Celi LA, Mark RG, Badawi O. The eICU Collaborative Research Database, a freely available multicenter database for critical care research. Sci Data. 2018;5:180178. 10.1038/sdata.2018.178.
29. Singer M, Deutschman CS, Seymour CW, Shankar-Hari M, Annane D, Bauer M, et al. The Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3). JAMA. 2016;315:801–10. 10.1001/jama.2016.0287.
30. van Buuren S, Groothuis-Oudshoorn K. mice: Multivariate Imputation by Chained Equations in R. J Stat Softw. 2011;45:1–67. 10.18637/jss.v045.i03.
31. Benchimol EI, Smeeth L, Guttmann A, Harron K, Moher D, Petersen I, et al. The REporting of studies Conducted using Observational Routinely-collected health Data (RECORD) statement. PLoS Med. 2015;12:e1001885. 10.1371/journal.pmed.1001885.
32. Morris TP, White IR, Crowther MJ. Using simulation studies to evaluate statistical methods. Stat Med. 2019;38:2074–102. 10.1002/sim.8086.
33. Monti S, Tamayo P, Mesirov J, Golub T. Consensus clustering: a resampling-based method for class discovery and visualization of gene expression microarray data. Mach Learn. 2003;52:91–118. 10.1023/A:1023949509487.
34. Kidney Disease: Improving Global Outcomes Acute Kidney Injury Work Group. KDIGO Clinical Practice Guideline for Acute Kidney Injury. Kidney Int Suppl. 2012;2:1–138.
35. Hartigan JA, Hartigan PM. The dip test of unimodality. Ann Stat. 1985;13:70–84. 10.1214/aos/1176346577.
36. Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. J Stat Softw. 2010;33:1–22. 10.18637/jss.v033.i01.
37. Bates S, Hastie T, Tibshirani R. Cross-validation: what does it estimate and how well does it do it? J Am Stat Assoc. 2024;119:1434–45. 10.1080/01621459.2023.2197686.
38. Li F, Morgan KL, Zaslavsky AM. Balancing covariates via propensity score weighting. J Am Stat Assoc. 2018;113:390–400. 10.1080/01621459.2016.1260466.
39. Li F, Thomas LE, Li F. Addressing extreme propensity scores via the overlap weights. Am J Epidemiol. 2019;188:250–7. 10.1093/aje/kwy201.
40. Denaxas S, Direk K, Gonzalez-Izquierdo A, Pikoula M, Cakiroglu A, Moore J, et al. Methods for enhancing the reproducibility of biomedical research findings using electronic health records. BioData Min. 2017;10:31. 10.1186/s13040-017-0151-7.

41. van de Schoot R, Sijbrandij M, Winter SD, Depaoli S, Vermunt JK. The GRoLTS-Checklist: guidelines for reporting on latent trajectory studies. Struct Equ Model. 2017;24(3):451–67. 10.1080/10705511.2016.1247646.
42. Sorgente A, Caliciuri R, Robba M, Lanz M, Zumbo BD. A systematic review of latent class analysis in psychology: examining the gap between guidelines and research practice. Behav Res Methods. 2025;57:301. 10.3758/s13428-025-02812-1.
43. Johnson AEW, Bulgarelli L, Pollard TJ, Gow B, Moody B, Horng S et al. MIMIC-IV (version 3.1). PhysioNet. 2024. 10.13026/kpb9-mt58
44. Pollard TJ, Johnson AEW, Raffa JD, Celi LA, Badawi O, Mark RG. eICU Collaborative Research Database (version 2.0). PhysioNet. 2019. 10.13026/C2WM1R
45. Goldberger AL, Amaral LAN, Glass L, Hausdorff JM, Ivanov PCh, Mark RG, et al. PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. Circulation. 2000;101:e215–20. 10.1161/01.CIR.101.23.e215.
46. Pollard T, Moody BE, Lehman LWH, Gow BJ, Fernandes C, Xie C, et al. PhysioNet as a global platform for biomedical research. Nat Health. 2026;1:792–5. 10.1038/s44360-026-00096-z.
47. Wang T, Zhang Y, Zang S, Wu X, Jin X, Chen J et al. Claims and evidence for reused data-derived patient labels: a specification, a survey of sepsis partitions, and a worked example: analysis code. Version 1.0.1. Zenodo; 2026. https://doi.org/10.5281/zenodo.23396216

Figures

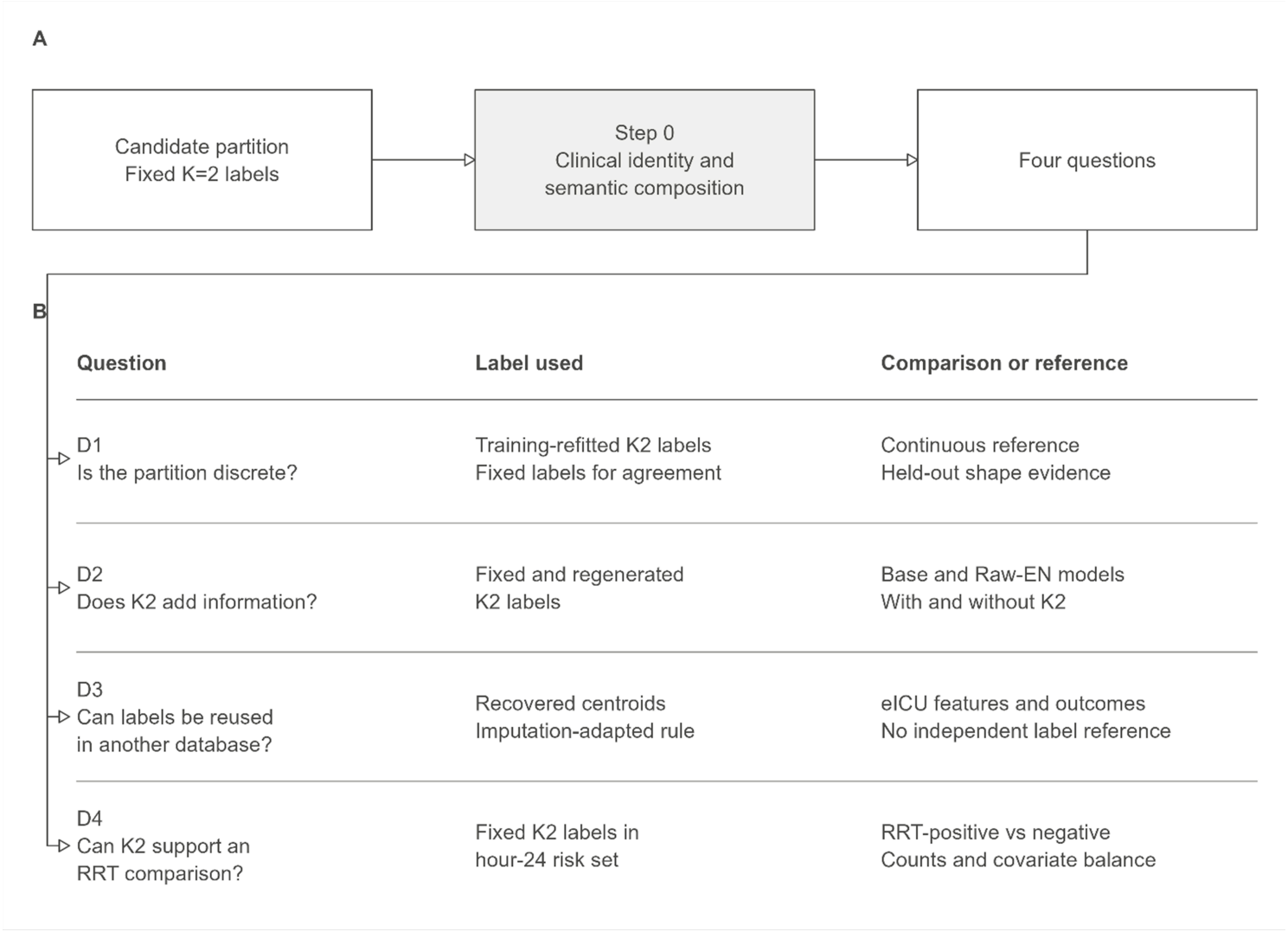

Figure 1

Study design links each question to its label version and comparison. (A) The fixed K=2 partition is characterized for clinical identity and semantic composition before four questions are examined. (B) Each row pairs a question with the labels used and the comparison or reference: training-refitted labels and held-out shape evidence for discreteness, fixed or regenerated labels and clinical or source-complete models for incremental information, an imputation-adapted rule for eICU assignment and mortality-risk updating in separate hospitals, and fixed labels in the hour-24 risk set for RRT comparison. K2, K=2 partition; RRT, renal replacement therapy; Raw-EN, elastic-net model containing the source features.

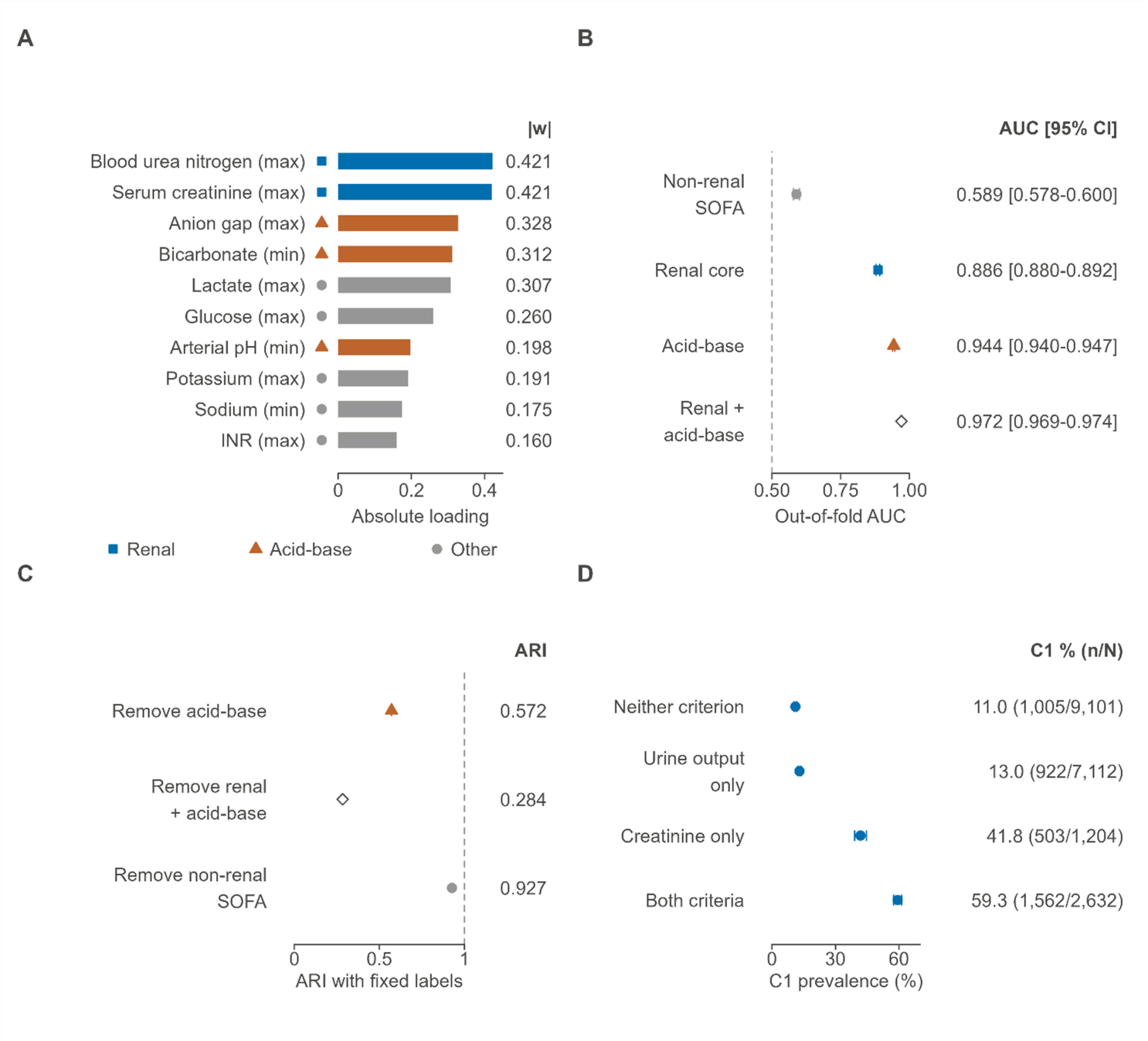

Figure 2

Renal and acid-base measurements account for much of the K2 partition. (A) The ten largest absolute standardized Fisher loadings are ranked, with symbols distinguishing feature blocks. (B) Out-of-fold label reconstruction is summarized by AUC and 95% DeLong confidence intervals for each feature block; the dashed line marks chance discrimination. (C) Points give mean adjusted Rand indices with the fixed labels after feature-block removal across 20 seeds, and bars span their minimum and maximum; the dashed line marks perfect agreement. (D) C1 prevalence is plotted by KDIGO criterion pattern, with Wilson 95% confidence intervals and group counts. AUC, area under the receiver-operating-characteristic curve; ARI, adjusted Rand index; INR, international normalized ratio; KDIGO, Kidney Disease: Improving Global Outcomes; SOFA, Sequential Organ Failure Assessment.

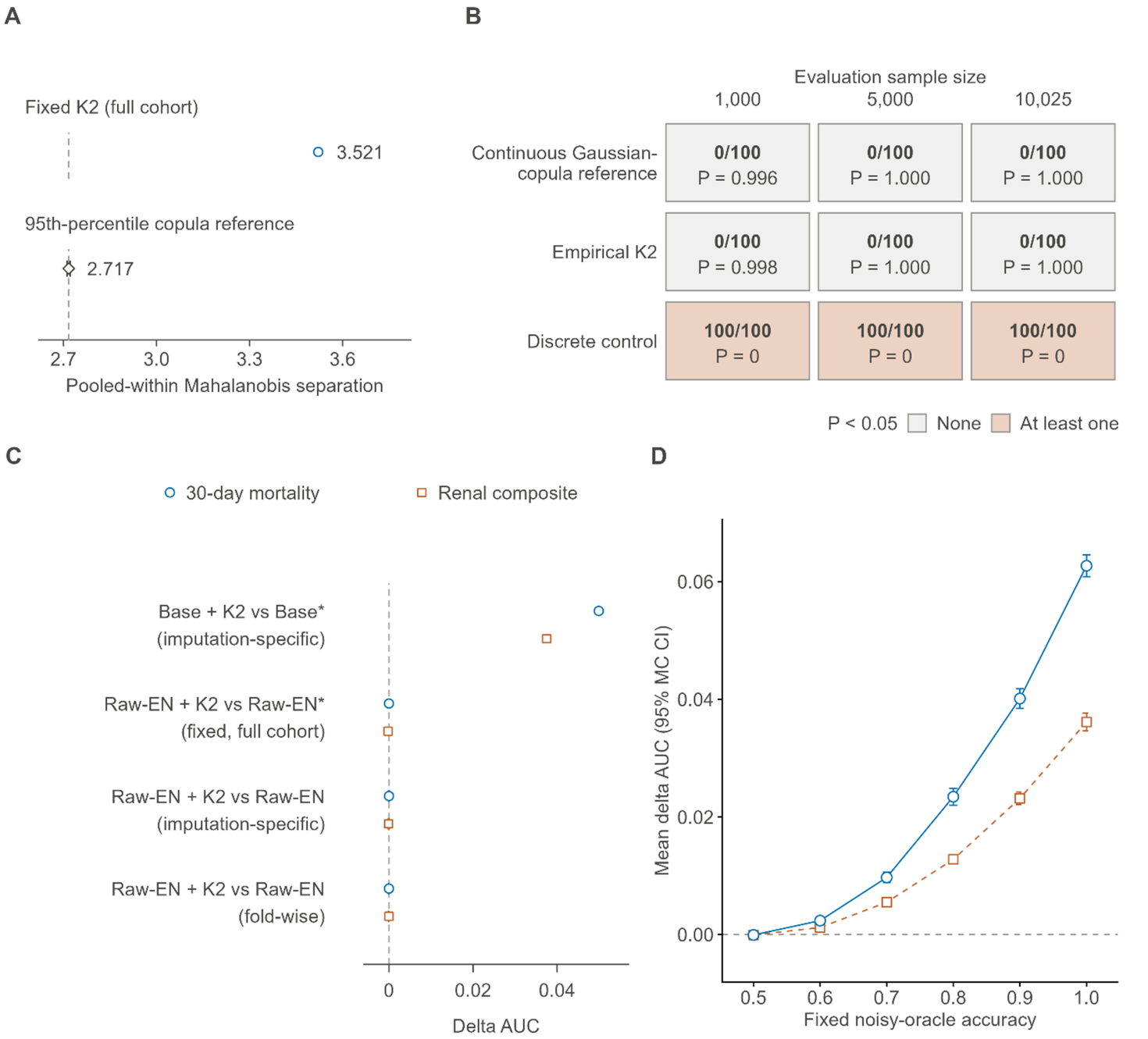

Figure 3

Separation exceeded the continuous reference, while predictive increments depended on the comparator. (A) Full-cohort Mahalanobis separation is compared with the Gaussian-copula reference's 95th percentile and its bootstrap 95% confidence interval. (B) Cells give counts below P=0.05 and median Fisher-axis dip-test P values after Benjamini-Hochberg correction across that axis and PC1-PC5, for each evaluation size. (C) AUC increments compare imputation-specific K2 added to Base and fixed, imputation-specific, or fold-wise K2 added to Raw-EN; asterisks mark point estimates without intervals. Imputation-specific Raw-EN intervals combine variances from 500 bootstrap resamples per imputation using Rubin-style pooling; fold-wise intervals are the 2.5th and 97.5th percentiles of 500 paired bootstrap

resamples of out-of-fold predictions. Both procedures hold predictions fixed. (D) Mean AUC increments and 95% Monte Carlo confidence intervals across repetitions are plotted against noisy-oracle accuracy; the outcome symbols in C also apply to D. AUC, area under the receiver-operating-characteristic curve; PC, principal component; Raw-EN, elastic-net model containing the source features.

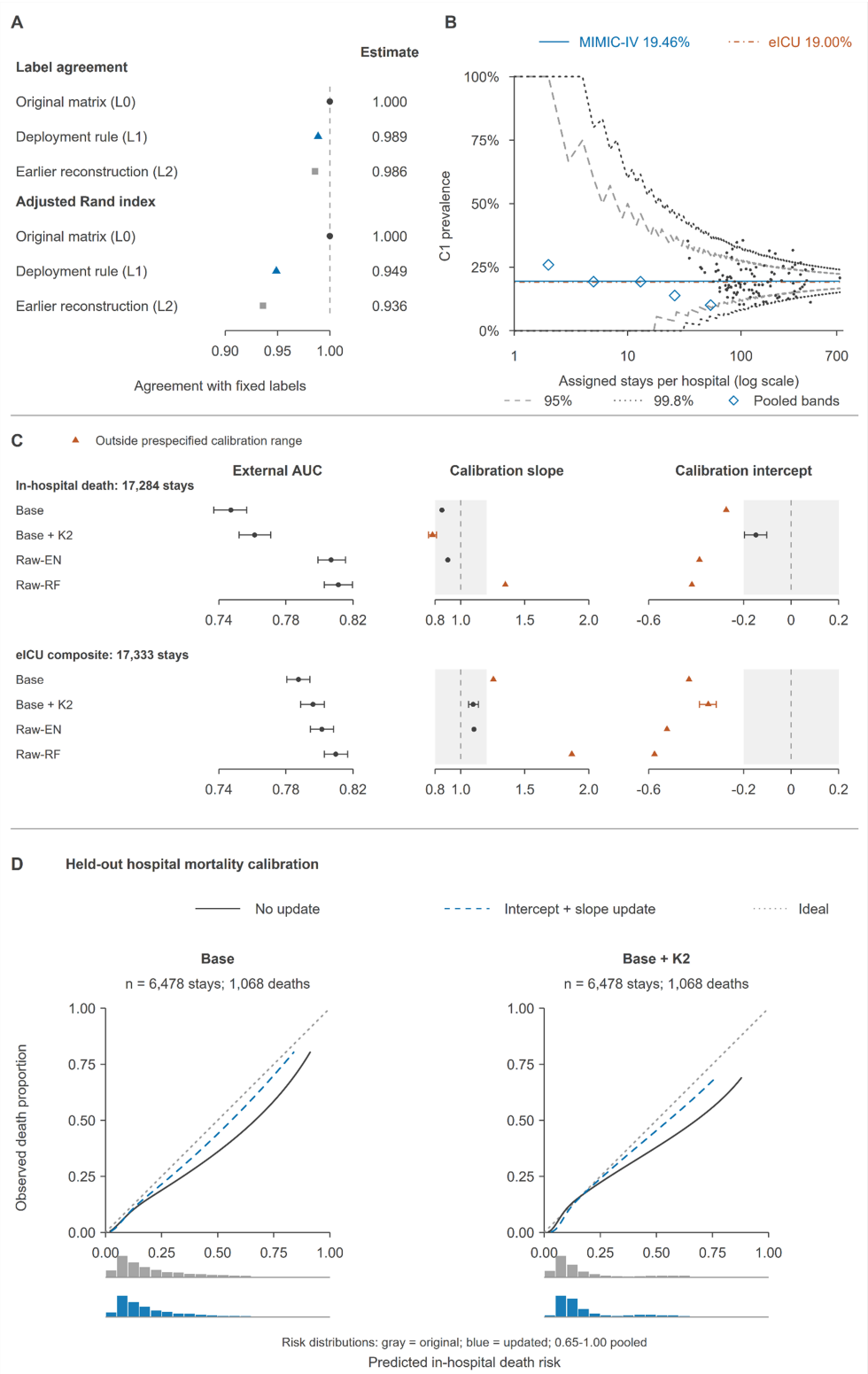

Figure 4

The recovered rule reproduced fixed labels, but hospital prevalence varied and mortality predictions needed updating. (A) Agreement and adjusted Rand index compare the original completed matrix (L0), the deployed rule with source-median imputation (L1), and the earlier reconstructed rule (L2) with fixed MIMIC-IV labels. (B) Under the external assignment with corrected temperature inputs, C1 prevalence is plotted against assigned stays per eICU hospital on a logarithmic scale. Ninety-six hospitals with at least 10 stays in each label group appear individually; open diamonds pool the other 103 into five size bands and are placed at each band's median hospital size. Pooled markers are descriptive and are not judged against single-hospital control limits. Solid and dot-dash lines show pooled MIMIC-IV and eICU prevalence; dashed and dotted curves are exact binomial 95% and 99.8% limits around the MIMIC-IV prevalence. (C) The original external evaluation procedure, repeated with corrected eICU temperature inputs, covers assignable eICU stays with a recorded outcome and the required Base-model predictors: 17,284 for in-hospital death and 17,333 for the eICU composite (Additional file 1: Section S6). Shaded areas mark the prespecified calibration ranges, and triangles mark estimates outside them. Whiskers show 95% confidence intervals for every AUC and for the Base+K2 calibration estimates. Intervals are percentile intervals from 1,000 patient-level clustered bootstrap replicates, conditional on frozen source-model predictions. (D) Mortality calibration in 6,478 stays (1,068 deaths) from 60 hospitals withheld from updating, before and after intercept-and-slope updates fitted in 139 other hospitals; dotted diagonals mark ideal calibration. Histograms show predicted-risk distributions before (gray) and after (blue) updating, with risks of 0.65-1.00 pooled for disclosure. Calibration curves are natural cubic spline smoothers with three degrees of freedom. AUC, area under the receiver-operating-characteristic curve; K2, K=2 partition; Raw-EN, elastic-net model containing the source features; Raw-RF, random-forest model containing the source features.

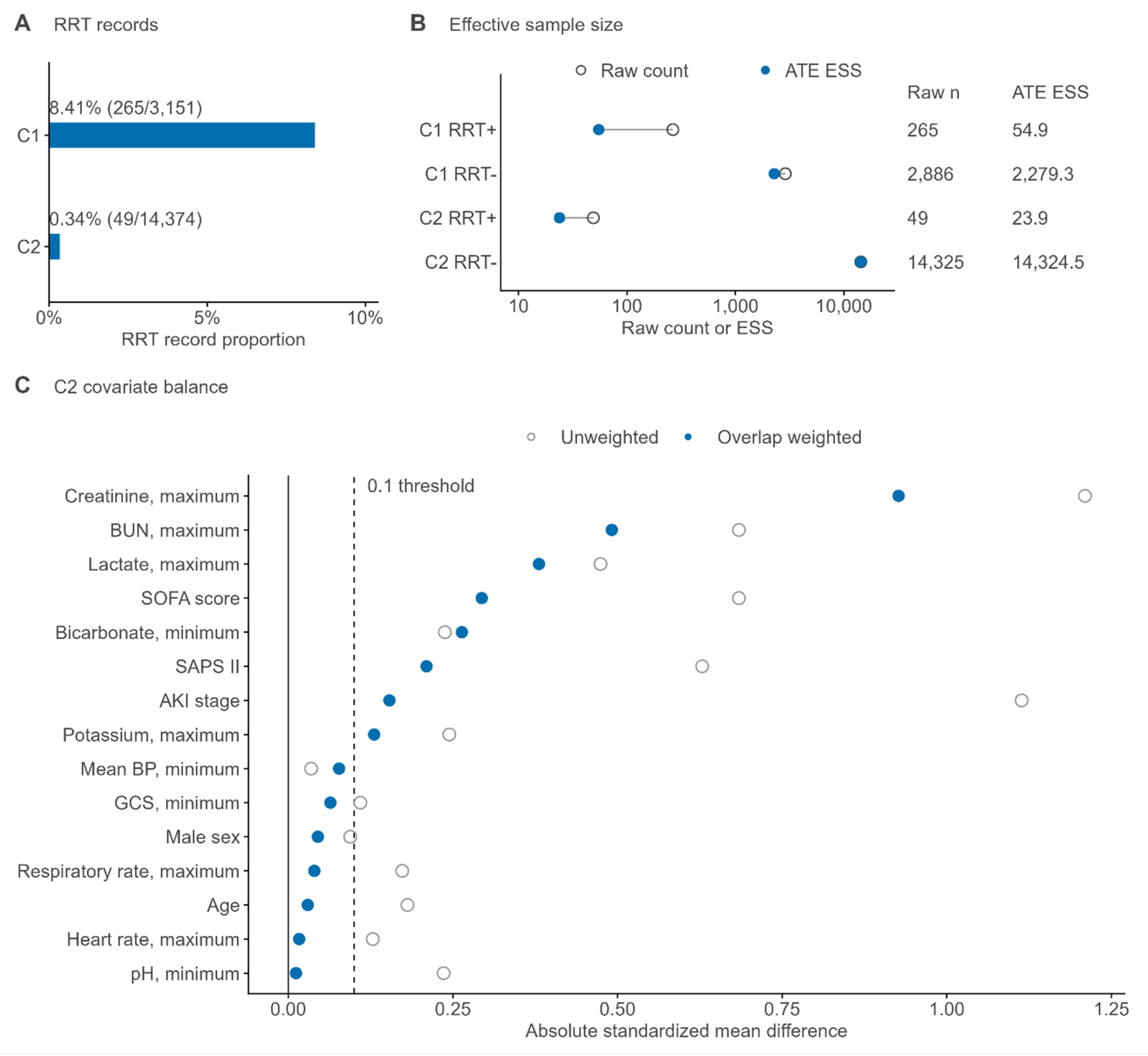

Figure 5

RRT records were sparse in C2, and weighting reduced effective exposed counts in both groups. (A) Proportions with a first RRT-positive record from ICU hour 24 to before hour 72 are shown for C1 and C2, with numerators and denominators. (B) Raw counts and arm-specific effective sample sizes under whole-cohort ATE weights are connected on a logarithmic scale and listed by RRT-record status. (C) Absolute standardized mean differences between C2 patients with and without records are compared before and after overlap weighting, ordered by the weighted values; the dashed line marks 0.1, and all differences use the fixed unweighted reference scale. Urine output is omitted because negative source values affected its interpretation. Neither group met the original support criteria (Additional file 1: Table S11). First-event composition and the C1/C2 balance display appear in Additional file 1: Figure S3. RRT,

renal replacement therapy; ATE, average treatment effect; ESS, effective sample size; BUN, blood urea nitrogen; AKI, acute kidney injury; BP, blood pressure; GCS, Glasgow Coma Scale; SAPS II, Simplified Acute Physiology Score II; SOFA, Sequential Organ Failure Assessment.

Supplementary Files

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

- supplementaryfile20261020.docx