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.

# Two ways an AKI prediction model can look better than it is: whole-stay feature leakage and marginal-only coverage, measured in 20,401 ICU patients

**Jeffery Opoku**

jeffery.opoku01@utrgv.edu

University of Texas Rio grande valley

**David Banahene**

Florida International University

**Research Article**

**Keywords:** acute kidney injury, conformal prediction, machine learning, uncertainty quantification, external validation, class-conditional coverage.

**Posted Date:** August 5th, 2026

**DOI:** https://doi.org/10.21203/rs.3.rs-10463539/v2

**License:** This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

**Additional Declarations:** The authors declare no competing interests.

# Two ways an AKI prediction model can look better than it is: whole-stay feature leakage and marginal-only coverage, measured in 20,401 ICU patients

Jeffery Opoku[1,*] David Banahene[2]

[1]School of Mathematical and Statistical Sciences, University of Texas Rio Grande Valley, Edinburg, TX 78539, USA

[2]Department of Biostatistics, Robert Stempel College of Public Health and Social Work, Florida International University, Miami, FL 33199, USA

[*]Corresponding author: opokujeffery5@gmail.com

**Abstract**

**Background.** Machine-learning models for acute kidney injury (AKI) in the intensive care unit (ICU) are increasingly reported with strong discrimination and, more recently, with conformal prediction sets that carry a formal coverage guarantee. Both headline numbers can be produced by a study design that is not wrong in any single step but that answers an easier question than the clinical one.

**Objective.** To measure, on one ICU cohort, how much two common evaluation choices inflate apparent performance: aggregating features over the whole ICU stay, and reporting conformal coverage marginally rather than per class.

**Methods.** We used a public two-hospital ICU dataset and built a landmark cohort of 20,401 patients: features from the first 24 hours only, outcome defined as incident KDIGO stage 1+ AKI arising *after* hour 24, and patients already meeting AKI criteria by hour 24 excluded. Models were developed at one hospital ($n = 10{,}830$, AKI 11.9%) and validated at a second, independent hospital ($n = 9{,}571$, AKI 10.2%). We compared six classifiers, applied split, class-conditional and Mondrian conformal prediction, and

ran a controlled ablation in which cohort, outcome, model, split and seed were held fixed and only the feature window varied.

**Results.** Under the landmark design the best model (XGBoost) reached an external AUC of 0.735 (95% CI: 0.720–0.751). Re-fitting the identical model on whole-stay features raised this to 0.919, so the feature window alone accounts for **+0.188 AUC**; a conventional whole-stay analysis of these data reports 0.935. Repeating the ablation at prediction times of 6, 12, 24 and 48 hours found the inflation at every one (+0.179 to +0.250), so it is a property of whole-stay aggregation rather than of a particular landmark. Conformal prediction showed the second inflation. A marginal rule met its target overall, 90.8% coverage, while covering only **14.3%** of patients who developed AKI. We prove this is permitted rather than anomalous: a valid marginal $1-\alpha$ rule bounds coverage of a class of prevalence $\pi_1$ only by $\max\{0,\, 1-\alpha/\pi_1\}$, here 1.7%, and recovering 90% AKI coverage marginally would require a 99.0% rule. Class-conditional calibration reached 88.4% AKI coverage directly, at a cost of 50.5% two-label sets, and we prove its guarantee is unaffected by the prevalence difference between sites, so the residual shortfall measures within-class shift ($\geq 1.6$ percentage points in total variation). Internal cross-validation gave AUC $0.802 \pm 0.013$ against 0.735 externally, a further gap that a single-site study would not have revealed.

**Conclusion.** Neither inflation requires a mistake in any individual step, and both are invisible in the numbers usually reported. A landmark feature window and per-class coverage should be the default for ICU prediction models, and we give the bound that says how much a marginal guarantee is worth for any given event rate.

**Keywords:** acute kidney injury, intensive care, clinical decision support, data leakage, landmark analysis, conformal prediction, external validation, class imbalance

# 1 Introduction

Acute kidney injury (AKI) is one of the most common and most serious organ problems in the intensive care unit (ICU). It affects 10–50% of critically ill patients, depending on the definition and the population studied.[1, 2] Patients who develop AKI are four to seven times more likely to die in hospital, stay longer, cost more to treat, and are more likely to go on to chronic kidney disease and kidney failure.[3, 4] Spotting these patients early therefore matters, because acting in time — managing fluids, avoiding drugs that harm the kidney, supporting blood pressure — can reduce or prevent the damage.[5]

Many machine-learning models for AKI have been built over the past ten years, using logistic regression,[6] random forests,[7] gradient boosting,[8] deep learning,[9] and recurrent neural networks.[10] These models usually report how well they separate cases from non-cases (AUC) and how well their probabilities are calibrated, and the best of them are used to drive alerts in real hospitals.[11] More recently, conformal prediction has been proposed as a way to attach a formal guarantee to such a model, so that a clinician is told not only what the model predicts but how often that prediction set contains the truth.[16, 17]

This paper is about two ways such a model can report an impressive number while answering an easier question than the clinical one. Neither involves a mistake in any single step, and neither shows up in the metrics that are conventionally reported.

**The first: what the model is allowed to see.** A very common design aggregates each variable over the patient's whole ICU stay — mean, minimum, maximum, first and last value — and defines the outcome from the whole stay as well. Every step is defensible on its own. Together they let the model see measurements taken *after* the outcome has begun. For AKI the channel is direct: blood urea nitrogen rises along with creatinine, so a last-value BUN carries the answer. The result is a model that detects AKI rather than predicting it, and there is no way to tell from the reported AUC that this has happened.[12] The standard remedy is a landmark design: fix a prediction time, use only what was observable

before it, and count only outcomes that occur after.[13] Reviews of clinical prediction models repeatedly find this step missing.[14, 15]

**The second: who the guarantee is for.** Split conformal prediction turns any model into one that outputs a set of labels satisfying $\Pr(Y_{n+1} \in \mathcal{C}(X_{n+1})) \geq 1-\alpha$, needing only that the data be exchangeable.[16] Exchangeability is the assumption that fails first in practice, and a body of work now adapts the construction when it does: by reweighting under covariate shift,[22] by exploiting spectral structure in dependent data,[23] and by tracking drift in streaming deployments.[24] Parallel extensions carry conformal *risk* control into settings where the deployment distribution moves away from the calibration one.[25, 26]

Our concern is different, and it is prior to all of these. Even with exchangeability holding exactly, the probability above is an average over the whole population. When the outcome is uncommon the average is supplied almost entirely by the patients who do not have it, so the guarantee can be met while the patients who do have it are covered hardly at all. Romano et al.[18] proposed class-conditional conformal prediction as the remedy. What has not been made precise is *how* badly the rare class can do while the marginal guarantee still holds — and, as we show, the answer for an ICU event is: arbitrarily badly.

We study both failures on a single cohort so that their sizes can be compared directly, and we fix both. This paper makes the following contributions.

1. We build a **leakage-free landmark cohort** of 20,401 ICU patients — features from the first 24 hours, incident AKI ascertained only after hour 24, prevalent AKI excluded — with genuine external validation at a second hospital.

2. We **measure the cost of the feature window** in a controlled ablation that holds cohort, outcome, model, split and seed fixed: whole-stay features give AUC 0.919 against 0.732 for the landmark design, an inflation of **+0.188 AUC**. Repeating this at four prediction times shows the inflation is present at every one, and stable at +0.179 to +0.188 wherever the cohort is large enough to support the estimate.

3. We prove a **sharp limit on what a marginal guarantee buys the rare class**: coverage of a class with prevalence $\pi_1$ can be as low as $\max\{0,\ 1-\alpha/\pi_1\}$, which is *zero* whenever the event is rarer than the miscoverage level $\alpha$. The bound is 1.7% here, and the observed 14.3% sits inside it.

4. We show **what protecting the rare class marginally would cost**: a guarantee of $1-\beta$ needs $\alpha \leq \beta\pi_1$, here a 99.0% rule, and prediction sets only grow as $\alpha$ falls.

5. We prove a **label-shift robustness result**: class-conditional conformal prediction keeps its per-class guarantee unchanged under any change in disease prevalence.

6. We turn the residual gap into a **measurement of cross-site shift**: observed coverage $c_k$ implies the two sites' class-$k$ score distributions differ in total variation by at least $(1-\alpha) - c_k$.

7. We report **Mondrian** group-conditional coverage, **SHAP** explanations, decision curve analysis, calibration comparison and five-fold cross-validation on the corrected design, and release all code.

# 2 Methods

## 2.1 Study population, landmark design and feature construction

**Data source.** We used the PhysioNet/Computing in Cardiology Challenge 2019 dataset,[19] a publicly available collection of hourly clinical data from ICU patients at two academic medical centers: Beth Israel Deaconess Medical Center (BIDMC; 20,336 patients) and Emory University Hospital (EUH; 20,000 patients), comprising 1,552,210 hourly records. For each patient the data provide vital signs, laboratory values and demographics. The dataset was assembled for a sepsis prediction challenge. We adopt it here because it is public, hourly resolved and spans two institutions, which are jointly the minimum requirements for a land-

mark design with external validation; the constraints this imposes are addressed in Section 4.5.

**Landmark design.** We fix a prediction time $T = 24$ hours after ICU admission. This is the design choice that separates prediction from detection, and every subsequent step follows from it. Features are computed exclusively from records with ICULOS $\leq T$, so that no measurement taken after hour 24 enters the model in any form. Baseline creatinine is the first creatinine observed within that window. The outcome is incident AKI, defined as KDIGO stage 1+ evaluated on creatinine measured strictly after hour 24:

$$\text{AKI} = \begin{cases} 1 & \text{if } \text{Cr}_{\text{max}}^{>T} \geq 1.5 \times \text{Cr}_{\text{baseline}}, \\ 1 & \text{if } \text{Cr}_{\text{max}}^{>T} - \text{Cr}_{\text{baseline}} \geq 0.3 \text{ mg/dL}, \\ 0 & \text{otherwise.} \end{cases} \tag{1}$$

Eligibility requires an ICU stay exceeding 24 hours, at least one creatinine measurement before $T$ to establish the baseline, and at least one after $T$ so that the outcome can be ascertained. Patients already satisfying Eq. (1) on creatinine measured *within* the first 24 hours are excluded, so that the target is incident disease rather than disease that has already declared itself. This exclusion is essential: retaining prevalent cases would allow a model to score well by recognising established injury, which is precisely the confusion the landmark is intended to prevent.

Of 40,336 patients, 30,212 had a stay longer than 24 hours and 21,634 also had creatinine on both sides of the landmark. Excluding 1,233 patients with prevalent AKI yields a final cohort of $N = 20{,}401$, of whom 2,262 (11.1%) developed incident AKI.

**Feature construction.** Within the first 24 hours we computed summary statistics for every vital sign and laboratory variable: mean, standard deviation, minimum and maximum, together with first and last value for laboratory measurements. To these we added shock

index (HR/SBP), pulse pressure (SBP − DBP), BUN/creatinine ratio, heart rate variability, the number of hours actually observed, and the count of creatinine measurements ordered before the landmark. Creatinine trajectory features enter the outcome definition and were therefore removed, with only baseline creatinine retained. Missing values were imputed using the column-wise median fitted on the training partition alone, so that no information from the calibration or external sets influenced imputation. The final feature set comprises 190 variables.

**Development and external validation.** We used the natural hospital-of-origin split. The development set (BIDMC) comprises $n$ = 10,830 patients at 11.9% AKI prevalence, partitioned by stratified sampling into a training set ($n$ = 7,581, 70%) and a conformal calibration set ($n$ = 3,249, 30%). The external validation set (Emory) comprises $n$ = 9,571 patients at 10.2% prevalence and was reserved entirely for evaluation, entering neither model fitting nor conformal calibration.

## 2.2 Prediction models, calibration and clinical utility

We trained six classifiers on the BIDMC training set only: $L_2$-regularized logistic regression ($C$ = 0.1); random forest (500 trees, depth 10, minimum leaf 20, balanced class weights); gradient boosting, XGBoost[20] and LightGBM[21] (300 trees, depth 4, learning rate 0.05, subsample 80%); and a multi-layer perceptron (256–128–64, ReLU, Adam, $\alpha = 10^{-3}$, batch 256, early stopping). Bootstrap confidence intervals for AUC use 1,000 resamples of the external set; confidence intervals for coverage, which is a simple proportion, are Wilson score intervals.

Beyond discrimination we assessed three further properties on the external set. Calibration was compared across raw probabilities, Platt scaling[30] and isotonic regression,[31] scored by the Brier statistic and expected calibration error (ECE) over ten bins. Explanations were obtained from SHAP values[29] computed with TreeExplainer on 2,000 external

patients. Clinical utility was assessed by decision curve analysis across threshold probabilities from 0.01 to 0.50. Stability was assessed by five-fold stratified cross-validation within the development hospital, refitting the model and recomputing both conformal thresholds independently within each fold.

## 2.3 The leakage ablation

To attribute the difference between this study and a conventional analysis, we re-fit the identical XGBoost model on the identical cohort, with the identical outcome, split and random seed, changing *only* the window over which features are aggregated:

- **Arm A**: features aggregated over the whole ICU stay.
- **Arm B**: features from hours 1–24 only (the landmark design used throughout the rest of this paper).

Because everything else is held fixed, any difference in external AUC is attributable to the feature window alone.

The design deserves brief justification, since a simpler comparison was available and would have been weaker. One could contrast our landmark result against AUC figures reported in the published AKI literature, but such a comparison confounds the feature window with every other difference between studies — cohort, outcome definition, feature set, model class, era and site. It would establish only that our number is lower, not why. By contrast, an internal ablation in which a single design choice is varied supports a counterfactual claim: had this analysis been conducted in the conventional way, on these same patients with this same outcome, it would have reported the inflated figure. That is the claim the paper needs, and it is available only from within a single dataset.

We repeat the ablation at four prediction times, $T \in \{6, 12, 24, 48\}$ hours, rebuilding the cohort from scratch at each, so that the finding cannot be an artefact of having chosen $T = 24$. Each $T$ defines a different eligible population, since eligibility requires a stay

exceeding $T$ with creatinine on both sides, and the comparison is therefore made within each $T$ rather than across.

## 2.4 Conformal prediction

We applied split conformal prediction.[17] For each calibration point $(x_i, y_i)$ the nonconformity score is $s_i = 1 - \hat{p}_{y_i}(x_i)$. The conformal quantile is

$$\hat{q} = \text{Quantile}\left(s_1, \ldots, s_m;\ \frac{\lceil (1-\alpha)(m+1) \rceil}{m}\right), \tag{2}$$

and the prediction set for a new point $x$ is $\mathcal{C}(x) = \{y \in \{0,1\} : 1 - \hat{p}_y(x) \leq \hat{q}\}$.

**Theorem 1** (Finite-sample coverage[16])**.** *If the calibration and test data are exchangeable, then* $1-\alpha \leq \Pr(Y_{n+1} \in \mathcal{C}(X_{n+1})) \leq 1-\alpha + \frac{1}{m+1}$.

With $m = 3{,}249$ calibration points the theoretical bound is $[0.900, 0.9003]$.

Three refinements of this construction are used throughout. *Class-conditional* conformal prediction[18] computes a separate threshold for each class,

$$\hat{q}_k = \text{Quantile}\left(\{s_i : y_i = k\};\ \tfrac{\lceil (1-\alpha)(m_k+1) \rceil}{m_k}\right), \tag{3}$$

thereby guaranteeing $\Pr(Y \in \mathcal{C}_{\text{CC}}(X) \mid Y = k) \geq 1 - \alpha$ for each class individually rather than on average. *Mondrian* conformal prediction[16] applies the same device to predefined subgroups, computing a separate quantile within each and so guaranteeing $1 - \alpha$ coverage group by group; we defined groupings by gender and by age ($< 50$, 50–65, $> 65$ years). Finally, we derive a two-level *conformal risk stratification* directly from the class-conditional sets,

$$\text{Risk} = \begin{cases} \textbf{Low Risk} & \text{if } \mathcal{C}_{\text{CC}}(x) = \{0\}, \\ \textbf{Elevated Risk} & \text{if } 1 \in \mathcal{C}_{\text{CC}}(x), \end{cases} \tag{4}$$

so that a Low Risk designation inherits the formal guarantee that the probability of missed AKI is at most $\alpha$.

## 2.5 How much a marginal guarantee can leave to a rare class

Theorem 1 controls coverage averaged over the whole population. It says nothing about either class on its own, and when one class is rare the difference is not a technicality. The following elementary bound makes it quantitative. Write $\pi_k = \Pr_P(Y = k)$ for the class prevalences in the calibration distribution and $F_k(t) = \Pr_P\big(s(X,k) \le t \mid Y = k\big)$ for the distribution of the nonconformity score within class $k$.

**Lemma 1** (Minority-class floor under a marginal guarantee)**.** *Suppose the marginal threshold $\hat{q}$ attains marginal coverage $1-\alpha$, that is $\pi_0 F_0(\hat{q}) + \pi_1 F_1(\hat{q}) = 1-\alpha$. Then the coverage of class 1 satisfies*

$$\Pr\big(Y \in \mathcal{C}(X) \mid Y = 1\big) \;=\; F_1(\hat{q}) \;\ge\; \max\Big\{0,\; 1 - \frac{\alpha}{\pi_1}\Big\},$$

*and this bound is attained. In particular, if $\pi_1 \le \alpha$ the bound is* 0*: a rule with perfectly valid $1-\alpha$ marginal coverage may cover* none *of class 1.*

*Proof.* Since $F_0(\hat{q}) \le 1$ and $\pi_0 = 1 - \pi_1$,

$$\pi_1 F_1(\hat{q}) \;=\; 1 - \alpha - \pi_0 F_0(\hat{q}) \;\ge\; 1 - \alpha - (1 - \pi_1) \;=\; \pi_1 - \alpha,$$

so $F_1(\hat{q}) \ge 1 - \alpha/\pi_1$, and a coverage probability is also $\ge 0$. Equality holds exactly when $F_0(\hat{q}) = 1$, that is when every majority-class score sits below the threshold — which is what a confident classifier on an easy majority class produces. □

The bound is elementary, and the phenomenon behind it — that marginal coverage need not transfer to a subgroup[16, 18] — is well known in qualitative form. Stating it sharply is what makes it useful, because the right-hand side involves only $\alpha$ and the prevalence.

The next consequence is the one that decides what a practitioner should actually do. If a marginal rule is all you have, the only lever available is $\alpha$ itself.

**Corollary 1** (The price of protecting a rare class marginally)**.** *To guarantee coverage of at least* $1-\beta$ *for a class of prevalence* $\pi_1$ *using a single marginal threshold, it is necessary that*

$$\alpha \;\leq\; \beta\,\pi_1.$$

*Moreover the prediction sets* $\mathcal{C}_\alpha$ *are nested in* $\alpha$*, so buying that protection can only make every set larger.*

*Proof.* Lemma 1 gives $F_1(\hat{q}) \geq 1-\alpha/\pi_1$ and shows the bound is attained, so a guarantee of $1-\beta$ forces $1-\alpha/\pi_1 \geq 1-\beta$, i.e. $\alpha \leq \beta\pi_1$. Nestedness is immediate: lowering $\alpha$ raises the quantile level in Eq. (2), hence raises $\hat{q}$, hence enlarges $\mathcal{C}(x) = \{y : 1-\hat{p}_y(x) \leq \hat{q}\}$. □

*Remark* 1 (What the bounds say for AKI). With $\alpha = 0.10$ and the calibration prevalence $\pi_1 = 386/3{,}249 = 0.1188$, the floor is $1-0.10/0.1188 = 15.8\%$. At the external site AKI runs at $\pi_1 = 974/9{,}571 = 0.1018$, giving a floor of **1.7%** — so a formally valid 90% marginal rule may cover fewer than one AKI patient in fifty. Going the other way, Corollary 1 says that promising 90% coverage of AKI patients through a marginal rule requires $\alpha \leq 0.10 \times 0.1018 = 0.0102$, a **99.0% rule**. Because both bounds use only $\alpha$ and the prevalence, they can be evaluated before a model is fitted and before any test data exist. More generally, no ICU event with prevalence below $\alpha$ is protected by a marginal conformal guarantee — which covers most of what an alerting system is built to catch.

## 2.6 Coverage under cross-site distribution shift

The exchangeability assumption behind Theorem 1 is violated when the calibration data (BIDMC) and the test data (Emory) come from different hospitals. The two sites differ in AKI prevalence (11.9% vs. 10.2%), and they may also differ in what cases and non-cases

look like physiologically. Separating these two kinds of shift is worth the effort, because they affect conformal coverage in completely different ways.

Two responses are available when exchangeability fails. One is to modify the construction so that coverage is restored — reweighting by a likelihood ratio,[22] adapting to spectral structure in the dependence,[23] or updating thresholds online as the stream drifts.[24] The other, which we take here, is to leave the standard construction alone and ask what the coverage it achieves *tells us* about the shift. The two are complementary: the first is what one does to deploy a model at a new site, the second is how one decides whether that is necessary.

We say the test distribution $Q$ is a *label-shifted* version of the calibration distribution $P$ if the class-conditional feature laws are preserved, $Q(X \in \cdot \mid Y = k) = P(X \in \cdot \mid Y = k)$ for each class $k$, while the class priors may differ, $Q(Y = k) \neq P(Y = k)$. A change in how common a disease is, with no change in the physiology of cases and non-cases, is exactly this situation.

**Proposition 1** (Label-shift robustness of class-conditional conformal prediction)**.** *Let the calibration sample be drawn i.i.d. from $P$ and the test point from a label-shifted distribution $Q$. Then, conditionally on the class counts $m_k$, the class-conditional conformal set $\mathcal{C}_{\mathrm{CC}}$ of Eq.* (3) *carries the same two-sided guarantee it has under exchangeability,*

$$1 - \alpha \;\leq\; \Pr_Q\big(Y \in \mathcal{C}_{\mathrm{CC}}(X) \mid Y = k\big) \;\leq\; 1 - \alpha + \frac{1}{m_k + 1}$$

*for each class $k$. Both bounds are unchanged by the shift: the prevalence of the classes does not enter. In contrast, the marginal conformal set with a single threshold $\hat{q}$ has class-$k$ coverage $F_k(\hat{q}) = \Pr_P\big(s(X,k) \leq \hat{q} \mid Y = k\big)$, which is not calibrated per class and, by Lemma 1, may fall to zero as that class becomes rarer than $\alpha$.*

*Proof.* Condition on the class counts $m_0, m_1$. The class-conditional threshold $\hat{q}_k$ is a function only of the calibration scores $\{s_i : y_i = k\}$, which are then i.i.d. draws from the law of $s(X,k)$

under $P(\cdot \mid Y = k)$. A test point with $Y = k$ has score $s(X, k)$ distributed as $Q(\cdot \mid Y = k)$, and this equals $P(\cdot \mid Y = k)$ by the label-shift assumption. The $m_k$ calibration scores in class $k$ and the test score are therefore exchangeable, so the standard split-conformal quantile argument[17] applies verbatim within the class and gives both bounds, with $m_k$ in place of $m$. Nothing in the argument refers to $\pi_k$, which is why the shift leaves the guarantee alone. For marginal conformal prediction $\hat{q}$ is the single $(1-\alpha)$-quantile of the $P$-mixture $\sum_j \pi_j^P F_j$, so class-$k$ coverage is $F_k(\hat{q})$, and Lemma 1 bounds how small this can be. □

**Corollary 2** (Where the external coverage gap comes from)**.** *Because the class-conditional guarantee is untouched by pure label shift, any class-conditional under-coverage seen at an external site must come from* within-class *shift — a change in $P(X \mid Y = k)$ — and not from the two sites having different amounts of disease.*

Corollary 2 says the residual gap is within-class shift. The next result says that the size of the gap measures how much.

**Proposition 2** (Coverage as a lower bound on cross-site shift)**.** *Let $d_k = \mathrm{TV}\big(Q(\cdot \mid Y = k),\, P(\cdot \mid Y = k)\big)$ be the total-variation distance between the class-$k$ score distributions at the two sites. If class-conditional conformal prediction calibrated on $P$ achieves coverage $c_k$ on $Q$, then*

$$d_k \;\geq\; (1 - \alpha) - c_k.$$

*Proof.* Let $A = \{s(X, k) \leq \hat{q}_k\}$. By construction $\Pr_P(A \mid Y = k) \geq 1-\alpha$, and total variation bounds the change in the probability of any event, so $c_k = \Pr_Q(A \mid Y = k) \geq \Pr_P(A \mid Y = k) - d_k \geq (1 - \alpha) - d_k$. Rearranging gives the claim. □

This inverts the usual reading of an external coverage shortfall. A gap is normally reported as a disappointment; here it is a measurement, and it is a measurement of something otherwise hard to observe — how far apart two hospitals' patients are once the difference in disease prevalence has been removed.

*Remark* 2 (Checks that need no test data). Lemma 1 and Corollary 1 need only $\alpha$ and the prevalence, so they give a worst case straight away. The actual minority coverage of a marginal rule, $F_1(\hat{q})$, depends only on the calibration data and not on the test set, so it too can be computed in advance; a large gap between $F_1(\hat{q})$ and $1 - \alpha$ is advance warning that a single-threshold rule will fail the rare class.

# 3 Results

## 3.1 Cohort characteristics

The landmark cohort comprised 20,401 ICU patients, of whom 2,262 (11.1%) developed incident AKI after hour 24. Table 1 summarizes characteristics measured within the first 24 hours, stratified by outcome. Patients who went on to develop AKI were older (median 68.0 vs. 63.8 years), more often male (59.9% vs. 54.7%), and already exhibited higher BUN (23.0 vs. 16.0 mg/dL) and higher baseline creatinine (1.3 vs. 0.9 mg/dL) before the landmark, alongside lower platelets and a markedly higher rate of sepsis (14.9% vs. 5.8%).

Two features of this table bear on what follows. First, the separation present *before* the landmark is real but modest: the interquartile ranges of BUN and baseline creatinine overlap substantially between groups, which is consistent with the discrimination subsequently achieved and inconsistent with the near-perfect separation a whole-stay analysis reports. Second, heart rate and oxygen saturation did not differ at all ($p = 0.56$ and $p = 0.84$), and mean arterial pressure differed only slightly. The pre-landmark signal is therefore carried almost entirely by renal and haematological laboratory values rather than by haemodynamic deterioration, which has not yet occurred at 24 hours in patients whose kidney injury is still to come. A design that permits late vital signs into the feature set imports exactly that later deterioration.

Table 1: Cohort characteristics by incident AKI status, measured in the first 24 hours. Values are median [IQR] or %. $p$-values from Mann–Whitney $U$ or chi-square tests.

| | **No AKI** ($n = 18{,}139$) | **Incident AKI** ($n = 2{,}262$) | $\boldsymbol{p}$ |
|---|---|---|---|
| Age (years) | 63.8 [51.4–74.0] | 68.0 [58.0–77.0] | $< 10^{-34}$ |
| Male sex (%) | 54.7 | 59.9 | $< 10^{-5}$ |
| Heart rate (bpm) | 83.8 [73.6–94.7] | 84.1 [75.5–92.8] | 0.56 |
| MAP (mmHg) | 80.6 [73.3–89.8] | 77.5 [70.4–87.0] | $< 10^{-25}$ |
| $SpO_2$ (%) | 97.7 [96.3–98.9] | 97.8 [96.4–98.9] | 0.84 |
| BUN (mg/dL) | 16.0 [11.0–25.0] | 23.0 [15.0–36.8] | $< 10^{-116}$ |
| Baseline Cr (mg/dL) | 0.9 [0.7–1.2] | 1.3 [0.9–2.4] | $< 10^{-149}$ |
| Lactate (mmol/L) | 1.7 [1.2–2.5] | 1.9 [1.3–2.8] | $< 10^{-4}$ |
| Platelets ($\times 10^3/\mu$L) | 198 [146–261] | 174 [128–233] | $< 10^{-30}$ |
| Hemoglobin (g/dL) | 10.6 [9.4–12.1] | 10.1 [9.1–11.3] | $< 10^{-29}$ |
| Total ICU LOS (hours) | 46 [41–52] | 48 [42–56] | $< 10^{-31}$ |
| Sepsis (%) | 5.8 | 14.9 | $< 10^{-57}$ |

## 3.2 How much the feature window inflates performance

Table 3 and Figure 1(a) report the ablation of Section 2.3. Holding the cohort, outcome, model, split and seed fixed and varying only the feature window, external AUC falls from 0.919 to 0.732. The feature window alone therefore accounts for **+0.188 AUC**. A conventional whole-stay analysis of these data — whole-stay features, outcome from the whole-stay creatinine peak, prevalent AKI retained — reports 0.935, so the cohort and outcome definition contribute a further +0.016 and the total inflation is +0.203.

The ablation arms use 188 features rather than the 190 of the main analysis: hours observed and the count of creatinine orders before the landmark are defined by the window itself, so they cannot be held fixed across arms and are dropped from both. This is why Arm B reads 0.732 against the 0.735 of Table 4 — a difference smaller than a third of either confidence interval, and immaterial to the attribution.

## 3.3 The inflation does not depend on the prediction time

Table 2 repeats the ablation at four prediction times, rebuilding the cohort at each. The inflation is present everywhere, between +0.179 and +0.250 AUC. At $T = 6$, 12 and 24 hours, where the cohorts are large, it is remarkably stable at +0.179 to +0.188 — so the

effect is a property of whole-stay aggregation, not of any particular landmark.

The $T = 48$ row warrants caution and is included for completeness. Only 2,738 patients had a stay beyond 48 hours with creatinine on both sides, giving roughly 235 events across two hospitals and fewer than 100 in the training split. Both arms are unreliable at that sample size, and the collapse of the landmark arm to 0.579 reflects overfitting rather than anything about 48-hour prediction. The three adequately powered rows carry the claim.

Table 2: Leakage ablation repeated at four prediction times. The cohort is rebuilt at each $T$; within each row the cohort, outcome, model, split and seed are identical across arms.

| $T$ (h) | $n$ | AKI | Whole-stay | Landmark | Inflation |
|---|---|---|---|---|---|
| 6 | 14,735 | 15.9% | 0.920 | 0.738 | +0.182 |
| 12 | 21,173 | 14.3% | 0.925 | 0.746 | +0.179 |
| 24 | 20,401 | 11.1% | 0.919 | 0.732 | +0.188 |
| *Under-powered, shown for completeness* | | | | | |
| 48 | 2,738 | 8.6% | 0.830 | 0.579 | +0.250 |

Table 3: Leakage ablation. Cohort, outcome, model hyperparameters, split and random seed are identical across arms; only the feature aggregation window differs.

| Feature window | External AUC | Difference |
|---|---|---|
| Whole ICU stay (Arm A) | 0.919 | — |
| Hours 1–24 only (Arm B, landmark) | **0.732** | −0.188 |
| *For reference* | | |
| Whole-stay features, outcome and cohort | 0.935 | −0.203 |

The mechanism is visible in which features each arm relies on. With whole-stay features the model leans on last-value phosphate and last-value BUN — laboratory results drawn late in the stay, by which time the creatinine rise defining the outcome is already under way. Under the landmark design the leading feature is baseline creatinine, a legitimate pre-landmark risk marker. Nothing about the whole-stay AUC of 0.919 reveals this; the number looks like a good model.

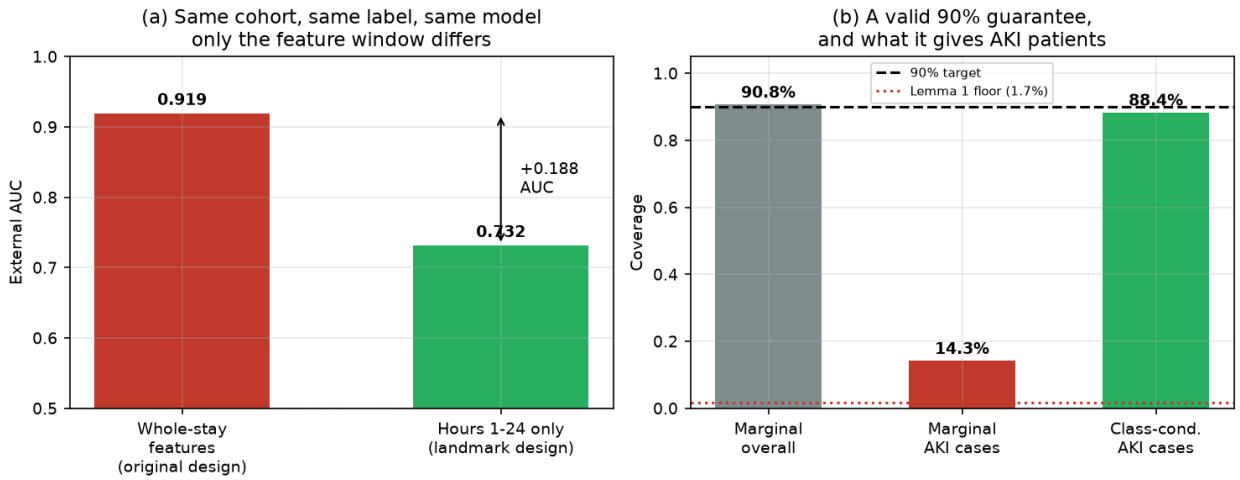



Figure 1: The two inflations. (a) External AUC with cohort, outcome, model, split and seed held fixed, varying only the feature window. (b) A marginal conformal rule meets its 90% target overall while covering 14.3% of AKI patients; the dotted line is the floor of Lemma 1 at the external prevalence (1.7%), and class-conditional calibration recovers 88.4%.

## 3.4 Model discrimination under the landmark design

Table 4 gives external performance on Emory ($n$ = 9,571). XGBoost was best (AUC 0.735, 95% CI: 0.720–0.751) and was used for all subsequent analyses. The three boosting methods and the random forest clustered between 0.718 and 0.735, logistic regression reached 0.690, and the MLP trailed at 0.666, consistent with evidence that gradient boosting outperforms deep learning on tabular clinical data.[28] Figure 2 shows the ROC and precision-recall curves; the latter make the difficulty of the task plain, with average precision of 0.250 against a prevalence of 0.102.

Table 4: External validation on Emory ($n$ = 9,571) under the landmark design. Models trained on BIDMC ($n$ = 7,581). AUC 95% CI from 1,000 bootstrap resamples. Best in bold.

| Model | AUC | 95% CI | AUPRC | Brier | F1 | Acc. |
|---|---|---|---|---|---|---|
| LR | 0.690 | 0.672–0.706 | 0.199 | 0.103 | 0.179 | 0.866 |
| RF | 0.724 | 0.708–0.740 | 0.219 | 0.144 | 0.298 | 0.813 |
| GB | 0.718 | 0.701–0.734 | 0.221 | 0.098 | 0.224 | 0.874 |
| XGBoost | **0.735** | **0.720–0.751** | **0.250** | **0.086** | 0.135 | 0.893 |
| LightGBM | 0.730 | 0.714–0.746 | 0.237 | 0.088 | 0.132 | 0.892 |
| MLP | 0.666 | 0.647–0.684 | 0.192 | 0.090 | 0.031 | 0.896 |

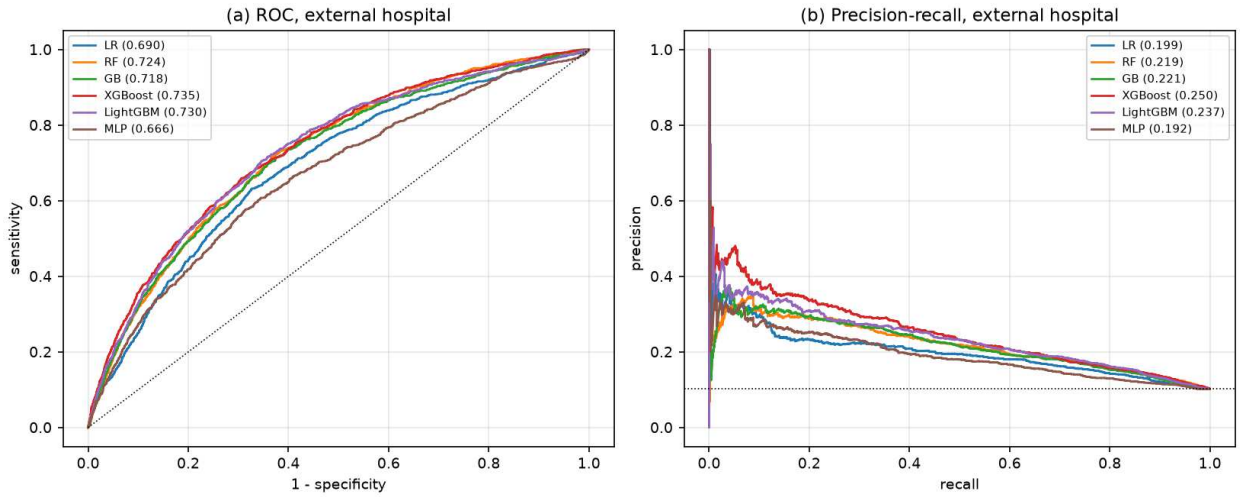



Figure 2: (a) ROC curves and (b) precision-recall curves on external validation (Emory) under the landmark design. Legends show AUC and average precision.

Calibration is reported alongside discrimination because a model whose probabilities are miscalibrated will produce misleading conformal thresholds even when its ranking is sound. Table 5 compares the three approaches for XGBoost on the external set, with reliability diagrams in Figure 3. The raw model was already the best calibrated (Brier 0.086, ECE 0.021). Platt scaling left it essentially unchanged, and isotonic regression degraded it slightly (ECE 0.027). This ordering is expected rather than anomalous: a flexible monotone recalibration fitted at one hospital and applied at another has enough freedom to absorb site-specific idiosyncrasy, and the additional flexibility costs more than it recovers. We therefore used raw probabilities for all conformal analyses, which also keeps the nonconformity score interpretable.

Table 5: Calibration comparison for XGBoost (external validation).

| **Method** | **Brier** | **ECE** |
|---|---|---|
| Raw | **0.086** | **0.021** |
| Platt scaling | 0.086 | 0.021 |
| Isotonic regression | 0.087 | 0.027 |

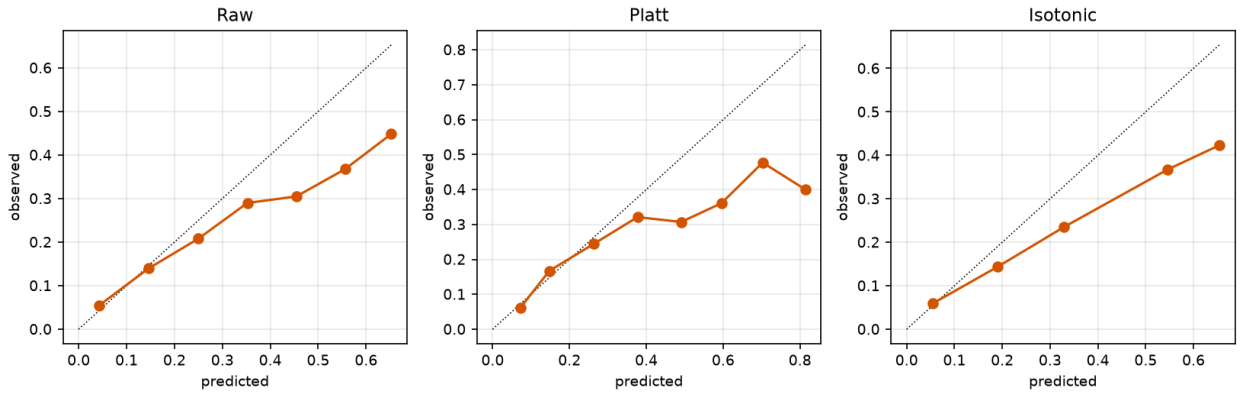



Figure 3: Reliability diagrams (external validation) for raw, Platt-scaled and isotonic-calibrated XGBoost probabilities.

## 3.5 Conformal prediction: the second inflation

Conformal thresholds were computed on the BIDMC calibration set ($m$ = 3,249) and applied to Emory. Table 6 compares marginal and class-conditional conformal prediction, and Figures 1(b) and 4 show set composition and coverage by class.

Table 6: Conformal prediction on external validation ($\alpha = 0.10$). Thresholds from the BIDMC calibration set.

| **Metric** | **Marginal** | **Class-cond.** |
|---|---|---|
| Threshold(s) | $\hat{q} = 0.599$ | $\hat{q}_0 = 0.228$, $\hat{q}_1 = 0.959$ |
| Coverage (overall) | 90.8% | 89.6% |
| Coverage (AKI = 1) | 14.3% | **88.4%** |
| Coverage (AKI = 0) | 99.4% | 89.8% |
| Singleton sets | 96.8% | 49.5% |
| Both-class sets | 3.2% | 50.5% |
| Empty sets | 0.0% | 0.0% |

Marginal conformal prediction met its target overall: 90.8% coverage (95% CI: 90.2%–91.3%). Coverage of the patients who developed AKI was 14.3% (95% CI: 12.2%–16.6%). By Lemma 1 this is permitted rather than pathological: at the external prevalence $\pi_1 = 0.102$ the floor is 1.7%, and the observed value sits inside the permitted range. Reported on its own, "90.8% coverage, guarantee satisfied" is a true statement that conveys the opposite of the clinical situation.

Class-conditional conformal prediction raised AKI coverage from 14.3% to 88.4% (95% CI: 86.2%–90.3%). The price is visible in Table 6: 50.5% of external patients receive a two-label set, against 3.2% under the marginal rule. That price is the one honestly incurred. By Corollary 1, obtaining the same protection through a marginal rule would require $\alpha \leq 0.0102$, a 99.0% rule, whose sets would be larger still.

The residual shortfall below 90% is informative. By Corollary 2 it cannot come from the prevalence difference between sites (11.9% vs. 10.2%), so it reflects within-class shift; by Proposition 2 the AKI score distributions at the two hospitals differ in total variation by at least $0.900 - 0.884 = 1.6$ percentage points. This is obtained from coverage alone, with no density estimation and no assumption about which features moved.

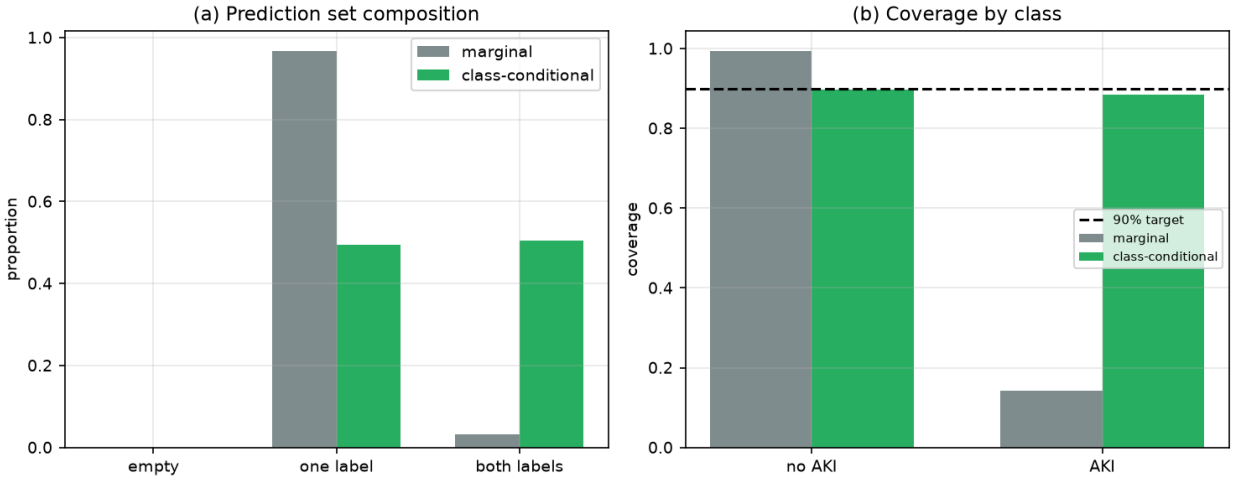



Figure 4: Conformal prediction on external validation. (a) Prediction set composition. (b) Coverage by class, comparing marginal and class-conditional approaches.

## 3.6 Group-conditional coverage and risk stratification

Class-conditional calibration guarantees coverage by outcome class but says nothing about coverage within demographic subgroups, which is a separate and equally deployable concern. Table 7 and Figure 5 give group-conditional coverage under the Mondrian construction. Both gender groups and the two older age groups sat within one point of target. The $< 50$ group reached 83.9%, well below target, on the smallest calibration subset ($m_g = 696$); we report

rather than omit this, since it instantiates the same lesson as the main result — a guarantee estimated on few patients is a weak guarantee, and subgroup coverage has to be checked rather than assumed.

Table 7: Mondrian conformal prediction: group-conditional coverage on external validation.

| Group | $m_g$ | $\hat{q}_g$ | Coverage | Avg size |
|---|---|---|---|---|
| *By gender* | | | | |
| Female | 1,385 | 0.533 | 91.3% | 1.01 |
| Male | 1,864 | 0.639 | 90.5% | 1.05 |
| *By age group* | | | | |
| $<50$ years | 696 | 0.186 | 83.9% | 0.89 |
| 50–65 years | 917 | 0.530 | 90.6% | 1.01 |
| $>65$ years | 1,636 | 0.691 | 90.7% | 1.08 |

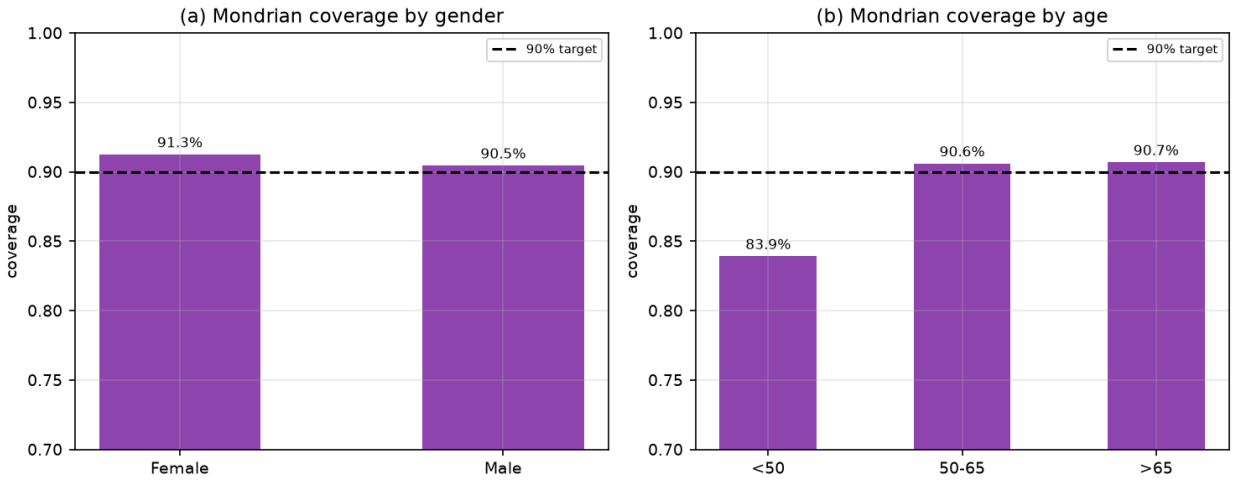



Figure 5: Mondrian conformal coverage on external validation by (a) gender and (b) age group. Dashed line is the 90% target.

Table 8 translates the class-conditional sets into the two-level risk stratification of Eq. (4).

The separation is real but modest, and much weaker than a whole-stay analysis of the same data would suggest. A rule that flags 63.4% of the ward is not an actionable alert, and we state this directly: at this level of discrimination, conformal risk stratification is an instrument for representing uncertainty faithfully rather than a triage tool. The Low Risk group is the useful half of the output — 36.6% of patients at a 3.2% AKI rate and 96.8%

Table 8: Conformal risk stratification (external validation).

| **Category** | $n$ | % | **AKI rate** | **NPV** |
|---|---|---|---|---|
| Low Risk $\{\mathcal{C} = \{0\}\}$ | 3,501 | 36.6% | 3.2% | 96.8% |
| Elevated Risk $\{1 \in \mathcal{C}\}$ | 6,070 | 63.4% | 14.2% | — |

NPV, carrying a formal guarantee that the missed-AKI probability is at most $\alpha$.

## 3.7 Explainability and clinical utility

Figure 6 shows the SHAP summary and Figure 7 the model-based importance ranking. Under the landmark design the leading predictors are baseline creatinine (mean |SHAP| = 0.49), ICU unit (0.22), last BUN within the first 24 hours (0.19), age (0.13) and last phosphate (0.12).

The ordering is clinically coherent, and the contrast with the whole-stay arm is itself diagnostic. When features are drawn from the entire stay the model relies most heavily on last-value phosphate and last-value BUN — results drawn late, by which time the creatinine rise defining the outcome is already established. Restricting to the first 24 hours moves baseline creatinine to the top, which is a legitimate pre-landmark risk marker rather than a consequence of the outcome. Explanation methods are frequently offered as a safeguard against precisely this failure, on the reasoning that a leaking model will betray itself through implausible attributions. Our whole-stay ranking shows why that reassurance is thin: phosphate and BUN are entirely plausible AKI predictors, and a reader shown only that ranking would have no reason for concern. Explanations describe what a model uses, not when the model was permitted to see it, and the distinction is invisible to SHAP.

Figure 8 shows the decision curve analysis. The model yields positive net benefit relative to both treat-all and treat-none across a narrow band of low threshold probabilities, which is what its modest discrimination and 10.2% event rate would predict. We read this as delimiting the clinical role available to a correctly evaluated model of this strength: useful for de-escalation decisions at low thresholds, and of little value for identifying patients for

intervention at higher ones.

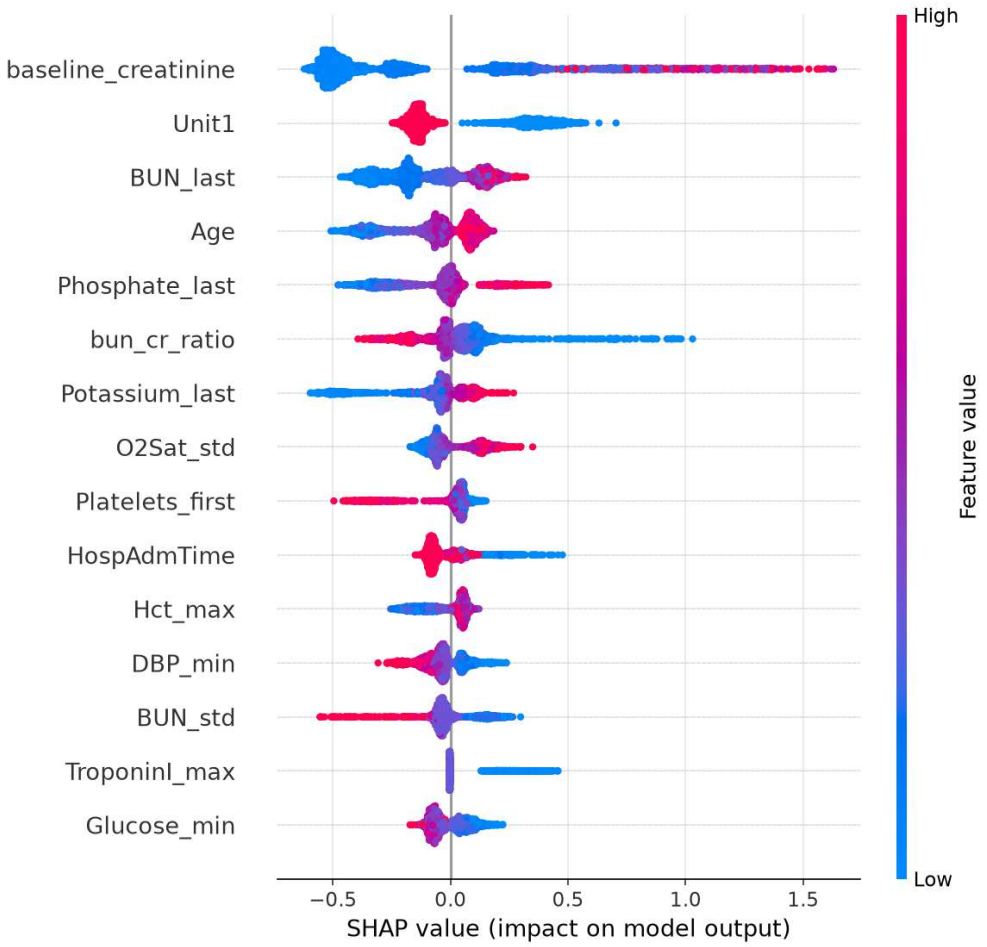



Figure 6: SHAP summary plot for XGBoost on external validation under the landmark design. Each dot is one patient; color indicates feature value.

## 3.8 Cross-validation and the cost of changing hospital

Five-fold cross-validation within BIDMC (Table 9, Figure 9) gave AUC $0.802 \pm 0.013$, marginal coverage $90.2\% \pm 0.6\%$ and class-conditional AKI coverage $90.5\% \pm 3.2\%$.

The comparison with the external result is the point. Internally the model achieves 0.802 and both conformal guarantees hold; at the second hospital discrimination falls to 0.735 and class-conditional AKI coverage falls to 88.4%. A study reporting only internal cross-validation would have reported a model 0.067 AUC better than it is, with coverage guarantees

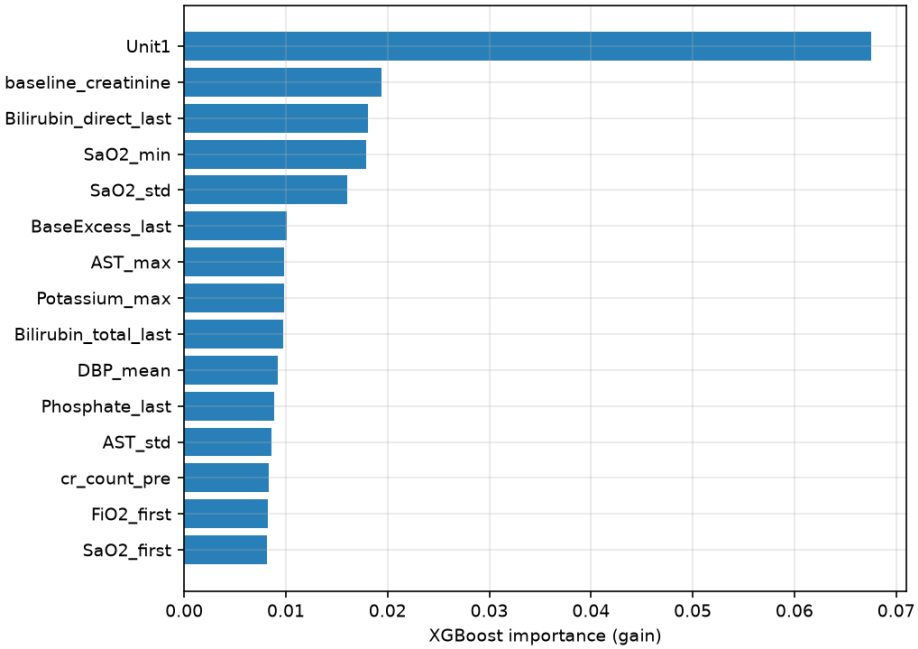



Figure 7: Top 15 features by XGBoost gain importance under the landmark design.

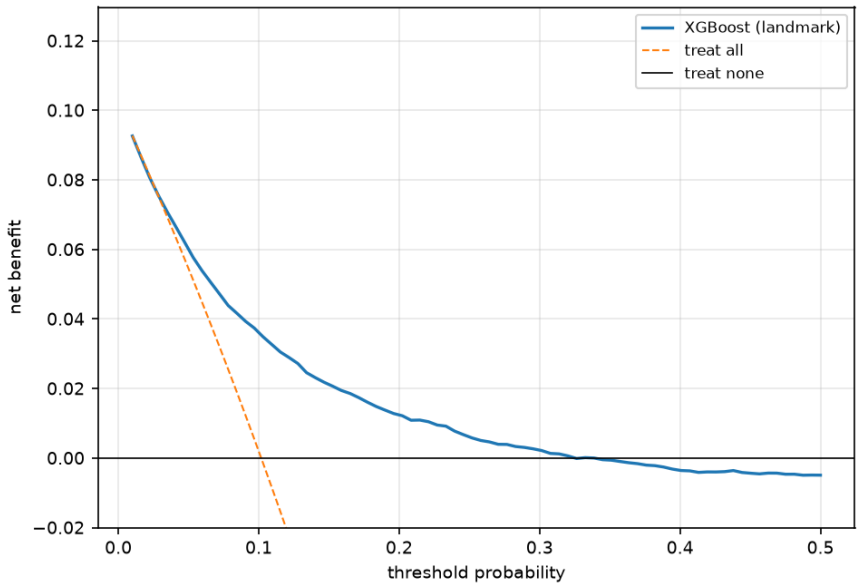



Figure 8: Decision curve analysis on external validation.

that appear to hold exactly. This is a third way for a number to be optimistic, and it is the one external validation exists to catch. Fold-to-fold spread is worth reporting alongside the mean for the same reason that motivates stability-calibrated analysis:[27] a point estimate that moves under resampling is a point estimate that will move under deployment. Here the class-conditional AKI coverage varies by 7.3 percentage points across folds (84.9%–92.2%) while the mean sits at 90.5%, so the mean alone would overstate how reliably the per-class guarantee is attained.

Table 9: Five-fold cross-validation within the development hospital (BIDMC).

| **Fold** | **AUC** | **Coverage** | **CC-AKI** |
|---|---|---|---|
| 1 | 0.805 | 90.4% | 91.8% |
| 2 | 0.803 | 90.6% | 91.4% |
| 3 | 0.808 | 90.7% | 92.2% |
| 4 | 0.780 | 89.2% | 84.9% |
| 5 | 0.814 | 90.0% | 92.2% |
| Mean ± SD | $0.802 \pm 0.013$ | $90.2\% \pm 0.6\%$ | $90.5\% \pm 3.2\%$ |
| *External (Emory)* | *0.735* | *90.8%* | *88.4%* |

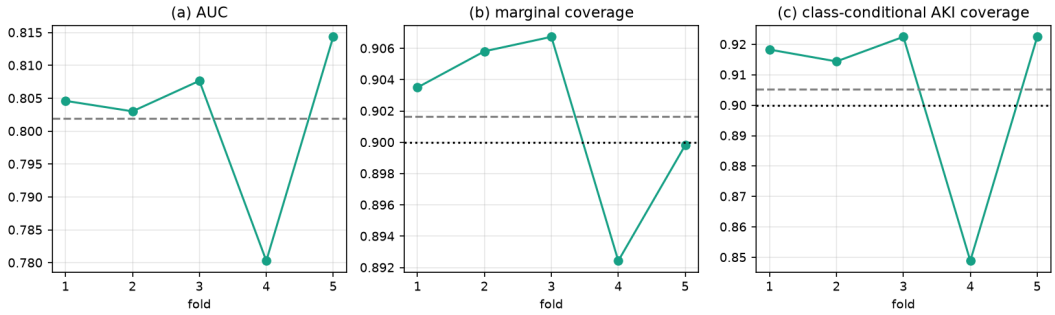



Figure 9: Cross-validation stability within BIDMC. (a) AUC, (b) marginal coverage, (c) class-conditional AKI coverage.

# 4 Discussion

## 4.1 Summary of findings

On one ICU cohort we measured two ways a prediction model can report a better number than it deserves. Aggregating features over the whole ICU stay raised external AUC from 0.732 to 0.919 with everything else held fixed, an inflation of 0.188 that recurred at every prediction time we tried. Reporting conformal coverage marginally produced a true statement — 90.8% coverage, guarantee satisfied — while the patients who developed AKI were covered 14.3% of the time. Class-conditional calibration recovered 88.4% coverage at the cost of two-label sets for half the ward, and the residual shortfall bounds the within-class difference between the hospitals at 1.6 percentage points.

## 4.2 Why the feature window matters more than it looks

The whole-stay design does not arise from an implementation error. Each step is individually reasonable: summarise each variable over the stay, define AKI by the KDIGO creatinine criteria, exclude the creatinine trajectory features that construct the label. That last exclusion is precisely the safeguard a careful analyst would think to apply, and it is insufficient, because BUN and phosphate track creatinine. The model is then rewarded for reading the outcome off correlated laboratory results drawn after it began.

What makes the design hazardous is that the resulting number appears entirely credible. An AUC of 0.919 on genuine external data, with a sensible SHAP ranking, is approximately what a competent AKI model would be expected to produce. The only reliable defence is structural: fix a prediction time and refuse to look past it. Reporting guidelines ask for this,[15] systematic reviews repeatedly find it absent,[14] and no reader can verify it from the metrics alone. Authors should state the landmark explicitly; reviewers should ask "prediction at what time point?" as a matter of routine.

## 4.3 Why marginal coverage is the wrong guarantee for the ICU

The 14.3% figure is the part of this study that generalises furthest. A marginal $1 - \alpha$ guarantee constrains coverage of a class with prevalence $\pi_1$ only down to $\max\{0, 1 - \alpha/\pi_1\}$, and once the event is rarer than $\alpha$ that floor is zero. Most acute events an ICU alerting system exists to catch — AKI, sepsis, arrest, unplanned transfer — occur in well under 10% of admissions, so a 90% marginal conformal rule offers them no protection in principle.

One might hope to fix this while keeping a single threshold, by demanding more coverage overall. Corollary 1 shows what that costs: here a 99.0% rule, with prediction sets that only grow. A method that answers "it could be either" for most of the ward is not a decision aid. This is the argument that makes per-class calibration necessary rather than merely preferable.

The resulting recommendation is straightforward. Coverage should be reported for the patients who have the condition. A marginal coverage figure for a rare outcome is not wrong, but it is close to uninformative, and it reads as reassurance.

The argument is not specific to prediction sets. Conformal risk control replaces coverage with the expectation of a bounded loss and controls it at a chosen level,[25, 26] but that expectation is again taken over the whole population. Whenever the loss is carried almost entirely by a rare subgroup, the same dilution applies and the same per-class or per-group construction is the remedy. Lemma 1 is stated for coverage because that is what we measure here; the mechanism it describes is a property of averaging, not of conformal prediction.

**Reading a coverage shortfall as a measurement.** Class-conditional coverage reached 88.4% rather than 90%. Reported in isolation this reads as a failure. Interpreted through Propositions 1 and 2 it becomes a quantity: the guarantee is immune to the prevalence difference between the hospitals, so the 1.6-point shortfall is a lower bound on the within-class distance between them. We would recommend this reading whenever a model crosses sites, because it distinguishes two situations that a raw coverage figure conflates — that the

receiving hospital has more disease, and that its patients are physiologically different — and only the second is an argument for recalibration. The decomposition costs nothing beyond the coverage numbers already being computed.

## 4.4 What this model is and is not

The clinical claim requires precision. At an external AUC of 0.735 this is not a deployable AKI early-warning system, and we do not present it as one. It is a correctly evaluated one, which here is the more useful thing: it is the honest baseline against which 0.919 and 0.935 can be read. The gap between those numbers is the contribution.

Two features of the data cap the clinical ceiling. The source dataset was assembled for a sepsis challenge, so its laboratory panel is sparse, and 24 hours of it is thin evidence on which to predict AKI. Richer sources such as MIMIC-IV or eICU would very likely support better discrimination. That would change the clinical claim; it would not touch either methodological finding, since both are properties of the evaluation design rather than of the model.

**Deep learning versus tree-based methods.** A secondary observation is consistent across both designs. The multi-layer perceptron (AUC 0.666) fell behind every tree-based method (0.718–0.735), reproducing benchmarks in which gradient boosting outperforms deep learning on tabular data,[28] particularly where columns are heterogeneous and missingness is heavy, as here. Tree ensembles accommodate mixed column types natively, are invariant to monotone rescaling and perform implicit feature selection, whereas a neural network must learn each of these from data that, at this sample size, cannot support it. For tabular AKI prediction, gradient boosting remains the appropriate default unless the hourly sequence itself is the object of modelling.

## 4.5 Limitations

1. **Modest discrimination.** External AUC 0.735 is below what published AKI models report, partly because those models are often evaluated without a landmark and partly because this dataset is sparse. The comparison that matters here is internal to the study: 0.735 against 0.919 on identical data.

2. **Repurposed dataset.** PhysioNet 2019 was built for sepsis prediction. It is public, hourly and two-site, which is what this design needs, but a purpose-built AKI cohort would be better.

3. **Creatinine-only KDIGO.** The urine-output criterion is unavailable in these data, so some AKI is missed.

4. **One dataset.** The ablation is repeated across four prediction times but on a single cohort. Whether the +0.18 figure transfers to MIMIC-IV or eICU is the obvious next test, and we would expect the direction but not necessarily the size to hold.

5. **Two US academic hospitals.** Community, international and pediatric validation is still needed.

6. **Baseline creatinine.** Taking the first in-window creatinine as baseline sets it too high for patients admitted with creatinine already raised.

7. **The floor is a worst case.** Lemma 1 bounds how little coverage a marginal rule *may* afford the rare class, not how little it will: our own model achieved 14.3% against a floor of 1.7%. The substantive point is that nothing in the marginal guarantee excludes the adverse case, and that a practitioner reading only the marginal figure has no way to tell which regime they are in.

# 5 Conclusion

We measured two inflations on one ICU cohort. Aggregating features over the whole stay was worth +0.188 AUC with everything else held fixed. Reporting conformal coverage marginally produced a satisfied 90% guarantee that covered 14.3% of the patients who developed AKI, which we show is permitted by the guarantee itself: a marginal $1-\alpha$ rule bounds coverage of a class of prevalence $\pi_1$ only by $\max\{0, 1-\alpha/\pi_1\}$, and buying back 90% coverage marginally would need a 99.0% rule. Class-conditional calibration recovers per-class coverage directly, keeps its guarantee when prevalence changes between sites, and turns whatever coverage it still loses into a lower bound on within-class shift.

Neither inflation requires an error in any individual step, and neither is visible in the numbers that are conventionally reported. Two defaults would prevent both: fix a prediction time and use only what was observable before it, and report coverage for the patients who have the condition rather than for the ward as a whole. All code and data are released so that both measurements can be checked and repeated elsewhere.

# Data availability statement

All code is publicly available at `https://github.com/jopoku16/Statistical-Paper-8`. The PhysioNet 2019 dataset is available at `https://physionet.org/content/challenge-2019/`.

# Declaration of generative AI in the writing process

AI tools were used for brainstorming ideas and debugging code. The authors reviewed all text, code, analyses, and conclusions.

## Declaration of conflicting interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

## Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

## Ethical approval

This study used the publicly available, de-identified PhysioNet/Computing in Cardiology Challenge 2019 dataset.[19] The data were collected under institutional review board approval at the originating institutions and released under the PhysioNet Credentialed Health Data Use Agreement. As a secondary analysis of fully de-identified public data, this study was exempt from additional ethical approval.

## References

[1] Hoste EAJ, Bagshaw SM, Bellomo R, et al. Epidemiology of acute kidney injury in critically ill patients: the multinational AKI-EPI study. *Intensive Care Med* 2015; 41(8): 1411–1423.

[2] Kellum JA, Romagnani P, Ashuntantang G, et al. Acute kidney injury. *Nat Rev Dis Primers* 2021; 7: 52.

[3] Chawla LS, Bellomo R, Bihorac A, 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): 241–257.

[4] Siew ED, Ikizler TA, Matheny ME, et al. Kidney disease awareness and knowledge among survivors of acute kidney injury. *Am J Nephrol* 2020; 51(6): 449–459.

[5] Meersch M, Schmidt C, Hoffmeier A, et al. Prevention of cardiac surgery–associated AKI by implementing the KDIGO guidelines in high risk patients identified by biomarkers. *Intensive Care Med* 2017; 43(11): 1749–1758.

[6] Koyner JL, Carey KA, Edelson DP, et al. The development of a machine learning inpatient acute kidney injury prediction model. *Crit Care Med* 2018; 46(7): 1070–1077.

[7] Kate RJ, Perez RM, Mazumdar D, et al. Prediction and detection models for acute kidney injury in hospitalized older adults. *BMC Med Inform Decis Mak* 2020; 20: 75.

[8] Churpek MM, Carey KA, Edelson DP, et al. Internal and external validation of a machine learning risk score for acute kidney injury. *JAMA Netw Open* 2020; 3(8): e2012892.

[9] Tomašev N, Glorot X, Rae JW, et al. A clinically applicable approach to continuous prediction of future acute kidney injury. *Nature* 2019; 572(7767): 116–119.

[10] Li Y, Yao L, Mao C, et al. Early prediction of acute kidney injury in critical care setting using clinical notes. In: *Proc IEEE BIBM* 2021; pp.683–686.

[11] Wilson FP, Martin M, Yamamoto Y, et al. Electronic health record alerts for acute kidney injury: multicentre, randomised clinical trial. *BMJ* 2021; 372: m4786.

[12] Kaufman S, Rosset S, Perlich C, et al. Leakage in data mining: formulation, detection, and avoidance. *ACM Trans Knowl Discov Data* 2012; 6(4): 15.

[13] van Houwelingen HC. Dynamic prediction by landmarking in event history analysis. *Scand J Stat* 2007; 34(1): 70–85.

[14] Wynants L, Van Calster B, Collins GS, et al. Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal. *BMJ* 2020; 369: m1328.

[15] Collins GS, Reitsma JB, Altman DG, et al. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *BMJ* 2015; 350: g7594.

[16] Vovk V, Gammerman A and Shafer G. *Algorithmic Learning in a Random World*. New York: Springer, 2005.

[17] Lei J, G'Sell M, Rinaldo A, et al. Distribution-free predictive inference for regression. *J Am Stat Assoc* 2018; 113(523): 1094–1111.

[18] Romano Y, Sesia M and Candès EJ. Classification with valid and adaptive coverage. In: *Adv Neural Inf Process Syst* 2020; 33: 3581–3591.

[19] Reyna MA, Josef CS, Jeter R, et al. Early prediction of sepsis from clinical data: the PhysioNet/Computing in Cardiology Challenge 2019. *Crit Care Med* 2020; 48(2): 210–217.

[20] Chen T and Guestrin C. XGBoost: a scalable tree boosting system. In: *Proc 22nd ACM SIGKDD* 2016; pp.785–794.

[21] Ke G, Meng Q, Finley T, et al. LightGBM: a highly efficient gradient boosting decision tree. In: *Adv Neural Inf Process Syst* 2017; 30.

[22] Tibshirani RJ, Foygel Barber R, Candès EJ, et al. Conformal prediction under covariate shift. In: *Adv Neural Inf Process Syst* 2019; 32.

[23] Opoku J and Banahene D. Spectral adaptive conformal prediction for structured non-exchangeable data. *arXiv preprint* arXiv:2606.15950, 2026.

[24] Opoku J and Banahene D. Drift-aware spectral conformal prediction for non-exchangeable streaming data. *arXiv preprint* arXiv:2606.15953, 2026.

[25] Opoku J and Banahene D. PromptShift-CRC: drift-aware conformal risk control for foundation models under prompt and domain shift. *arXiv preprint* arXiv:2606.15964, 2026.

[26] Opoku J and Banahene D. ToolChain-CRC: conformal risk control for agentic AI under retrieval and tool-use drift. *arXiv preprint* arXiv:2606.18467, 2026.

[27] Opoku J and Banahene D. Signed evidence flow: conflict-aware and stability-calibrated data analysis. *arXiv preprint* arXiv:2606.21875, 2026.

[28] Grinsztajn L, Oyallon E and Varoquaux G. Why do tree-based models still outperform deep learning on typical tabular data? In: *Adv Neural Inf Process Syst* 2022; 35.

[29] Lundberg SM and Lee SI. A unified approach to interpreting model predictions. In: *Adv Neural Inf Process Syst* 2017; 30.

[30] Platt JC. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In: *Advances in Large Margin Classifiers*. Cambridge, MA: MIT Press, 1999; pp.61–74.

[31] Zadrozny B and Elkan C. Transforming classifier scores into accurate multiclass probability estimates. In: *Proc ACM SIGKDD* 2002; pp.694–699.