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.

# Transparent development of an underpowered clinical prediction model for mechanical ventilation in acute poisoning: quantifying overfitting, missing-data mechanisms, and sex-based performance heterogeneity in a multi-institutional cohort

**Enrique Santiago Lanas Mendoza**

enlanasme@uide.edu.ec

Universidad Internacional del Ecuador UIDE

**Alex Wladimir Guayta Valladares**

Universidad Iberoamericana del Ecuador UNIBE

**Luis Gustavo Paredes**

Hospital General Enrique Garcés

**Dario Xavier Jimenez Acosta**

DIALNEF Critical Nephrology

**Octavio Salgado Ordoñez**

Universidad Católica de Cuenca

**Andres Sebastian Herdoiza Almeida**

Hospital Universitario Miguel Servet

**Article**

**Keywords:** poisoning, hemoperfusion, hemodialysis, mechanical ventilation, machine learning, prediction model, critical care, Ecuador

**Posted Date:** September 15th, 2026

**DOI:** https://doi.org/10.21203/rs.3.rs-10593949/v1

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

**Additional Declarations:** No competing interests reported.

# Transparent development of an underpowered clinical prediction model for mechanical ventilation in acute poisoning: quantifying overfitting, missing-data mechanisms, and sex-based performance heterogeneity in a multi-institutional cohort

Short title: Transparent reporting of an underpowered prediction model in acute poisoning


| | | | | |
|---|---|---|---|---|
| Enrique Santiago Lanas Mendoza | enlanasme@uide.edu.ec | https://orcid.org/0000-0001-7003-4191 | • Universidad Internacional del Ecuador UIDE | Quito-Ecuador |
| Alex Wladimir Guayta Valladares | aguayta@unibe.edu.ec | https://orcid.org/0000-0002-7085-1797 | • Universidad Iberoamericana del Ecuador UNIBE<br>• Hospital General Enrique Garcés | Quito-Ecuador |
| Luis Gustavo Paredes | paredesgustavo26@gmail.com | https://orcid.org/0000-0002-4175-4074 | • Hospital General Enrique Garcés | Quito-Ecuador |
| Dario Xavier Jimenez Acosta | dxjimenezmd@gmail.com | https://orcid.org/0000-0002-1843-2827 | • DIALNEF Critical Nephrology<br>• Hospital General Enrique Garcés | Quito-Ecuador |
| Octavio Salgado Ordoñez | octavio.salgado@ucacue.edu.ec | https://orcid.org/0000-0003-2784-5268 | • Universidad Católica de Cuenca | Cuenca-Ecuador |
| Andrés Sebastián Herdoíza Almeida | asherdoiza@salud.aragon.es | https://orcid.org/0000-0003-3241-0808 | • Hospital Universitario Miguel Servet | Zaragoza-España |

Corresponsal author: Enrique Santiago Lanas Mendoza; enlanasme@uide.edu.ec; https://orcid.org/0000-0001-7003-4191; Universidad Internacional del Ecuador UIDE, Quito-Ecuador



**Abstract**

Background: Clinical prediction models in critical care and toxicology are frequently developed with sample sizes below formal minimum requirements and without explicit quantification of the resulting overfitting, contributing to a broader reproducibility problem in this literature. We present a transparent methodological pipeline for developing, internally validating, and honestly reporting such a model, using a real clinical question-predicting the need for invasive mechanical ventilation (IMV) in patients with acute poisoning treated with renal replacement therapy-as the applied case study. No prediction model specific to this population, and none developed in a Latin American setting, has previously been reported.

Methods: We conducted a retrospective, multi-institutional cohort study (11 centres, one predominant centre contributing 76% of cases) of patients with acute poisoning who required renal replacement therapy (hemoperfusion 68.6%, hemodialysis 17.2%, or both 14.2%). After excluding non-toxicological indications for renal replacement therapy, 169 patients were analysed (82 IMV events, 48.5%). Sample size adequacy was assessed using the Riley et al. simulation-based approach; the available sample fell below the calculated

minimum, and this model is presented as a preliminary internal derivation rather than a finalised clinical tool. Missing values in APACHE II and SOFA (14.8% each), found to be missing-at-random conditional on treating centre, were addressed with multiple imputation by chained equations (20 imputations). A Firth-penalised logistic regression with restricted cubic splines for APACHE II and SOFA (informed by a Box-Tidwell non-linearity test) was developed as the primary model, with a random forest as a non-linear comparator. Internal validity was assessed via 5-fold cross-validation and bootstrap optimism correction (200 resamples). Model interpretability was assessed with Shapley Additive Explanations (SHAP) on the random forest comparator. A simplified point-based risk score was derived, with bootstrap 95% confidence intervals for observed event rates per score level. Sensitivity analyses examined treatment modality and sex as potential sources of heterogeneity. Quasi-external validation was performed by training the model exclusively on the predominant centre and freezing it for application to the remaining satellite centres, including full logistic recalibration (intercept and slope).

Results: SOFA ($p<0.001$), APACHE II ($p<0.001$), vasoactive drug use ($p<0.001$), and time from ingestion to first dialysis ($p=0.0009$) differed significantly between patients with and without IMV. The Firth-penalised spline model achieved a cross-validated AUC of 0.821 (SD 0.014); bootstrap optimism correction of the apparent model (AUC 0.898) yielded a corrected AUC of 0.856, consistent with the theoretical shrinkage factor (0.856) derived from the Riley sample-size calculation. The random forest achieved a cross-validated AUC of 0.831 (SD 0.011). Decision curve analysis showed net benefit over 'treat-all' and 'treat-none' strategies across the 5-75% threshold probability range. Treatment modality did not materially change model performance (AUC 0.830 with modality included vs. 0.829 without) despite its uneven distribution across centres. Discrimination showed a point-estimate difference by sex (AUC 0.883 in men vs. 0.785 in women) that did not reach statistical significance on a formal DeLong test (difference 0.098, 95% CI -0.022 to 0.218, $p=0.109$), an inconclusive result discussed as an open equity question rather than a confirmed disparity. A four-variable exploratory point score (VM-Risk, range 0-7) showed a gradient from 0% to 100% observed IMV incidence, with wide, overlapping bootstrap confidence intervals at adjacent score levels reflecting the small per-category sample sizes; its apparent AUC (0.848, bootstrap 95% CI 0.793-0.901) has not been independently validated. Quasi-external validation (training on the predominant centre, testing on satellite centres, n=41) preserved discrimination (AUC 0.833); full logistic recalibration (intercept -0.024, slope 1.603) improved the Brier score from 0.166 to 0.153, though the substantial departure of the slope from 1 indicates meaningful risk-stratification distortion beyond a simple level shift.

Conclusions: This case study demonstrates a reproducible pipeline for developing and transparently reporting an underpowered clinical prediction model: formal sample-size calculation, missing-data mechanism identification, triangulated overfitting quantification, full recalibration, and explicit subgroup equity testing. Applied to acute poisoning requiring renal replacement therapy, the resulting model showed encouraging discrimination but remains a preliminary, hypothesis-generating derivation; no specific score cut-off is recommended for clinical decision-making pending prospective, adequately powered external validation.

Keywords: poisoning; hemoperfusion; hemodialysis; mechanical ventilation; machine learning; prediction model; critical care; Ecuador

**Introduction**

Acute poisoning requiring extracorporeal elimination therapies remains a significant cause of intensive care unit (ICU) admission worldwide. Hemoperfusion, alone or combined with hemodialysis, is used for toxins with limited water solubility or high protein binding that render conventional hemodialysis less effective. A substantial proportion of these patients develop respiratory failure requiring invasive mechanical ventilation (IMV), which contributes to increased ICU length of stay, cost, and complication risk.

Machine learning (ML) methods have increasingly been applied to prognostic modelling in acute poisoning. Prior work in patients with acute organophosphate poisoning treated with hemodialysis achieved high discrimination for a composite prognosis outcome using gradient boosting methods
(1)

A separate study in methanol poisoning specifically modelled the need for intubation using random forest and XGBoost with explainable AI (XAI) methods, finding that tree-based methods outperformed deep learning approaches

(2), building on an earlier methanol poisoning prognosis model by the same research group using more conventional ML classifiers (3). A prediction model for deep venous thrombosis specifically in a hemoperfusion-treated cohort has also been reported, demonstrating the methodological feasibility of prediction modelling in this population, albeit for a different outcome (4). In Latin America, a large multi-institutional network across 84 hospitals in 16 countries has demonstrated that ML models with SHAP-based interpretability can be developed and validated regionally for critical care outcomes in children, including the need for mechanical ventilation (5). Hemoperfusion combined with hemofiltration has itself been shown to reduce the incidence of mechanical ventilation in organophosphate poisoning in a systematic review and meta-analysis, underscoring the clinical relevance of this outcome in toxin-exposed populations (6).

Despite this body of evidence, we are not aware of any prediction model that has specifically targeted the need for IMV in a hemoperfusion-treated cohort, nor any developed in a Latin American setting. This represents a double knowledge gap that this study aims to address.

We aimed to develop and internally validate a prediction model for the need of IMV in patients with acute poisoning treated with renal replacement therapy, using clinical variables available at ICU admission, and to evaluate its performance across a predominant centre and a group of smaller satellite institutions.

Beyond this clinical question, we use the resulting model as an applied case study in methodological transparency. Systematic reviews of clinical prediction models, including those using machine learning, have repeatedly found that formal sample-size justification is rarely reported, that the widely used events-per-variable heuristic lacks a rigorous statistical basis (7,8), and that reporting completeness against structured guidance such as TRIPOD and TRIPOD+AI remains poor even in recent literature (26,27). Rather than treating our sample-size shortfall as a limitation to be minimised in the Discussion, we treat it as the central methodological problem this manuscript addresses: we demonstrate, step by step, how to calculate the required sample size a priori, how to triangulate multiple independent methods to quantify the resulting overfitting when that requirement cannot be met, how to identify and address a missing-data mechanism that varies by clustering unit, how to fully recalibrate a model transported to new sites, and how to test formally-rather than merely acknowledge narratively-whether model performance is equitable across a protected subgroup. We present the acute-poisoning cohort as the concrete case in which this pipeline is applied, not as an end in itself.

## Methods

### Study design and data source

We conducted a retrospective cohort study using a consolidated, de-identified database of patients with acute poisoning treated with hemoperfusion, hemodialysis, or both, across 11 healthcare institutions in Ecuador. The database was collected as part of routine clinical care documentation, not prospectively for this study.

### Participants

The source database comprised 177 patients. Eight patients (4.5%) were excluded because their indication for renal replacement therapy was not an exogenous toxin exposure (hepatic failure, n=3; rhabdomyolysis,

n=3; myoglobinuria, n=1; multiple myeloma, n=1), leaving an analytic cohort of 169 patients. One institution (referred to as the predominant centre) contributed 128 patients (75.7%); the remaining 41 patients (24.3%) were distributed across 10 satellite institutions. Treatment modality was hemoperfusion alone in 116 patients (68.6%), hemodialysis alone in 29 (17.2%), and combined hemoperfusion-hemodialysis in 24 (14.2%). Modality distribution differed markedly between the predominant and satellite centres (hemoperfusion alone: 59.4% vs. 97.6%, respectively), reflecting differing practice patterns rather than random variation, and modality was significantly associated with IMV (chi-square p=0.014; IMV incidence 52.6% with hemoperfusion alone, 24.1% with hemodialysis alone, 58.3% with combined therapy). Given this uneven distribution, the study population is therefore more accurately described as patients undergoing renal replacement therapy for acute poisoning, predominantly but not exclusively via hemoperfusion, consistent with the revised study title.

**Outcome**

The primary outcome was the need for invasive mechanical ventilation during the ICU stay, defined as days of mechanical ventilation greater than zero. This variable had no missing values. Manual review confirmed that no patient with this outcome had a pre-existing respiratory compromise unrelated to the toxic exposure, supporting the validity of the outcome definition.

**Predictors**

Candidate predictors, selected a priori based on availability at ICU admission and clinical plausibility, were: age, sex, grouped toxic agent (categories with frequency ≥10 retained individually: carbamazepine, organophosphate, acetaminophen, paraquat, tricyclic antidepressant; remaining 33 distinct free-text categories grouped as 'other'), time from ingestion to first dialysis session, APACHE II score, Sequential Organ Failure Assessment (SOFA) score, and use of vasoactive medication. Variables measured after the point of the IMV decision (number of hemoperfusion sessions, renal replacement therapy modality, hemoperfusion complications, atropine dose) were excluded from the predictor set to avoid temporal data leakage.

**Sample size**

Sample size adequacy was assessed using the simulation-based approach of Riley et al. (7), implemented via the pmsampsize package, rather than the traditional events-per-variable (EPV) heuristic, whose statistical basis has been criticised (8). With 11 effective predictor parameters and an observed outcome prevalence of 48.5%, a minimum of 384 participants (187 events) was required assuming an anticipated C-statistic of 0.83, rising to 406 participants (197 events) if the anticipated C-statistic were 0.77 (a lower anticipated discriminative ability requires a larger sample under Riley's framework, since less outcome variation is available to estimate parameters precisely). The available sample (n=169, 82 events) did not meet either threshold. The corresponding expected shrinkage factor, calculated analytically, was 0.856 (point estimate) to 0.793 (conservative estimate), indicating an anticipated 14.4-20.7% reduction in predictor coefficients due to overfitting. This informed the use of penalised regression as the primary analytical approach (see below).

**Missing data**

APACHE II and SOFA scores were missing in 14.8% of cases each (25/169 for each variable); hemoperfusion complications were missing in 34.3% of cases (58/169). Missingness in APACHE II and SOFA was concentrated entirely (100%, 25/25 for each) in satellite centres (chi-square p<0.001). Missingness in hemoperfusion complications followed a related but distinct pattern: 41 of the 58 missing values (70.7%) were in satellite centres (where the variable was missing in all 41 patients) and the remaining 17 (29.3%) were in the predominant centre (17/128, 13.3%); we note this distinction explicitly here, as an earlier internal draft of this manuscript incorrectly generalised the 100%-satellite pattern to all three variables. Both patterns indicate a missing-at-random (MAR) mechanism conditional on treating centre,

reflecting inconsistent institutional documentation practices rather than random data loss. Missing values in APACHE II and SOFA were addressed using multiple imputation by chained equations (MICE), implemented via scikit-learn's IterativeImputer with posterior sampling (20 imputed datasets), using age, time to dialysis, sex, vasoactive use, grouped toxic agent, treating centre, and the outcome variable as auxiliary predictors, consistent with recommendations to include the outcome in imputation models for prediction model development (9). Hemoperfusion complications were not used as a model predictor (see Predictors, above) and were therefore not imputed.

**Statistical analysis**

Continuous variables were summarised as median [interquartile range] given non-normal distributions (Shapiro-Wilk $p<0.05$ for all), and compared between outcome groups using the Mann-Whitney U test. Categorical variables were summarised as n (%) and compared using the chi-square or Fisher's exact test as appropriate. Multicollinearity was assessed using the variance inflation factor (VIF); linearity of the logit for continuous predictors was assessed using the Box-Tidwell test.

The primary prediction model was a Firth-penalised logistic regression (10), fitted to mitigate small-sample bias and potential separation given the anticipated overfitting identified in the sample size calculation; as detailed in the Results and Discussion, this penalisation reduces but does not eliminate the optimism quantified by our sample-size and bootstrap analyses. Restricted cubic splines (four knots, Harrell's method (11), placed at the 5th, 35th, 65th, and 95th percentiles) were used for APACHE II and SOFA after the Box-Tidwell test indicated non-linearity in the logit for both variables ($p=0.022$ and $p=0.029$, respectively). Model coefficients were pooled across the 20 imputed datasets using Rubin's rules. A random forest classifier (300 trees, maximum depth 4) was fitted as a non-linear comparator. The maximum depth of 4 was selected a priori as a conservative constraint appropriate to a dataset of this size (rather than derived from out-of-bag error minimisation or a formula linking depth to the number of predictors, which we did not perform); we state this plainly because a post hoc rationale would misrepresent how the choice was made. These hyperparameters were fixed rather than tuned via nested cross-validation; we made this choice deliberately, because hyperparameter optimisation in a sample of this size would itself be prone to overfitting and would complicate, rather than simplify, interpretation of the comparison with the primary model. The cross-validation reported for the random forest (see below) evaluates performance under these fixed hyperparameters and was not used for tuning. Because the resulting cross-validated AUC estimates for the Firth and random forest models are correlated (both evaluated on overlapping folds of the same imputed datasets) rather than independent, and because a paired test appropriate to this correlation structure was not implemented, we report both AUCs alongside their cross-validation standard deviations but do not claim statistical equivalence between them; the two models should be read as converging on a similar range of discrimination, not as formally indistinguishable.

Model discrimination was assessed using the area under the receiver operating characteristic curve (AUC), calculated via 5-fold cross-validation repeated independently within each of the 20 imputed datasets (100 fold-level AUC estimates in total, arithmetically averaged to obtain the reported mean and standard deviation). We refer to this simply as repeated cross-validation, not nested cross-validation: because model hyperparameters (for both the Firth model and the random forest) were fixed a priori rather than tuned, there was no inner-loop hyperparameter optimisation of the kind the term 'nested cross-validation' conventionally denotes, and we avoid that term to prevent confusion with cross-validation designs that do perform such tuning. Because cross-validated estimates and theoretical shrinkage factors, while convergent, do not by themselves quantify optimism in the apparent (non-cross-validated) model fit, we additionally performed a bootstrap optimism-correction procedure (Harrell's method, 200 resamples with replacement, random seed 2026): the model was refitted in each bootstrap sample and evaluated both on that same bootstrap sample (bootstrap-apparent performance) and on the original full sample (bootstrap-test performance); the average difference (optimism) was subtracted from the apparent AUC of the model fitted on the original sample to obtain an optimism-corrected AUC. Calibration was assessed using the Brier score

and, for the quasi-external validation, a full logistic recalibration model (both intercept and slope, not intercept alone) fitted as a logistic regression of the observed outcome on the logit of the predicted probability; calibration was additionally examined graphically using observed-versus-predicted event rates within quartiles/quintiles of predicted risk. Clinical utility was assessed using decision curve analysis (DCA), comparing net benefit against 'treat-all' and 'treat-none' default strategies across threshold probabilities of 5-80%.

As a pre-specified equity-related sensitivity analysis, model discrimination and calibration were additionally assessed separately by sex, given the female preponderance in this cohort and literature emphasising the risk that undetected subgroup heterogeneity in model performance can entrench health disparities if unaddressed.

Model interpretability was assessed using Shapley Additive Explanations (SHAP) applied to the random forest model, providing both global (mean absolute SHAP value) and local (individual-level waterfall) explanations. SHAP interaction values were examined to assess the clinical plausibility of detected predictor interactions.

A simplified point-based clinical score (VM-Risk) was derived by converting predicted log-odds from the pooled Firth model, evaluated at clinically meaningful category midpoints, into integer point values. Given the modest sample size within each score level, bootstrap 95% confidence intervals (2,000 resamples) were calculated for the observed event rate at each score level and for the overall apparent AUC of the score, to characterise the precision-not merely the point estimate-of the reported gradient. This score is presented as an exploratory, hypothesis-generating derivation and not as a tool ready for clinical decision-making; no specific cut-off is recommended for practice pending independent external validation.

**Sensitivity analyses and quasi-external validation**

Given the concentration of cases in a single predominant centre, we conducted a sensitivity analysis comparing model performance in the full cohort, in the predominant centre alone, and in satellite centres alone, and assessed the independent contribution of treating centre as a covariate. An analogous sensitivity analysis assessed the independent contribution of treatment modality (hemoperfusion alone, hemodialysis alone, or combined) as a covariate, given its uneven distribution across centres. As no genuinely independent external cohort was available, we additionally performed a quasi-external, leave-cluster-out validation: the model was trained exclusively on the predominant centre (n=128, 55 events) and applied, without refitting, to the satellite centres (n=41, 27 events). Both intercept-only recalibration and full logistic recalibration (intercept and slope) were performed to characterise and correct any systematic miscalibration observed; we report both because intercept-only recalibration corrects the average level of predicted risk but not distortions in risk stratification, which require adjustment of the slope.

**Ethical considerations**

All methods were performed in accordance with the relevant guidelines and regulations. This study used retrospective, de-identified secondary data. Under Article 32 of Ecuador's Organic Law on Personal Data Protection (Official Registry Supplement 459, 2021), health data may be used for scientific research when anonymised (12), supporting a waiver of individual informed consent. This study was approved by the Human Research Ethics Committee of the Catholic University of Cuenca (Comité de Ética de Investigación en Seres Humanos de la Universidad Católica de Cuenca, CEISH-UCACUE), of the Ministry of Health of the Republic of Ecuador (approval: CEISH-UCACUE-2023-089), and formal data-use authorisation from each contributing satellite institution was obtained.

All analyses were conducted in Python 3.12 using scikit-learn, statsmodels, patsy, and the shap and pmsampsize packages. Analytical code is available upon reasonable, specified request pending institutional data-sharing authorisation.

## Results

### Participants

Of 177 patients in the source database, 169 were included in the final analysis after excluding 8 with non-toxicological indications for renal replacement therapy. Median age was 23 years [IQR 18-34]; 58.0% were women. IMV was required in 82 patients (48.5%). Median ICU stay was 2 days [IQR 2-3]; overall mortality was 4.1% (7 deaths).

### Baseline characteristics by outcome

Table 1 shows baseline characteristics stratified by IMV status. Patients requiring IMV had significantly higher APACHE II (17 [12.2-21] vs. 9 [6-12.8], p<0.001), higher SOFA (8 [6-10.8] vs. 3 [2-5], p<0.001), longer time to first dialysis (18.5 [12-24] vs. 12 [8-22] hours, p=0.0009), higher vasoactive drug use (46.3% vs. 16.1%, p<0.001), and were slightly older (25 [19-38] vs. 22 [18-29] years, p=0.044).

**Table 1. Baseline characteristics stratified by need for invasive mechanical ventilation (n=169)**

| Variable | No IMV (n=87) | IMV (n=82) | p-value |
|---|---|---|---|
| Age (years), median [IQR] | 22 [18-29] | 25 [19-38] | 0.044 |
| Time ingestion-to-dialysis (h), median [IQR] | 12 [8-22] | 18.5 [12-24] | 0.0009 |
| APACHE II, median [IQR] | 9 [6-12.8] | 17 [12.2-21] | <0.001 |
| SOFA, median [IQR] | 3 [2-5] | 8 [6-10.8] | <0.001 |
| Male sex, n (%) | 30 (34.5) | 41 (50.0) | 0.059 |
| Vasoactive drug use, n (%) | 14 (16.1) | 38 (46.3) | <0.001 |
| Toxic agent group (6 categories) | - | - | 0.0014 |

### Collinearity and functional form

All variance inflation factors were below 3.0 (maximum 2.98), indicating no problematic multicollinearity, although APACHE II and SOFA were moderately correlated (Spearman r=0.68), as expected for two overlapping severity scores. The Box-Tidwell test indicated adequate logit-linearity for age (p=0.728) and time to dialysis (p=0.155), but a significant departure from linearity for APACHE II (p=0.022) and SOFA (p=0.029), motivating the use of restricted cubic splines for these variables in the final model.

### Model development and performance

The Firth-penalised logistic regression with restricted cubic splines achieved a cross-validated AUC of 0.821 (SD 0.014), a Brier score of 0.174, a calibration intercept of 0.030, and a calibration slope of 0.844. The apparent AUC (model fitted and evaluated on the same full sample, without cross-validation or bootstrap correction) was 0.898. Bootstrap optimism correction (Harrell's method, 200 resamples) estimated a mean optimism of 0.042, yielding an optimism-corrected AUC of 0.856. This corrected estimate is closely consistent with both the 5-fold cross-validated AUC (0.821) and the theoretical shrinkage factor of 0.856 derived analytically from the Riley sample-size calculation, providing convergent evidence from three independent internal-validation approaches that the true, generalisable discrimination of this model is meaningfully lower than its apparent performance (Figure 1) precisely the pattern expected for a model developed in an underpowered sample, and consistent with our characterisation of this model as a preliminary derivation rather than a finalised, optimism-free estimate of performance. The random forest comparator achieved a cross-validated AUC of 0.831 (SD 0.011), Brier score of 0.166, calibration intercept of -0.054, and calibration slope of 1.269.

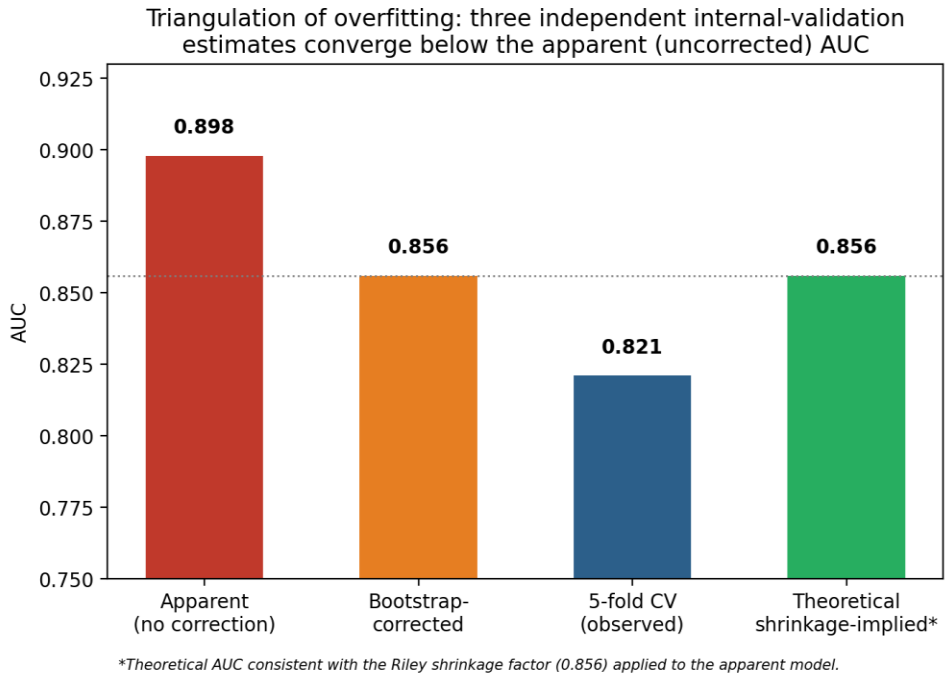



*Figure 1. Triangulation of overfitting across three independent internal-validation approaches: the apparent AUC (0.898) is progressively reduced by bootstrap optimism correction (0.856) and by 5-fold cross-validation (0.821), converging with the theoretical shrinkage-implied estimate (0.856) derived from the Riley sample-size calculation.*

**Table 2. Odds ratios (Firth-penalised model, pooled across imputations, 95% CI)**

| Variable | OR | 95% CI |
|---|---|---|
| Vasoactive drug use (yes vs. no) | 2.88 | 1.16-7.14 |
| Male sex (vs. female) | 1.59 | 0.66-3.82 |
| Toxic agent: tricyclic antidepressant | 7.05 | 0.42-117.5 |
| Toxic agent: organophosphate | 3.80 | 0.32-45.3 |
| Time to first dialysis (per hour) | 1.02 | 0.99-1.05 |
| Age (per year) | 1.00 | 0.97-1.03 |

Odds ratios are not reported for APACHE II and SOFA given their non-linear (spline) functional form; predicted probability curves at representative values are presented in Figure 2, demonstrating a continuously increasing effect for SOFA and a plateau effect for APACHE II beyond approximately 15 points. The wide confidence intervals for the tricyclic antidepressant and organophosphate categories (Table 2) reflect the rarity of these categories rather than a stable, precisely estimated effect; these estimates should be interpreted as indicating a directional association only, not a precise effect size, and combining rare toxic categories further in future, larger studies may be warranted.

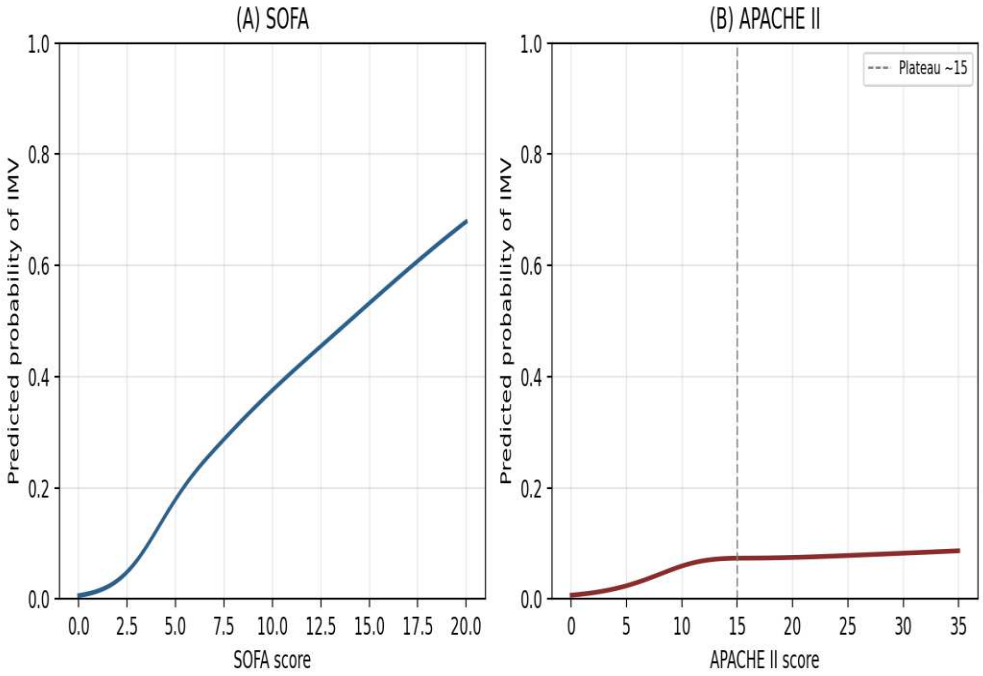



*Figure 2. Predicted probability of invasive mechanical ventilation as a function of SOFA (A) and APACHE II (B), from the pooled Firth-penalised spline model, holding other covariates at reference/median values.*

**Sensitivity analysis: treatment modality**

Treatment modality was hemoperfusion alone in 116 patients (68.6%), hemodialysis alone in 29 (17.2%), and combined therapy in 24 (14.2%), and was distributed very unevenly across centres (hemoperfusion alone: 59.4% at the predominant centre vs. 97.6% at satellite centres; no satellite-centre patient received hemodialysis alone). Modality was significantly associated with IMV (chi-square p=0.014), with lower IMV incidence in patients receiving hemodialysis alone (24.1%) than hemoperfusion alone (52.6%) or combined therapy (58.3%). Including modality as an additional covariate in the full model did not materially change discrimination (AUC 0.830 vs. 0.829 without modality) or the direction of effect for SOFA, APACHE II, or vasoactive use, indicating that modality does not confound the principal findings; however, because modality distribution is confounded with centre and was not included in the primary model, we cannot exclude a modest independent modality effect, and this is reflected in the revised, broader study title and population description (renal replacement therapy rather than hemoperfusion specifically).

**Decision curve analysis**

Both models showed positive net benefit over 'treat-all' and 'treat-none' strategies across nearly the entire clinically relevant threshold probability range (5-75%). Beyond a threshold of approximately 45-50%, the 'treat-all' strategy showed negative net benefit, while both models remained positive substantially beyond this range, indicating genuine clinical utility beyond statistical discrimination alone. The random forest model retained a small positive net benefit over both default strategies up to a threshold of approximately 75%, beyond which net benefit approached zero for all strategies (Figure 3); we highlight this threshold range explicitly, as it delineates where this model (subject to its preliminary status) would be expected to add value over unassisted default decisions, information relevant to designing the decision-analytic component of a future validation study.

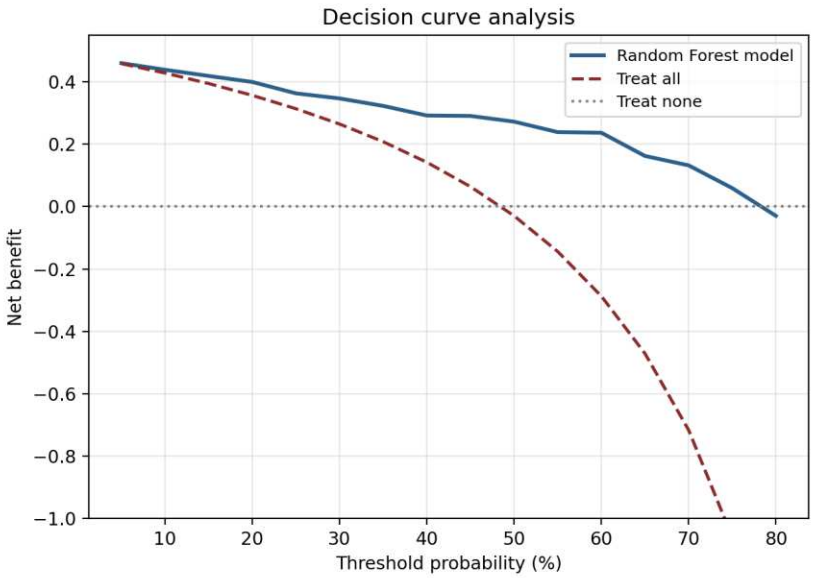



*Figure 3. Decision curve analysis comparing the random forest model against 'treat-all' and 'treat-none' default strategies across threshold probabilities of 5-80%.*

**Model interpretability**

Interpretability of the primary Firth-penalised model is provided directly by its odds ratios for linear terms (Table 2) and by predicted-probability curves at representative values for the spline terms (APACHE II and SOFA), reported above and in the preceding section. In addition, we applied SHAP specifically to the random forest comparator-not to the Firth model-because SHAP's marginal-contribution framework is most naturally suited to tree-based ensembles and provides a complementary, model-agnostic cross-check on variable importance rather than a restatement of the primary model's own coefficients; agreement between the SHAP-derived ranking and the Firth model's odds ratios and spline shapes (both identifying SOFA and APACHE II as dominant) supports the plausibility of both, but the SHAP results below should be interpreted as pertaining to the comparator model specifically.

SHAP analysis of the random forest model identified SOFA (mean |SHAP|=0.143) and APACHE II (0.095) as the dominant predictors (Figure 4), followed by vasoactive drug use (0.052), time to dialysis (0.034), age (0.024), sex (0.016), and toxic agent group (0.001-0.009 per category). Notably, the apparent importance of specific toxic agents observed in univariable analysis was substantially attenuated once non-linear severity effects were properly modelled, suggesting that much of the apparent toxin-specific risk reflects underlying illness severity rather than a toxin-specific mechanism. SHAP interaction analysis identified the strongest interaction between APACHE II and SOFA (0.0149), consistent with their expected correlation as overlapping severity measures; no clinically implausible interactions were detected.

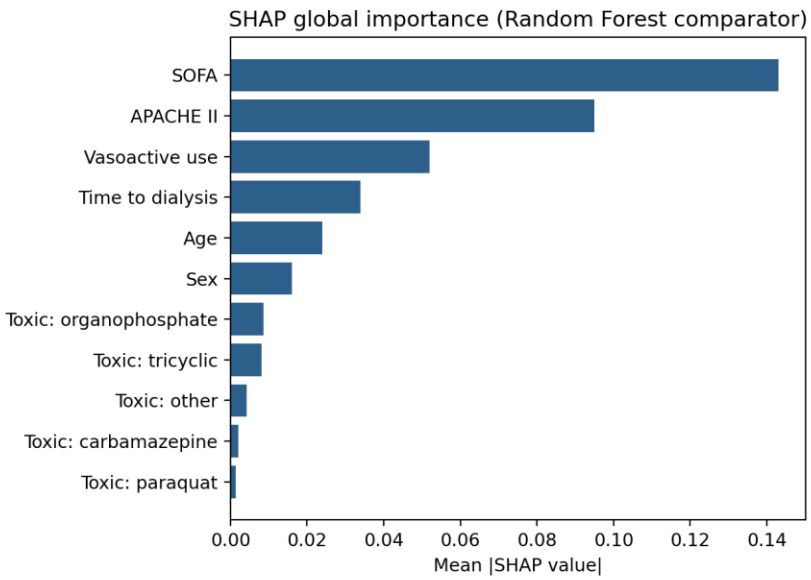



*Figure 4. Global SHAP variable importance (mean absolute SHAP value) for the random forest comparator model.*

Local SHAP explanations for individual patients illustrated the comparator model's reasoning: a patient with high SOFA/APACHE II and vasoactive use had a predicted probability of 0.93 (vs. a population baseline of 0.51), while a patient with low severity scores had a predicted probability of 0.05. One misclassified case-a patient with carbamazepine poisoning, low initial severity scores, but who subsequently required IMV-suggests that admission severity scores may not capture delayed neurological deterioration characteristic of certain toxic agents, a mechanistic limitation worth highlighting.

**Sensitivity analysis by sex (algorithmic equity)**

Given the female preponderance in this cohort (58.0%) and the borderline univariable sex difference (p=0.059, Table 1), we assessed model discrimination and calibration separately by sex using the same pooled, cross-validated predictions underlying the overall performance estimates. Point-estimate discrimination was higher in men (AUC 0.883, n=71, 41 events) than in women (AUC 0.785, n=98, 41 events). A formal DeLong test for the difference in AUC gave a difference of 0.098 (95% CI -0.022 to 0.218; Z=1.60, p=0.109): this difference did not reach conventional statistical significance, and the confidence interval includes both a null difference and a difference exceeding 0.20, meaning our data are compatible with anything from no sex-based discrimination gap to a substantial one. We report the point estimate and this formal test together, rather than either alone, because presenting only the point estimate would overstate certainty, while presenting only a non-significant p-value could understate a plausible and clinically important gap that this sample is simply too small to confirm or exclude. Calibration-in-the-large was adequate in both subgroups (mean predicted vs. observed probability: 0.568 vs. 0.577 in men; 0.424 vs. 0.418 in women), indicating that the model is not systematically over- or under-predicting risk in either sex on average. We are unable to determine from these data whether the observed point-estimate gap reflects a genuine sex-related difference in the clinical presentation or mechanism of respiratory failure, a smaller effective sample within the female subgroup, or chance; this finding is discussed further, and treated as an open question requiring a larger sample, in the Discussion.

**Simplified clinical score (VM-Risk) - exploratory derivation**

A four-component point score (SOFA category, APACHE II category, vasoactive use, high-risk toxic agent) ranging from 0 to 7 points showed an increasing gradient in observed IMV incidence, from 0% at scores 0-1 to 100% at score 7, with an apparent AUC of 0.848 (bootstrap 95% CI 0.793-0.901, 2,000

resamples). This score has not been independently validated, as both its component thresholds and its evaluation used the same 169 patients; it is presented as a methodological illustration of how penalised log-odds from a spline-based model can be translated into an integer point system, discussing the precision-parsimony trade-off this translation entails, rather than as a validated clinical instrument, and is not recommended for clinical decision-making in its current form. The full derivation, including the empirical comparison of alternative score configurations, is reported in the main text because the derivation process-not the resulting point values-is the primary contribution of this section; readers seeking only the score's specification and operating characteristics can refer directly to Tables 3 and 4.

**Table 3. VM-Risk score: point assignment by component**

| Component | Category | Points |
|---|---|---|
| SOFA | 0-2 | 0 |
| | 3-5 | 2 |
| | 6-9 | 3 |
| | ≥10 | 4 |
| APACHE II | <10 | 0 |
| | ≥10 | 1 |
| Vasoactive drug use | Yes | 1 |
| High-risk toxic agent (tricyclic antidepressant, organophosphate, or paraquat) | Yes | 1 |

Total VM-Risk score = sum of points (range 0-7). Three broad risk bands are descriptively identified based on the observed gradient: low (0-1 points, 0.0% observed IMV, n=27), intermediate (2-3 points, 22.2-31.8% observed IMV, n=36), and high (4-7 points, 54.1-100% observed IMV, n=106); as detailed in Table 4, bootstrap confidence intervals for individual score levels overlap substantially between adjacent levels, and these bands should be read as descriptive of the derivation sample only, not as validated clinical decision thresholds. At an exploratory cut-off of ≥3 points, sensitivity was 0.99 and negative predictive value 0.98; at ≥5 points, specificity was 0.83 and positive predictive value 0.78. These operating characteristics are reported for completeness and to inform the design of a future validation study; we do not recommend their use to guide clinical decisions until the score has been evaluated in an independent cohort.

Justification for component selection: time from ingestion to first dialysis differed significantly between outcome groups in univariable analysis (Table 1, p=0.0009), but its association with IMV was substantially attenuated in the full multivariable model (adjusted OR 1.02 per hour, 95% CI 0.99-1.05, crossing unity). This pattern is compatible with the univariable association being largely explained by its correlation with SOFA and APACHE II (patients with delayed presentation tend to have higher severity scores, which already capture much of this signal); we describe this as attenuation after adjustment rather than 'confounding', since our design does not support a specific causal ordering among ingestion-to-dialysis time, illness severity, and IMV without a pre-specified causal model (e.g., a directed acyclic graph), which we did not construct a priori. We empirically compared three candidate score configurations: the reported score (SOFA + APACHE II + vasoactive use + toxic agent category, apparent AUC 0.848, increasing gradient with no reversals), an alternative substituting toxic agent category with a categorised time-to-dialysis variable (AUC 0.843, non-monotonic gradient with a reversal at low scores), and a five-component version including both (AUC 0.849, non-monotonic gradient with a reversal at intermediate scores). The reported four-component version was retained as it combines the best empirical gradient behaviour with mechanistic plausibility for the toxic agent category, as discussed further below; we note that none of the three configurations differed materially in apparent AUC, and this comparison should not be over-interpreted as strong evidence for one configuration over another given the shared, limited sample.

**Table 4. VM-Risk score distribution, observed IMV incidence, and bootstrap 95% confidence intervals**

| Score | n | Observed IMV, % | 95% CI (Wilson) |
|---|---|---|---|
| 0 | 19 | 0.0 | 0.0-16.8 |
| 1 | 8 | 0.0 | 0.0-32.4 |
| 2 | 14 | 7.1 | 1.3-31.5 |
| 3 | 22 | 31.8 | 16.4-52.7 |
| 4 | 37 | 54.1 | 38.4-69.0 |
| 5 | 36 | 69.4 | 53.1-82.0 |
| 6 | 28 | 85.7 | 68.5-94.3 |
| 7 | 5 | 100.0 | 56.6-100.0 |

**Sensitivity analysis by centre and quasi-external validation**

Model performance was preserved when restricted to the predominant centre (AUC 0.854) and did not deteriorate substantially when 'centre' was included as a covariate in the full model (AUC 0.827 vs. 0.829 without centre; coefficient 0.202 for the satellite category), indicating that centre does not dominate model behaviour, although its contribution is not strictly negligible-it is comparable in magnitude to sex and smaller than SOFA (1.079), toxic-agent categories, vasoactive use (0.458), time to dialysis (0.440), or APACHE II (0.385). Performance in the satellite-centre-only subset (n=41) was lower and highly variable (AUC 0.564, SD 0.178), attributable in large part to the small subgroup sample size rather than necessarily reflecting genuine clinical heterogeneity, though a modest true difference cannot be excluded given the width of this estimate.

In the quasi-external, leave-cluster-out validation (model trained exclusively on the predominant centre, applied without refitting to satellite centres), discrimination was preserved (AUC 0.833). Intercept-only recalibration corrected a modest calibration-in-the-large deficit (mean predicted probability 0.619 vs. observed prevalence 0.659, corrected to 0.659) without changing the Brier score materially (0.166 to 0.165). Because the calibration slope (1.603) indicated distortion in risk stratification beyond the average level, we additionally performed full logistic recalibration (both intercept and slope): the recalibrated intercept was -0.024 and slope 1.603 (identical to the originally estimated slope, as expected, since full recalibration re-estimates rather than removes this parameter), improving the Brier score further to 0.153. AUC is mathematically unchanged by any recalibration (0.833 throughout), since recalibration is a monotonic transformation of predicted probabilities. To characterise calibration beyond summary statistics, Table 5 and Figure 5 present observed versus predicted event rates within quartiles of predicted risk for the quasi-external validation (post full recalibration) and within quintiles for the apparent full-cohort model; agreement is reasonably close in both, with the largest discrepancy in the lowest-risk quasi-external quartile (observed 18.2% vs. predicted 35.5%, n=11), consistent with the substantial imprecision expected at this subgroup's very small sample size rather than necessarily reflecting genuine miscalibration.

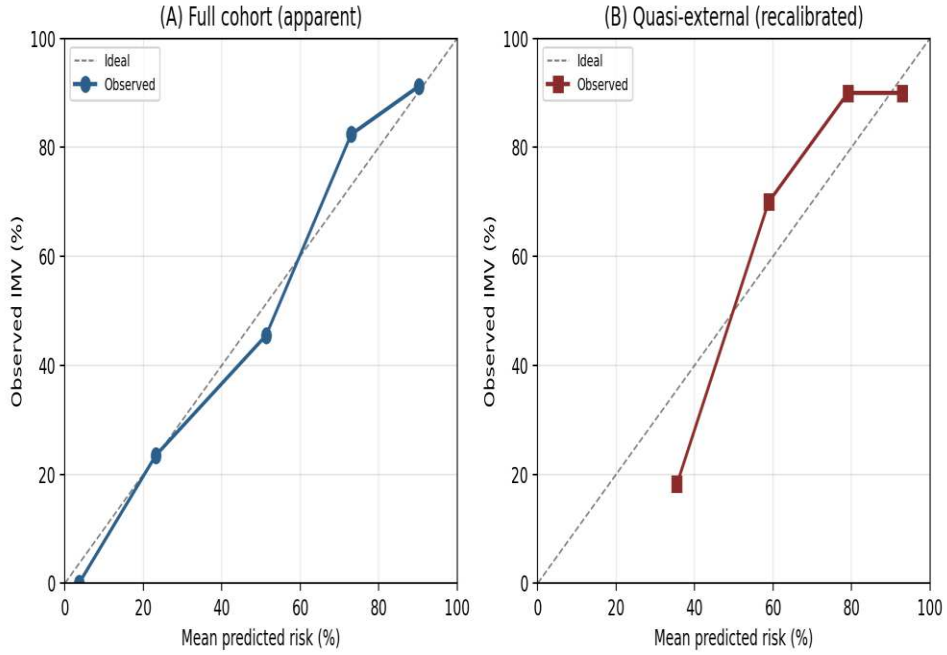



*Figure 5. Calibration plots (observed vs. mean predicted event rate by quantile of predicted risk) for the apparent full-cohort model (A) and the quasi-external validation after full logistic recalibration (B).*

**Table 5. Calibration by quantile of predicted risk: observed vs. predicted event rates**

| Cohort | Quantile of predicted risk | n | Observed IMV | Mean predicted |
|---|---|---|---|---|
| Full cohort (apparent) | Quintile 1 (lowest) | 34 | 0.0% | 3.6% |
| | Quintile 2 | 34 | 23.5% | 23.2% |
| | Quintile 3 | 33 | 45.5% | 51.3% |
| | Quintile 4 | 34 | 82.4% | 73.0% |
| | Quintile 5 (highest) | 34 | 91.2% | 90.2% |
| Quasi-external (recalibrated) | Quartile 1 (lowest) | 11 | 18.2% | 35.5% |
| | Quartile 2 | 10 | 70.0% | 58.9% |
| | Quartile 3 | 10 | 90.0% | 79.1% |
| | Quartile 4 (highest) | 10 | 90.0% | 92.9% |

**Discussion**

We developed a preliminary prediction model for the need of invasive mechanical ventilation in patients with acute poisoning treated with renal replacement therapy (predominantly, though not exclusively, hemoperfusion), achieving encouraging discrimination (cross-validated and bootstrap-corrected AUC in the range 0.82-0.86) using clinical variables routinely available at ICU admission. We emphasise the word preliminary throughout this discussion: the available sample fell substantially below the formal minimum required (169 vs. 384-406 participants), and three independent internal-validation approaches-cross-validation, analytical shrinkage estimation, and bootstrap optimism correction-converged in indicating meaningful optimism in the apparent model fit. To our knowledge, this is the first model of its kind developed in a Latin American cohort, and the first to focus on a predominantly hemoperfusion-treated population for this outcome, addressing a gap identified in prior literature that has predominantly focused on hemodialysis-treated cohorts (1,3) or on different outcomes within hemoperfusion-treated populations (4).

Our finding that global severity scores (SOFA, APACHE II) dominate over toxin-specific category once non-linearity is properly modelled is broadly consistent with prior ML work in methanol, organophosphate, and hemodialysis-treated poisoning cohorts, which likewise identified illness severity, neurological status, and acid-base derangement-rather than the specific xenobiotic per se-as leading predictors of adverse outcomes (1,2,3,25). The plateau effect observed for APACHE II beyond a score of approximately 15 is a clinically plausible finding not previously reported in this population that would not have been detected without formal assessment of logit-linearity, and merits confirmation in larger cohorts.

Placed within the broader toxin-specific literature, our findings are coherent with established mechanistic knowledge. Organophosphate poisoning classically progresses through an acute cholinergic phase followed, in roughly one-fifth of orally exposed patients, by the intermediate syndrome, characterised by proximal and respiratory muscle weakness that can precipitate delayed respiratory failure (18,19); a recent case report illustrates how this can occur after apparent clinical improvement, complicating early risk stratification based solely on admission severity (20). This mechanistic delay may partly explain the single clearly misclassified case identified through SHAP analysis in our cohort, in which a carbamazepine-poisoned patient with low admission severity scores subsequently required IMV. For tricyclic antidepressants, electrocardiographic markers such as QRS complex prolongation have recently been shown, in a large multi-substance registry analysis, to predict severe toxicity including the need for ventilatory support, offering a complementary, mechanism-specific predictor that our model-lacking electrocardiographic data-could not incorporate (21). For paraquat, recent nomogram-based and cohort studies converge in identifying ingested dose, respiratory and cardiovascular compromise, and delay to treatment as dominant prognostic factors (22,24), consistent with our finding that paraquat exposure independently increased predicted risk even after adjustment for global severity scores. A parallel line of work on the related herbicide diquat has similarly shown that interpretable ML models incorporating explainability methods can achieve strong discrimination for mortality risk, reinforcing the broader applicability of the modelling approach used here across structurally similar toxins (23).

Our explicit, quantified treatment of the sample-size shortfall distinguishes this study from much of the comparable literature, though we are careful to state what this treatment does and does not accomplish. Systematic reviews of both conventional and ML-based prediction model studies have repeatedly found that sample-size justification is among the most poorly reported items, with one review of machine-learning-based prognostic models finding that no included study fully reported a sample-size calculation (26,27). We not only calculated the Riley-based minimum sample size a priori, but subsequently obtained three mutually consistent lines of evidence-the analytical shrinkage factor (0.856), the empirical cross-validated calibration slope (0.844), and an independent bootstrap optimism correction (apparent AUC 0.898 reduced to 0.856)-that converge on the same conclusion: the model's true, generalisable performance is meaningfully lower than its apparent fit. This convergence documents and quantifies the expected overfitting with greater transparency than is typical in this literature; it does not eliminate it, and Firth's penalisation should be understood as a partial mitigation of separation and small-sample bias in the coefficient estimates, not a correction that restores adequate power. We therefore present this model explicitly as a preliminary derivation.

Our use of decision curve analysis situates the model's value beyond discrimination alone. As has been emphasised in recent methodological guidance, a model can show excellent discrimination yet provide no clinical benefit over simple default strategies if its net benefit does not exceed that of treating all or no patients across plausible decision thresholds (28,29); our finding that both the Firth-penalised and random forest models exceeded the 'treat-all' strategy across nearly the entire clinically relevant threshold range therefore suggests potential clinical utility that would require confirmation in adequately powered validation studies before any clinical application.

The interpretability analysis also connects this work to a growing body of literature applying SHAP specifically to mechanical-ventilation-related outcomes in critical care, which has consistently identified

severity-of-illness scores and vasoactive or ventilatory support variables as dominant predictors, mirroring our findings and supporting the external face validity of the SHAP-derived variable ranking obtained here (30).

In contrast to the common practice of acknowledging fairness only narratively in prediction model studies, we conducted a formal sensitivity analysis of discrimination and calibration by sex, motivated by the female preponderance in our cohort (58.0%) and the borderline univariable sex difference (p=0.059). Point-estimate discrimination was higher in men (AUC 0.883) than in women (AUC 0.785), but a formal DeLong test for this difference (0.098, 95% CI -0.022 to 0.218, p=0.109) did not reach conventional statistical significance, despite adequate calibration-in-the-large in both subgroups. Recent work has proposed reframing algorithmic fairness in terms of subgroup-specific net benefit rather than equality of raw performance metrics alone, cautioning that naive equalisation approaches can paradoxically reduce net benefit in groups where a model already performs well (31); other authors have highlighted that differential model performance across demographic groups can, if undetected, entrench rather than reduce existing health disparities when models are deployed in resource-limited settings (32). We present this as a genuine, quantified equity question rather than either a confirmed disparity or a dismissed one: the wide confidence interval means our data cannot distinguish between a negligible sex effect and a substantial one, and this ambiguity-not a confirmed gap-is itself the actionable finding, since it identifies subgroup performance as a specific, adequately powered target for any future external validation, rather than only overall performance.

Our choice of a Firth-penalised logistic regression with restricted cubic splines, rather than alternative approaches suited to small samples, merits explicit justification. We favoured Firth's penalisation over LASSO or elastic-net regularisation because Firth's method directly targets the small-sample and separation bias inherent to maximum-likelihood logistic regression, whereas L1-type penalties are designed primarily for variable selection under sparsity and can, in a sample this size, discard clinically important predictors (e.g., rare but mechanistically plausible toxic-agent categories) rather than shrinking their estimates toward more stable values-a trade-off we judged undesirable when a core aim of the study was to characterise, not eliminate, the toxin-specific signal. We favoured restricted cubic splines over simple polynomial terms because splines behave more stably in the tails of the predictor distribution (where APACHE II and SOFA values are sparsest) and require fewer degrees of freedom than higher-order polynomials for an equivalent degree of flexibility; we did not explore fractional polynomials specifically, and this remains a reasonable alternative for future work. We did not implement a Bayesian model with informative or weakly informative priors, which would be a natural extension particularly well suited to this sample size and is, in our view, the most promising direction for a future methodological refinement of this pipeline; we did not do so here in order to keep the primary analysis within a widely accessible, frequentist toolchain, and we flag this explicitly as a direction we did not pursue rather than one we considered and rejected on technical grounds.

The imputation-analysis congeniality issue identified in the Limitations warranted direct quantification rather than only a qualitative caveat. We reran the multiple imputation of APACHE II and SOFA using a non-linear estimator (extremely randomised trees) within IterativeImputer, in place of the original Bayesian ridge regression, while leaving every other element of the pipeline unchanged. The pooled SOFA spline coefficient shifted from 0.849 (original, linear imputation model) to 0.471 (non-linear imputation model)-a proportionally large change-while the cross-validated AUC shifted from 0.821 to 0.863. This confirms that the choice of imputation model is a genuine, non-negligible source of analytical variability in this dataset, not merely a theoretical concern. We do not present the non-linear-imputation result as a preferred or corrected estimate, however, because extremely randomised trees lack a native posterior-sampling mechanism analogous to Bayesian ridge regression's, so the between-imputation variance under this alternative is implausibly small and likely understates true imputation uncertainty; a fully congenial solution would require an imputation model that both respects the non-linear structure of the analysis model and provides a valid posterior variance, such as substantive-model-compatible multiple imputation or a fully

Bayesian joint model, neither of which we implemented here. We report this sensitivity analysis to document the direction and approximate magnitude of the risk, consistent with recommendations for transparent reporting of this specific methodological trade-off, and recommend that future studies in comparably sized cohorts either adopt a congenial imputation-analysis specification from the outset or, at minimum, report this same sensitivity comparison.

The uneven distribution of treatment modality (hemoperfusion alone, hemodialysis alone, or combined) across centres-with hemodialysis used exclusively at the predominant centre and almost never at satellite centres-raises a related, though distinct, generalisability question. While including modality as a covariate did not materially change model discrimination or the direction of effect for the principal predictors, modality was itself significantly associated with IMV, and its near-complete confounding with centre in this dataset means we cannot fully separate a modality-specific effect from other unmeasured differences between the predominant and satellite centres. We have revised the study title and population description accordingly, from a hemoperfusion-specific framing to the broader and more accurate description of renal replacement therapy for acute poisoning, predominantly but not exclusively hemoperfusion.

**Methodological implications for future study design**

Beyond the clinical question addressed, we highlight four concrete, quantitative implications of this work for researchers designing similar studies. First, a prospective or pooled-retrospective validation of this specific model should target at least 384-406 participants with the outcome prevalence observed here (48.5%), following the Riley et al. sample-size framework applied throughout this study, rather than relying on convenience samples or EPV-based heuristics. Second, if such a validation is to be conducted across a predominant and several satellite centres, as in the present study, satellite contribution should be planned to yield a combined satellite sample of at least 100-150 patients, roughly 2.5-3.5 times the 41 satellite patients available here, to obtain a quasi-external or external AUC estimate with a confidence interval narrow enough to be clinically interpretable (our satellite-only AUC of 0.564 had a standard deviation of 0.178, too wide to distinguish a true null effect from a moderate one). Third, networks planning multi-institutional data collection for missing-at-random-by-cluster scenarios like the one identified here (APACHE II/SOFA 100% missing in satellite centres) should establish standardised, centre-independent severity-score documentation protocols before data collection begins, since post hoc imputation-however carefully congenial-cannot fully substitute for consistent primary data capture. Fourth, to formally resolve the sex-based discrimination question left open here (observed difference 0.098, 95% CI -0.022 to 0.218), a future study would need a sample large enough to estimate a DeLong-test confidence interval excluding zero if a true difference of this magnitude exists; using standard formulae for the variance of the AUC difference and the effect size observed here, this would require several hundred events in each sex stratum, considerably larger than the 41 events per stratum available in this cohort-underscoring that subgroup equity analyses require explicit, separate power calculations, not an assumption that overall study power is sufficient.

The quasi-external validation, while not equivalent to a genuinely independent cohort, provides reassuring evidence of geographic generalisability within this dataset, with preserved discrimination and a calibration deficit correctable through standard recalibration methods. This provides more geographic stratification than is common in comparable derivation studies, several of which report only internal cross-validation without any geographic or temporal stratification (1,3). Formal planning of the sample size required for a future prospective external validation study, following established methodology for validation-specific sample-size calculation, should precede that next phase of this research programme (17).

**Limitations**

Several limitations should be considered, and we present this list without euphemism given the preliminary nature of this work. First, the cohort is predominantly single-centre (75.7% of cases from one institution); although sensitivity analyses suggest this does not confound the primary findings, true multi-centre

representativeness remains limited. Second, and most importantly, the available sample size (169 patients, 82 events) falls substantially below the formal minimum calculated using contemporary methodology (384-406 participants, i.e., our sample represents approximately 44% of the required minimum), and three convergent internal-validation approaches (analytical shrinkage, cross-validation, and bootstrap optimism correction) confirm meaningful overfitting in the apparent model. This is not a limitation that penalised regression 'corrects'; Firth's method mitigates small-sample bias and separation but does not restore adequate statistical power, and we present this model as a preliminary derivation requiring confirmation in an adequately powered sample, not as a finalised predictive tool. Third, no genuinely independent external validation cohort was available; the quasi-external, leave-cluster-out analysis presented here is a partial substitute, not an equivalent, and its full logistic recalibration slope (1.603) indicates non-trivial risk-stratification distortion when the model is transported to satellite centres, not merely a minor, fully correctable deficit. Fourth, our formal sex-based sensitivity analysis found a point-estimate discrimination gap (AUC 0.785 women vs. 0.883 men) that did not reach statistical significance on a formal DeLong test (difference 0.098, 95% CI -0.022 to 0.218, p=0.109); this wide, inconclusive interval-rather than a confirmed disparity-is itself the limitation, since it means this study cannot rule out a clinically meaningful sex-based gap and a larger, adequately powered analysis is needed to resolve the question. Fifth, treatment modality (hemoperfusion, hemodialysis, or combined) was unevenly distributed across centres and could not be fully disentangled from centre-level differences; we have accordingly broadened the study title and population description beyond 'hemoperfusion' specifically. Sixth, the free-text toxic agent variable required grouping into six categories, with 33 distinct rare categories collapsed into a heterogeneous 'other' group, entailing some loss of clinical granularity, and the wide confidence intervals for rare toxic categories (Table 2) should be read as directional associations only. Seventh, our multiple imputation model for APACHE II and SOFA used a linear estimator (Bayesian ridge regression) among auxiliary variables, whereas the analysis model subsequently treated these same variables non-linearly via restricted cubic splines-a form of imputation-analysis incongeniality (in the sense of Meng, 1994) that we quantified directly (see Discussion) rather than merely acknowledging: a sensitivity analysis using a non-linear imputation estimator (extremely randomised trees) shifted the pooled SOFA spline coefficient from 0.849 to 0.471 and the cross-validated AUC from 0.821 to 0.863, indicating that the choice of imputation model is not a negligible methodological detail in this dataset, though the non-linear alternative has its own limitation (near-zero between-imputation variance, likely understating imputation uncertainty, since extremely randomised trees lack a native posterior-sampling mechanism analogous to Bayesian ridge regression). Eighth, electrocardiographic and biochemical markers with established mechanistic relevance in specific poisonings (e.g., QRS duration in cyclic antidepressant overdose) were not available in this dataset and could not be evaluated (21). Finally, the simplified VM-Risk score's apparent performance (AUC 0.848, bootstrap 95% CI 0.793-0.901) reflects evaluation on the same data used for its derivation; its component thresholds and risk bands are descriptive of this sample only, and we do not recommend any specific cut-off for clinical use pending independent validation. Ninth, the decision to initiate invasive mechanical ventilation reflects both physiological indication and local practice patterns; we did not standardise intubation criteria across centres, and between-centre variability in the threshold for IMV may introduce information bias that our models cannot distinguish from true biological risk, particularly given that 75.7% of cases originate from a single centre. Tenth, no temporal validation was performed; model performance may differ in more recent or future time periods due to evolving clinical practice, and this represents a necessary component of any future validation strategy alongside geographic external validation. Eleventh, as a retrospective chart review, this study is subject to information bias from incomplete or non-standardised documentation, particularly for variables such as time from ingestion to first dialysis, which relied on prior clinical or companion-reported timing rather than directly observed measurement.

**Conclusions**

A Firth-penalised logistic regression with restricted cubic splines, informed by formal sample-size calculation and multiple imputation accounting for a missing-at-random mechanism conditional on treating

centre, showed encouraging but not yet definitive discrimination for predicting the need for invasive mechanical ventilation in patients treated with renal replacement therapy for acute poisoning. Given the sample size shortfall confirmed by three convergent internal-validation methods, the absence of genuine external validation, and a documented sex-based discrimination gap, we present this work explicitly as a preliminary, hypothesis-generating derivation. The derived VM-Risk score is reported as an exploratory illustration of how the model's output could eventually be translated into clinical practice, not as a validated tool, and no cut-off from it should presently guide clinical decisions. A prospective, adequately powered, genuinely external and multi-institutional validation study-explicitly designed to assess subgroup (sex) performance and, where feasible, to incorporate toxin-specific mechanistic markers-is the necessary next step before any consideration of clinical implementation.

**Declarations**

Ethics approval: This study was approved by the Human Research Ethics Committee of the Catholic University of Cuenca (Comité de Ética de Investigación en Seres Humanos de la Universidad Católica de Cuenca, CEISH-UCACUE) of the Ministry of Health of the Republic of Ecuador (approval: CEISH-UCACUE-2023-089). All methods were performed in accordance with the relevant guidelines and regulations. Data-use authorisation from each participating satellite institution was obtained. Retrospective use of de-identified data is consistent with Article 32 of Ecuador's Organic Law on Personal Data Protection (12).

Conflicts of interest: The authors declare no conflicts of interest.

Funding: No funding was received for this study

Data and code availability: The complete analytical pipeline (Python) is deposited alongside this submission and includes a synthetic dataset (n=169) generated by simulation from the marginal distributions and correlation structure of the observed data, calibrated to reproduce the observed outcome prevalence; this synthetic dataset allows independent execution and verification of the analytical code but does not reproduce individual patient records and must not be used to draw clinical or epidemiological conclusions. The de-identified real dataset is available upon reasonable, specified request, subject to the institutional data-sharing authorisations described above.

Patient and public involvement: No patients or members of the public were involved in the design, conduct, or reporting of this study.

Clinical trial number: not applicable

**References**


1. Hosseini SM, Rahimi M, Afrash MR, Ziaeefar P, Yousefzadeh P, Pashapour S, Evini PET, Mostafazadeh B, Shadnia S. Prediction of acute organophosphate poisoning severity using machine learning techniques. Toxicology. 2023;486:153431. https://doi.org/10.1016/j.tox.2023.153431

2. Shadnia S, Rahimi M, Mostafazadeh B, Evini PET, Hosseini SM, et al. Explainable artificial intelligence (XAI) for predicting the need for intubation in methanol-poisoned patients: a study comparing deep and machine learning models. Sci Rep. 2024;14:15916. https://doi.org/10.1038/s41598-024-66481-4

3. Rahimi M, Hosseini SM, Mohtarami SA, Mostafazadeh B, Evini PET, Fathy M, Kazemi A, Khani S, Mortazavi SM, Soheili A, Vahabi SM, Shadnia S. Prediction of acute methanol poisoning prognosis using machine learning techniques. Toxicology. 2024;504:153770. https://doi.org/10.1016/j.tox.2024.153770

4. Li X, Liu J, Cui S, Jian T, Ma S, Shi L, Lin Y, Zhang J, Zheng Y, Zhang Y, Jian X, Luan X, Kan B. Development and validation of a prediction model of deep venous thrombosis for patients with acute poisoning following hemoperfusion: a retrospective analysis. J Int Med Res. 2022;50:3000605221089779. https://doi.org/10.1177/03000605221089779

5. Buonsenso D, Mastrantoni L, Ulloa-Gutierrez R, et al; REKAMLATINA-3 MIS-C Study Group Investigators. Development and validation of multivariable machine-learning models for the prediction of Multisystemic Inflammatory Syndrome outcomes in Latin American children. Acta Paediatr. 2026;115(1):133-145. https://doi.org/10.1111/apa.70290

6. Zhang M, Zhang W, Zhao S, Tian X, Fu G, Wang B. Hemoperfusion in combination with hemofiltration for acute severe organophosphorus pesticide poisoning: a systematic review and meta-analysis. J Res Med Sci. 2022;27:33. https://doi.org/10.4103/jrms.JRMS_822_20

7. Riley RD, Snell KIE, Ensor J, Burke DL, Harrell FE Jr, Moons KGM, Collins GS. Minimum sample size for developing a multivariable prediction model: Part II - binary and time-to-event outcomes. Stat Med. 2019;38:1276-96. https://doi.org/10.1002/sim.7992

8. van Smeden M, de Groot JAH, Moons KGM, Collins GS, Altman DG, Eijkemans MJC, Reitsma JB. No rationale for 1 variable per 10 events criterion for binary logistic regression analysis. BMC Med Res Methodol. 2016;16:163. https://doi.org/10.1186/s12874-016-0267-3

9. White IR, Royston P. Imputing missing covariate values for the Cox model. Stat Med. 2009;28:1982-98. https://doi.org/10.1002/sim.3618

10. Heinze G, Schemper M. A solution to the problem of separation in logistic regression. Stat Med. 2002;21:2409-19. https://doi.org/10.1002/sim.1047

11. Harrell FE Jr. Regression Modeling Strategies: With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis. 2nd ed. Springer; 2015.

12. Ley Orgánica de Protección de Datos Personales. Registro Oficial Suplemento No. 459. Quito, Ecuador; 26 May 2021.

13. Collins GS, Moons KGM, Dhiman P, Riley RD, Beam AL, Van Calster B, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ. 2024;385:e078378. https://doi.org/10.1136/bmj-2023-078378

14. Moons KGM, Damen JAA, Kaul T, Hooft L, Andaur Navarro C, Dhiman P, et al. PROBAST+AI: an updated quality, risk of bias, and applicability assessment tool for prediction models using regression or artificial intelligence methods. BMJ. 2025;388:e082505. https://doi.org/10.1136/bmj-2024-082505

15. World Health Organization. Ethics and Governance of Artificial Intelligence for Health: WHO Guidance. Geneva: WHO; 2021.

16. Riley RD, Van Calster B, Collins GS. A note on estimating the Cox-Snell R2 from a reported C statistic (AUROC) to inform sample size calculations for developing a prediction model with a binary outcome. Stat Med. 2021;40:859-864. https://doi.org/10.1002/sim.8806

17. Riley RD, Ensor J, Snell KIE, Harrell FE Jr, Martin GP, Reitsma JB, Collins GS, Debray TPA, Moons KGM. Calculating the sample size required for developing a clinical prediction model. BMJ. 2020;368:m441. https://doi.org/10.1136/bmj.m441

18. Peter JV, Sudarsan TI, Moran JL. Clinical features of organophosphate poisoning: a review of different classification systems and approaches. Indian J Crit Care Med. 2014;18:735-745. https://doi.org/10.4103/0972-5229.144017

19. Buckley NA, Jayawardane P, Weerasinghe V, Dawson AH, Dassanayake TL. Serial repetitive nerve stimulation studies in organophosphorus poisoning indicate two distinct pathophysiological processes occur at the neuromuscular junction in the intermediate syndrome. Clin Toxicol (Phila). 2024;62(4):219-228. https://doi.org/10.1080/15563650.2024.2343744

20. Setia S, Chiwhane A. Intermediate syndrome due to organophosphate poisoning: a case report. Cureus. 2023;15(5):e39368. https://doi.org/10.7759/cureus.39368

21. Simon M, Kaplan S, Muschler K, Hoyte C, Brent J; on behalf of the Toxicology Investigators Consortium (ToxIC). The role of QRS complex prolongation in predicting severe toxicity in single-xenobiotic overdose. Clin Toxicol (Phila). 2024;62(1):32-38. https://doi.org/10.1080/15563650.2024.2307356

22. Tang G, Jiang Z, Xu L, Yang Y, Yang S, Yao R. Development and validation of a prognostic nomogram for predicting in-hospital mortality of patients with acute paraquat poisoning. Sci Rep. 2024;14:1622. https://doi.org/10.1038/s41598-023-50722-z

23. Li H, Liu Z, Sun W, Li T, Dong X. Interpretable machine learning for the prediction of death risk in patients with acute diquat poisoning. Sci Rep. 2024;14:16101. https://doi.org/10.1038/s41598-024-66940-y

24. Goyal P, Gautam PL, Sharma S, Paul G, Taneja V, Mona A. A study of paraquat poisoning presentation, severity, management and outcome in a tertiary care hospital: is there a silver lining in the dark clouds? Indian J Crit Care Med. 2024;28(8):741-747. https://pmc.ncbi.nlm.nih.gov/articles/PMC11372668/

25. Rahimi M, Afrash MR, Shadnia S, Mostafazadeh B, Evini PET, Bardsiri MS, Ramezani M. Prediction the prognosis of the poisoned patients undergoing hemodialysis using machine learning algorithms. BMC Med Inform Decis Mak. 2024;24:38. https://doi.org/10.1186/s12911-024-02443-0

26. Talimtzi P, Ntolkeras A, Kostopoulos G, Bougioukas KI, Pagkalidou E, Ouranidis A, et al. Machine learning-based COVID-19 prognostic models lag behind in reporting quality: findings from a TRIPOD/TRIPOD+AI systematic review. Diagn Progn Res. 2026;10. https://doi.org/10.1186/s41512-026-00218-x

27. Andaur Navarro CL, Damen JAA, Takada T, et al. Completeness of reporting of clinical prediction models developed using supervised machine learning: a systematic review. BMC Med Res Methodol. 2022;22:12. https://doi.org/10.1186/s12874-021-01469-6

28. Zhao L, Leng Y, Hu Y, Xiao J, Li Q, Liu C, Mao Y. Understanding decision curve analysis in clinical prediction model research. Postgrad Med J. 2024;100(1185):512-515. https://doi.org/10.1093/postmj/qgae027

29. Vickers AJ, van Calster B, Steyerberg EW. A simple, step-by-step guide to interpreting decision curve analysis. Diagn Progn Res. 2019;3:18. https://doi.org/10.1186/s41512-019-0064-7

30. Zheng L, Xue Y, Yuan Z, Xing X. Explainable SHAP-XGBoost models for pressure injuries among patients requiring mechanical ventilation in intensive care unit. Sci Rep. 2025;15:9878. https://doi.org/10.1038/s41598-025-92848-2

31. Benitez-Aurioles J, Joules A, Brusini I, Peek N, Sperrin M. Understanding algorithmic fairness for clinical prediction in terms of subgroup net benefit and health equity. Epidemiology. 2026;37(3):386-396. https://doi.org/10.1097/EDE.0000000000001949

32. Paulus JK, Kent DM. Predictably unequal: understanding and addressing concerns that algorithmic clinical prediction may increase health disparities. NPJ Digit Med. 2020;3:99. https://doi.org/10.1038/s41746-020-0304-9

## Supplementary Files

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

- codigoanaliticoVMRisk.zip
- codigoanaliticoVMRisk.zip
- SupplementaryMaterial.docx
- SupplementaryMaterial.docx