

Exposure-outcome window overlaps in biomarker studies: a simulation study and reporting checklist

Zhixia Bai

General Hospital of Ningxia Medical University

Yang Zhao

YZHAO@nxmu.edu.cn

General Hospital of Ningxia Medical University

Research Article

Keywords: landmark analysis, immortal time bias, prognostic biomarkers, outcome ascertainment, composite endpoints, interval censoring, acute kidney injury

Posted Date: September 25th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10653502/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.

A biomarker summarised over a window does not exist until that window closes, yet the outcome it is meant to predict can meet its study definition inside it. Analyses that count every event then mix contemporaneous severity with prediction, and nothing in the usual reporting makes the mixture visible. We define this exposure-outcome window overlap and measure the cost.

Methods.

We specify five quantities on one clock: time origin, exposure closing time, latent onset, first detection, and end of observation. We evaluated three analysis strategies in seven simulation scenarios, with all estimators and the oracle target expressed as post-window log hazard ratios. The framework was then applied to the single-centre INSPIRE dataset, using peak arterial lactate over 24 hours after cardiopulmonary bypass and a creatinine-only modified KDIGO outcome.

Results.

Under a global null, the rejection rates were 4.0%, 3.6%, and 2.6% for the naive, landmark, and interval-censored analyses, respectively. Overlap alone did not create bias. When the exposure was associated with onset only before the window closed, the biases were 0.369, 0.212, and 0.223; severity-dependent sampling increased them to 0.501, 0.267, and 0.151. Across scenarios, 28.6% to 44.6% of post-landmark detections had latent onset before the landmark. No method dominated. In the case study, 531 of 727 patients meeting the AKI definition (73.0%) had done so before the lactate summary existed. Among patients entering the landmark risk set, criterion-specific point estimates differed: the top-stratum hazard ratio was 3.36 for events involving the absolute-rise criterion and 0.56 for the ratio criterion alone, with the latter estimate imprecise.

Conclusions.

Window overlap changes the estimand, but bias depends on how associations and measurement differ before and after the exposure window. A landmark removes detected concurrent cases, not latent onsets that remain undetected at that time. Studies of early biomarkers should report how many events already met criteria when the summary became available, which components of a composite outcome remain open during follow-up, and how observation was maintained.

Background

A biomarker measured repeatedly is usually analysed as a summary: a peak, a mean, an area under a curve, a value at a nominated time. That summary does not exist until its window closes. Suppose the outcome it is meant to predict can meet its study definition inside that same window. Then some of the events being counted had already occurred by the time the predictor became available, and the estimate mixes contemporaneous severity with prediction in proportions that the reader cannot see.

We call this **exposure-outcome window overlap**. It is not a subtle problem, and it is not confined to one field. Any design in which a windowed summary is paired with an outcome ascertained from routine measurement is exposed to it: troponin and myocardial injury, procalcitonin and infection, ventilator settings and lung injury, and the windowed features that supply most electronic health record prediction models. What makes it easy to miss is that nothing in the usual reporting reveals it. Two studies with identical methods sections can differ entirely in how much of their association is prospective, depending on where each placed its time origin.

The problem belongs to a family that epidemiology already knows. Immortal time bias arises when a stretch of follow-up is misallocated because the outcome could not have occurred during that period [1], and it remains common enough that a recent survey found that the reported consequences varied systematically between methodological and empirical papers [2]. Window overlap is a sibling and not the same thing. Immortal time concerns person-time during which an event was impossible; window overlap concerns events that happened and were counted before the exposure that supposedly predicted them existed. The two share a cause, which is a time origin chosen without asking what was knowable when, and they call for the same discipline about clocks.

Landmark analysis is the standard approach to the problem: define one origin, restrict to patients who are event-free at a fixed time, and count only what follows. Anderson and colleagues introduced it in 1983 to prevent responders in cancer trials from being credited with the survival time they needed to respond [3], and its relation to time-dependent Cox regression has since been worked out in detail [4, 5]. It is not new, and this paper does not propose a new estimator. What has been missing is an account of what the design choice actually costs, what a landmark does and does not repair, and what a study must report for a reader to judge the question for themselves.

This paper has three parts. We first define the quantities to be separated: the time origin, the moment when the exposure summary closes, the latent onset of the outcome, the first measurement that detects it, and the end of observation. We then evaluate three analysis strategies by simulation, because latent onset is never observed in real data and no single cohort can reveal what any method is getting wrong. Finally, we work the framework through a case study with an unusually clean instance of the problem: early postoperative lactate and acute kidney injury after cardiac surgery, in the publicly available INSPIRE dataset [6].

That case study is worth stating plainly. Hyperlactataemia after cardiopulmonary bypass has been associated with acute kidney injury, prolonged intensive care stay and death in many reports [7–26], although the evidence is not uniformly positive. Lactate is attractive as an early marker because it is

readily available and changes more rapidly than serum creatinine. Most studies report it peaks on the first postoperative day. Acute kidney injury defined by the KDIGO creatinine criteria [27] can be met during exactly that interval. The overlap is therefore near-total, and the outcome has a further complication: its components cease to be satisfiable at different times.

A second distinction runs through the case study. Association is not added prognostic information. A marker may track physiological stress while contributing nothing beyond age, baseline renal function, and operative severity, and an uncertain estimate of incremental discrimination does not establish that the increment is absent [28, 29]. We treat those as separate questions throughout.

Methods

Reported in accordance with STROBE [30]; completed checklist in Additional file 1. The three parts below correspond to the three parts of the paper: the framework, its simulation-based evaluation, and the case study that instantiates it.

Part 1. The framework

The estimand, and the clocks it depends on

A study of this shape needs five quantities defined on one clock, and confusion between them is what the overlap problem consists of.

Time origin $t = 0$ —the event from which every window is measured.

Exposure closing time W . The biomarker summary is a function of data on $[0, W]$ and does not exist for $W < 0$. Any claim that it predicts something is a claim about the interval after W .

Latent onset T . When the patient's state actually changes. It is never observed.

First detection D . The first measurement that satisfies the outcome definition. $D \geq T$ always, and the gap is set by when someone chose to measure, so T is interval censored in $[previous\ measurement, D]$.

Observation end C . When the patient is no longer measurable, through discharge, death, or the end of the study window, events after C are never recorded.

The **landmark risk set** is then the set of patients with $D > W$ and $C > W$, and the estimand we and others intend is the effect of the exposure on the outcome process after W within that set. Two things follow immediately. A design that counts events with $D \leq W$ is not estimating that quantity. And because the risk set is defined by D rather than by T , patients whose onset preceded W but whose first qualifying measurement came afterwards remain in it: a landmark on detection cannot remove prevalent cases, only detected ones.

Part 2. Evaluating the designs

Simulation study

We simulated the mechanism described above to estimate the cost of the design choice, since no real cohort reveals T . The study is specified and reported following ADEMP, which separates the aims, the data-generating mechanism, the estimands, the methods compared and the performance measures [31]. Additional file 7 sets out all five in full; what follows is a summary.

Severity was latent and drove the biomarker level, the number of biomarker draws, the outcome-sampling rate and onset. The exposure was the maximum of the observed draws, so sampling intensity was mechanically applied to it. Onset followed a piecewise-exponential hazard whose association with the biomarker was set separately before and after W . Detection occurred at the first outcome measurement at or after onset, and discharge ended observation. The outcome definition was held fixed across all scenarios so that window overlap, association timing, and observation could be varied without changing the event rule.

Seven scenarios varied the association before and after W , whether sampling depended on severity, and the outcome-sampling rate. Each used 500 replicates of 2,000 patients. Three analyses were compared on the same log hazard-ratio scale. The **naive Cox** analysis started at the original time zero and counted all detected events. The **landmark Cox** analysis began at W among patients without a detected event and still under observation. The **interval-censored Weibull** model used the same landmark risk set and the interval from the last negative measurement to the first positive one for each event, following the proportional-hazards treatment of interval-censored failure times [32]. The oracle target was the post-window Cox coefficient for the same observed exposure in a reference population of 200,000 patients, using exact latent onset and the same landmark risk-set definition. We report bias, Monte Carlo error, empirical and model-based standard errors, coverage of nominal 95% intervals, rejection rates, fit failures and the fraction of post-landmark detections whose onset preceded W . All scenarios used a fixed outcome definition. Code and full results are in Additional file 7.

Part 3. The case study

Everything from here to the end of the Results concerns one cohort. It illustrates the framework in a clinical dataset. The simulation supplies a mechanism-specific evaluation rather than a claim of universal performance.

Data source and ethics

INSPIRE (v1.4.2) is a de-identified single-centre perioperative dataset of 130,960 operations in 99,886 patients, distributed through PhysioNet under a data use agreement [6, 38]. It contains operation-level timing, laboratory results, vital signs, medications and registry-linked mortality. The source study was

approved by the Seoul National University Hospital Institutional Review Board (H-2210-078-1368), which waived the requirement for informed consent, and public release was approved by its Data Review Board (BD-R-2022-11-02) [6]. This secondary analysis used only de-identified records and involved no participant contact, recruitment or new consent.

This analysis was **not preregistered**. The estimands, outcome definitions, and interpretive gates were defined for the present reanalysis and recorded before computing the corresponding effect estimates. The record has no independent timestamp and is not evidence of priority. Additional file 2 distinguishes the initial analysis plan, analyses added during an internal methodological audit, and diagnostics performed after results were known. Code provenance and aggregate outputs are provided with the submission (see Availability of data and materials).

Generative AI tools, including OpenAI Codex and Anthropic Claude, assisted with code review, language editing and assembly of submission files. The authors specified the scientific question and estimands, executed or independently reproduced the analysis code, checked quantitative outputs against their sources, resolved discrepancies, and approved the final text. The tools were not treated as authors, references or independent statistical decision-makers.

Participants and unit of analysis

We included operations performed by the cardiothoracic service in patients aged ≥ 18 years, with CPB and bypass initiation, separation, end of surgery, and anesthesia start all recorded. Patients also had to be alive at the 24-hour landmark and to have a peak lactate value, a preoperative baseline creatinine and at least one postoperative creatinine.

The unit of analysis is the patient, represented by their first eligible operation. Among 2,468 eligible operations from 2,365 patients, 98 patients contributed more than one operation (103 additional operations). Eighteen adjacent pairs were separated by ≤ 2 days and 22 by ≤ 7 days, with a minimum interval of 4.5 hours. Because laboratory records are linked at the patient level while exposure and outcome windows are defined for each operation, closely spaced operations can share the same laboratory values. In five patients, one death could also be attributed to two operations. This is an attribution problem, not merely within-patient correlation, so a cluster-robust standard error cannot solve it. We therefore used the first eligible operation in the primary analysis. We then repeated the model using all operations with cluster-robust standard errors and, separately, after excluding 21 first operations followed by another eligible operation within their 7-day renal window.

Throughout, counts are explicitly labelled as patients or operations.

Time origin

All times were referred to CPB separation. Bypass separation and end of surgery are not the same instant: their interval had a median of 100 minutes (IQR 75–135, range 10–620). We evaluated a unified bypass clock (primary), a unified end-of-surgery clock, and the hybrid clock used in earlier work, the last reported only to quantify reclassification.

Exposure

Peak arterial lactate between the origin and the 24-hour landmark. INSPIRE reports lactate on a grid of 17 values top-coded at 8.6 mmol/L, a property of de-identification, not of the assay. We used four ordered strata aligned with that grid (≤ 1.9 , 2.1–3.0, 3.8–5.9, 8.6, top-coded). The top category is a mixture of all patients above the cap, so its estimate is an average over that mixture and not a bound on any specific contrast.

Because a peak based on more measurements is more likely to be extreme, we prespecified alternative exposure definitions. These included peaks within 12 and 6 hours, the first value after the origin, and the 24-hour peak restricted to patients with 4–9 lactate results.

A peak is an order statistic, so adding a draw can only raise it. During the internal methodological audit, we added two summaries without that property, computed on the same window and clock. The **time-weighted average** is the integral of the lactate curve over the window divided by 24 hours, using a step function that holds each value until the next draw. The **lactate load** is the integral of the excess above 2.0 mmol/L [8]. Both carry values to the window edges, so for sparsely sampled patients they are largely extrapolation. Both are therefore reported with the proportion of the window actually observed (Additional file 6, Section 12a).

Outcomes

Newly detected AKI. We used a creatinine-only modification of the KDIGO definition [27]: a rise to $\geq 1.5\times$ baseline within 7 days or to ≥ 0.3 mg/dL within 48 hours. Baseline was defined as the most recent creatinine recorded at or before the start of anesthesia and no more than 180 days before that. That cap was never bound in practice; the oldest baseline actually used was 89.6 days, and the median was 2.0 days (IQR 1.0–3.7). Urine output was unavailable, and renal replacement therapy was not reliably captured. Because a remote baseline can misrepresent preoperative renal function, we reported the age at baseline and repeated the primary model among patients with a baseline recorded within 7 days before surgery.

We say "**newly detected**" where others say "incident", and "**first met the creatinine criterion**" where others say "occurred". The recorded time is the time of the first qualifying creatinine sample, not the onset of injury, which lies somewhere between that sample and the preceding one. The outcome is therefore interval-censored, and its detection depends on the sampling scheme.

AKI was classified as **concurrent** if first met at or before the landmark and **newly detected** if first met afterwards among patients free of concurrent AKI. A patient without an observed event was not

assumed event-free: follow-up ended at the last creatinine measurement, and death before AKI was a competing event.

Death is the only competing event in the primary analysis, and hospital discharge is not modelled separately. Once a patient goes home, creatinine testing usually stops, so discharge and the end of renal observation collapse into a single censoring time. The primary analysis treats that censoring as non-informative conditional on covariates. We tested this assumption using a censoring-hazard model, inverse probability weighting and a tipping-point analysis. We then recoded discharge as a competing event and, in a separate analysis, recoded all censoring times accordingly (Additional file 6, Sections 8-12b). We did not use discharge as a competing event in the primary model because it stops detection without biologically preventing AKI, and treating it as a competing event changes the estimand to the cumulative incidence among patients who remain hospitalised.

For the fixed-window binary analysis, **outcome status was treated as ascertainable** when the event was observed or when no event occurred, and the creatinine follow-up reached the horizon. This rule does not require a late creatinine from a patient whose qualifying event has already been recorded; doing so would discard events that are known to have occurred.

Mortality. The secondary contrast outcome was all-cause death after the landmark and by day 90 from the origin, based on registry-linked death time. INSPIRE does not provide an explicit administrative end date, so the assumption of complete 90-day ascertainment remains. Deaths were recorded out to 4,390 days across the cohort, placing the selected horizon well within the observed linkage span but not establishing that vital status was complete for every patient; a death not captured by the registry would be indistinguishable from survival.

Statistical analysis

Case-study outputs were generated in Python 3.14.5 (pandas 2.3.3, statsmodels 0.14.6, lifelines 0.30.3 and SciPy 1.18.0) and R 4.5.2 (cmprsk). The final simulation and submission package validation was performed using Python 3.12.13 with NumPy 2.4.6, pandas 2.3.3, matplotlib 3.11.0, SciPy 1.17.1, statsmodels 0.14.6, and lifelines 0.30.1. Continuous variables are median (IQR). Baseline balance is summarised by the maximum pairwise standardised difference rather than by P values, because Table 1 is descriptive.

Table 1

Baseline characteristics of the primary cohort by peak lactate stratum 2,365 patients, first eligible operation each. Continuous variables median (IQR); categorical n (%). Missing is the number of patients with no recorded value; all summaries and percentages are based on patients with a recorded value. Balance summarised by the maximum standardised difference among the six pairwise stratum comparisons; values above 0.10 indicate imbalance.

Variable	≤ 1.9	2.1-3.0	3.8-5.9	8.6 (capped)	Overall	Missing	Max SMD
Patients	815	570	765	215	2365		
Age, years	65 (55-70)	65 (55-75)	65 (55-75)	60 (40-70)	65 (55-70)	0	0.46
Female sex, n (%)	257 (31.5)	245 (43.0)	328 (42.9)	83 (38.6)	913 (38.6)	0	0.24
Weight, kg	65.0 (55.0-70.0)	60.0 (55.0-70.0)	60.0 (50.0-70.0)	60.0 (55.0-70.0)	60.0 (55.0-70.0)	4	0.18
ASA physical status ≥ 3, n (%)	540 (67.0)	379 (66.8)	531 (70.6)	144 (68.6)	1594 (68.3)	30	0.08
Emergency operation, n (%)	83 (10.2)	45 (7.9)	97 (12.7)	25 (11.6)	250 (10.6)	0	0.16
CPB duration, min	190 (140-230)	190 (141-235)	200 (155-255)	220 (180-262)	195 (150-240)	0	0.64
CPB separation to end of surgery, min	95 (75-110)	90 (75-120)	110 (80-155)	140 (95-200)	100 (75-135)	0	0.86
Preoperative creatinine, mg/dL	0.88 (0.74-1.15)	0.88 (0.74-1.05)	0.88 (0.74-1.05)	0.92 (0.79-1.05)	0.88 (0.74-1.05)	0	0.09
Lactate results in the 0-24 h exposure window, n	3 (1-7)	6 (2-8)	8 (6-10)	10 (8-13)	7 (2-9)	0	1.62
Creatinine results after the landmark, n	6 (5-8)	6 (5-9)	7 (5-11)	9 (5-13)	7 (5-10)	4	0.58
Concurrent AKI at the landmark, n (%)	101 (12.4)	94 (16.5)	224 (29.3)	112 (52.1)	531 (22.5)	0	0.94
ASA physical status is the only variable in this table that is also an adjustment covariate and the only covariate with any missing values. Weight is descriptive and does not enter the model.							

Primary effect. We estimated the cause-specific hazard of newly detected AKI with a Cox model; patients who died first left the cause-specific risk set. The event time is the time of a qualifying creatinine draw,

not of biological onset, so this model estimates the hazard of detection under the observed sampling process. Cumulative incidence was estimated with Aalen-Johansen, with death as a competing event, and compared using Gray's k-sample test [33]. Subdistribution hazard ratios [34] were estimated with `cmprsk::crr`; the Python implementation used for sensitivity analyses was validated against it (largest relative coefficient difference 0.0045%; Additional file 4). Proportional hazards was assessed using scaled Schoenfeld residuals.

Secondary effects. We fitted a logistic model in the fixed-window ascertainable subset. A model that codes incompletely followed patients as event-free changes the outcome definition and is reported only as a sensitivity analysis. Mortality was analysed using Firth penalised likelihood [35] with profile intervals, together with separation diagnostics and events per variable [36].

Covariate sets. A parsimonious set (age, preoperative creatinine, $\text{ASA} \geq 3$) and an expanded set (adding sex, emergency status, CPB duration; for mortality, also concurrent AKI). **These are the covariates available in this dataset, not a validated clinical risk model**; INSPIRE does not contain surgical category, ventricular function, cross-clamp time, transfusion or haemodynamic support, so no comparison with EuroSCORE II or STS models was possible.

Missing data. Covariate completeness is reported for each variable in Table 1. ASA physical status was missing for 30 of 2,365 patients (1.3%) and was the only covariate with any missing values; age, sex, emergency status, CPB duration and preoperative creatinine were complete. Missing ASA was coded to the reference category ($\text{ASA} < 3$) rather than excluded, so that the analysis set is identical across models. That is an assumption, not a neutral default, so the primary model was refitted under two alternatives: either excluding those patients (17 of the 1,831 at risk, carrying 4 events) or adding a missing indicator. The top-stratum cause-specific hazard ratio was 2.13 (95% CI 1.22–3.74) as reported, 2.07 (1.16–3.68) under complete-case analysis and 2.13 (1.22–3.73) with an indicator; the per-stratum trend was 1.25 under all three. No values were imputed.

Incremental value. The discrimination measure is the area under the curve of the Wilcoxon-Mann-Whitney statistic for a binary outcome, defined over a fixed window, computed from a Firth-penalised logistic model. It is not Harrell's censoring-aware C-index, not a time-dependent AUC at a nominated horizon, and not a competing-risk discrimination index; for a binary outcome, the AUC and the concordance statistic coincide, and we use the term C-index only in that sense. Decision curves are on the same fixed-window binary basis, so they carry no separate prediction horizon.

Uncertainty came from 500 patient-level bootstrap resamples. Models were refitted within each resample and scored only on the patients omitted by that resample. The interval is the 2.5th to 97.5th percentile of the resulting out-of-bag differences. A replicate was discarded if the resample held fewer than five events, if it left no out-of-bag patients, if the out-of-bag set was single-class, or if either fit failed to converge. None of these occurred; all 500 replicates contributed to every reported interval. We did not define an equivalence margin, so an interval containing zero means an increment was not established, not that it was shown to be absent. Because the comparator is the covariate set this dataset provides

rather than EuroSCORE II, an STS model, or any other validated instrument, these estimates address whether lactate adds to those covariates and cannot address whether it adds to a clinical risk model.

Sensitivity analyses. We varied the exposure definition, modelled the censoring hazard and applied inverse probability of censoring weights to the corrected ascertainable indicator. A tipping-point analysis assigned each censored patient a common probability δ of an undetected event. This explores a narrow, missing-not-at-random mechanism and does not rule out exposure-specific hidden risk. We also repeated the primary model after excluding operations with a subsequent eligible operation inside the renal window and after requiring a baseline creatinine within 7 days before surgery.

Results

Part 2. What the overlap costs

Simulation: what each design estimates when the truth is known

Table 6 and Fig. 2 give the simulation results. All 10,500 model fits succeeded. With 500 replicates, the largest Monte Carlo standard error for a reported proportion was 2.2 percentage points.

Table 6

Simulation: what each design estimates when the truth is known Each scenario used 500 replicates, each with 2,000 patients. The oracle target is the post-window Cox coefficient for the observed exposure in a 200,000-patient reference population with exact onset. Values in the three method columns are mean estimate [bias]; coverage of nominal 95% intervals. Rejection rates are shown only where the target is approximately zero. Monte Carlo standard errors and all standard-error diagnostics are reported in Additional file 7. No fit failed.

Scenario	Target	Naive Cox	Landmark Cox	Interval-censored Weibull	Earlier onset among post-landmark detections, %
N. Global null	-0.002	0.000 [+0.003]; 95.4%, reject 4.0%	-0.001 [+0.002]; 96.2%, reject 3.6%	0.001 [+0.003]; 97.4%, reject 2.6%	28.6
A. Strong early-only association	-0.002	0.367 [+0.369]; 0.0%, reject 100.0%	0.210 [+0.212]; 0.0%, reject 100.0%	0.220 [+0.223]; 0.0%, reject 100.0%	30.9
B. Early and post-window association	0.330	0.539 [+0.209]; 0.0%	0.441 [+0.111]; 9.0%	0.325 [-0.005]; 95.6%	32.4
C. Weak early-only association	0.001	0.118 [+0.116]; 0.2%, reject 100.0%	0.070 [+0.069]; 35.2%, reject 67.4%	0.075 [+0.074]; 31.2%, reject 70.6%	29.2
D. Early-only association, severity-dependent sampling	-0.001	0.501 [+0.501]; 0.0%, reject 100.0%	0.267 [+0.267]; 0.0%, reject 100.0%	0.150 [+0.151]; 0.0%, reject 100.0%	29.2
E. Post-window effect, sparse outcome measurement	0.328	0.400 [+0.071]; 34.0%	0.345 [+0.017]; 90.6%	0.332 [+0.004]; 95.4%	44.6
F. Constant association across the window	0.357	0.352 [-0.005]; 93.8%	0.349 [-0.009]; 92.4%	0.214 [-0.143]; 0.4%	30.2
Window overlap was present in the global-null scenario without material bias. Earlier latent onset became consequential when the association changed across the landmark or when sampling depended on severity. No estimator dominated across all mechanisms.					
Box 1. Reporting checklist for studies of a windowed biomarker and a measured outcome					
Each item costs a sentence or a number. Together they let a reader decide how much of a reported association could be prospective, which no current reporting guideline asks for.					

The global-null scenario is the necessary negative control. Mean estimates were 0.000, -0.001 and 0.001 for the naive Cox, landmark Cox and interval-censored Weibull analyses. Rejection rates were 4.0%, 3.6%,

and 2.6%, with 95% interval coverage of 95.4%, 96.2%, and 97.4%, respectively. Window overlap and delayed detection were present, yet none of the methods generated a material association. Overlap alone is therefore not a sufficient cause of bias.

Bias appeared when the exposure-onset association differed across the window. In the strong early-only scenario, where the oracle post-window target was -0.002 , mean estimates were 0.367 for the naive analysis, 0.210 for the landmark and 0.220 for the interval-censored model. Their biases were 0.369, 0.212 and 0.223, and coverage was 0% for all three. With a weaker early-only association, bias fell to 0.116, 0.069 and 0.074, but the null was still rejected in 100.0%, 67.4% and 70.6% of replicates. Severity-dependent sampling amplified the mismatch: corresponding biases were 0.501, 0.267 and 0.151.

A landmark improved estimation in those scenarios but did not remove latent prevalent cases. Across the seven mechanisms, 28.6% to 44.6% of events first detected after the landmark had their onset before the landmark. This quantity is not the bias itself. The global-null scenario had 28.6% such detections, with no meaningful bias. It became consequential when pre-window and post-window associations differed, or when measurement depended on severity.

No method was uniformly best. When a true post-window effect coexisted with an earlier association, the interval-censored model closely recovered the target in the standard- and sparse-sampling scenarios, with biases of -0.005 and 0.004 and coverages of 95.6% and 95.4%, respectively. In the constant-association scenario, however, the naive and landmark estimates were close to the target. In contrast, the interval-censored model was biased downward by 0.143 and covered the target in only 0.4% of replicates. The simulation therefore supports mechanism-specific diagnosis rather than a fixed ranking of estimators.

Part 3. The case study

Cohort

Of 130,960 operations, 2,512 met the source-population criteria, and 44 were then excluded: 35 had no lactate result inside the exposure window, seven died before the 24-hour landmark, one had no preoperative baseline creatinine, and one met none of those three requirements. The final operation-level cohort contained 2,468 operations in 2,365 patients; the primary cohort retained each patient's first eligible operation (Fig. 3B).

Two baseline patterns frame the analysis (Table 1). Concurrent AKI rose from 12.4% in the lowest lactate stratum to 52.1% in the highest. Lactate sampling also increased sharply, from a median of 3 results to 10. Because the exposure is a maximum, more measurements create more opportunities to enter a higher stratum. The observed lactate gradient therefore combines physiology, clinical concern and measurement opportunity.

Reconciling the time origin

Bypass separation preceded the end of surgery by a median of 100 minutes (IQR 75–135). The hybrid clock and the unified bypass clock classified AKI identically in 2,464 of 2,468 operations; four changed class, and none moved between concurrent and newly detected AKI. Using the end of surgery as the origin was not a like-for-like sensitivity analysis. It removed all exposure measurements in 810 operations and therefore changed the eligible population (Additional file 3; Table 2).

Table 2
Classification of renal events and reconciliation of the time origin

Units analysed	Primary cohort (patients)	All operations
	2,365 patients	2,468 operations (2,365 patients)
Met AKI criteria at any point	727	777
of which concurrent at the landmark	531 (73.0%)	570 (73.4%)
of which newly detected after the landmark	196	207
At risk after the landmark	1,834	1,898
Time-origin reconciliation (all operations)		
Hybrid vs unified CPB clock: eligible under both		2,468
of which AKI class changed		4 (0.16%)
of which switched between concurrent and newly detected		0
End-of-surgery clock: operations retaining an exposure value		1,662 of 2,476
of which lost because all lactate fell before end of surgery		810 (33%)
Repeat-operation diagnostics		
Patients with > 1 eligible operation		98 (103 extra operations)
Adjacent pairs ≤ 2 days / ≤ 7 days apart		18 / 22 (minimum 4.5 h)
Deaths attributable to two operations		5

Most AKI cases met the definition before the landmark

Among 2,365 patients, 727 met the creatinine definition at some point. By the 24-hour landmark, 531 (73.0%) had already done so. The proportion was similar when all eligible operations were retained (570 of 777, 73.4%). This left 1,834 patients in the post-landmark risk set, of whom 196 later had a qualifying creatinine result.

As expected, the concurrent fraction depended on where the exposure window ended. It was 6.7% at 6 hours, 34.1% at 12 hours, 73.0% at 24 hours and 91.5% at 48 hours (Additional file 5). These values describe the timing of case detection in this cohort. They are not estimates of the proportion of the association between an exposure and an outcome attributable to window overlap.

Association with newly detected AKI after the landmark

The landmark risk set is a conditioned population, and the conditioning depends on the exposure. Removing concurrent cases retained 87.6% of the lowest-lactate stratum but only 47.9% of the highest, so the strata entering the model are not the strata described in Table 1 (Table 4). Baseline characteristics of the risk set are given alongside it. Most covariates are distributed much as in the full cohort, and preoperative creatinine is, if anything, better balanced after conditioning (maximum standardised difference of 0.05 versus 0.09); the exposure-sampling imbalance remains unchanged (1.68 versus 1.62). Of the 1,834 patients at risk, 1,831 contributed post-landmark renal follow-up. The other three had no post-landmark creatinine at all: two in the lowest stratum and one in the 3.8–5.9 stratum. They contribute no time at risk, rather than being counted as event-free. There were 196 newly detected AKI events and seven deaths before AKI. Day 5 was the last horizon with roughly half of each stratum still under observation. At that point, cumulative incidence increased from 9.0% (95% CI 6.8–11.1) in the lowest stratum to 15.6% (8.6–22.6) in the highest (Fig. 4; Gray P = 0.038).

Table 4

The landmark risk set: who enters it, and what they look like Conditioning on being alive and free of AKI at 24 hours removes patients at a rate that rises steeply with the exposure, so the modelled population differs from the cohort in Table 1.

Lactate stratum	Primary cohort, n	Concurrent AKI, n (%)	Entering the risk set, n (%)
≤ 1.9	815	101 (12.4)	714 (87.6)
2.1-3.0	570	94 (16.5)	476 (83.5)
3.8–5.9	765	224 (29.3)	541 (70.7)
8.6 (capped)	215	112 (52.1)	103 (47.9)
All	2,365	531 (22.5)	1,834 (77.5)
Baseline characteristics of the 1,834 patients who enter it. Continuous variables median (IQR); categorical n (%). Percentages are over patients with the value recorded. Max SMD is the largest standardised difference among the six pairwise stratum comparisons, computed as for Table 1.			

Relative to the lowest stratum, adjusted cause-specific hazard ratios were 1.16 (95% CI, 0.79–1.69), 1.48 (95% CI, 1.04–2.10), and 2.13 (95% CI, 1.22–3.74). These describe the hazard of a qualifying creatinine being **recorded** in patients who are still event-free and under observation at 24 hours. They are not causal effects on the occurrence of kidney injury, and they are averaged over a follow-up period during which the outcome definition itself changes (see below). The per-stratum hazard ratio was 1.25 (1.08–1.45). Subdistribution estimates were nearly identical because competing deaths were uncommon.

When the analysis was restricted to 1,109 patients with ascertainable fixed-window outcome status, the corresponding adjusted odds ratios were 1.15 (0.76–1.74), 1.57 (1.06–2.31), and 1.77 (0.93–3.37) (Table 3). The wider interval in the highest stratum is important. The time-to-event and binary models address related but different estimands, and the latter discards patients without complete fixed-window ascertainment.

Table 3

Newly detected AKI by peak lactate stratum, primary cohort Adjusted models include age, preoperative creatinine, ASA ≥ 3 , sex, emergency status and CPB duration.

At risk (time-to-event set)	≤ 1.9	2.1–3.0	3.8–5.9	8.6 (capped)	Test
	712	476	540	103	
Newly detected AKI, n	62	49	69	16	
Cumulative incidence, day 1, %	5.6	6.6	9.1	13.6	
Cumulative incidence, day 5, % (95% CI)	9.0 (6.8–11.1)	9.9 (7.1–12.6)	12.5 (9.7–15.4)	15.6 (8.6–22.6)	Gray P = 0.038
Competing-death cumulative incidence, day 5, %	< 1	< 1	< 1	< 1	
Adjusted cause-specific HR (95% CI)	1.00	1.16 (0.79–1.69)	1.48 (1.04–2.10)	2.13 (1.22–3.74)	trend 1.25 (1.08–1.45), P = 0.003
Adjusted subdistribution HR, cmprsk	1.00	1.16 (0.79–1.69)	1.48 (1.04–2.11)	2.12 (1.19–3.76)	
All operations, cluster-robust HR	1.00	1.16 (0.80–1.68)	1.54 (1.09–2.17)	2.04 (1.17–3.58)	
Ascertainable subset (n = 1,109, 196 events)					
Absolute risk, % (95% CI)	14.6 (11.3–18.3)	16.9 (12.8–21.7)	21.4 (17.0–26.2)	22.9 (13.7–34.5)	
Adjusted OR (95% CI)	1.00	1.15 (0.76–1.74)	1.57 (1.06–2.31)	1.77 (0.93–3.37)	

Closely spaced operations did not drive the result. After excluding 21 patients who had another eligible operation within the renal window, the highest-stratum hazard ratio was 1.97 (95% CI 1.11–3.51), with a per-stratum estimate of 1.23 (1.06–1.43). Baseline creatinine was recorded a median of 2.0 days before

surgery (IQR 1.0-3.7); 59 patients had a baseline more than 7 days before surgery. Restricting the cohort to baselines within 7 days gave a highest-stratum hazard ratio of 2.15 (1.23–3.77) and a per-stratum estimate of 1.25 (1.07–1.46).

Detection timing and which criterion was met

Of the 196 events, 134 were first detected within 24 hours after the landmark, 28 between 24 and 48 hours, and 34 later (Fig. 5). The absolute-rise criterion contributed to 131 of the earliest 134 events. In contrast, every event detected more than 24 hours after the landmark met the ratio criterion alone, because the 48-hour absolute-rise window had closed.

The rule generating the composite outcome therefore changed during follow-up, and the single hazard ratio reported above is an average across that change. **This restriction is KDIGO's, not ours.** The guideline defines the absolute-rise criterion as a rise of at least 0.3 mg/dL *within 48 hours*, so a patient whose creatinine rises by that amount on day four does not meet it and can qualify only through the ratio criterion [27]. Our operational definition inherits that boundary; it does not impose it, which is why the problem is not confined to this dataset. Separating the two criteria shows how much the pooled estimate conceals.

Analysed as two cause-specific outcomes, with each criterion treating the other as a competing event, the point estimates differed sharply. For the 131 events involving the absolute-rise criterion, adjusted hazard ratios were 1.18 (95% CI 0.73–1.90), 1.80 (1.17–2.77) and **3.36 (1.80–6.27)**. For the 65 events satisfying the ratio criterion alone, they were 1.11 (0.60–2.05), 1.05 (0.57–1.91) and **0.56 (0.13–2.42)**. The ratio-only estimates were imprecise, so they do not establish absence of an association.

Splitting follow-up at the moment the absolute-rise criterion closes gives a similar contrast. While both criteria are open, the top-stratum hazard ratio is 3.24 (1.74–6.02) across 134 events; thereafter, with only the ratio criterion available, it is 0.59 (0.14–2.53) across 62 events (Table 5). The pooled Cox coefficient of 2.13 is a risk-set- and event-time-weighted log hazard contrast across these periods. It is not the arithmetic average of 3.24 and 0.59.

That reframes what had looked like a fading biological signal. The observed attenuation coincides with closure of one arm of the composite definition. It may therefore reflect the endpoint rule, component-specific aetiology, selection, confounding or chance rather than a biological decay in the lactate association. The data do not identify which explanation dominates. This is why criterion-specific estimates are more informative here than the pooled hazard ratio.

The interval-censored analysis, applied to the cohort

Because the simulation showed that event-time treatment can matter under sparse measurement, we compared three treatments in the cohort. Every one of the 196 events could be bracketed by the creatinine draw that produced it and the one before it. Those brackets were narrow: median 0.48 days (IQR 0.28–0.73), with 13.8% wider than a day and none wider than three days.

At that sampling density, the choice of event-time treatment makes almost no difference. The top-stratum hazard ratio was 2.13 (95% CI 1.22–3.74) treating the qualifying draw as the event time, 2.11 (1.20–3.70) using the interval midpoint, and 2.13 (1.20–3.80) from an interval-censored Weibull model, with the same pattern in the other strata.

The similarity is consistent with the short intervals in this cohort, but it does not prove that recorded detection times can be substituted for onset in other settings. In simulation, the interval-censored model was accurate for some mechanisms and biased for others. Investigators should report interval widths, examine the observation process, and justify their event-time model rather than choosing an estimator solely based on measurement density.

Renal observation process

Post-landmark creatinine sampling was more similar across strata than lactate sampling. Patients had medians of 6, 6, 7 and 7 creatinine results, while the median time to the first post-landmark result ranged from 6.9 to 9.1 hours. The rate of testing per day of follow-up was also close to constant, at 1.3, 1.3, 1.4, and 1.4 results per day across the strata from lowest to highest. Detection intensity per unit of time at risk, which is the quantity a cause-specific hazard depends on, was therefore comparable on average. An equal average rate does not make the observation process ignorable. The same average is consistent with quite different temporal patterns, such as dense early sampling in the higher strata against sparse late sampling in the lower ones. And the timing of a blood draw remains a clinical decision that may depend on unmeasured severity. Total observation time was not higher in the higher-lactate strata, which tended to remain under renal observation longer, creating more opportunity for an event to be recorded and, in particular, for matters to be included in the fixed-window analysis. The net direction and magnitude of that contribution cannot be identified from these data.

Several analyses tested specific observation assumptions. Inverse probability weighting moved the highest-stratum fixed-window odds ratio from 1.77 to 1.59 (95% CI 0.91–2.76). Treating discharge as a competing event produced a cause-specific hazard ratio of 2.16, compared with 2.13 in the primary model. Treating all censoring times as competing events gave a subdistribution hazard ratio of 2.18 (1.25–3.82). These estimates were numerically stable, but none proved non-informative censoring. Most censoring occurred when a patient was still in hospital, but creatinine testing had stopped, and the reason for stopping was unavailable (Additional file 6).

Exposure definitions with differing measurement opportunity

Alternative exposure summaries pointed the same way, though several estimates were imprecise (Additional file 6). In quartile analyses, highest-versus-lowest cause-specific hazard ratios were 1.92 (95% CI 1.29–2.87) for peak lactate, 1.88 (1.25–2.82) for the time-weighted average and 1.81 (1.24–2.64) for lactate load. Replacing the maximum reduced, but did not remove, its correlation with the number of measurements. Peak lactate and time-weighted average were themselves highly correlated

(Spearman $\rho = 0.932$), and 36.2% of time-weighted averages relied mainly on values carried forward from early samples. These analyses make a purely mathematical maximum effect less likely. They do not separate lactate physiology from the clinical concern that prompted repeated testing. Because all three summaries are built from the same clinician-initiated draws, their agreement cannot exclude a bias common to that testing process.

An outcome that does not depend on someone drawing blood

Ninety-day mortality is included for one reason: it is ascertained by registry linkage, so it is not subject to the detection process that shapes everything above. It shows a strong gradient, from 1.4% (95% CI, 0.7–2.4) in the lowest-lactate stratum to 10.2% (95% CI, 6.5–15.1) in the highest-lactate stratum. Adding lactate to the available covariates nonetheless produced no demonstrable gain in discrimination for either outcome. The out-of-bag change in the discrimination measure was 0.024 (95% CI, -0.019 to 0.057) for mortality and 0.001 (95% CI, -0.035 to 0.027) for newly detected AKI, both relative to the expanded comparator. Every interval we computed contained zero, while leaving upper bounds that would matter clinically. Association and incremental performance are separate results; neither outcome establishes the other. Effect estimates, model specifications, internal validation and decision curves are in Additional file 6; they are not reproduced here because the mortality analysis is a contrast, not a second study.

Discussion

The finding that lactate is not about transfers. When an exposure summary and an outcome definition share an interval, counting every event mixes associations before and after the summary becomes available. The global-null simulation is an important qualification: overlap and delayed detection did not create bias when the association was null throughout. Large errors arose when the early association differed from the post-window target, especially when sampling depended on severity. The case study shows why the distinction matters in practice, since 73.0% of patients meeting the creatinine definition did so before the summary existed.

Conditioning at the landmark improved the early-only scenarios, reducing bias from 0.369 to 0.212, from 0.116 to 0.069 and, with severity-dependent sampling, from 0.501 to 0.267. It did not recover the oracle target in those mechanisms. The reason is specific. A landmark defined on *detection* removes detected cases, not every case whose latent onset occurred earlier. Across the scenarios, 28.6% to 44.6% of post-landmark detections had pre-landmark onset. That fraction is not observable in a real cohort, and it is not itself a measure of bias. Its effect depends on the association structure and observation process.

This concurrent fraction should not be mistaken for a bias fraction. It describes when cases first became detectable, whereas the exposure association also depends on how those cases are distributed across lactate strata and on the population included in each model. In this cohort, a conventional analysis that counted all AKI cases yielded a highest-versus-lowest odds ratio of 6.31 (95% CI 4.46–8.93), whereas

the landmark analysis yielded 2.18 (1.16–4.07). The ratio between these estimates is not a correction factor. The models use different risk sets and estimate different contrasts.

An association persisted after adjusting for concurrent AKI. Patients in the highest lactate stratum had about twice the cause-specific hazard of a later qualifying creatinine result. This finding is reasonably stable across exposure summaries and the two new sensitivity analyses. Its interpretation is narrower than the conventional prognostic claim. The model follows detection, not biological onset, and patients with higher lactate remained under renal observation longer. Lactate may mark haemodynamic stress that also injures the kidney. It may also rise when renal clearance is already impaired [37]. Two versions of that second pathway need to be separated. Pre-existing impairment is constrained by these data. Median preoperative creatinine was 0.88 mg/dL in the three lower strata versus 0.92 in the highest; the maximum standardized difference across strata was 0.09 in the primary cohort and 0.05 among patients at risk at the landmark, and in every adjusted model conditioned on it. We report standardized differences here instead of P values because the corresponding Kruskal-Wallis tests yield 0.063 and 0.66 for the two populations, a difference in interpretation driven by the analysis set rather than the data. Impaired clearance caused by injury already underway when the lactate was drawn is not constrained, and cannot be with an outcome that only becomes visible hours later. Reverse causation therefore remains plausible even when the creatinine criterion is first recorded after the landmark.

The composite outcome is not a complication to be noted and set aside. Twenty-four hours after the landmark, the absolute-rise arm of the modified KDIGO definition can no longer generate a case, and all later detections arise through the ratio criterion. The component-specific point estimates differed, with a top-stratum hazard ratio of 3.36 for events involving the absolute-rise arm and 0.56 for ratio-only events. The period-specific estimates were 3.24 and 0.59. The ratio-only intervals were wide, so these contrasts should be interpreted as heterogeneity in the observed estimates, not as proof of no association beyond the boundary.

This matters beyond the present cohort for a reason unrelated to lactate. A composite endpoint whose components have different open intervals can produce time-varying observed associations even when the exposure definition is fixed. Endpoint closure is one possible explanation, not a sufficient causal account. Component-specific biology, risk-set selection, confounding and chance remain alternatives. Studies using a composite renal endpoint should state which component can generate a case at each follow-up time point and report component-specific estimates when components close at different times.

We conducted a purposive methodological audit of 20 published studies [7–26], not a systematic review or meta-analysis (Additional file 5; record-level extraction, derived ratings, summary counts and exclusions are Additional files 8–11). The selected reports span several lactate windows and designs, but they cannot estimate prevalence. None explicitly reported the concurrent fraction or a landmark exclusion for AKI already meeting criteria during the exposure window. Ten studies measured lactate after surgery. Four had a stated exposure period that extended into the renal ascertainment period,

whereas six lacked sufficient endpoint-timing detail for a bounded overlap judgment. Two illustrate the reporting problem. One measured creatinine at intensive care admission and at 6, 12, and 24 hours, using 24-hour lactate as the predictor [7]. Another diagnosed AKI within 48 hours of admission while using the 24-hour peak lactate [25]. Neither reported how many events preceded the close of the exposure window. These observations identify a gap in the included reports. They do not quantify bias in any individual study or frequency across the literature.

The mortality analysis clarifies a different point. Lactate had a strong association with 90-day death, an outcome obtained through registry linkage and not through serial creatinine testing. Yet adding lactate did not show a clear gain in discrimination for death or newly detected AKI. Association and incremental performance are therefore separate results. Because all change-in-C-index intervals crossed zero and had positive upper bounds, our data support neither a clinically useful increment nor equivalence to no increment.

Two dataset features matter for readers who may reuse INSPIRE. First, renal observation thins markedly near day 7, making late cumulative incidence unstable. Observation ended within the final 12 hours before the 7-day window closed for 855 of the 1,834 patients at risk (46.6%), and 846 of those were censored rather than reaching an event. Day 5 was the latest horizon with a usable risk set in each lactate stratum. Second, creatinine is discretised to 21 values and top-coded at 5.54 mg/dL. The nominal 0.3 mg/dL rise is therefore not equally attainable across all baseline values. Fifty-nine patients could not meet either creatinine arm because both required values above the dataset ceiling. Excluding them changed the highest-stratum hazard ratio from 2.13 to 2.11, but this small change does not render the measurement defect harmless in other cohorts (Additional file 3).

Limitations

This was a retrospective, single-centre secondary analysis without preregistration or external validation. The covariates available in INSPIRE do not form a validated cardiac surgical risk model. Surgical category, ventricular function, aortic cross-clamp time, transfusion and vasoactive support were unavailable. Each could influence both lactate and renal injury, so residual confounding is a central limitation here rather than a routine caveat.

Of everything in this section, the limitation we can least dispose of is that exposure and outcome were shaped by the same clinical judgement. Peak lactate is partly determined by how often lactate is measured, and post-landmark AKI is detectable only while creatinine continues to be ordered. A clinician worried about a patient draws more of both. One unmeasured worry, therefore, raises both exposure and the chance of detecting the outcome. That is confounding by indication, working through the measurement process instead of through treatment.

We did not attempt to model that process away, and the reason is worth stating outright. Inverse probability of sampling weights or a marginal structural model would require a model of the visit process itself. The variables that drive a decision to draw blood are mostly the clinician's read of how the patient

looks, and none of them are in this dataset. Fitting such a model on the covariates INSPIRE does contain would produce weights that appear to adjust for the visiting process while in fact adjusting for age, baseline creatinine and bypass duration a second time. So we stayed with standard covariate adjustment. We report the sampling intensity openly and vary the exposure across summaries that depend on it in different ways. That seemed more honest than a weighting scheme whose assumptions we could not check. Inverse probability of censoring weighting is used, but for the end of observation rather than its frequency, and it addresses a different mechanism.

What follows is that the alternative exposure summaries and the censoring analyses test particular assumptions without removing the problem. They cannot recover AKI that was never measured, and they cannot separate lactate physiology from the concern that prompted the measurement.

The most recent preoperative creatinine was occasionally remote. Restricting baseline measurements to the preceding 7 days did not materially change the result, although this does not validate the selected baseline for every patient. De-identification further coarsened both lactate and creatinine. The event time was interval-censored, urine output was unavailable, and renal replacement therapy could not be used reliably. These features mean that our modified creatinine-only outcome is not the full KDIGO definition.

Repeat operations created a potential attribution problem. The primary cohort used one operation per patient, and exclusion of 21 patients who underwent another eligible operation within the renal window produced a similar estimate. The highest lactate stratum nevertheless contained only 103 patients at risk and 17 renal events, leaving the upper-stratum estimate imprecise.

Finally, the literature work is purposive and non-exhaustive by construction. It should be read only as a design audit of the included reports. Searching was limited to PubMed Central, retrieval counts were not preserved, and no second database or duplicate-removal audit was used. Extraction was performed by one author with automated assistance. Six of the 20 external-study rows underwent a second source-text check; the other 14 were not independently double-extracted. Selection and extraction error therefore remain possible. Additional file 8 provides the record-level fields and verification status.

What to do about it

Showing that the standard remedy is incomplete obliges us to say what should replace it. Five things, in descending order of how much they buy for what they cost; Box 1 gives them as a checklist with the reason for each.

1. **Report the concurrent fraction.** How many patients meeting the outcome definition did so before the exposure summary existed? It is a single number, computable from any dataset that records when criteria were first met, and it bounds how much of an association could possibly be prospective. Its absence across all 20 studies we examined is the clearest reporting gap we found.
2. **Put every window on one axis, from one origin.** Most of the confusion we encountered stemmed from studies that used different anchors for exposure and outcome without stating so. A figure costs nothing and makes the overlap either visible or provably absent.

3. **3. Use a landmark, and describe it accurately.** It is the correct response, not a correction. Report it as conditioning on being event-free at W , never as recovering a true effect, and acknowledge that patients whose onset preceded W remain in the risk set whenever detection depends on measurement.
4. **4. State which component of a composite outcome is available at each point in follow-up.** Where components close at different times, report component-specific and period-specific estimates. Our case study produced markedly different, but partly imprecise, point estimates across this boundary. The contrast may reflect component-specific aetiology, selection, confounding or chance; closure alone is not a sufficient causal explanation.
5. **5. Report measurement intervals and justify the event-time model.** Interval-censored analysis was accurate in two scenarios with a post-window effect, but it was biased when the association was constant before and after the window. It is not a default replacement for a landmark Cox model. The choice should follow the observation mechanism, interval widths and model assumptions.

For clinicians, the case study says something narrower than the literature it re-examines. A high peak lactate marks patients more likely to have AKI detected later and more likely to die by 90 days. It does not validate real-time triage during the first postoperative day, and adding it to routinely available variables produced no demonstrable gain in discrimination. Acting on it would require external calibration, a prespecified threshold and a prospective test of benefit, none of which this study provides.

What this study establishes, and what it does not

This paper often hedges, and the hedges are load-bearing rather than decorative. It is worth separating them from what the work does support.

Established. Exposure-outcome window overlap changes the estimand. In scenarios with an early association but little or no post-window association, the naive Cox analysis remained biased after effect scales were aligned. A landmark-on-detection removed recorded concurrent cases but retained patients whose latent onset preceded the landmark and was detected later. In the case study, component-specific and period-specific estimates differed sharply depending on when the absolute-rise criterion closed.

Not established. No method was uniformly best across the simulated mechanisms. The interval-censored fit was parametric and was favoured by parts of the data-generating process. The simulation does not establish that endpoint-component closure alone creates a time-varying association. The literature audit establishes neither prevalence nor the size of bias in another study. The case-study estimates are associations in the observed detection process, not causal effects on kidney injury.

Not answered, and answerable. Whether a landmark on *the first negative measurement after W , rather than on the absence of detection at W , would remove the prevalent residue is a question our framework*

poses that our data cannot settle. So is the size of the residue in fields with different measurement habits. Both are tractable with the design we describe.

Conclusions

A biomarker summary cannot prospectively predict outcomes that have already met their study definition before the summary becomes available. Window overlap therefore changes the question being answered. It does not, by itself, determine the direction or size of bias. A landmark defines a more prospective risk set, while remaining tied to the recorded detection process. Studies of windowed biomarkers should report the concurrent fraction, place all windows on a single clock, state which components of a composite outcome remain available, and describe how outcomes were observed. Without those details, readers cannot distinguish prediction from contemporaneous association.

Abbreviations

ADEMP: Aims, data-generating mechanism, estimands, methods, performance measures

AKI: Acute kidney injury

ASA: American Society of Anesthesiologists

AUC: Area under the receiver operating characteristic curve

CI: Confidence interval

CPB: Cardiopulmonary bypass

DRB: Data Review Board

HR: Hazard ratio

IQR: Interquartile range

IRB: Institutional Review Board

KDIGO: Kidney Disease: Improving Global Outcomes

OR: Odds ratio

SMD: Standardised mean difference

STROBE: Strengthening the Reporting of Observational Studies in Epidemiology

STS: Society of Thoracic Surgeons

Declarations

Ethics approval and consent to participate. This work is a secondary analysis of de-identified perioperative clinical data released for public research use. It involved no participant contact, recruitment, intervention or new data collection. The source study was approved by the Institutional Review Board of Seoul National University Hospital (IRB No. H-2210-078-1368), which waived the requirement for informed consent, and public release was approved by that institution's Data Review Board (DRB No. BD-R-2022-11-02). The present analysis complied with the PhysioNet data use agreement and relies on the source approval and waiver; no new consent was sought.

Consent for publication. Not applicable.

Clinical trial number: Not applicable.

Availability of data and materials. INSPIRE v1.4.2 is available to credentialed users from PhysioNet (<https://doi.org/10.13026/1eay-yc85>) [38]. Additional file 12 is the authoritative code-and-results archive supplied with this submission. It contains the analysis protocol, executable code, aggregate outputs and a manifest of withheld files. The same archive is deposited at Zenodo under <https://doi.org/10.5281/zenodo.21804476>, which resolves to its most recent version. Restricted source data and patient-level derivatives are not redistributed; credentialed users can regenerate them from the INSPIRE source tables.

Competing interests. The authors declare no competing interests.

Funding. This work was supported by the Ningxia Natural Science Foundation (grant numbers 2026AAC031176, 2026AAC031199). The funder had no role in the design, analysis, interpretation or reporting of this work.

Authors' contributions. ZB: Conceptualization, Data Curation, Formal Analysis, Investigation, Methodology, Validation, Visualization, Writing-Original Draft, Writing-Review & Editing. YZ: Conceptualization, Data Curation, Formal Analysis, Investigation, Methodology, Visualization, Writing-Review & Editing. Both authors read and approved the final manuscript.

Acknowledgements. The authors thank the investigators who curated and released INSPIRE.

Declaration of generative AI use. The use of generative AI tools and the authors' verification responsibilities are described in the Methods section. The authors reviewed and approved all tool-assisted output and take responsibility for the final article.

References

1. Suissa S. Immortal time bias in pharmacoepidemiology. *Am J Epidemiol.* 2008;167(4):492–9. [10.1093/aje/kwm324](https://doi.org/10.1093/aje/kwm324).

2. Zhang X, Xu J, He Q, Wang Y, Xie S, Zhang X, et al. Immortal time bias tends to be more pronounced in methodological studies than in empirical studies: a metaepidemiological study. *J Clin Epidemiol*. 2025;187:111936. 10.1016/j.jclinepi.2025.111936.
3. Anderson JR, Cain KC, Gelber RD. Analysis of survival by tumor response. *J Clin Oncol*. 1983;1(11):710–9. 10.1200/JCO.1983.1.11.710.
4. Dafni U. Landmark analysis at the 25-year landmark point. *Circ Cardiovasc Qual Outcomes*. 2011;4(3):363–71.
5. Putter H, van Houwelingen HC. Understanding landmarking and its relation with time-dependent Cox regression. *Stat Biosci*. 2017;9(2):489–503. 10.1007/s12561-016-9157-9.
6. Lim L, Lee H, Jung CW, Sim D, Borrat X, Pollard TJ, et al. INSPIRE, a publicly available research dataset for perioperative medicine. *Sci Data*. 2024;11:655. 10.1038/s41597-024-03517-4.
7. Lopez-Delgado JC, Esteve F, Torrado H, Rodriguez-Castro D, Carrio ML, Farrero E, et al. Influence of acute kidney injury on short- and long-term outcomes in patients undergoing cardiac surgery: risk factors and prognostic value of a modified RIFLE classification. *Crit Care*. 2013;17(6):R293. 10.1186/cc13159.
8. Zhang Z, Ni H. Normalized lactate load is associated with development of acute kidney injury in patients who underwent cardiopulmonary bypass surgery. *PLoS ONE*. 2015;10(3):e0120466. 10.1371/journal.pone.0120466.
9. Radovic M, Bojic S, Kotur-Stevuljevic J, Lezaic V, Milicic B, Velinovic M, et al. Serum lactate as reliable biomarker of acute kidney injury in low-risk cardiac surgery patients. *J Med Biochem*. 2019;38(2):118–25. 10.2478/jomb-2018-0018.
10. Yang HH, Chang JC, Jhan JY, Cheng YT, Huang YT, Chang BS, et al. Prognostic value of peak lactate during cardiopulmonary bypass in adult cardiac surgeries: a retrospective cohort study. *Tzu Chi Med J*. 2020;32(4):386–91. 10.4103/tcmj.tcmj_215_19.
11. Zhang K, Shang J, Chen Y, Huo Y, Li B, Hu Z. The prognosis and risk factors for acute kidney injury in high-risk patients after surgery for type A aortic dissection in the ICU. *J Thorac Dis*. 2021;13(7):4427–37. 10.21037/jtd-21-823.
12. Wang Z, Xu J, Kang Y, Liu L, Zhang L, Wang D. Early dynamic behavior of lactate in predicting continuous renal replacement therapy after surgery for acute type A aortic dissection. *Front Cardiovasc Med*. 2022;9:948672. 10.3389/fcvm.2022.948672.
13. He Q, Tan Z, Chen D, Cai S, Zhou L. Association between intraoperative hyperglycemia/hyperlactatemia and acute kidney injury following on-pump cardiac surgery: a retrospective cohort study. *Front Cardiovasc Med*. 2023;10:1218127. 10.3389/fcvm.2023.1218127.
14. Lyu Y, Liu Y, Xiao X, Yang Z, Ge Y, Jiang H. High level of intraoperative lactate might predict acute kidney injury in aortic arch surgery via minimally invasive approach in patients with type A dissection. *Front Cardiovasc Med*. 2023;10:1188393. 10.3389/fcvm.2023.1188393.
15. Yan Y, Gong H, Hu J, Wu D, Zheng Z, Wang L, et al. Perioperative parameters-based prediction model for acute kidney injury in Chinese population following valvular surgery. *Front Cardiovasc Med*.

- 2023;10:1094997. 10.3389/fcvm.2023.1094997.
16. Yu R, Liang T, Li L, Bi Y, Meng X. Predictive role of arterial lactate in acute kidney injury associated with off-pump coronary artery bypass grafting. *Front Surg.* 2023;10:1089518. 10.3389/fsurg.2023.1089518.
 17. Flores-Salinas HE, Zambada-Gamboa AdJ, Garcia-Garduno TC, Rodriguez-Zavala G, Valle Y, Chavez-Herrera JC, et al. Association of postoperative serum lactate levels with acute kidney injury in Mexican patients undergoing cardiac surgery. *Clin Pract.* 2024;14(3):1100–9. 10.3390/clinpract14030087.
 18. Hadipourzadeh F, Rastravan R, Totonchi Z, Heydarpur E, Faritous Z. Evaluating the relationship between lactate levels during coronary artery bypass graft surgery and postoperative renal dysfunction. *J Cardiovasc Thorac Res.* 2024;16(2):129–34. 10.34172/jcvtr.33051.
 19. Wang XD, Bao R, Lan Y, Zhao ZZ, Yang XY, Wang YY, et al. The incidence, risk factors, and prognosis of acute kidney injury in patients after cardiac surgery. *Front Cardiovasc Med.* 2024;11:1396889. 10.3389/fcvm.2024.1396889.
 20. Chen DX, Zhang YY, Xiong XL, Zhou L, Shi J. Association between intraoperative lactate levels and acute kidney injury after on-pump cardiac surgery: a retrospective cohort study across two centers. *BMC Surg.* 2025;25:324. 10.1186/s12893-025-03082-6.
 21. Peng L, Zheng Y, Guo F, Su M, Wei W. Near-infrared spectroscopy combined with vascular occlusion test to predict acute kidney injury in patients undergoing cardiac surgery: a prospective observational study. *J Cardiothorac Surg.* 2025;20:48. 10.1186/s13019-024-03312-7.
 22. Wang D, Yang C, Wang H, et al. High perioperative lactate levels as a potential predictor for severe acute kidney injury following aortic arch surgery. *Front Med (Lausanne).* 2025;11:1495502. 10.3389/fmed.2024.1495502.
 23. Wang C, Sun X, Chen K, et al. Lactate-to-albumin ratio index correlates with the occurrence and prognosis of acute kidney injury complicated by cardiac surgery. *Int J Cardiol Heart Vasc.* 2025;59:101734. 10.1016/j.ijcha.2025.101734.
 24. Fu Y, Qian J, Zhang Y, Ji F. Development and validation of a machine learning model for predicting serum creatinine-defined acute kidney injury in older adults after cardiac surgery. *Clin Interv Aging.* 2026;21:612843. 10.2147/CIA.S612843.
 25. Marianello D, Puddu A, Biuzzi C, Fogagnolo A, Spadaro S, Galasso L, et al. Prognostic value of red blood cell distribution width in predicting acute kidney injury after cardiac surgery: a retrospective cohort study. *J Clin Med.* 2026;15(6):2403. 10.3390/jcm15062403.
 26. Tang X, Li C, Ren J, Lin C, Chen Z, Zhu J, et al. Prediction of acute kidney injury after total aortic arch replacement with postoperative serum lactate: a retrospective cohort study. *Ren Fail.* 2026;48(1):2671444. 10.1080/0886022X.2026.2671444.
 27. KDIGO Acute Kidney Injury Work Group. KDIGO clinical practice guideline for acute kidney injury. *Kidney Int Suppl.* 2012;2:1–138.

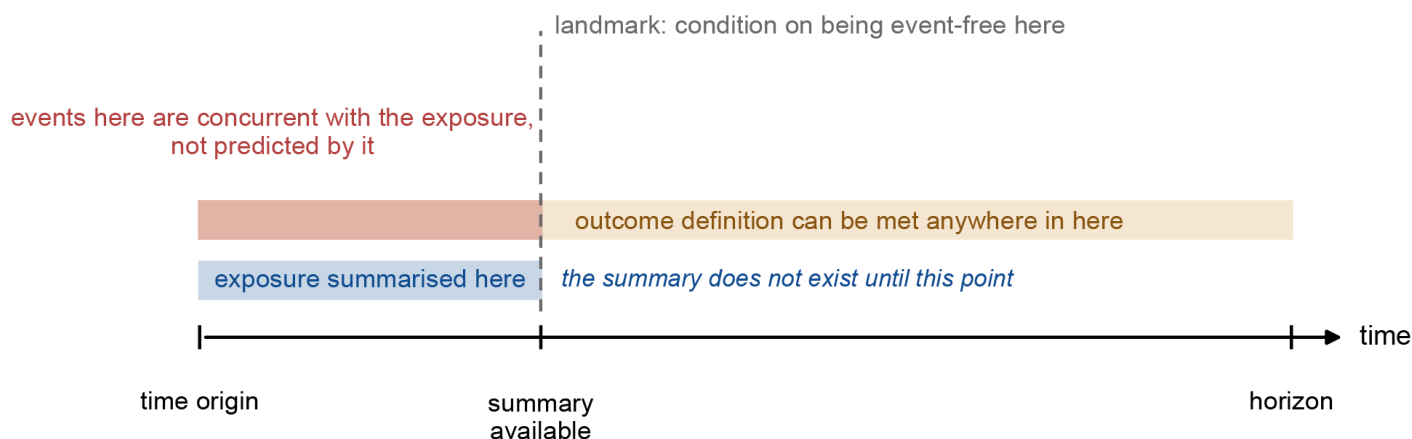
28. Harrell FE Jr, Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Stat Med*. 1996;15(4):361–87.
29. Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. *Med Decis Mak*. 2006;26(6):565–74.
30. Vandenbroucke JP, von Elm E, Altman DG, Gøtzsche PC, Mulrow CD, Pocock SJ, et al. Strengthening the Reporting of Observational Studies in Epidemiology (STROBE): explanation and elaboration. *Epidemiology*. 2007;18(6):805–35. 10.1097/EDE.0b013e3181577511.
31. Morris TP, White IR, Crowther MJ. Using simulation studies to evaluate statistical methods. *Stat Med*. 2019;38(11):2074–102. 10.1002/sim.8086.
32. Finkelstein DM. A proportional hazards model for interval-censored failure time data. *Biometrics*. 1986;42(4):845–54. 10.2307/2530698.
33. Gray RJ. A class of K-sample tests for comparing the cumulative incidence of a competing risk. *Ann Stat*. 1988;16(3):1141–54.
34. Fine JP, Gray RJ. A proportional hazards model for the subdistribution of a competing risk. *J Am Stat Assoc*. 1999;94(446):496–509.
35. Firth D. Bias reduction of maximum likelihood estimates. *Biometrika*. 1993;80(1):27–38.
36. Peduzzi P, Concato J, Kemper E, Holford TR, Feinstein AR. A simulation study of the number of events per variable in logistic regression analysis. *J Clin Epidemiol*. 1996;49(12):1373–9.
37. Guzzi J, Salter R, Pettit S, Kelting T, Falter F, Perrino AC Jr. Rethinking lactate in cardiac surgery and intensive care. *J Cardiothorac Vasc Anesth*. 2026;40(7):2177–90. 10.1053/j.jvca.2026.03.010.
38. Lim L, Lee H. INSPIRE, a publicly available research dataset for perioperative medicine, version 1.4.2 [dataset]. *PhysioNet*. 2026. 10.13026/1eay-yc85

Box 1

Box 1 is available in the Supplementary Files section.

Figures

A The overlap



B The four clocks behind one recorded event time

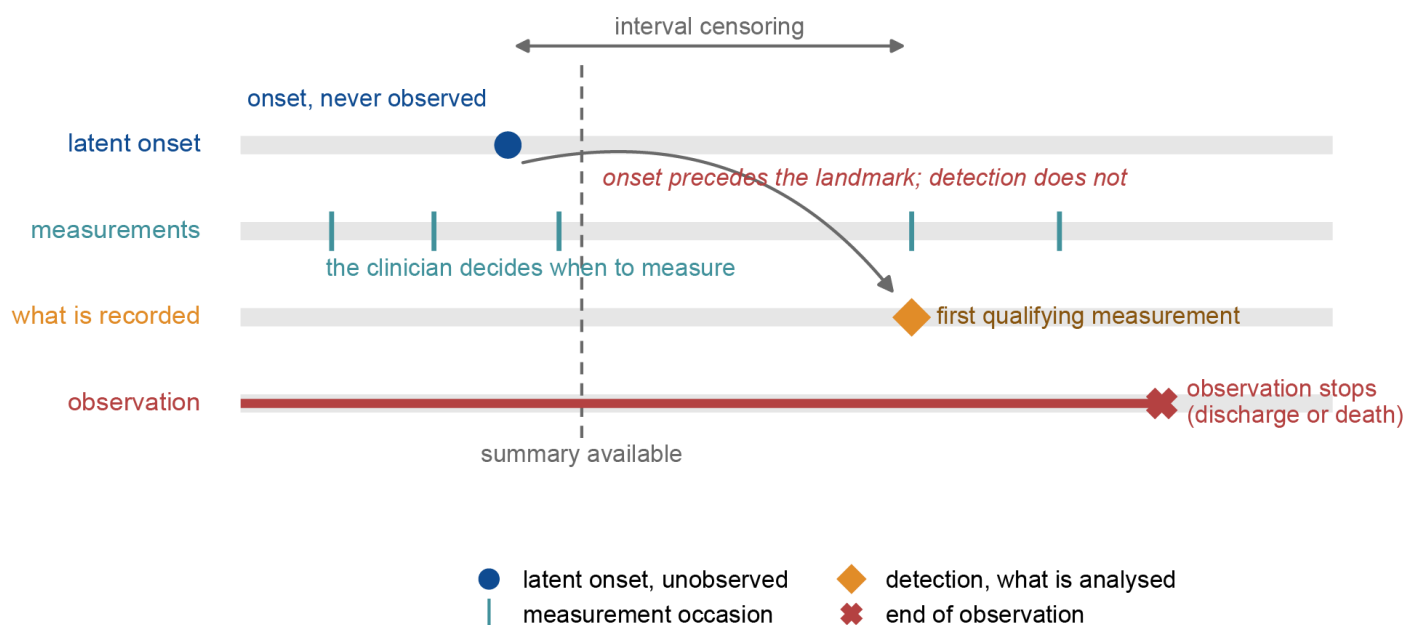


Figure 1

The problem, stated without reference to any particular biomarker or outcome. (A) The exposure summary does not exist until its window closes, while the outcome definition can be met from the time origin onwards. Events in the shaded region are concurrent with the exposure rather than predicted by it. A landmark condition is being event-free at the time the summary becomes available. (B) Four clocks lie behind a single recorded event time: latent onset, which is never seen; the measurement occasions, which the clinician chooses; the first qualifying measurement, which is what the analysis uses; and the end of observation. The gap between onset and detection is interval-censored, which is why a landmark on detection cannot remove cases whose onset preceded it.

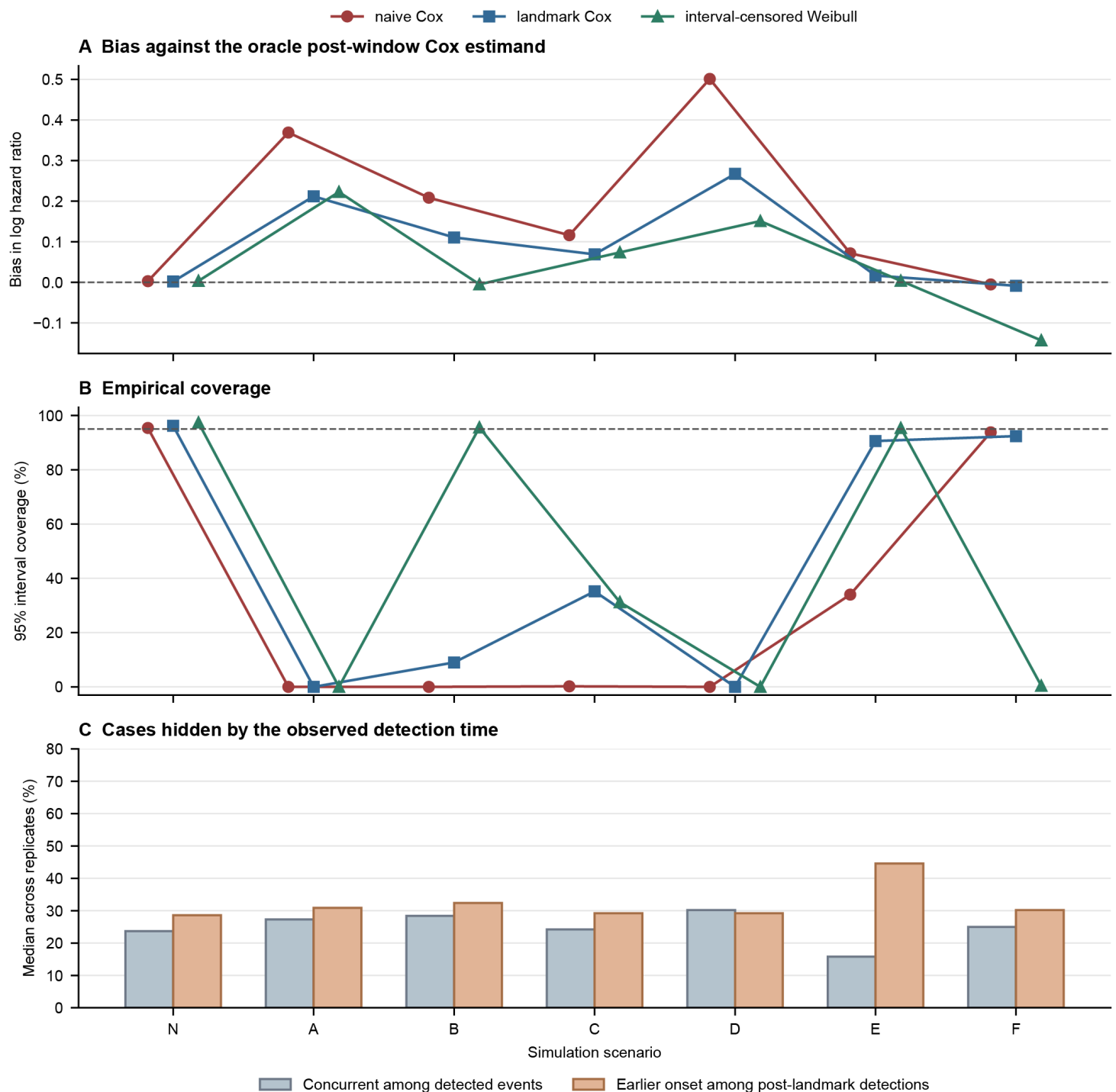
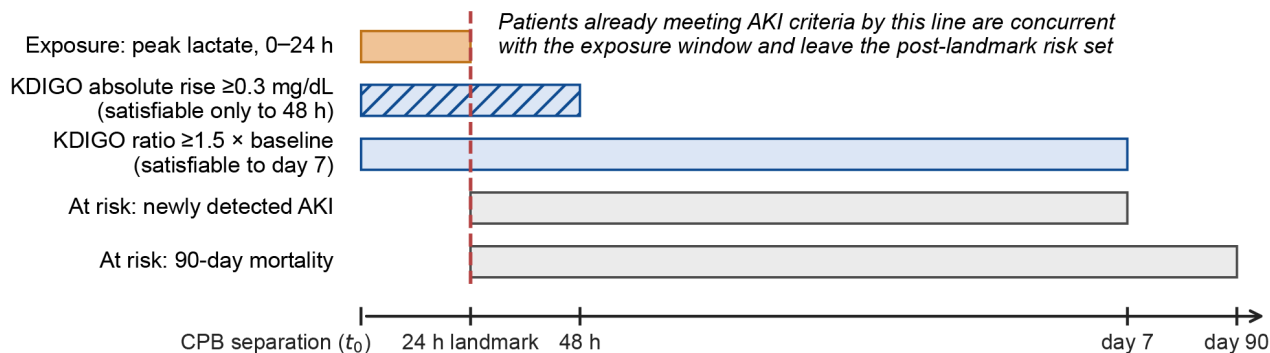


Figure 2

Simulation performance across seven observation mechanisms. (A) Bias of the mean log hazard-ratio estimate against the oracle post-window Cox target; the horizontal line marks zero bias. (B) Empirical coverage of nominal 95% intervals; the horizontal line marks 95%. (C) Median percentage of events detected by the landmark and percentage of post-landmark detections whose latent onset occurred earlier. The three methods share the same effect scale, and each scenario used 500 replicates. Scenario

N is the global-null negative control. Full estimates, standard errors, rejection rates, Monte Carlo errors and fit counts are reported in Table 6 and Additional file 7.

A One clock: exposure, landmark and outcome windows



B Cohort flow and classification of renal events

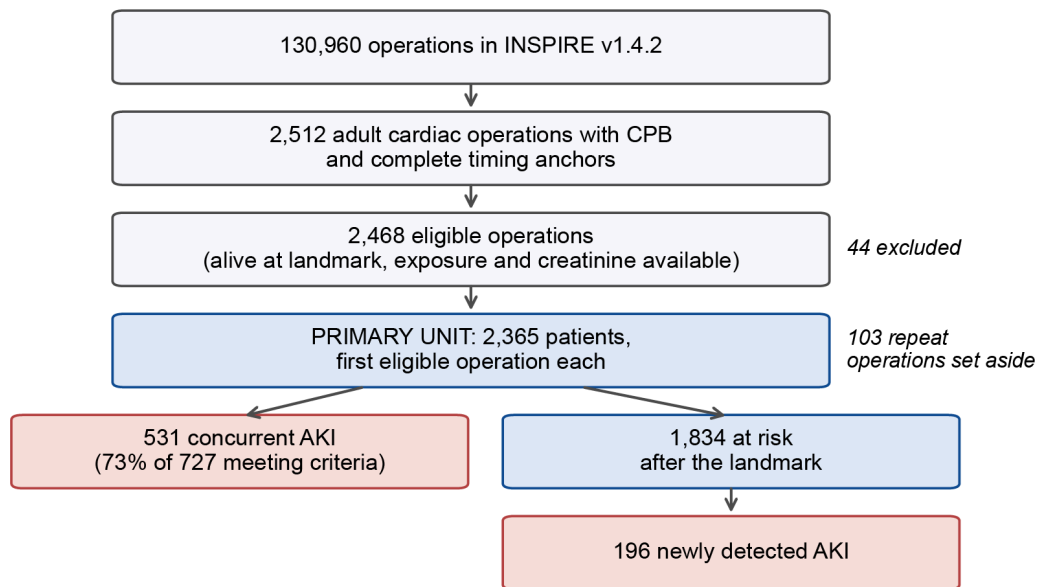


Figure 3

Estimand and cohort flow. (A) All quantities use separation from cardiopulmonary bypass as the time origin. The exposure window closes at 24 hours, which defines the landmark. Patients who have already met the AKI criterion are classified as concurrent cases and are removed from the risk set. The absolute-rise criterion closes at 48 hours from the origin, while the ratio criterion remains open to day 7. (B) Cohort selection, restriction to one operation per patient, and classification of renal events. Alternative time origins and their effect on eligibility are reported in Additional file 3.

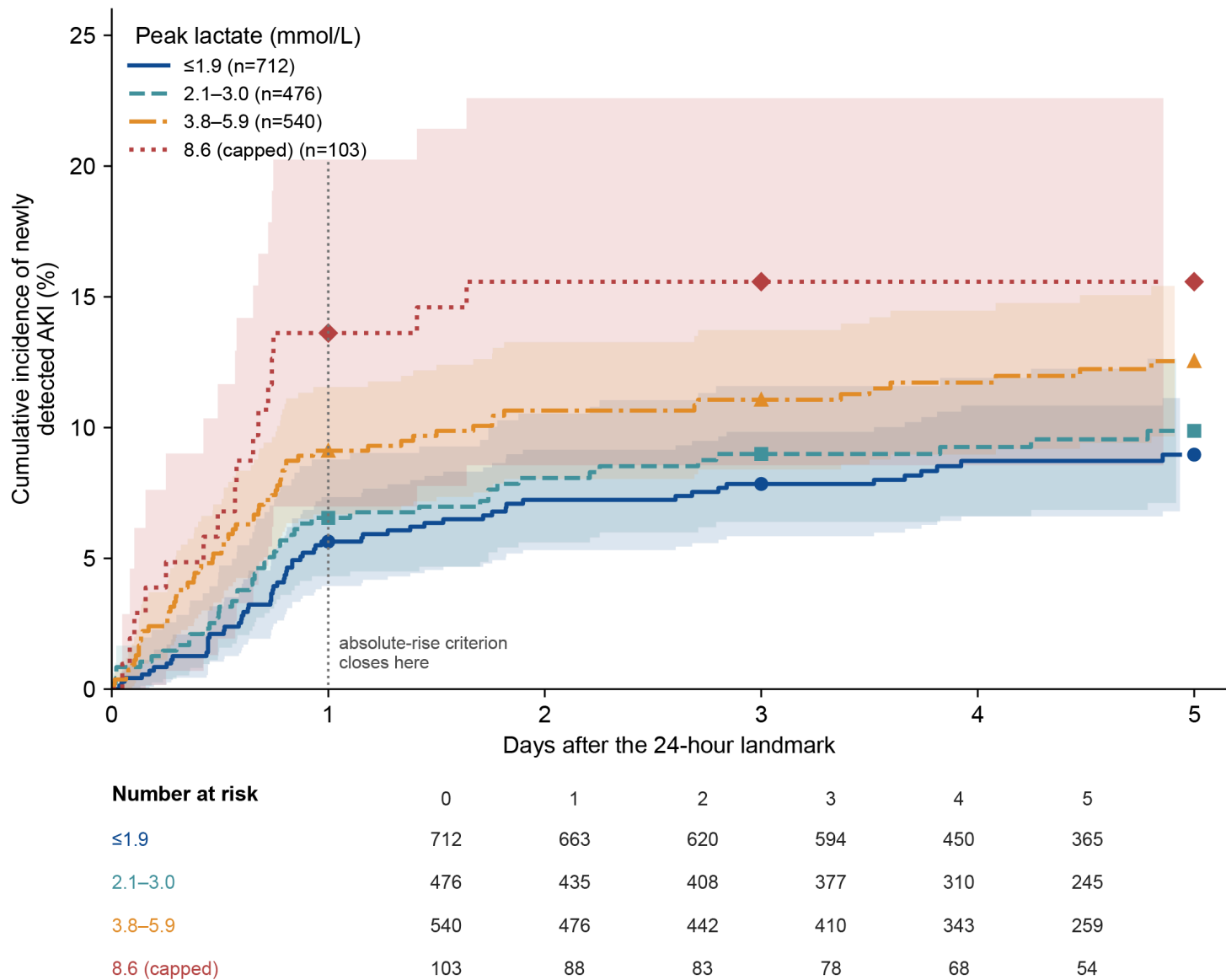


Figure 4

Cumulative incidence of newly detected acute kidney injury after the 24-hour landmark. Aalen-Johansen estimates treat death as a competing event and are truncated at day 5, the last horizon with a usable risk set in each stratum. Shaded bands are pointwise 95% confidence intervals. Colour, line style and markers identify lactate strata. The dotted vertical line marks the point 24 hours after the landmark at which the absolute-rise criterion closes, 48 hours from the time origin. That line is the boundary whose consequences are quantified in Figure 5 and Table 5: every event to its right could be generated only by the ratio criterion. Numbers at risk are shown below the plot.

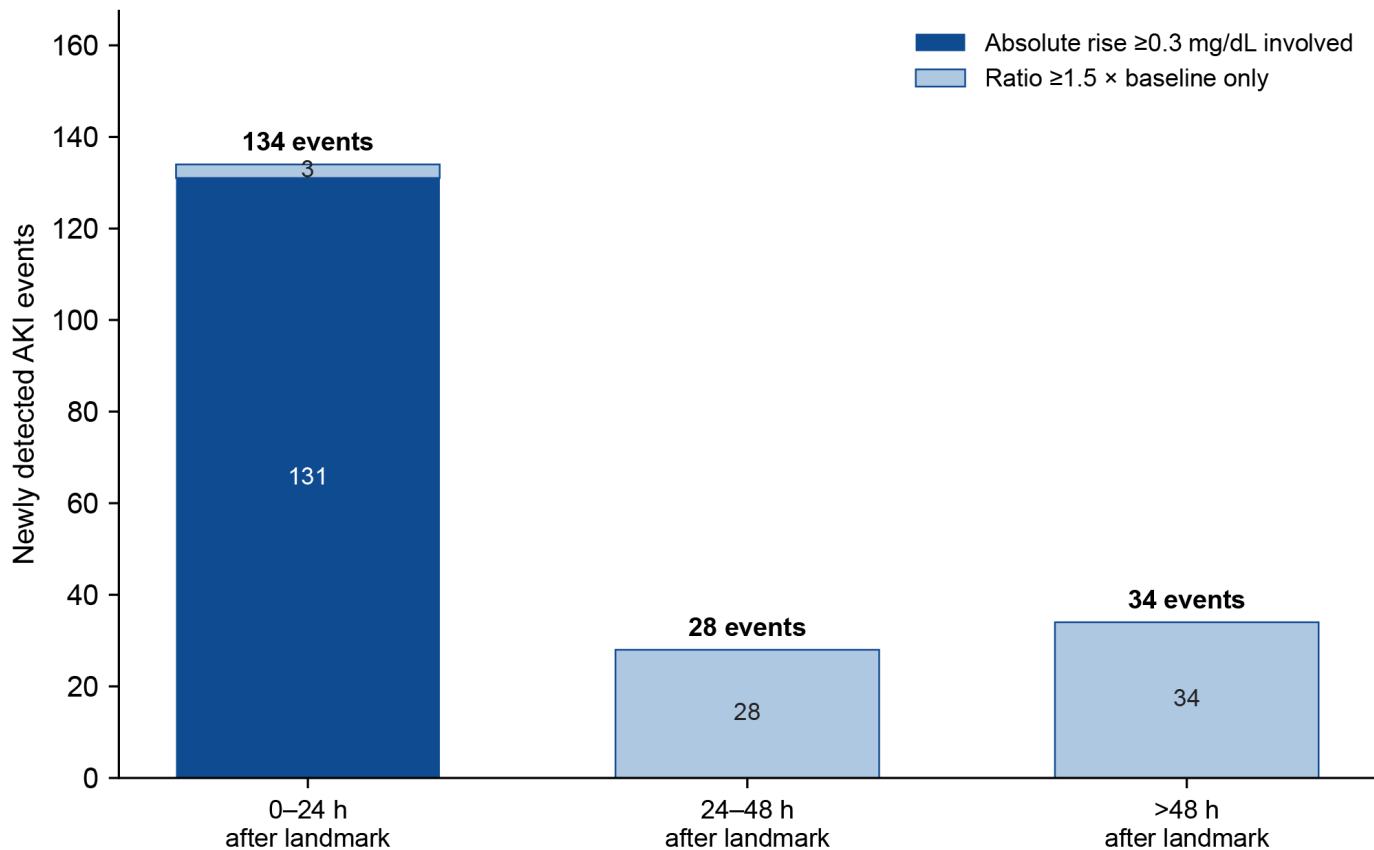


Figure 5

When newly detected AKI was first detected, and which criterion it met. Of 196 events, 134 were first detected within 24 hours of the landmark and 131 of those involved the absolute-rise criterion. All 62 events detected later were met by the ratio criterion alone, because the absolute-rise criterion is satisfiable only to 48 hours from the time origin. The rule generating the outcome therefore changes during follow-up, a fact that must be considered before interpreting the timing of events as a biological property of the exposure.

Supplementary Files

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

- [Additionalfile1STROBEchecklist.docx](#)
- [Additionalfile2analysisprovenance.docx](#)
- [Additionalfile3timeorigindiagnosics.docx](#)
- [Additionalfile4finegrayvalidation.docx](#)
- [Additionalfile5windowoverlapaudit.docx](#)
- [Additionalfile6mortalityandsensitivity.docx](#)
- [Additionalfile7simulation.docx](#)

- [Additionalfile8extractionsheet.csv](#)
- [Additionalfile9derivedratings.csv](#)
- [Additionalfile10summarycounts.csv](#)
- [Additionalfile11exclusionslog.csv](#)
- [Additionalfile12codeandresults.zip](#)
- [FigureS1.png](#)
- [Box1.docx](#)