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.

Sample size determines the algorithm: a two-cohort study of preoperative prediction models for postoperative acute kidney injury after noncardiac surgery

Zhen Dai
The Fourth Affiliated Hospital of Anhui Medical University
Xiaoyu Zhao
The Fourth Affiliated Hospital of Anhui Medical University
Shujiang Xia
xsj23669090@163.com

The Fourth Affiliated Hospital of Anhui Medical University

Research Article

Keywords: acute kidney injury, noncardiac surgery, prediction model, machine learning, sample size, learning curve

Posted Date: November 27th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11777639/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

Postoperative acute kidney injury (AKI) complicates 5 to 10% of noncardiac surgery. Whether machine learning outperforms logistic regression remains unsettled, yet nearly all comparisons use a single sample size. We tested whether algorithm performance depends on sample size.

Methods

Using an identical protocol, KDIGO outcome definition and pipeline (paired logistic regression vs XGBoost, bootstrap testing, out-of-fold recalibration, decision curves), we modelled preoperative AKI in two same-institution cohorts differing 37-fold in size: VitalDB (n = 3,408) and INSPIRE (n = 125,587). Within INSPIRE we drew a five-point subsampling learning curve; transportability and robustness (seeds, hyperparameter grid, inverse-probability weighting) were examined.

Results

Rankings reversed. VitalDB favoured logistic regression (AUROC 0.845 vs 0.741; paired difference 0.104 [0.050 to 0.165]); INSPIRE favoured XGBoost (0.866 vs 0.832; 0.034 [0.029 to 0.038]). The learning curve showed logistic regression plateauing by about 30,000 patients while XGBoost overtook it at 2,000 to 5,000 and kept rising. Small-sample models needed recalibration (slope 0.78 to 0.99); the large-sample model was nearly self-calibrated (slope 1.014). Transport failed (AUROC 0.6175) owing to unstable small-stratum coefficients; local re-estimation on INSPIRE recovered performance (AUROC 0.832). Rankings were preserved under a prespecified stringent outcome definition in both cohorts.

Conclusions

Algorithm choice should follow the volume of local data: a parsimonious model sufficed at routine single-centre sample sizes, whereas gradient boosting paid off at large scale. Small-sample models did not transport across databases; because transport broke ranking itself, recovery required full local re-estimation rather than intercept-and-slope updating alone.

Introduction

Acute kidney injury (AKI) develops in roughly one in ten patients after noncardiac surgery[1], bringing higher short-term mortality, progression to chronic kidney disease and rising costs[2]. The preventive measures that work, such as fluid planning, avoidance of nephrotoxins and adjusted monitoring[3], must start before induction, which makes a reliable preoperative risk tool the first link in the chain.

Which algorithm should build that tool is still contested. Machine learning has produced impressive discrimination in some reports[4, 5], yet an equally large body of work shows logistic regression matching or beating complex models on tabular clinical data[6]. One premise behind these comparisons has gone largely unexamined: almost all of them have been run at a single sample size. Learning theory nevertheless predicts a crossover as data accrue[7], with simple models winning on variance when data are scarce and flexible models winning on bias when data are abundant. Whether this crossover occurs in real perioperative data, with methods held strictly constant, has not been tested.

This study had an unusual opportunity: one hospital has released two databases of very different size with consistent variable definitions[8, 9]. We compared logistic regression and XGBoost under an identical protocol in both, testing the algorithm-by-sample-size interaction directly and mapping its boundaries with a subsampling learning curve, a transportability experiment and several robustness analyses.

Methods

We conducted a retrospective prognostic modelling study with internal validation. The protocol was locked and timestamped before any data extraction. The report follows TRIPOD + AI 2024[10], and risk of bias was appraised with PROBAST[11]. We used two databases from Seoul National University Hospital: VitalDB v1.0.0 (2016 to 2017; 6,388 operations; open access) and INSPIRE v1.4.2 (2011 to 2020; 130,960 operations; credentialed access). Both are de-identified. The original data collections were approved with a waiver of informed consent by the institutional review board of Seoul National University Hospital (IRB No. H-1408-101-605); the present study is a secondary analysis of the publicly released, de-identified versions only. Because VitalDB (2016 to 2017) falls within the INSPIRE accrual period (2011 to 2020) and patient identifiers are not linkable across the two releases, partial patient overlap cannot be excluded; the transport analysis below is therefore interpreted as cross-database rather than strictly cross-population.

Eligible adults (age 18 years or older) had at least one serum creatinine within 7 days before anaesthesia. The baseline value was the most recent such measurement. We excluded cardiac surgery, CKD stage 4 to 5 (baseline creatinine 4.0 mg/dL or higher, approximating eGFR below 30), ASA physical status 6 (organ donors) and implausible age or BMI. The unit of analysis was the operation. In INSPIRE, 22% of patients underwent more than one operation. Each counted separately. Cohort derivation is shown in Fig. 1 for VitalDB and Figure S1 for INSPIRE.

The primary outcome was AKI within 7 days after surgery or before discharge, whichever came first, by KDIGO creatinine criteria[1]: an increase of at least 0.3 mg/dL within 48 h, or to at least 1.5 times baseline within 7 days. Values outside 0.2 to 25.0 mg/dL were discarded as probable artefacts. A single qualifying measurement defined AKI. A prespecified stringent definition required two concordant measurements. Patients without postoperative creatinine were classified as no AKI. We tested this assumption with inverse-probability weighting. The definition was identical in both cohorts. The algorithm was implemented independently in SQL and pandas with case-by-case agreement.

All predictors were required to be available before induction. These were age, sex, weight, height, ASA class, emergency surgery, department and the most recent preoperative values for albumin, urea, sodium, potassium, haemoglobin, platelets, glucose, PT-INR, APTT, AST, ALT, bilirubin and CRP. Categories with fewer than 10 events or 50 cases were merged to prevent separation. Harmonisation between databases was not routine. The data dictionary diverged from the released data in several respects, including an unreleased field, naming mismatches and recorded scales at odds with the dictionary. We verified and documented every discrepancy. The anaesthesia-type variable, present in VitalDB only, was excluded from cross-cohort analyses.

Each cohort was split 70/30 with stratified sampling (seed 42). Model 1 was L2-penalised logistic regression. The penalty was chosen by 5-fold cross-validation within the training set from 20 candidates. Model 2 was XGBoost (300 trees, max depth 4, learning rate 0.05). Hyperparameters were fixed a priori based on pilot tuning in the VitalDB training set to avoid overfitting in the small cohort. Categorical variables were one-hot encoded, continuous variables standardised, and missing values were imputed by median imputation for continuous variables and mode imputation for categorical variables within the pipeline to prevent leakage. We report AUROC, AUPRC and Brier score with 500 bootstrap resamples of the test set. Models were compared by paired bootstrap. Calibration was assessed by intercept, slope and a decile plot[12]. We fitted recalibration on out-of-fold training predictions and applied it to the test set, never using test information. Clinical utility used decision curve analysis across thresholds of 1 to 50%.

To map the algorithm-by-sample-size relation in full, we drew stratified subsamples of the INSPIRE training set at 2,000, 5,000, 10,000, 30,000 and 87,910. We trained both models with identical settings and evaluated them on the same test set. Robustness analyses comprised five re-splits with different seeds, a hyperparameter grid for XGBoost (depth 3, 4 and 6 crossed with learning rate 0.03 and 0.1), inverse-probability weighting by the propensity of having postoperative creatinine measured, and multiple imputation by chained equations (20 sets).

A harmonised 14-variable logistic regression model trained on the full VitalDB cohort was applied directly to the INSPIRE test set. We report discrimination and calibration and dissect the failure mode at the level of coefficients and single variables. When the same 14-variable model was re-estimated on the INSPIRE training set, we report its performance on the INSPIRE test set to assess whether local re-estimation recovers transport failure. VitalDB contributed 3,408 patients with 144 events. INSPIRE contributed 125,587 operations with 9,323 events (EPV about 4.8 and over 300). Analyses used Python 3.11 (scikit-learn, XGBoost, pandas). AI assistance was used to draft and edit the analysis code for the stringent-definition sensitivity analyses and to polish the English language. All content was reviewed and verified by the authors, who take full responsibility for the integrity of the work.

Results

After exclusions (Fig. 1, Figure S1), the VitalDB cohort comprised 3,408 patients (AKI 144, 4.2%) and INSPIRE 125,587 operations (AKI 9,323, 7.42%). In INSPIRE, 33.5% of patients had no postoperative creatinine. Among those measured, AKI incidence was 10.9%. The all-comers rate (7.42%) underestimates true incidence because unmonitored mild AKI is missed; the measured rate (10.9%) overestimates it because monitored patients are selectively sicker. The true burden lies between these bounds. Baseline patterns agreed across cohorts. Emergency surgery was at least twice as common in AKI patients (24.3% vs 12.5% in VitalDB; 19.4% vs 8.3% in INSPIRE). Male sex and urological operations were similarly enriched, and albumin was about 0.4 g/dL lower. In INSPIRE, non-bypass cardiothoracic surgery carried the highest proportion of AKI (8.3% vs 4.7%), an easily overlooked high-risk group. Baseline characteristics by cohort are shown in Table 1; full distributions are in Supplement S7.

Table 1
Baseline characteristics by AKI status and cohort

Variable	VitalDB AKI (n = 144)	VitalDB no AKI (n = 3,264)	INSPIRE AKI (n = 9,323)	INSPIRE no AKI (n = 116,264)
Age, yr	60 (51–68)	61 (51–69)	65 (52.5–70)	55 (45–70)
Weight, kg	—	—	60 (55–70)	60 (55–70)
ASA class	2 (2–3)	2 (2–3)	2 (2–3)	2 (1–2)
Baseline creatinine, mg/dL	0.8 (0.7–1.1)	0.8 (0.7–0.9)	0.8 (0.7–1.1)	0.8 (0.7–0.9)
Albumin, g/dL	3.6 (2.9–4.1)	4.1 (3.8–4.3)	3.8 (3.2–4.2)	4.2 (4.0–4.4)
Haemoglobin	11.2 (9.2–13.6)	12.8 (11.5–14.0)	15.0 (10.0–21.0)	13.0 (10.0–17.0)
Urea	16 (12–22)	14 (11–18)	18 (13–30)	17 (13–27)
Emergency surgery	35 (24.3%)	407 (12.5%)	1,810 (19.4%)	9,613 (8.3%)
Male sex	113 (78.5%)	1,771 (54.2%)	5,715 (61.3%)	49,157 (42.3%)
Dept: general surgery	—	—	3,173 (34.0%)	30,514 (26.2%)
Dept: cardiothoracic	—	—	778 (8.3%)	5,410 (4.7%)
Dept: urology	—	—	2,206 (23.7%)	9,898 (8.5%)

Values are median (IQR) or n (%). In INSPIRE, recorded scales for sodium, potassium, glucose and platelets deviate from the data dictionary (Table S4) and are shown as recorded. The 37 anaesthesia-department operations were minor day cases with no AKI.

The two cohorts returned opposite rankings (Table 2, Fig. 2). In VitalDB, logistic regression reached an AUROC of 0.845 (95% CI 0.784 to 0.899) against 0.741 (0.656 to 0.829) for XGBoost. The paired difference was 0.104 (0.050 to 0.165). The wide bootstrap CI reflects the variance expected at an EPV of

about 5; the paired comparison with XGBoost, which shares the same training set and is evaluated on the same test set, is nevertheless internally valid. In INSPIRE, XGBoost led with 0.866 versus 0.832. The difference was 0.034 (0.029 to 0.038). AUPRC and Brier scores pointed the same way.

Table 2
Model performance

Metric	VitalDB LR	VitalDB XGB	INSPIRE LR	INSPIRE XGB
AUROC (95% CI)	0.845 (0.784–0.899)	0.741 (0.656–0.829)	0.832 (0.830–0.834)	0.866 (0.864–0.868)
AUPRC	0.274	0.194	0.320	0.390
Brier (raw to recalibrated)	0.0359 to 0.0355	0.0389	0.0587	0.0551 to 0.055
Calibration (raw to recalibrated)	slope 0.78 to 0.99; intercept − 0.13	—	—	intercept 0.036; slope 1.014
Paired difference (95% CI)	LR minus XGB + 0.104 (0.050–0.165)	—	—	XGB minus LR + 0.034 (0.029–0.038)
Learning-curve behaviour	plateau about 30k (ceiling 0.833)	—	near-plateau at 2,000	overtake at 2k to 5k; rising at 87,910

Recalibration was applied only to models that showed miscalibration (slope below 0.9 or above 1.1) on out-of-fold training predictions. Brier scores are not directly comparable across cohorts because of differing event rates; within-cohort comparisons (raw vs recalibrated) remain valid. INSPIRE bootstrap 95% CIs had width below 0.005 and are shown rounded to three decimals.

The subsampling curve completed the picture (Fig. 3). Logistic regression started at 0.813 with 2,000 training cases and had levelled off near 0.833 by 30,000, its bias ceiling. XGBoost began lower, at 0.808, crossed logistic regression between 2,000 and 5,000 cases, and then climbed steadily: 0.835, 0.849, 0.859, 0.866, still rising at 87,910. The curve endpoints reproduced the primary analysis to three decimals (0.8320 and 0.8655 vs 0.832 and 0.866). VitalDB, at 3,408 patients with logistic regression ahead, falls inside the 2,000-case region of this curve. A single learning curve thus explains both cohorts.

Calibration tracked sample size the same way (Fig. 4; INSPIRE counterparts in Figure S2). The small-sample logistic model arrived optimistic (slope 0.78, intercept − 0.66). It needed out-of-fold recalibration to become usable: slope 0.99, intercept − 0.13, Brier 0.0359 to 0.0355. The large-sample XGBoost was nearly calibrated as trained (intercept 0.036, slope 1.014; Brier 0.0551 vs 0.055 after recalibration). Its decile calibration lies on the diagonal across the full risk range (Figure S2A). Decision curves in both cohorts showed net benefit across thresholds of 1 to 50% over treating all or none.

A harmonised 14-variable logistic regression model trained on the full VitalDB cohort (3,408 patients, 144 events; internal CV AUROC 0.7591) collapsed on INSPIRE. AUROC was 0.6175. Calibration slope was

0.114, intercept − 2.721. The predicted mean risk was 0.598 against an observed event rate of 7.4% (Fig. 5). When the same model was re-estimated on the INSPIRE training set, AUROC recovered to 0.832, showing that local re-estimation restores performance. The mechanism is apparent from the coefficients. The largest belonged to urological surgery: +2.009, an odds ratio above 7. That estimate rests on 16 operations with 6 events. Applied to 12,104 urological patients in INSPIRE, it generated mass false positives and flattened the ranking. Albumin behaved differently across cohorts. It acted additively in VitalDB (OR 0.89). In INSPIRE its quartile AKI rates were flat (7.3%, 8.0%, 7.5%, 7.4%). Its predictive value lives in non-linear patterns that a linear transport model cannot capture.

Under the prespecified stringent definition (two concordant qualifying measurements), 104 operations (3.0%) met the AKI criterion. The between-model ranking was preserved: the logistic model again outperformed XGBoost (AUROC 0.821 vs 0.798; paired bootstrap difference 0.023, 95% CI -0.026 to 0.075), although with fewer events the paired difference was no longer statistically significant (Table S3b). When the primary single-measurement definition was re-run on the rebuilt VitalDB extraction, the original ranking was likewise preserved (logistic 0.735 vs XGBoost 0.728); absolute values differ from Table 2 for the reasons documented in the Table S4 audit. In the rebuilt INSPIRE extraction (n = 106,775; primary events 4,309, stringent 2,545), the ranking under both definitions again matched Table 2: XGBoost ahead of logistic regression (primary AUROC 0.858 vs 0.809, paired bootstrap 95% CI -0.057 to -0.040; stringent 0.895 vs 0.882, CI -0.019 to -0.007), with absolute values slightly lower than Table 2, for reasons documented in the Table S4 audit.

Across five seeds, XGBoost AUROC ranged 0.8572 to 0.8655 (mean 0.8622), always ahead of logistic regression. All six hyperparameter combinations scored 0.8601 to 0.8673; the ranking never changed. IPW reweighting toward monitored patients raised AUROC from 0.8655 to 0.8947. Sparse monitoring dilutes measured performance rather than distorting it. In the monitored population the model discriminates substantially better than the whole-cohort estimate suggests. Multiple and median imputation agreed (0.842 vs 0.843; full outputs in Supplement S3).

The predictor structure agreed closely across cohorts (Fig. 6). Urological surgery ranked first in both: INSPIRE importance 0.181, VitalDB OR 4.8. Albumin ranked second (0.153). Baseline creatinine sat ninth (0.030). ICD-10 comorbidities added nothing. With diabetes 10.5%, hypertension 10.3%, heart failure 0.9%, COPD 2.2% and malignancy 41.7% defined from full history, AUROC stayed at 0.866. Comorbidity information is evidently captured by ASA class and preoperative tests.

Discussion

In two surgical cohorts from the same institution, differing 37-fold in size and analysed under strictly identical methods, the algorithm ranking reversed. Logistic regression won by 0.104 AUROC at 3,408 patients; XGBoost won by 0.034 at 125,587. The five-point subsampling curve turned this two-point observation into a complete picture: logistic regression plateaus near 0.833 and stops learning, whereas XGBoost overtakes between 2,000 and 5,000 cases and keeps gaining. Variable definitions, outcome

adjudication and the pipeline were locked in advance, the subsampling drew from one data pool, and the ranking was invariant across the hyperparameter grid, making confounding or tuning artefacts unlikely explanations for the reversal. Within a given prediction task and predictor set, the practical message is that algorithm choice should follow the volume of local data rather than allegiance to a modelling school. Most single-centre studies number in the thousands and are best served by parsimonious models; hospital-scale data in the tens of thousands justify gradient boosting. The exact crossover location will vary with the outcome, predictor set and pipelines, but the qualitative rule should not.

Sample size also governed calibration. Small-data models arrive optimistic (slope 0.78) and need recalibration before their probabilities are usable, whereas with enough data flexible models are nearly calibrated as trained. Calibration is, after all, a matter of averaging error away, which requires events in every predicted stratum; volume therefore governs both discrimination and the accuracy of risk estimates. When a model's ranking survives transport and only its overall level is off, moving it to a new population compresses into two numbers, an intercept and a slope, and develop-once-update-locally is a realistic route, as implementation experience in the British Journal of Anaesthesia has begun to show[13]. Our transport experiment also shows where this route ends: when transport breaks ranking itself, as it did for the small-sample model (calibration slope 0.114), intercept-and-slope updating cannot help, and full local re-estimation is the only repair. The two-number update is therefore a large-data tool, not a small-sample one. This dovetails with the learning-curve result: data volume decides both which algorithm to use and how to deploy it.

The transport experiment adds independent evidence from the opposite direction. Even after harmonisation, the small-cohort model failed on INSPIRE: AUROC 0.62, with predicted mean risk inflated eightfold. The reason is readable in its coefficients. An unstable urology coefficient, built on 16 operations, misfired across 12,104, and recalibration can correct overall level but cannot repair ranking. When a predictor effect changes form, as albumin did between the cohorts, linear transport cannot carry it at all. Local re-estimation, in contrast, recovers performance entirely (AUROC 0.832 when the same model was retrained on the target data). For small-sample models, then, the defensible route is not transport but local re-estimation.

The signal structure points clinical attention in specific directions. Urological surgery and preoperative albumin head both lists. Baseline creatinine and ICD-10 comorbidities play minor roles. In the all-comers INSPIRE cohort, non-bypass cardiothoracic surgery carried the highest proportion of AKI (8.3%), a high-risk group that current risk awareness tends to overlook. Predictable preoperative risk appears to reside chiefly in surgical burden and immediate physiology. Nutrition optimisation, relief of urinary obstruction and targeted assessment before major thoracic surgery may matter more than managing a comorbidity list.

Set against the literature, years of apparently contradictory findings resolve into one frame. The subgroup signal from a 76,314-patient cohort that preoperative models generalise best[14], the recent first independent external validation[15], and the many single-centre reports of machine learning failing to

beat logistic regression each occupy a different segment of one learning curve. Our contribution is to make the algorithm-by-sample-size interaction itself the object of study: we pinned it down with a reversal in both directions plus a complete curve. The practical rule we leave behind is simple: report sample size right alongside AUROC and calibration, not tucked away in the methods.

Two practical implications follow. Below roughly five thousand cases we would default to a parsimonious model, allowing for recalibration; gradient boosting earns its keep only above tens of thousands. In between, the honest option is to run both and let a paired test decide. For clinicians, the deployable artefact can be simple, provided it is re-estimated locally: a parsimonious model whose coefficients are refreshed on local data, reduced to an intercept-and-slope update only where ranking survives transport. Emergency surgery would be a natural first use case: emergency status was the strongest single predictor in our data, and emergency decision-making is the most time-pressured.

Four limitations deserve plain statement. First, both cohorts come from one Korean tertiary hospital. Cross-institutional and cross-ethnic transport awaits other data; access requests are in progress. The single-institution design strengthens the causal reading of the interaction but limits generalisability. Because cohort size is entangled with era and pipeline in the two-cohort contrast, the within-pool INSPIRE subsampling curve, not the cross-cohort contrast, carries the causal weight of the sample-size claim. Partial overlap between the two releases cannot be excluded, but shared patients would have made transport easier, not harder. Second, reverse transport (INSPIRE to VitalDB) was not attempted. A 144-event target cohort cannot meaningfully evaluate a model trained on 88,000 cases. Third, postoperative creatinine monitoring in INSPIRE was sparse (33.5% unmeasured). Patients without postoperative creatinine were classified as no AKI, which underestimates true incidence; we tested this assumption with inverse-probability weighting, but unmeasured confounders may remain. The IPW analysis indicates dilution rather than distortion, and the between-cohort comparison is unaffected in direction. Fourth, comorbidity capture reflects EMR coding habits. Hypertension 10.3% is under-ascertained, malignancy 41.7% appears inflated by rule-out codes. The null incremental value of comorbidities should be read in that context. Everything here is retrospective. Prospective impact on outcomes remains to be shown. Finally, the VitalDB logistic AUROC (0.845) exceeding the INSPIRE plateau (0.833) reflects pipeline and era differences rather than small data beating large data; our inferences rest on within-cohort paired comparisons, not on absolute levels.

In summary, holding methods strictly constant, we found that the ranking of algorithms for preoperative AKI prediction reverses with sample size, and that the recalibration burden dissolves as data accumulate. Small-sample models cannot be transported across databases, but full local re-estimation restores them. For routine sample sizes the parsimonious model remains the better tool; machine learning pays off only with large-scale local data.

Abbreviations

AKI

acute kidney injury
ASA
American Society of Anesthesiologists
AUROC
area under the receiver operating characteristic curve
AUPRC
area under the precision-recall curve
BMI
body mass index
CI
confidence interval
CKD
chronic kidney disease
CRP
C-reactive protein
DCA
decision curve analysis
eGFR
estimated glomerular filtration rate
EMR
electronic medical record
EPV
events per variable
ICD-10
International Classification of Diseases,Tenth Revision
IPW
inverse-probability weighting
IQR
interquartile range
KDIGO
Kidney Disease: Improving Global Outcomes
LR
logistic regression
ML
machine learning
OR
odds ratio
PROBAST
Prediction model Risk Of Bias ASsessment Tool
TRIPOD

Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis
XGBoost
extreme gradient boosting.

Declarations

Ethics approval and consent to participate

This study is a secondary analysis of two de-identified, publicly available datasets (VitalDB and INSPIRE). The original data collections were approved with a waiver of informed consent by the institutional review board of Seoul National University Hospital (IRB No. H-1408-101-605); this secondary analysis used only publicly released, de-identified data.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Authors' contributions

ZD: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Visualization, Writing – original draft. XZ: Conceptualization, Methodology, Supervision, Writing – review & editing. SX: Conceptualization, Methodology, Supervision, Writing – review & editing, Guarantor. All authors read and approved the final manuscript.

Use of AI tools

During the preparation of this manuscript, the authors used an AI to assist with drafting and editing the analysis code for the stringent-definition sensitivity analyses and with English language polishing. All

data, analyses, results and interpretations were reviewed and verified by the authors, who take full responsibility for the integrity and accuracy of the work.

Reporting guideline

This report follows the TRIPOD + AI 2024 statement; the completed checklist is provided as Table S6 and as a separate file. Risk of bias was assessed with PROBAST (Table S5).

Funding:

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Author Contribution

ZD conceived the study, curated and analysed the data, and wrote the manuscript. XZ supervised the study and revised the manuscript critically for important intellectual content. SX contributed to the study design, interpreted the data, revised the manuscript critically, and is the guarantor. All authors read and approved the submitted version.

Acknowledgement

The authors thank the teams maintaining VitalDB and INSPIRE for open and credentialed data access. During the preparation of this manuscript, the authors used AI to assist with drafting and editing the analysis code for the stringent-definition sensitivity analyses and with English language polishing; all content was reviewed and verified by the authors, who take full responsibility for the integrity of the work.

Data Availability

VitalDB is openly available at https://vitaldb.net. INSPIRE is available under credentialed access; applications are reviewed by the data custodians. Analysis code, the locked protocol and derived tables are available at https://github.com/zhendai123/postop-aki-vitaldb (to be made made public upon acceptance; interim access from the corresponding author on reasonable request).

Code availability

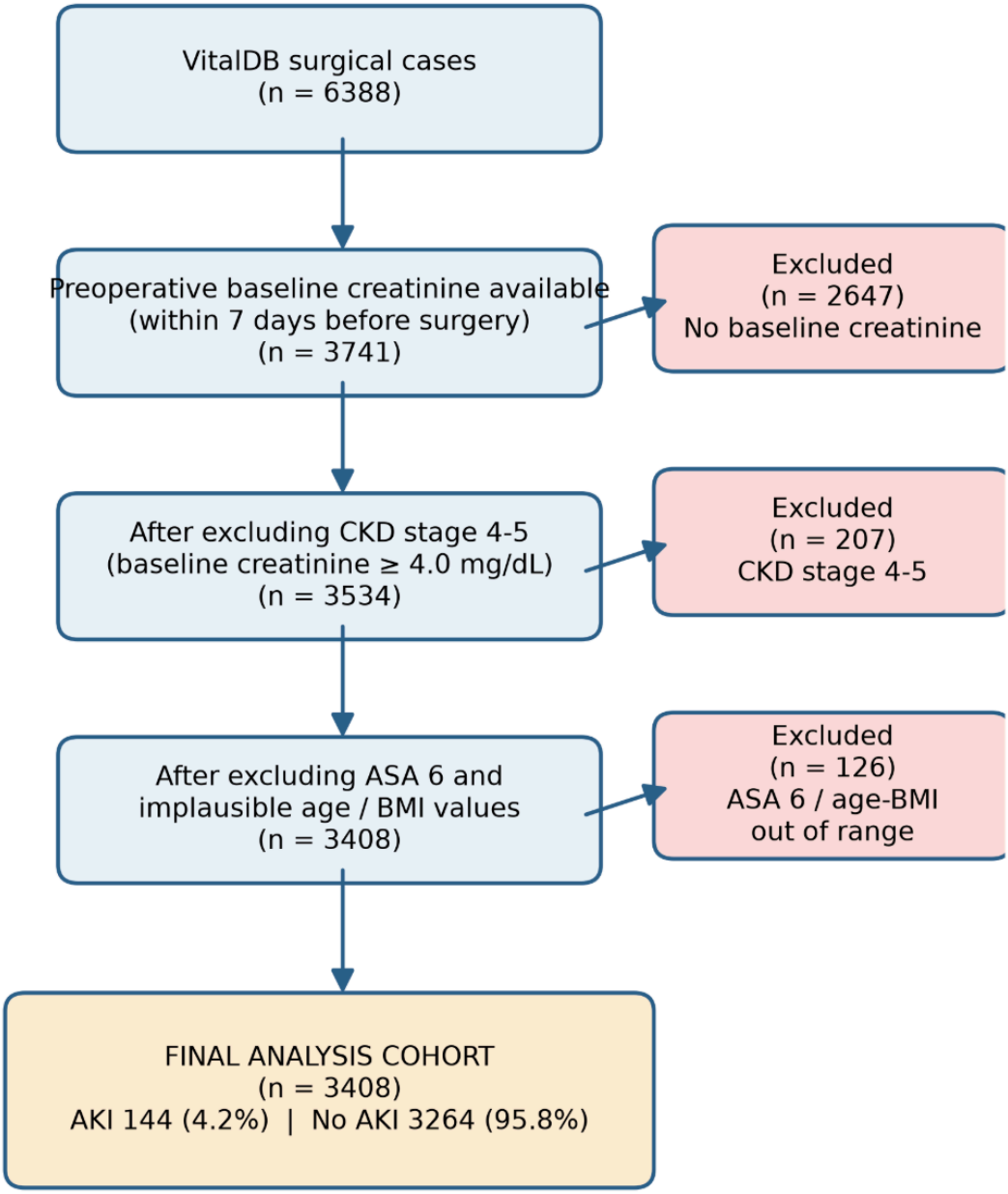
Analysis code (Python 3.11; scikit-learn, XGBoost, pandas) is available at https://github.com/zhendai123/postop-aki-vitaldb.

References

1. Khwaja A. KDIGO clinical practice guidelines for acute kidney injury. Nephron Clin Pract. 2012;120(4):c179–84.
2. Salmasi V, Maheshwari K, Yang D, et al. Relationship between intraoperative hypotension, defined by either reduction from baseline or absolute thresholds, and acute kidney injury and myocardial injury after noncardiac surgery: a retrospective cohort analysis. Anesthesiology. 2017;126(1):47–65.
3. Silver SA, Nadim MK, O'Donoghue DJ, et al. Community health care quality standards to prevent acute kidney injury and its consequences. Am J Med. 2020;133(5):552–e603.
4. Bihorac A, Ozrazgat-Baslanti T, Ebadi A, et al. MySurgeryRisk: development and validation of a machine-learning risk algorithm for major complications and death after surgery. Ann Surg. 2019;269(4):652–62.
5. Fritz BA, Cui Z, Zhang M, et al. Deep-learning model for predicting 30-day postoperative mortality. Br J Anaesth. 2019;123(5):688–95.
6. Damen JAAG, Hooft L, Schuit E, et al. Prediction models for cardiovascular disease risk in the general population: systematic review. BMJ. 2016;353:i2416.
7. Hastie T, Tibshirani R, Friedman J. The Elements of Statistical Learning. 2nd ed. New York: Springer; 2009.
8. Lee HC, Park Y, Yoon SB, et al. VitalDB, a high-fidelity multi-parameter vital signs database in surgical patients. Sci Data. 2022;9(1):279.
9. Lim L, Lee H, Jung C-W, et al. INSPIRE, a publicly available research dataset for perioperative medicine. Sci Data. 2024;11(1):655.
10. Collins GS, Moons KGM, Dhiman P, et al. TRIPOD + AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ. 2024;385:e078378.
11. Wolff RF, Moons KGM, Riley RD, et al. PROBAST: a tool to assess the risk of bias and applicability of prediction model studies. Ann Intern Med. 2019;170(1):51–8.
12. Van Calster B, McLernon DJ, van Smeden M, et al. Calibration: the Achilles heel of predictive analytics. BMC Med. 2019;17(1):230.
13. King CR, Fritz BA. Machine learning models in anaesthesiology: bridging the gap from model training to implementation. Br J Anaesth. 2026;136(6):1723–6.
14. Zhuo XY, Lei SH, Sun L, et al. Preoperative risk prediction models for acute kidney injury after noncardiac surgery: an independent external validation cohort study. Br J Anaesth. 2024;133(3):508–18.

15. Lee SW, Jang J, Seo WY, et al. Internal and external validation of machine learning models for predicting acute kidney injury following non-cardiac surgery using open datasets. J Pers Med. 2024;14(6):587.

Figures



Note: 610 patients (17.9%) had no postoperative creatinine measurement within 7 days and were classified as no AKI (verification bias; see Limitations).

Figure 1

Flow of VitalDB participants (STROBE). 6,388 operations screened to a 3,408-patient analysis cohort.

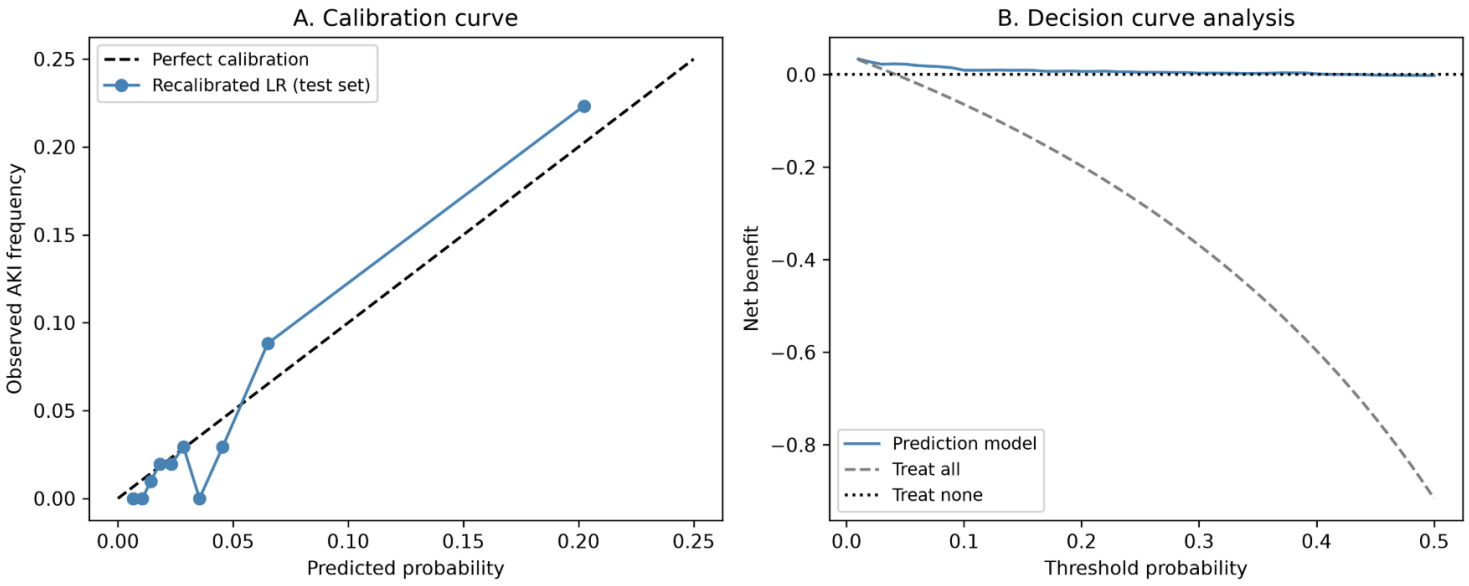


Figure 4

Calibration and utility of the recalibrated VitalDB logistic model. A: decile calibration. B: decision curve analysis.

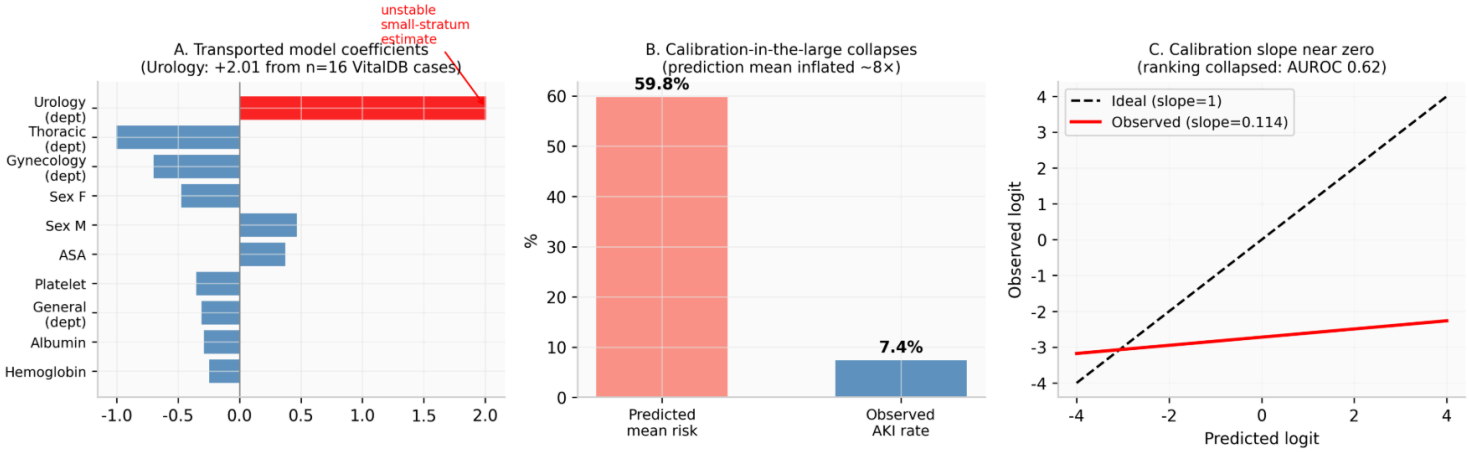


Figure 5

Anatomy of the transport failure. A: coefficients. The urology dummy (+2.009) was built on 16 cases. B: calibration-in-the-large collapse (predicted mean 0.598 vs 7.4%). C: calibration slope 0.114, indicating ranking lost.

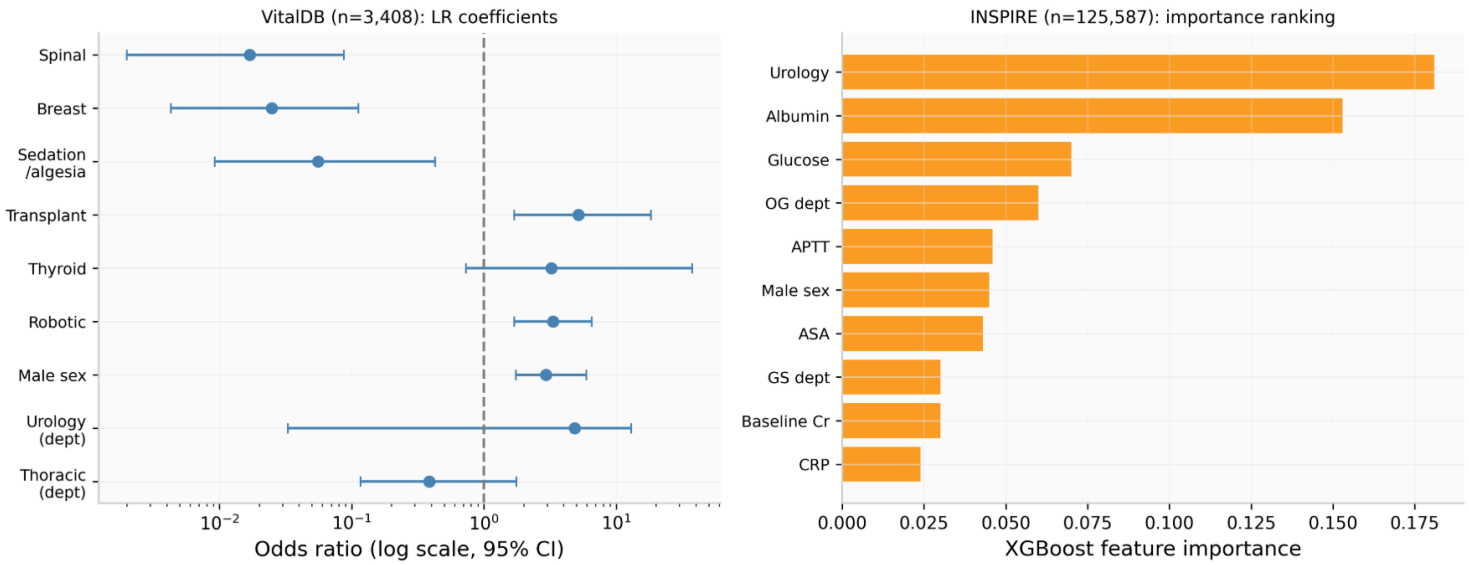


Figure 6

Cross-cohort signal comparison. Left: VitalDB odds ratios (log scale). Right: INSPIRE XGBoost importance. Urology and albumin head both lists.

Supplementary Files

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

- SupplementaryMaterialBMCMIDM.docx
- TRIPODAIChecklist.docx