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.

A retrospective evaluation of temporal aggregation and creatinine reuse in critical care records

Tianyi Yu

tianyi.1.yu@kcl.ac.uk

King's College London

Research Article

Keywords: electronic health records, data quality, critical care, temporal aggregation, creatinine, missing data, crystalloid, reproducibility

Posted Date: November 18th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11261752/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

Research using intensive care records often represents treatment and physiology in fixed time bins and reuses earlier laboratory measurements. Whether these choices preserve the intended measurements requires empirical evaluation. We examined crystalloid exposure reconstruction, recorded-panel availability and agreement between reused and subsequently measured creatinine.

Methods

This retrospective methodological study used MIMIC-IV 3.1 records from adults with coded diabetes, hypertension and early physiological stress. A fixed 6–12-hour crystalloid dose was reconstructed from four-hour grids with four phase offsets and compared with recorded-event interval overlap. Physiological-panel eligibility required timely documentation and was evaluated under creatinine age limits of 0–24 hours. At observed repeat-test times, the new creatinine was withheld and compared with the latest previously available value. Uncertainty used 2000 patient-cluster bootstrap resamples. A separate eICU 2.0 analysis assessed population-specific interface coverage.

Results

The master cohort contained 12175 stays in 10183 people. Among 10717 stays observed through 15 hours, fixed-target dose-category disagreement ranged from 5.66% (95% interval 5.21–6.12) to 16.01% (15.30–16.67) across grid phases. Four consecutive eligible panels were available in 146 stays (1.20%) without creatinine carry-forward, 6456 (53.03%) with a 12-hour limit and 8755 (71.91%) with a 24-hour limit. In 35081 paired repeat tests, mean absolute creatinine differences across prior-value age groups of 0–4, > 4–8, > 8–12 and > 12–24 hours were 0.155, 0.169, 0.177 and 0.188 mg/dL. Corresponding proportions differing by at least 0.3 mg/dL were 16.43%, 21.03%, 21.78% and 24.10%; severe creatinine-stage proxy disagreement was nonmonotonic at 3.12%, 3.85%, 3.63% and 3.11%.

Conclusions

Four-hour aggregation did not reliably preserve a fixed six-hour exposure. Reusing creatinine expanded operational panel availability but showed imperfect agreement at clinically selected retesting times. These complementary analyses do not establish accuracy in unmeasured intervals or a safe universal carry-forward limit. Measurement audits should accompany temporal dataset construction before treatment-response modelling.

Clinical trial registration

Not applicable. This study did not assign an intervention.

Background

How much does the representation of routinely collected data change the exposure and physiological information available for critical care research? We address this question through three distinct measurements: the reconstruction of a fixed crystalloid dose window, the availability of consecutive physiological panels, and agreement between an earlier creatinine result and a subsequently observed result. The intended contribution is an empirical evaluation of data-processing assumptions. It is not an estimate of the benefit of fluid treatment.

Fluid research makes these assumptions consequential. Crystalloid composition, the amount delivered and the reason for administration describe different aspects of treatment [1]. Trials comparing balanced crystalloids with saline and restrictive with liberal fluid strategies have studied different populations, interventions and outcomes [2, 3, 4, 5, 6, 7, 8]. Fluid responsiveness is also distinct from an indication to administer fluid [9, 10]; management guided by dynamic responsiveness assessment incorporates information that a retrospectively reconstructed dose alone cannot capture [11]. Temporal representation is therefore part of defining the research exposure, not merely a computational convenience.

MIMIC-IV and the eICU Collaborative Research Database support reproducible critical care research [12, 13], building on earlier resources and the broader PhysioNet infrastructure [14, 15, 16]. Their detail does not remove the need to establish fitness for a particular analysis. Data-quality frameworks distinguish completeness, conformance and plausibility [17, 18], while empirical work shows that the healthcare process influences what is measured and when [19]. These concerns become acute when an irregular clinical record is converted into a regular sequence suitable for statistical learning.

Two apparently simple operations can alter that sequence. First, summing or weighting fixed-width bins cannot generally recover an exposure window whose boundaries cut across those bins. Second, carrying a laboratory result forward makes a panel numerically complete without observing the patient's current value. Collection time and documentation time further determine whether a result was actually available at the intended boundary. This distinction is relevant to leakage prevention [20] and to specifying the temporal ordering of an observational study [21, 22]. Correct temporal ordering is necessary for many analyses, although it cannot by itself establish causal identification.

Recent work has made this problem more concrete. The PRESERVE study illustrates how data-quality testing can be tailored to a study's variables and temporal requirements [23]. Repeated extraction of a clinical data warehouse has also shown that records can change after their initial appearance [24]. Reproducible end-to-end EHR frameworks make preprocessing easier to inspect [25], while recent methodological guidance explicitly distinguishes when a clinical event occurs from when its information becomes available [26]. These approaches motivate a use-specific evaluation rather than assuming that a reproducible transformation preserves the intended clinical measurement.

The study was motivated by a proposed investigation of individualized fluid therapy among patients with diabetes and hypertension. Feasibility work showed that measurement assumptions required evaluation before treatment-response modelling. We therefore compare alternative representations against recorded-event timing and observed repeat laboratory values. The repeat-test analysis supplies evidence about agreement where a later value exists; it cannot reveal what happened during unmeasured intervals. A complementary hospital-level analysis asks whether broad interface coverage persists when the denominator is restricted to the intended chronic-disease population.

Methods

Study design and sources

We conducted a retrospective measurement study using local MIMIC-IV version 3.1 and eICU version 2.0 data [27, 28]. The within-stay experiments used MIMIC-IV; eICU supplied a separate interface-coverage analysis. We followed applicable STROBE and RECORD reporting principles [29, 30]. Additional file 1 contains detailed specifications, denominators, supplementary results and a reporting checklist. Additional file 2 supplies aggregate results and analysis code.

The original feasibility assessment preceded the measurement analysis plan. Following internal critical review, we added fixed-target grid-phase reconstruction and observed-repeat creatinine experiments, corrected the weight-availability rule and tested common follow-up and alternative weighting. The amendments record when these decisions were made relative to execution and inspection of results. The work was not prospectively registered. The original urine-derived composite was excluded from the revised analysis because review found that its rolling-window implementation could attribute inter-record durations beyond the intended window. Historical outputs and the correction audit were retained; they do not contribute to the present renal findings.

Population and analysis units

The MIMIC-IV candidate population comprised adults with admission-level codes for type 2 diabetes and hypertension, excluding coded type 1 diabetes and chronic dialysis. Early physiological stress required at least one signal in the first six hours: mean arterial pressure (MAP) below 65 mmHg, lactate at least 2 mmol/L, vasopressor documentation, a creatinine-based acute kidney injury signal, or low recorded urine volume relative to weight and six hours. The urine signal was a documentation-dependent screen, not an adjudicated oliguria diagnosis. The first eligible intensive care unit (ICU) stay per admission was retained. Multiple admissions per person were allowed.

Admission codes identify a retrospective disease population and need not have been known at ICU admission. Recalculation of early creatinine reference minima using only results documented before six hours changed two non-AKI stages to unknown among 19711 candidates, but changed neither early-

stress eligibility nor membership of the 12175-stay master cohort. The correction audit is supplied in Additional file 2.

The original 12-hour landmark contained 11773 stays; requiring an admissible pre-exposure weight and no recorded renal replacement therapy (RRT) through that landmark yielded 10939 exposure-eligible stays. For the principal four-phase comparison, we additionally required the trajectory to continue through 15 hours, the latest endpoint of any source bin. All phases were therefore compared in the same 10717 stays. Trajectories ended at the earliest of ICU exit, recorded death or documented comfort care. This common-follow-up restriction avoids unequal source-window support, but selects a longer-observed population; the original landmark and a 14-hour restriction were retained as sensitivity analyses.

Panel availability used complete four-hour bins within the first 48 hours of the master cohort. The repeat-test analysis used actual creatinine collections during the same first 48 hours, before trajectory termination and before the first recorded RRT event. Its unit was a laboratory measurement rather than a bin or stay. These denominators are reported separately in Table 1.

Table 1
Analysis populations and measurement units

Population or unit	Number	Purpose
Master ICU stays	12175	First eligible stay per admission
People in master cohort	10183	Bootstrap clustering unit
Twelve-hour landmark stays	11773	Trajectory support through landmark
Exposure-eligible stays	10939	Weight and RRT eligibility
Common fifteen-hour stays	10717	Identical support for four phase grids
Complete four-hour bins	114929	First 48-hour panel availability
Repeat-test targets	39807	Observed creatinine before RRT
No admissible prior result	1901	Excluded from agreement
Prior value older than 24 hours	2825	Excluded from agreement
Evaluated repeat-test pairs	35081	11620 stays in 9772 people

Recorded crystalloid exposure and fixed-target reconstruction

Crystalloid mapping included validated MIMIC labels for 0.9% sodium chloride and lactated Ringer's solution. It did not represent all intravenous intake, all balanced solutions or net fluid balance. Rewritten orders were excluded. For administrations longer than 60 seconds, recorded volume was distributed uniformly over the documented administration interval and allocated according to interval overlap.

Administrations lasting at most 60 seconds were assigned to their start time. This was the reference representation of recorded delivery, not a pump-validated measure of instantaneous flow.

The target remained [6,12) hours after ICU admission throughout the principal experiment. Four-hour grids began at phase offsets of 0, 1, 2 or 3 hours. We first allocated events to each complete source bin, then reconstructed the target by multiplying each bin's dose by the proportion of its duration overlapping [6,12). The four reconstructions were: one-half of [4,8) plus [8,12); three-quarters of [5,9) plus three-quarters of [9,13); [6,10) plus one-half of [10,14); and one-quarter of [3,7) plus [7,11) plus one-quarter of [11,15). Each represents six hours under an assumed uniform distribution within bins. The comparison isolates information lost by aggregation while holding the target window, fluid mapping, stay and weight fixed. Because some source bins extend beyond 12 hours, it is a retrospective resolution experiment, not an online reconstruction at the landmark.

Weight was restricted to 30–300 kg, measured between 24 hours before ICU admission and six hours after admission, and documented before six hours. The earliest admissible measurement was used. For descriptive selection summaries, baseline creatinine was the closest pre-ICU value within seven days, documented before six hours; missing baseline values were not imputed. Dose comparisons included signed and absolute differences, absolute differences greater than 1 mL/kg, and disagreement across the descriptive categories 0, > 0–5, > 5–10 and > 10 mL/kg. These categories were implementation probes, not recommended doses. Sensitivities compared start-time allocation with interval allocation for the same target, shifted six-hour windows on a common 14-hour population, and the deliberately mismatched [4,12) window. The latter was retained as a duration-error example, rather than the main evidence for temporal-resolution loss.

Operational panel availability and documentation clocks

An eligible four-hour panel required MAP, heart rate, a nonnegative recorded urine sum normalized by weight, creatinine and a computable retrospectively reconstructed crystalloid dose. Qualifying physiological records had collection or chart times within the panel and documentation times before its end. The fixed weight also had to have been measured and documented before that bin ended. A recorded urine sum establishes available documentation; it does not prove complete urine collection over the interval. Ventilation, vasopressor intensity and net fluid balance were not components of this panel.

Creatinine was either observed within the current bin or carried forward from a measurement collected no more than 4, 8, 12 or 24 hours before the bin ended. The result had to have been documented before the end of the bin. Eligible pre-ICU measurements could contribute within these age limits. No future value was backfilled. No-carry and four-hour limits coincide for the full four-hour bins under study. We measured eligible bins and the proportion of master stays containing at least four consecutive eligible panels; this does not require complete observation throughout 48 hours.

Observation reuse was treated as an assumption rather than evidence that physiology remained unchanged. Methods designed for missing time-series values [31] and multiple-imputation procedures [32, 33, 34] address related statistical problems but do not establish the validity of a carried measurement. We separately compared MAP minima and lactate maxima in the same [0,6) window using chart time alone versus documentation available before six hours. Value and threshold changes used pairs known in both versions; losses of availability used the entire master cohort.

Agreement at observed repeat creatinine tests

At every eligible target creatinine collection, we withheld its measured value and retrieved the most recent earlier result whose documentation time preceded that collection. Prior records had to have a valid collection-to-documentation order. Multiple results at the same collection time were represented by their mean, with availability at the latest documentation time in that group, matching the frozen extraction. Targets without an admissible prior value or whose latest prior value was more than 24 hours old were excluded from agreement calculations and counted in the flow. The target was used only for evaluation.

We computed the carried-minus-observed difference, its absolute value, the proportion with an absolute difference of at least 0.3 mg/dL, and disagreement in creatinine-stage proxies. Creatinine is reported in source units; 1 mg/dL corresponds to approximately 88.4 µmol/L. Prior-value age was measured from collection, not from result availability. Nonoverlapping age groups were 0–4, > 4–8, > 8–12 and > 12–24 hours. Cumulative limits were reported separately.

Both the carried and observed values were classified against identical reference minima derived from earlier results already documented at the target collection time. The reference windows were seven days and 48 hours. Stage 3 was assigned for a value at least three times the seven-day minimum, or at least 4 mg/dL together with either an increase of at least 0.3 mg/dL over the 48-hour minimum or a value at least 1.5 times the seven-day minimum. Stage 2 required at least twice the seven-day minimum; stage 1 required at least 1.5 times that minimum or an increase of at least 0.3 mg/dL over the 48-hour minimum. Severe proxy status meant stage 2 or 3. These thresholds were adapted from the creatinine component of KDIGO guidance [35]; reference choice is consequential in acute kidney injury research [36]. These creatinine-only proxies exclude urine criteria, clinical adjudication and dialysis-based staging [37]; they do not establish persistent kidney injury or renal recovery [38, 39]. A difference of 0.3 mg/dL between two values is an agreement threshold here, not by itself an AKI diagnosis.

Primary summaries weighted each observed repeat test equally. Sensitivities gave each stay equal weight within an age group, retained only the first eligible stay per person, and described groups defined by sex, age below versus at least 75 years, and coded chronic kidney disease (CKD). All repeat analyses excluded targets after recorded RRT initiation. These were descriptive comparisons without interaction testing. Agreement conditional on remeasurement was kept separate from panel availability: the retested measurements do not identify the unknown values in newly completed panels.

Hospital interfaces

The eICU analysis used all 200859 source stays across 208 hospitals and a separately coded chronic-disease candidate population of 18737 stays across 198 hospitals. The populations were not harmonized with the MIMIC cohort. Coverage was the fraction of stays with core vital signs, creatinine, urine documentation or a generic intake interface. Loose, base and strict screens required at least 50, 50 and 100 stays and respective coverage thresholds of 70/50/40/20%, 80/60/50/30% and 90/80/70/50% for these four domains. Thresholds were fixed feasibility screens, not validated measures of hospital quality. Generic intake documentation was distinguished from validated interval-specific crystalloid doses; no external replication of the MIMIC measurement experiments was claimed.

Statistical analysis and reproducibility

We used all eligible local records without outcome-based sample-size selection. New estimates received percentile 95% intervals from 2000 bootstrap resamples of people, retaining all their included stays and tests; the seed was 20260930. For repeat tests, the resampled statistic was the ratio of total error or event count to total test count, preserving the test-weighted estimand. Equal-stay sensitivities first averaged within stay and age group. Confidence intervals quantify sampling variation within this cohort, not uncertainty about source measurement validity or transportability. We did not conduct significance tests, optimize a carry limit or estimate treatment effects.

Selection was described using standardized differences between observable and other stays, with the difference in means divided by the square root of the average group variance. Outcomes were not used to choose observation rules. All age limits and grid phases were retained, consistent with exposing consequential analytic choices [40]. Code-level checks covered temporal ordering, unique target keys, unchanged cohort membership, dose conservation and independent recomputation of target pairs and dose windows. DuckDB 1.5.6 processed restricted Parquet extracts locally; public outputs contain aggregates only.

Use of generative AI

OpenAI Codex assisted with code development and debugging, literature checking, figure construction, and manuscript grammar editing. Numerical results were calculated from local records. Tianyi Yu is responsible for the scientific interpretation and final submitted content.

Results

Cohort and measurement flow

The master cohort contained 12175 stays from 10183 people. Mean age was 70.84 years (standard deviation 11.63); 41.93% of stays were in women, 35.10% had coded CKD and 42.65% had coded heart failure. The original exposure population comprised 10939 stays from 9302 people. Requiring trajectory support through 15 hours retained 10717 stays from 9141 people. Figure 1 summarizes the design and Table 1 distinguishes the units used in each experiment.

The repeat-test audit identified 39807 eligible target collections in 11923 stays. Of these, 1901 had no admissible earlier result and 2825 had an earlier result more than 24 hours old. Agreement analyses therefore used 35081 pairs in 11620 stays from 9772 people; 8650 pairs used a prior measurement collected before ICU admission. The 555 master stays without an evaluated pair had higher descriptive in-hospital mortality than stays contributing pairs, 24.32% versus 13.19%. This selection difference limits extrapolation of repeat-test agreement to the whole cohort.

Temporal resolution changed a fixed exposure

On the common 15-hour population, the four grid phases produced mean absolute dose differences of 1.416, 1.410, 0.580 and 1.498 mL/kg, respectively, relative to recorded-event interval overlap. Dose-category disagreement was 13.81% (95% interval 13.16–14.45), 12.60% (11.98–13.22), 5.66% (5.21–6.12) and 16.01% (15.30–16.67) (Fig. 2; Table 2). The target exposure duration remained six hours in every comparison. The phase with the smallest discrepancy was not an exact reconstruction and does not identify a universally preferable grid origin.

Table 2
Fixed-target reconstruction in the common fifteen-hour population

Phase (h)	Signed difference (mL/kg)	Absolute difference (mL/kg)	Category disagreement (%)
0	0.350 (0.292–0.405)	1.416 (1.365–1.463)	13.81 (13.16–14.45)
1	0.180 (0.124–0.233)	1.410 (1.358–1.459)	12.60 (11.98–13.22)
2	-0.103 (-0.137–-0.071)	0.580 (0.547–0.611)	5.66 (5.21–6.12)
3	0.119 (0.061–0.174)	1.498 (1.450–1.546)	16.01 (15.30–16.67)

Values are estimates with 95% patient-cluster intervals. All four phases use 10717 stays in 9141 people. Signed difference is reconstructed minus recorded-event interval dose. Categories are 0, > 0–5, > 5–10 and > 10 mL/kg.

Results were similar in the original 10939-stay population, where phase-specific category disagreement ranged from 5.62% to 16.02%. Assigning all volume to administration start time instead of prorating the recorded interval changed categories in 24.17% (23.35–25.00) of these stays. The historical [4,12) comparator changed categories in 20.73%; its larger volume reflected the additional two hours and was not evidence of treatment differences. Shifted-window, allocation and common-follow-up results are supplied in Additional file 1.

Timely documentation and record reuse determined panel availability

The first 48 hours contained 114929 complete four-hour bins. Without carry-forward, 23293 bins met the strict operational rule and 146 master stays contained four consecutive eligible panels, 1.20% (95% interval 1.01–1.40). At eight, 12 and 24 hours, the corresponding stay counts were 2908, 6456 and 8755, or 23.89%, 53.03% and 71.91% (Fig. 3). The 12-hour interval was 52.15–53.93%; the 24-hour interval was 71.12–72.70%. Requiring timely weight availability corrected the preceding implementation's no-carry and 12-hour counts from 147 to 146 and from 6468 to 6456.

Observable and other stays differed in baseline composition, and those differences changed with the rule. Under no carry, the standardized difference in baseline creatinine was approximately 0.46; it was approximately 0.13 under the 12-hour rule. These are selection descriptions, not adjusted comparisons of clinical outcomes. The increase in eligible trajectories reflects permission to reuse information, not the acquisition of new measurements.

Within the same first-six-hour window, requiring timely documentation changed the MAP minimum in 9.91% of the 11891 stays with both summaries available and changed the below-65-mmHg classification in 1.81% (95% interval 1.56–2.04). The availability requirement removed MAP summaries from 68 stays and lactate summaries from 37 stays. Lactate threshold classification changed in 0.08% of 7754 paired summaries. Thus, the consequences of the documentation clock differed across variables and cannot be inferred merely from the number of delayed records.

Reused creatinine showed imperfect agreement at repeat testing

There were 3464, 11700, 9889 and 10028 evaluated pairs in the four prior-value age groups. Mean absolute differences were 0.155 (95% interval 0.143–0.171), 0.169 (0.164–0.174), 0.177 (0.172–0.182) and 0.188 (0.184–0.193) mg/dL. Absolute differences of at least 0.3 mg/dL occurred in 16.43% (15.07–17.86), 21.03% (20.11–21.93), 21.78% (20.88–22.68) and 24.10% (23.24–24.94) (Fig. 4; Table 3). Median absolute differences were 0.1 mg/dL in all groups; the 95th percentiles were 0.4, 0.5, 0.5 and 0.6 mg/dL. Large values were retained under the frozen extraction limits rather than removed because they increased error.

Table 3
Agreement with observed repeat creatinine by prior-value age

Age (h)	Pairs	Absolute difference (mg/dL)	Difference ≥ 0.3 mg/dL (%)	Severe-proxy disagreement (%)
0–4	3464	0.155 (0.143–0.171)	16.43 (15.07–17.86)	3.12 (2.53–3.77)
> 4–8	11700	0.169 (0.164–0.174)	21.03 (20.11–21.93)	3.85 (3.45–4.25)
> 8–12	9889	0.177 (0.172–0.182)	21.78 (20.88–22.68)	3.63 (3.26–4.01)
> 12–24	10028	0.188 (0.184–0.193)	24.10 (23.24–24.94)	3.11 (2.76–3.46)

Values are estimates with 95% patient-cluster intervals. Each observed repeat test receives equal weight. The target value is withheld from the carried value and both reference minima. The four age strata contain different retesting populations; results do not estimate agreement during unmeasured intervals.

Mean signed differences were small and did not provide reassurance about individual agreement: increases and decreases partly cancelled. Any-stage disagreement rose descriptively from 12.64% in the youngest age group to 17.21% in the oldest. Severe-proxy disagreement was 3.12%, 3.85%, 3.63% and 3.11%, showing no monotonic pattern. Across all 35081 pairs, there were 914 instances in which the observed value indicated severe status and the carried value did not, and 315 instances with the reverse classification. These are joint counts among evaluated tests, not conditional false-negative or false-positive rates. Full stage confusion counts are provided in Additional file 1.

Under a cumulative 12-hour limit, 25053 of 39807 target collections had an admissible prior value, and 20.69% of those pairs differed by at least 0.3 mg/dL. Under a 24-hour limit, the corresponding counts were 35081 and 21.67%. These laboratory-target coverage measures have different denominators from the panel-availability percentages and cannot be combined into a validated accuracy–coverage frontier.

Equal-stay weighting retained the broad pattern of increasing absolute-difference frequency across the age groups (Fig. 5). First-stay, sex and age-group summaries are supplied in Additional file 1. Coded CKD was associated descriptively with larger mean absolute differences, particularly at older prior-value ages. These comparisons reflect differing patients and testing patterns; they do not isolate an effect of waiting longer or establish subgroup-specific safe limits.

Interface coverage depended on the population denominator

In eICU, 96, 76 and 57 of 208 hospitals passed the loose, base and strict all-stay screens. Applying the same screens within the chronic-disease candidates yielded 54, 41 and 22 of 198 hospitals (Fig. 6). The base candidate screen covered 8183 of 18737 candidate stays. Interface presence did not resolve whether fluid entries represented cumulative totals or interval delivery, so these results provide a transportability warning about data support rather than an external validation of dose reconstruction.

Discussion

This study identifies three distinct ways that a usable-looking critical care dataset can differ from its underlying records. Aggregation altered a fixed crystalloid exposure even when the target duration and patients were unchanged. Permission to reuse creatinine greatly expanded operational panel eligibility. Where a subsequent measurement existed, the reused value sometimes differed enough to change a renal-stage proxy. The conclusion is not that one representation is universally invalid; it is that completeness, temporal resolution and agreement require separate evidence.

Why the fixed-target comparison matters

Comparing an eight-hour total with a six-hour total largely demonstrates that additional time can add fluid. The grid-phase experiment asks a more informative question: how much of a fixed six-hour target survives compression into four-hour bins? Its disagreement persisted on a common population observed through the latest source-bin endpoint. The result therefore cannot be explained solely by adding target duration or using different patients for different phases. Nevertheless, the reference is itself based on uniform delivery within recorded administration intervals. Pump traces or independent clinical records would be needed to establish physical delivery accuracy.

The practical implication is to preserve event-level timestamps until the exposure window is specified, then calculate interval overlap directly. If only preaggregated bins are available, a sensitivity analysis should describe how reconstruction depends on bin origin and allocation assumptions. A single convenient grid may hide information loss that is consequential for discretized treatment variables. This matters for sequential decision research, including approaches that learn from discretized clinical trajectories [41, 42, 43], without implying that our results invalidate any particular policy or estimate its clinical performance. Versioned pipelines help make these choices reproducible [25]; the present comparisons address the additional question of what their temporal representation preserves.

Availability and agreement answer different questions

Increasing the carry limit from no carry to 12 hours changed four-consecutive-panel eligibility from 1.20% to 53.03%. That increase is a property of the observation rule. The repeat-test analysis adds an empirical check: a prior value no more than 12 hours old differed from the observed repeat value by at least 0.3 mg/dL in approximately one-fifth of eligible pairs. These findings should be read together, but they cannot be multiplied or plotted as if they measured the same population. Repeat tests occur when clinical care generates them; newly completed missing panels have no observed comparison value.

The nonmonotonic severe-proxy disagreement is particularly instructive. Older values had larger mean absolute differences across the observed age groups, yet severe-status disagreement did not rise in parallel. Threshold crossings depend on starting values, prior minima and the frequency of severe observations, in addition to numerical error. More frequently tested patients may also differ from those tested less often [19]. Formal analyses of informative presence show why the observation process can

induce selection and why its consequences depend on the relationship between measurement and clinical state [44, 45]. Selecting a universal carry limit from the smallest severe-disagreement percentage would therefore confuse a selected comparison with a validated operating rule.

The results suggest reporting two audit tables during dataset construction: one describing how an observation rule changes eligible patients and time points, and another assessing agreement at observed repeat measurements under that rule. The latter should disclose its retesting population, reference construction and weighting. This extends the use-specific emphasis of PRESERVE [23] to explicit within-stay reconstruction contrasts. A stronger study of unmeasured intervals would require a sampling design with scheduled measurements or another defensible source of contemporaneous reference information. Statistical imputation can be useful, but its assumptions require separate justification [32, 33, 34].

Implications for clinical research and transportability

Our findings concern the reliability of inputs to a clinical research question, rather than a bedside fluid recommendation. A subsequent treatment-response analysis would still need a clearly specified estimand, adequate confounding control, treatment indication and valid outcomes [21, 22]. Prediction would introduce additional requirements for transparent evaluation and calibration [46, 47]. Differences in who becomes observable also deserve attention: a model can inherit selection from a measurement rule before any modelling algorithm is fitted. Work on algorithmic inequity illustrates why the clinical meaning of a recorded proxy matters [48], but our descriptive sex and age comparisons do not constitute a fairness evaluation.

The eICU analysis adds a related caution at a different level. Fewer hospitals met the joint sample-size and coverage screens in the chronic-disease candidate population. Because restricting the denominator also reduces per-hospital sample size, these counts do not isolate a decline in interface coverage. Screening should use the intended denominator, and a recorded fluid field should not be equated with a harmonized dose [13, 18]. Because the present eICU definitions and interval semantics were not fully aligned with MIMIC, we have not treated this second database as a replication cohort. Establishing such replication would require harmonized timing and exposure semantics before applying the same measurement experiments.

Strengths and limitations

Strengths include paired exposure comparisons with a fixed target, explicit collection and documentation clocks, withholding of target creatinine values, shared prior-only reference minima, patient-cluster uncertainty intervals and preservation of all evaluated specifications. Internal review led to numerical corrections and removal of an inadequately supported urine-derived composite. Reporting these changes makes the final evidence chain more interpretable than retaining an appealing but unverified endpoint.

Several limitations constrain interpretation. The detailed experiments come from a single MIMIC source institution and a retrospectively coded population with diabetes and hypertension; other populations and information systems may differ. Documentation time is an operational proxy for availability in the extract and does not prove exactly what a clinician could see at the bedside. Event timing and laboratory values may contain source errors. The age restriction and prior-result rules exclude some target collections, while retesting and survival determine which comparisons are observed. Even an equal-stay analysis cannot remove this selection.

The chosen dose categories and agreement threshold aid interpretation but are not clinical decision thresholds validated by this study. Creatinine-stage proxies omit urine output, RRT-based severity and clinical adjudication; their disagreement cannot be equated with diagnostic error or harm. The panel definition omits important aspects of fluid tolerance and contains retrospectively reconstructed treatment. It is therefore not a validated real-time decision state. Finally, the review-driven analyses are exploratory methodological evaluations, not preregistered confirmatory tests. Their reproducibility supports examination and extension, rather than a claim that every source measurement or clinical construct is correct.

Conclusions

In this critical care measurement study, four-hour aggregation changed dose classification for a fixed six-hour crystalloid target, and creatinine reuse greatly increased operational panel availability. At observed retesting times, carried values showed nontrivial numerical and stage-proxy disagreement, with patterns that depended on the metric and retesting population. Temporal resolution, information availability and repeat-measurement agreement should be audited separately before the resulting trajectories are used for treatment-response research. The evidence does not establish a universally safe carry-forward interval or the effectiveness of a fluid strategy.

Abbreviations

AKI: acute kidney injury; CKD: chronic kidney disease; EHR: electronic health record; ICU: intensive care unit; MAP: mean arterial pressure; RECORD: REporting of studies Conducted using Observational Routinely-collected health Data; RRT: renal replacement therapy; STROBE: Strengthening the Reporting of Observational Studies in Epidemiology.

Declarations

Ethics approval and consent to participate

This study was a secondary analysis of deidentified MIMIC-IV and eICU records, with no participant recruitment, contact or intervention. The source database publications describe the relevant institutional oversight, deidentification and consent arrangements [12, 13, 27, 28]. The author states that no

additional ethics approval was required for this secondary analysis. No new ethics approval or exemption certificate was obtained for the present work. Access and use remain subject to the respective PhysioNet credentialing and data-use agreements.

Consent for publication

Not applicable. No identifiable individual information is reported.

Availability of data and materials

MIMIC-IV version 3.1 and eICU version 2.0 are available through PhysioNet to credentialed users who complete the required training and data-use agreements [27, 28]. Restricted patient-level data and derived individual records are not redistributed. Additional file 2 provides the aggregate results, analytic specifications and code required to inspect the reported findings and reproduce the analyses with authorized source access. No public repository deposition is claimed.

Competing interests

The author declares no competing interests.

Funding

This research received no external funding.

Authors' contributions

Tianyi Yu is the sole author and is responsible for the study conception, methodology, analysis, interpretation and manuscript. No other individual is claimed as an author.

Acknowledgements

The author thanks the patients, care teams and database contributors whose records and curation made MIMIC-IV and eICU available for research.

Clinical trial registration

Not applicable. This study did not assign an intervention.

References

1. Myburgh JA, Mythen MG. Resuscitation fluids. N Engl J Med. 2013;369(13):1243–51. 10.1056/nejmra1208627.
2. Semler MW, Self WH, Wanderer JP, Ehrenfeld JM, Wang L, Byrne DW, et al. Balanced Crystalloids versus Saline in Critically Ill Adults. N Engl J Med. 2018;378(9):829–39. 10.1056/nejmoa1711584.
3. Self WH, Semler MW, Wanderer JP, Wang L, Byrne DW, Collins SP, et al. Balanced Crystalloids versus Saline in Noncritically Ill Adults. N Engl J Med. 2018;378(9):819–28. 10.1056/nejmoa1711586.
4. Meyhoff TS, Hjortrup PB, Wetterslev J, Sivapalan P, Laake JH, Cronhjort M, et al. Restriction of Intravenous Fluid in ICU Patients with Septic Shock. N Engl J Med. 2022;386(26):2459–70. 10.1056/nejmoa2202707.
5. National Heart, Lung, and Blood Institute Prevention and Early Treatment of Acute Lung Injury Clinical Trials Network, Shapiro NI, Douglas IS, Brower RG, et al. Early Restrictive or Liberal Fluid Management for Sepsis-Induced Hypotension. N Engl J Med. 2023;388(6):499–510. 10.1056/nejmoa2212663.
6. Finfer S, Micallef S, Hammond N, Navarra L, Bellomo R, Billot L, et al. Balanced Multielectrolyte Solution versus Saline in Critically Ill Adults. N Engl J Med. 2022;386(9):815–26. 10.1056/nejmoa2114464.
7. Zampieri FG, Machado FR, Biondi RS, Freitas FGR, Veiga VC, Figueiredo RC, et al. Effect of Intravenous Fluid Treatment With a Balanced Solution vs 0.9% Saline Solution on Mortality in Critically Ill Patients: The BaSICS Randomized Clinical Trial. JAMA. 2021;326(9):818–29. 10.1001/jama.2021.11684.
8. National Heart, Lung, and Blood Institute Acute Respiratory Distress Syndrome (ARDS) Clinical Trials Network, Wiedemann HP, Wheeler AP, Bernard GR, et al. Comparison of two fluid-management strategies in acute lung injury. N Engl J Med. 2006;354(24):2564–75. 10.1056/nejmoa062200.
9. Monnet X, Marik PE, Teboul JL. Prediction of fluid responsiveness: an update. Ann Intensive Care. 2016;6(1):111. 10.1186/s13613-016-0216-7.
10. Marik PE, Cavallazzi R. Does the central venous pressure predict fluid responsiveness? An updated meta-analysis and a plea for some common sense. Crit Care Med. 2013;41(7):1774–81. 10.1097/ccm.0b013e31828a25fd.
11. Douglas IS, Alapat PM, Corl KA, Exline MC, Forni LG, Holder AL, et al. Fluid Response Evaluation in Sepsis Hypotension and Shock: A Randomized Clinical Trial. Chest. 2020;158(4):1431–45. 10.1016/j.chest.2020.04.025.
12. 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):1. 10.1038/s41597-022-01899-x.
13. Pollard TJ, Johnson AEW, Raffa JD, Celi LA, Mark RG, Badawi O. The eICU Collaborative Research Database, a freely available multi-center database for critical care research. Sci Data. 2018;5:180178. 10.1038/sdata.2018.178.

14. Goldberger AL, Amaral LA, Glass L, Hausdorff JM, Ivanov PC, Mark RG, et al. PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. Circulation. 2000;101(23):E215–20. 10.1161/01.cir.101.23.e215.
15. Johnson AE, Pollard TJ, Shen L, Lehman LW, Feng M, Ghassemi M, et al. MIMIC-III, a freely accessible critical care database. Sci Data. 2016;3:160035. 10.1038/sdata.2016.35.
16. Pollard T, Moody BE, Lehman LH, Gow BJ, Fernandes C, Xie C, et al. PhysioNet as a global platform for biomedical research. Nat Health. 2026;1(8):792–5. 10.1038/s44360-026-00096-z.
17. Weiskopf NG, Weng C. Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. J Am Med Inf Assoc. 2013;20(1):144–51. 10.1136/amiajnl-2011-000681.
18. Kahn MG, Callahan TJ, Barnard J, Bauck AE, Brown J, Davidson BN, et al. EGEMS (Wash DC). 2016;4(1):1244. 10.13063/2327-9214.1244. A Harmonized Data Quality Assessment Terminology and Framework for the Secondary Use of Electronic Health Record Data.
19. Agniel D, Kohane IS, Weber GM. Biases in electronic health record data due to processes within the healthcare system: retrospective observational study. BMJ. 2018;361:k1479. 10.1136/bmj.k1479.
20. Kapoor S, Narayanan A. Leakage and the reproducibility crisis in machine-learning-based science. Patterns (N Y). 2023;4(9):100804. 10.1016/j.patter.2023.100804.
21. Hernán MA, Robins JM. Using Big Data to Emulate a Target Trial When a Randomized Trial Is Not Available. Am J Epidemiol. 2016;183(8):758–64. 10.1093/aje/kwv254.
22. Hernán MA, Dahabreh IJ, Dickerman BA, Swanson SA. Ann Intern Med. 2025;178(3):402–7. 10.7326/annals-24-01871. The Target Trial Framework for Causal Inference From Observational Data: Why and When Is It Helpful?.
23. Razzaghi H, Goodwin Davies A, Boss S, Bunnell HT, Chen Y, Chrischilles EA, et al. PLOS Digit Health. 2024;3(6):e0000527. 10.1371/journal.pdig.0000527. Systematic data quality assessment of electronic health record data to evaluate study-specific fitness: Report from the PRESERVE research study.
24. Liu J, Pandya S, Coppi A, Young HP, Krumholz HM, Schulz WL, et al. Assessment of the integrity of real-time electronic health record data used in clinical research. PLoS ONE. 2026;21(1):e0340287. 10.1371/journal.pone.0340287.
25. Heumos L, Ehmele P, Treis T, Upmeier Zu Belzen J, Roellin E, May L, et al. An open-source framework for end-to-end analysis of electronic health record data. Nat Med. 2024;30(11):3369–80. 10.1038/s41591-024-03214-0.
26. Hoffmann S, Morris T, Herrmann M, Heinze G, Wynants L, Van Calster B, et al. Using routinely collected data for research purposes: challenges and mitigation strategies. BMJ. 2026;393:e087812. 10.1136/bmj-2025-087812.
27. Johnson A, Bulgarelli L, Pollard T, Gow B, Moody B, Horng S et al. MIMIC-IV (version 3.1). PhysioNet. 2024. 10.13026/kpb9-mt58

28. Pollard T, Johnson A, Raffa J, Celi LA, Badawi O, Mark R. eICU Collaborative Research Database (version 2.0). PhysioNet. 2019. 10.13026/C2WM1R
29. von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP, et al. Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. BMJ. 2007;335(7624):806–8. 10.1136/bmj.39335.541782.ad.
30. 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(10):e1001885. 10.1371/journal.pmed.1001885.
31. Che Z, Purushotham S, Cho K, Sontag D, Liu Y. Recurrent Neural Networks for Multivariate Time Series with Missing Values. Sci Rep. 2018;8(1):6085. 10.1038/s41598-018-24271-9.
32. Pedersen AB, Mikkelsen EM, Cronin-Fenton D, Kristensen NR, Pham TM, Pedersen L, et al. Missing data and multiple imputation in clinical epidemiological research. Clin Epidemiol. 2017;9:157–66. 10.2147/clep.s129785.
33. Sterne JA, White IR, Carlin JB, Spratt M, Royston P, Kenward MG, et al. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. BMJ. 2009;338:b2393. 10.1136/bmj.b2393.
34. White IR, Royston P, Wood AM. Multiple imputation using chained equations: Issues and guidance for practice. Stat Med. 2011;30(4):377–99. 10.1002/sim.4067.
35. Kidney Disease: Improving Global Outcomes (KDIGO) Acute Kidney Injury Work Group. KDIGO Clinical Practice Guideline for Acute Kidney Injury. Kidney Int Suppl. 2012;2:1–138. Available from: https://kdigo.org/wp-content/uploads/2016/10/KDIGO-2012-AKI-Guideline-English.pdf
36. Siew ED, Matheny ME. Choice of Reference Serum Creatinine in Defining Acute Kidney Injury. Nephron. 2015;131(2):107–12. 10.1159/000439144.
37. Kellum JA, Sileanu FE, Murugan R, Lucko N, Shaw AD, Clermont G. Classifying AKI by Urine Output versus Serum Creatinine Level. J Am Soc Nephrol. 2015;26(9):2231–8. 10.1681/asn.2014070724.
38. Chawla LS, Bellomo R, Bihorac A, Goldstein SL, Siew ED, Bagshaw SM, et al. Acute kidney disease and renal recovery: consensus report of the Acute Disease Quality Initiative (ADQI) 16 Workgroup. Nat Rev Nephrol. 2017;13(4):241–57. 10.1038/nrneph.2017.2.
39. Forni LG, Darmon M, Ostermann M, Oudemans-van Straaten HM, Pettilä V, Prowle JR, et al. Renal recovery after acute kidney injury. Intensive Care Med. 2017;43(6):855–66. 10.1007/s00134-017-4809-x.
40. Steegen S, Tuerlinckx F, Gelman A, Vanpaemel W. Increasing Transparency Through a Multiverse Analysis. Perspect Psychol Sci. 2016;11(5):702–12. 10.1177/1745691616658637.
41. Komorowski M, Celi LA, Badawi O, Gordon AC, Faisal AA. The Artificial Intelligence Clinician learns optimal treatment strategies for sepsis in intensive care. Nat Med. 2018;24(11):1716–20. 10.1038/s41591-018-0213-5.
42. Gottesman O, Johansson F, Komorowski M, Faisal A, Sontag D, Doshi-Velez F, et al. Guidelines for reinforcement learning in healthcare. Nat Med. 2019;25(1):16–8. 10.1038/s41591-018-0310-5.

43. Liu S, See KC, Ngiam KY, Celi LA, Sun X, Feng M. Reinforcement Learning for Clinical Decision Support in Critical Care: Comprehensive Review. J Med Internet Res. 2020;22(7):e18477. 10.2196/18477.
44. McGee G, Haneuse S, Coull BA, Weisskopf MG, Rotem RS. On the Nature of Informative Presence Bias in Analyses of Electronic Health Records. Epidemiology. 2022;33(1):105–13. 10.1097/ede.0000000000001432.
45. Harton J, Mitra N, Hubbard RA. Informative presence bias in analyses of electronic health records-derived data: a cautionary note. J Am Med Inf Assoc. 2022;29(7):1191–9. 10.1093/jamia/ocac050.
46. 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.
47. Van Calster B, McLernon DJ, van Smeden M, Wynants L, Steyerberg EW. Topic Group 'Evaluating diagnostic tests and prediction models' of the STRATOS initiative. Calibration: the Achilles heel of predictive analytics. BMC Med. 2019;17(1):230. 10.1186/s12916-019-1466-7.
48. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. Science. 2019;366(6464):447–53. 10.1126/science.aax2342.

Figures

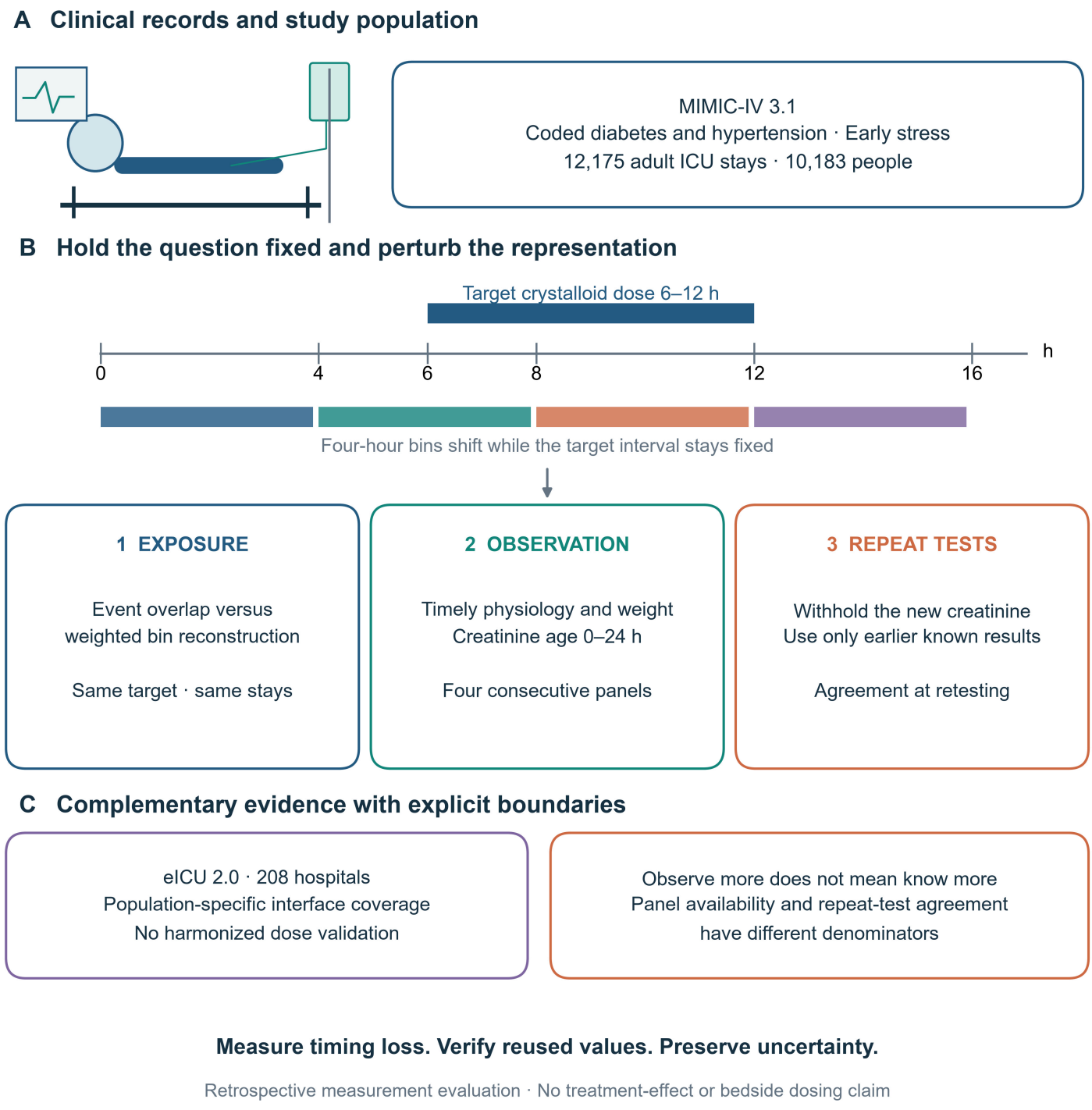


Figure 1

Study design and complementary measurement questions. The original schematic shows the clinical record source, fixed crystalloid target and three within-stay audits, with eICU interface coverage as a separate analysis. Panel availability and repeat-test agreement use different denominators. The patient illustration is schematic and contains no patient image.

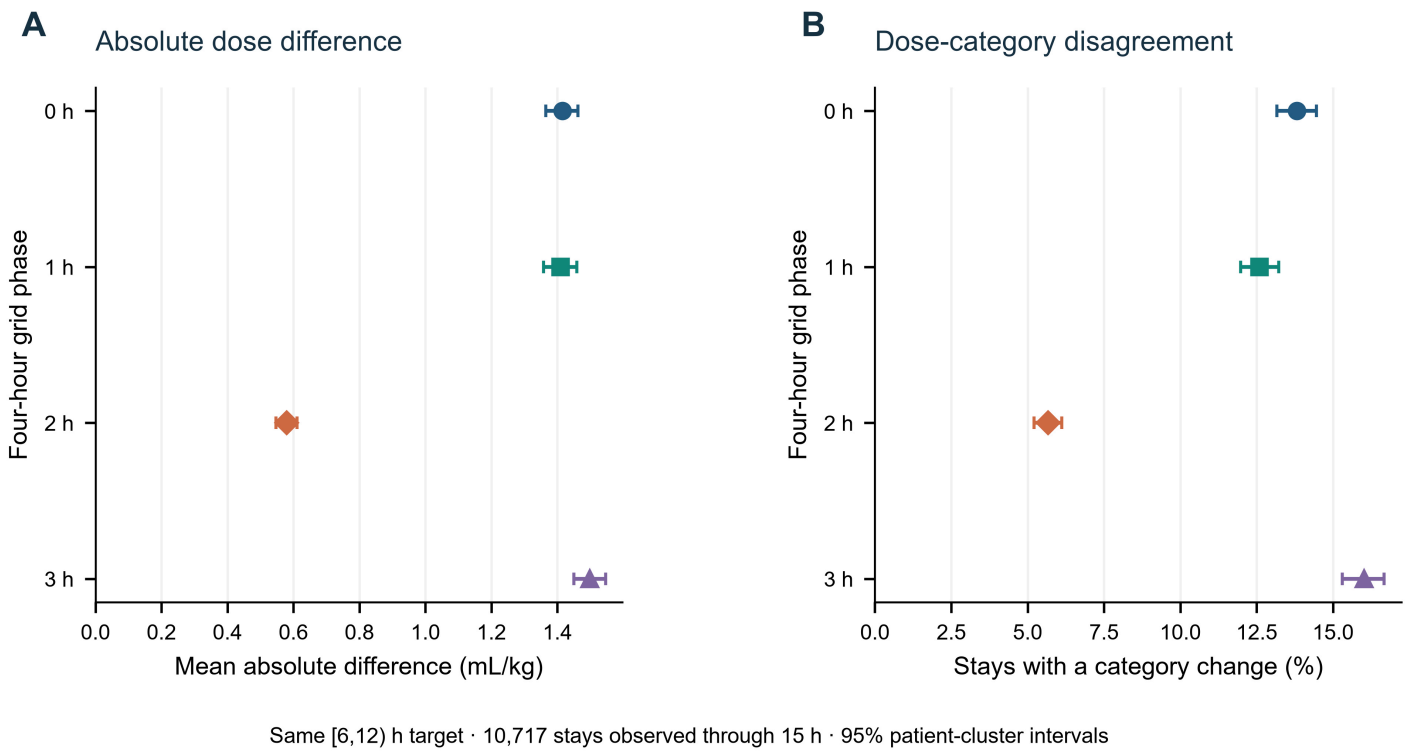


Figure 2

Fixed-target dose reconstruction across four-hour grid phases. All reconstructions target [6,12) hours and use the same 10717 stays observed through 15 hours. A, mean absolute difference from recorded-event interval overlap. B, disagreement in descriptive dose categories. Symbols identify phases redundantly with colour. Horizontal intervals are 95% patient-cluster bootstrap intervals. Neither representation is pump-validated truth.

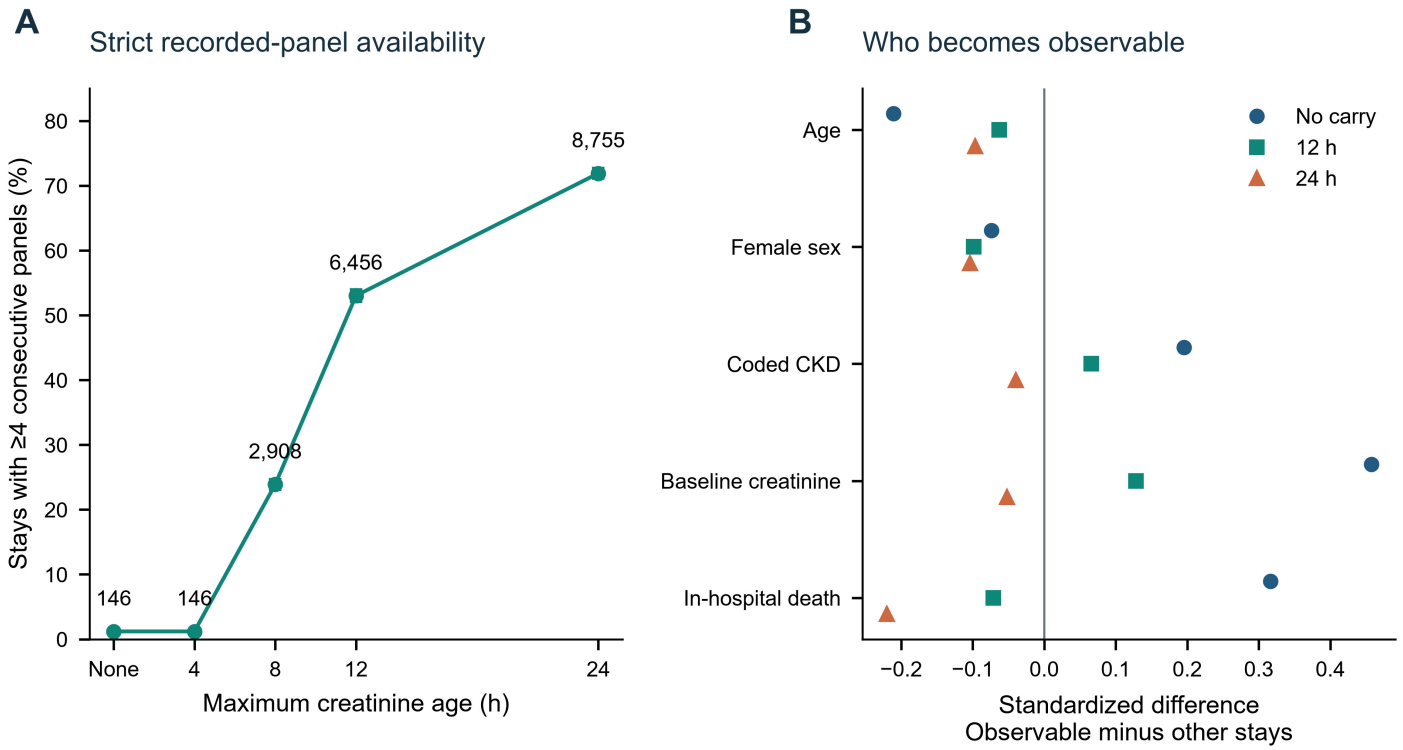


Figure 3

Operational availability and selection under creatinine age limits. A, percentage of 12175 master stays with at least four consecutive eligible four-hour panels; labels show stay counts. B, standardized differences between observable and other stays. Timely weight, vital signs, urine documentation and creatinine are required, whereas fluid dose is reconstructed retrospectively. Intervals in A use patient-cluster bootstrap resampling. Recorded urine presence does not establish complete collection.

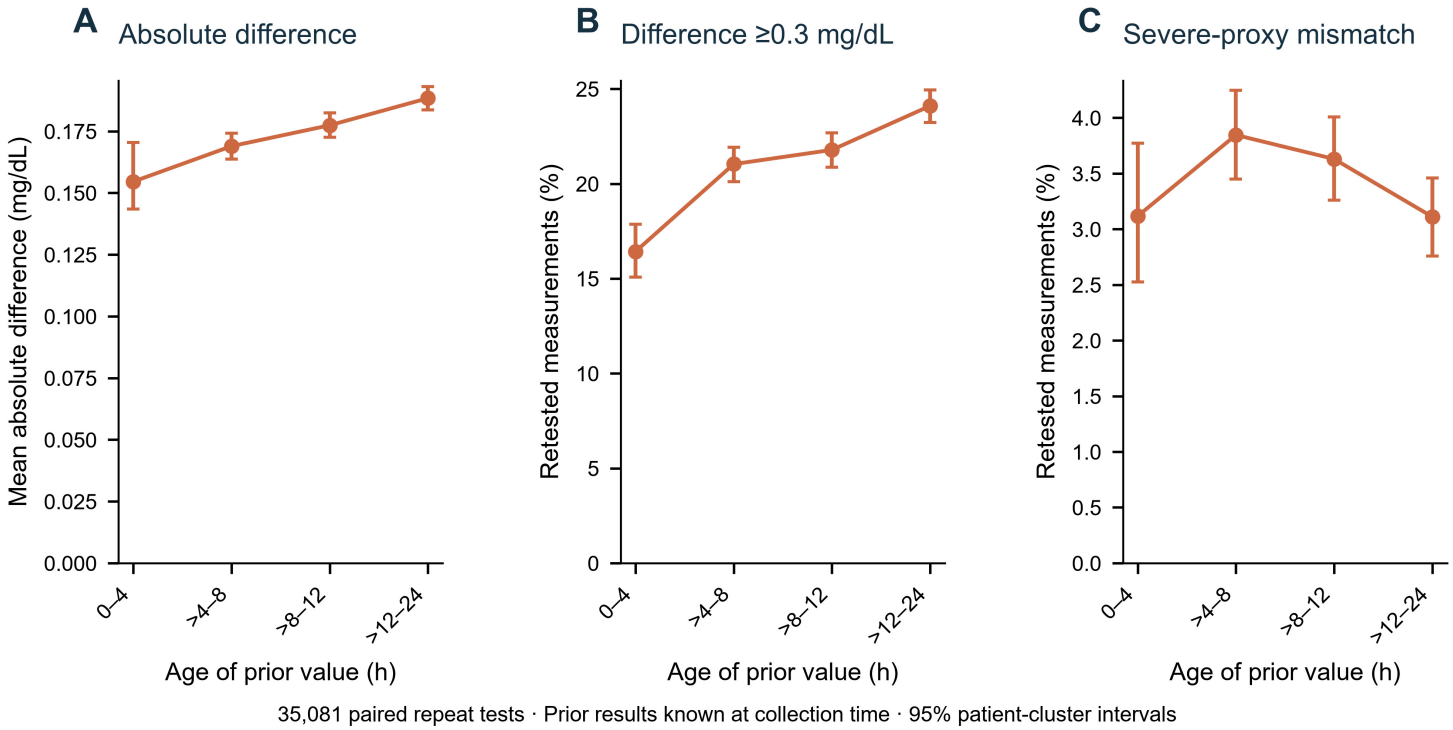


Figure 4

Agreement with observed repeat creatinine values. A, mean absolute difference. B, proportion differing by at least 0.3 mg/dL. C, severe creatinine-stage proxy disagreement using the same prior-only reference minima. The four age groups contain 3464, 11700, 9889 and 10028 paired measurements. Intervals are 95% patient-cluster bootstrap intervals. Lines connect descriptive strata, not a causal response to delaying testing; all findings are conditional on remeasurement.

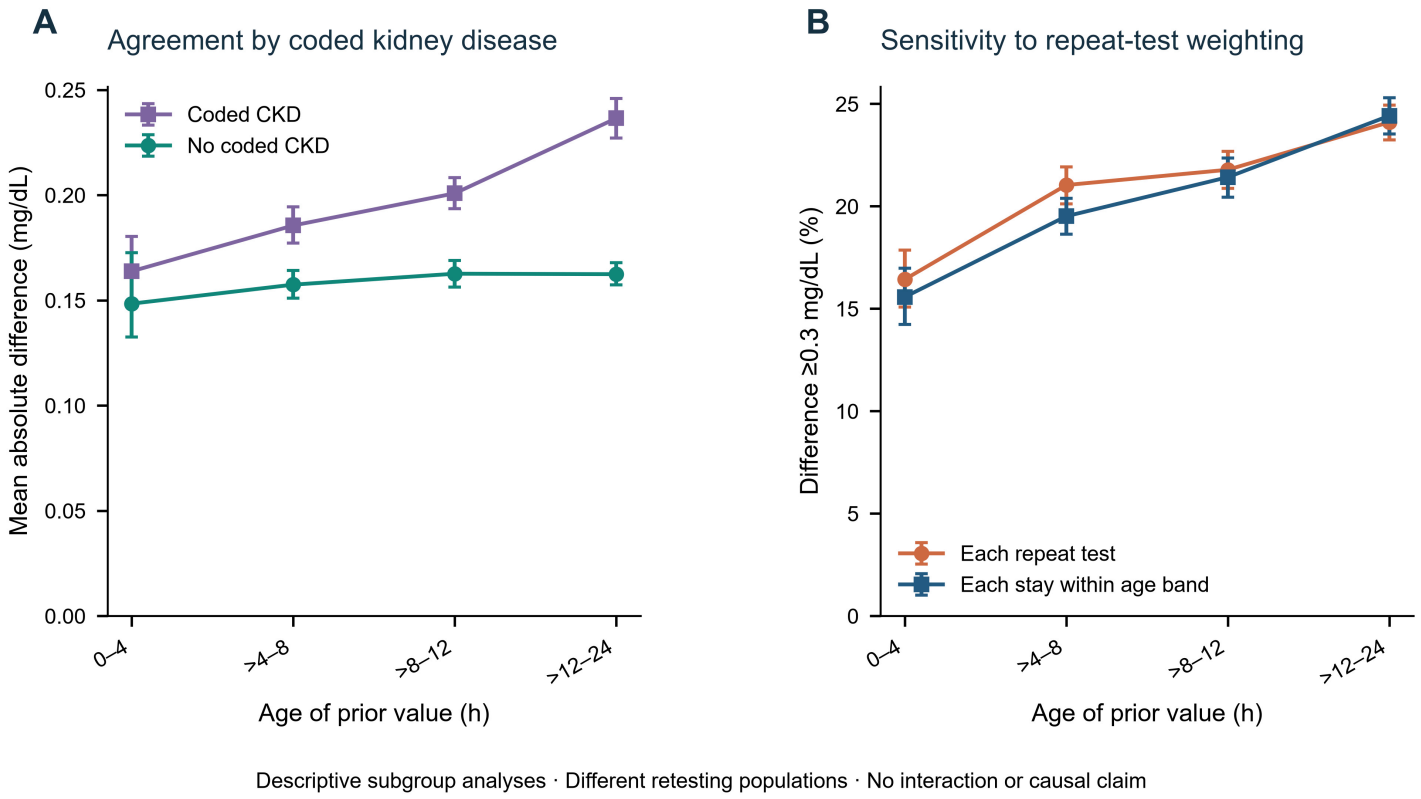


Figure 5

Repeat-test agreement in sensitivity analyses. A, mean absolute difference by coded CKD. B, frequency of absolute differences of at least 0.3 mg/dL with equal-test versus equal-stay-within-age-band weighting. Intervals are 95% patient-cluster bootstrap intervals. Subgroups differ in clinical composition and testing patterns; no interaction or causal effect is estimated.

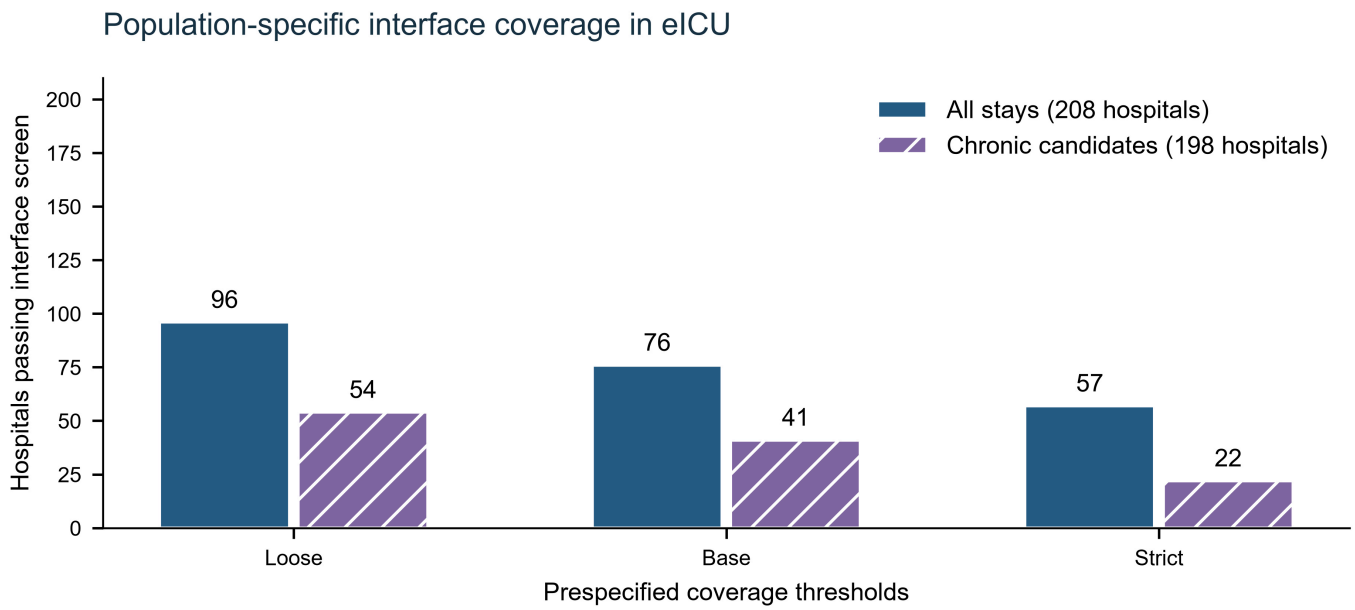


Figure 6

Population-specific hospital interface coverage in eICU. Bars show hospitals passing each fixed coverage screen in all source stays or chronic-disease candidates. Colour and hatching distinguish populations. Counts are descriptive and do not represent hospital performance, harmonized fluid measurement or replication of the MIMIC experiments.

Supplementary Files

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

- additionalfile1.docx