STATISTICAL ANALYSIS PLAN

UCSRS ATLAS Trial

Assessment of Transnational Performance of the Universal Cardiac Surgical Risk Score Across International Sites

Atiq Rehman MD — Union Health System, Terre Haute IN — atiqmd@aol.com

Field	Entry
SAP version	4.2
SAP date	24 October 2026 (amends v4.1: §1.1 layer structure, §1.5 and §1.5.1 calibration, §8.2, and the §16 calibration-status and MELD rows, for the v3.2 scoring changes)
Algorithm specification	UCSRS v3.2 (SPEC_VERSION 3.2.0), Algorithm Specification of Record v3.2
Status	Pre-specified prior to first patient enrollment. Coefficients and the calibration constant are those of v3.2 (24 October 2026); the constant is registry-anchored (Section 1.5.1).
Published algorithm	J Cardiothorac Surg 2026. DOI: 10.1186/s13019-026-04433-x
Calculator	https://ucsrs-calculator.netlify.app
Source code	https://github.com/atiqmdai/UCSRS (MIT license)
Registration	ClinicalTrials.gov, NCT number pending

PRE-SPECIFICATION STATEMENT

This Statistical Analysis Plan is finalized and dated before the first patient is enrolled at any participating site. It is submitted to ClinicalTrials.gov as part of the study registration record.

No analysis described here may be added, removed, or modified after enrollment begins without a dated amendment, which will be recorded in the registration and disclosed in any resulting publication. Analyses not pre-specified here are labeled exploratory and reported as hypothesis-generating only.

1. Algorithm Under Evaluation

The specification evaluated is UCSRS v3.2 (SPEC_VERSION 3.2.0), which departs materially from the published v1.0 architecture and from the v1.1/v2.x specifications previously governing this trial. The prior 50/50 blend of a physiology-derived baseline with an internally computed EuroSCORE II value is retired: Layer 1 is now a single log-odds baseline computed from the patient's own clinical variables, with EuroSCORE II published data as a comparator, and EuroSCORE II does not enter UCSRS score. The Layer 2b frailty instrument is the Frailty Test (FT, 0-6), extending the published 0-5 instrument with a second hemoglobin point below 8.0 g/dL. The published v1.0 worked cases do not reproduce under v3.0

or later and are superseded; the model constants (including the intercept) were set by the registry-anchored calibration described in Section 1.5.1.

1.1 Layer structure

```
Layer 1   Single log-odds physiology baseline (age, renal, EF, procedure, sternotomy
depth, aorta, PASP, pulmonary status), clamped 0.40-50%
Layer 2a  + MELD correction (log-odds, two-segment: 0.086/point 9-20, 0.038/point
above 20, capped at MELD 40)   additive, optional
Layer 2b  x FT odds multiplier (0-6 scale, x1.00 to x4.00, applied to the odds)
multiplicative on odds, required
Layer 2c  + LV geometry (LVESVI/LVEDD) + SYNTAX correction (log-odds)
additive, optional
Layer 3   + RHC hemodynamic corrections
additive, Tier 3 only
Final     UCSRS capped at 70%. EuroSCORE II is computed in parallel as a
comparator/reporting value only; it never enters any UCSRS layer.
```

1.2 Layer 2b —FT scoring and multipliers

Frailty Test (FT) is scored 0 to 6. Chair rise contributes 1 point at 15 seconds or longer and 2 points where the patient is unable to complete the maneuver; cognition contributes 1 point; albumin contributes 1 point below 3.5 g/dL, plus a second point below 3.0 g/dL; hemoglobin contributes 1 point below 13.0 g/dL (male) / 12.0 g/dL (female), plus a second point below 8.0 g/dL.

FT score	Multiplier	Component definitions
0	x1.00	Chair rise: 5x sit-to-stand >=15 s = 1 point; unable to complete = 2 points; <15 s = 0
1	x1.25	Cognition: Visual cognitive assessment: clinician eyeball judgment of impaired vs. normal = 1 point
2	x1.60	Hemoglobin: <13.0 g/dL (men) or <12.0 g/dL (women) = 1 point; a second point is scored below 8.0 g/dL
3	x2.10	Albumin: <3.5 g/dL = 1 point; a second point is scored below 3.0 g/dL
4	x2.60	Hemoglobin and albumin are required components
5	x3.10	Partial FT permitted where chair rise and/or cognition are not assessable
6	x4.00	Second hemoglobin point (<8.0 g/dL); scale extends the published 0-5 FT

***partial* FT rule: Hemoglobin and albumin are mandatory. Where chair rise or cognition are not assessable — typically urgent or emergency presentation — the available components are scored and the missing components recorded. Frailty may be underestimated in these patients. Partial FT status is captured as a binary variable and is a pre-specified stratification factor (Section 9).**

Application of the ladder (v3.2). The multiplier is applied to the odds of the pre-frailty risk (equivalent to adding the natural logarithm of the multiplier on the log-odds scale), not to the percentage. The ladder values are unchanged from v3.1 and may be re-estimated as ATLAS outcome data accrue.

1.5 Baseline construction (Layer 1)

Layer 1 is a single logistic (log-odds) model. For a given patient, an applicable-coefficient sum S is formed by adding the log-odds increment for every factor level present in that patient, drawn from the locked coefficient table (age per decade over 60 and per year over 75; female sex; renal function via serum creatinine or the dialysis substitution; pulmonary status at its highest applicable of four grades; ventricular/NYHA/recent-MI/PASP/atrial fibrillation terms; circulatory support at its highest applicable grade only, never additive; sternotomy depth; urgency; aortic extent; procedure increment; and comorbidity terms including BMI, diabetes, endocarditis, arteriopathy, neurological dysfunction, untreated valve lesions, immunosuppression and prior mediastinal radiation). Anemia, albumin and mobility carry zero weight in Layer 1 by design, being scored once only, in FT (Section 1.2), under the governing rule that no variable is double-counted.

The Layer 1 predicted probability is p = 1 / (1 + exp(-(intercept + S))), expressed as a percentage and clamped to the range 0.40-50%.

1.5.1 The intercept: registry-anchored calibration, executed; re-solved at v3.2

The intercept actually governing UCSRS as specified in this amendment (SPEC_VERSION 3.2.0) was set exclusively by the registry-anchored method of Section 1.5.1 as it reads below, executed; no provisional or interim intercept was ever applied to set a shipped constant.

Method, unchanged from the v2.1 precedent and now executed: for each of thirteen published surgical registries, a target mean predicted risk is calculated as that registry's published observed mortality divided by its published EuroSCORE II observed/expected ratio, and a synthetic-cohort severity parameter is tuned until the synthetic cohort's own mean EuroSCORE II equals that target -- the only point at which real-world information enters the exercise. No simulated patient is assigned an outcome at any stage, and no outcome model is used. The run was executed twice against the SPEC_VERSION 3.1.0 engine, at 30,000,000 simulated patients each time (13 anchored cohorts): a confirmatory run, and a locking run executed after final dead-code removal, which independently re-solved the intercept from scratch to 0.431392 against the shipped 0.433039 -- a difference of 0.0016 log-odds, within the solver's own tolerance -- and reproduced the result on two further hold-out realisations (median O/E 0.9976 and 0.9971) not used in any solve. The locking run's own result: median O/E 0.9965 across the thirteen registries, mean absolute log O/E 0.2662, 7 of 13 registries within the pre-specified band. In-band count is not stable across realisations (it has run 7 or 8 of 13); the median O/E and mean absolute log O/E are the primary quotes for this reason.

The v3.1 constant (0.433039) was set by the locking run above. By investigator instruction that run was the last calibration permitted before ATLAS reached 5,000 enrolled patients; the instruction was set aside by the investigator for the v3.2 release, which is made before any enrolment (v3.2 Specification of Record, Section 3). The complete evidence for the v3.1 calibration (run logs, generator, scoring code, and results files) is archived in the repository calibration folder and in UCSRS_v3.1_Calibration_and_Architecture_Review.md, both dated 20 September 2026, consistent with Section 17. From 24 October 2026 the instruction applies to v3.2: no coefficient or constant moves before ATLAS reaches 5,000 enrolled patients.

v3.2 re-solve (24 October 2026). The v3.2 scoring changes (aortic disease counted once; MELD slope 0.086 per point to MELD 20 and 0.038 above; frailty ladder applied to the odds) lower the

score for patients carrying MELD, frailty or aortic disease, so the calibration shift was re-solved by the same method: 325,000 simulated patients (25,000 for each of the 13 anchored cohorts), bisection on the shift to a median O/E of 1.00, giving 0.581146, adopted as 0.581. A hold-out check on a different random seed (520,000 simulated patients, 40,000 per registry) gave median O/E 1.0012, mean absolute log O/E 0.2503, and 7 of 13 registries within the pre-specified band. The same solver applied to the v3.1 engine returned 0.430 against the shipped 0.433039. The v3.2 constant was then confirmed by a 30,000,000-patient locking run (29,999,996 scored) on the same cohorts and dials, scored at nine shifts from 0.50 to 0.66: the shift at which the median O/E equals 1.00 is 0.5824 against the adopted 0.581 (0.0014 log-odds, within solver tolerance, so 0.581 stands); at 0.581 the median O/E is 1.0011, the mean absolute log O/E 0.2499, and 7 of 13 registries are within the band (0.80–1.25). Two further hold-out realisations of 2,600,000 patients each gave median O/E 0.9968 and 0.9936. As before, no simulated patient is assigned an outcome and no performance or discrimination claim follows.

NOTE ON LAYER 1

Layer 1 is a single log-odds baseline computed from the patient's own clinical variables. It does not take STS-PROM or EuroSCORE II as an input at any step. The prior 50/50 blend described in earlier versions of this trial's documents is retired as of UCSRS v3.0.

STS-PROM and EuroSCORE II remain required fields, recorded solely as external comparator scores for the trial's head-to-head analysis. Layer 1's age term is fitted to EuroSCORE II's own published log-odds age curve (a design choice, not a per-patient input), so UCSRS's claim to added value rests on Layers 2a (MELD), 2b (FT), 2c (LV geometry/SYNTAX) and 3 (RHC), not on Layer 1 independence from EuroSCORE II. STS has stated that adult cardiac surgery risk model coefficients are not currently available to investigators; STS-PROM is recorded from the site's own official calculator and is never reconstructed centrally. Similarly EuroScore is calculated separately from site’s own official calculator and recorded.

2. Objectives and Endpoints

2.1 Endpoint hierarchy

Endpoints are ordered by the strength of prior expectation. Published registry data document large, systematic calibration failure of the parent models — observed-to-expected ratios of 0.18 in low-risk

SWEDEHEART patients and 3.60 in a Brazilian CABG cohort — whereas discrimination is already moderate at AUC 0.74 to 0.84. Adding physiologic domains to a model reliably improves calibration and improves discrimination more modestly. The hierarchy reflects this.

Rank	Endpoint	Metric	Success criterion
Co-primary 1	Calibration (30-day mortality)	O/E ratio, calibration slope and intercept, Brier score	O/E within 20% of 1.0 (0.80 to 1.20) overall and in each risk stratum
Co-primary 2	Discrimination (30-day MACCE)	AUC by DeLong test with 95% CI, UCSRS vs each parent model	Statistically significant improvement over both parent models
Secondary 1	Clinical utility	Decision curve analysis, net benefit at 4% and 8% thresholds	Net benefit exceeding both parent models and treat-all/treat-none
Secondary 2	Reclassification	NRI at <4%, 4–7.9%, ≥8% strata; IDI	Positive NRI with CI excluding zero
Secondary 3	Morbidity	Composite of stroke, new dialysis, prolonged ventilation, deep sternal infection, reoperation, major bleeding	Descriptive; association with UCSRS reported
Secondary 4	Resource use	ICU and total hospital length of stay	Descriptive; association with UCSRS reported

2.2 Primary outcome definition

The study has two primary outcomes. The first is 30-day MACCE, as defined in the ATLAS Study Protocol, used for discrimination: a composite of all-cause mortality, permanent stroke, renal failure (per the current STS Adult Cardiac Surgery Database definition) and prolonged ventilation (>24 hours), with STS standard definitions applied uniformly at all sites. All-cause mortality is defined as death within 30 days of the index procedure, or in-hospital death occurring after day 30 where the patient has remained continuously admitted since the index procedure, whichever is later. Source documents are the hospital record, operative and anesthesia notes, intensive care records, discharge summary, and death certificate.

The second primary outcome is 30-day all-cause mortality, used for calibration, because the parent models (STS-PROM and EuroSCORE II) are calibrated to mortality. Mortality is a component of the MACCE composite. Discrimination (Co-primary 2) is therefore evaluated on 30-day MACCE and calibration (Co-primary 1) on 30-day all-cause mortality. The individual morbidity components of MACCE are reported under Secondary 3.

3. Sample Size

Target enrollment is 5,000 patients pooled across all participating sites. Justification proceeds from the number of outcome events rather than the number enrolled, since statistical power in a mortality prediction model derives from events.

- Weighted mean operative mortality across 14 international registries is approximately 3%.
- At 3%, 5,000 patients yields approximately 150 outcome events.
- Approximately 150 events supports stable estimation of calibration slope, Brier score, and decision curve analysis (Riley et al., BMJ 2020). Note: v3.0 dissolves the eight domain weights of earlier

versions into a flat vector of approximately forty individual log-odds coefficients. Forty parameters against ~150 pooled events is approximately four events per parameter, below conventional guidance for full re-estimation; full re-estimation of the v3.0 vector at the 5,000-patient pooled target is not achievable regardless of how data-sharing routes resolve. Section 12.2 recalibration is accordingly limited to intercept/slope adjustment and coefficient-block multipliers, not full per-coefficient re-estimation, unless enrollment materially exceeds target.

- The primary DeLong AUC comparison requires a minimum of 1,840 patients (DeLong et al., Biometrics 1988).

Sites are not individually powered. Site-level analyzes are descriptive except where a site independently reaches 1,000 patients, at which point local recalibration becomes estimable (Section 12).

4. Analysis Populations

Population	Definition	Used for
Full analysis set	All enrolled patients meeting eligibility criteria with a computable UCSRS and ascertained 30-day outcome	All primary and secondary analyzes
Complete-data set	Full analysis set restricted to patients with all optional domains entered (MELD, LV, SYNTAX)	Sensitivity 8
Full-FT set	Patients with all FT components assessed	Sensitivity 9, partial-FT comparison
Hepatic subset	Patients in whom MELD was calculated (bilirubin, INR, and creatinine all recorded)	Sensitivity 10
Tier 3 set	Patients with formal right heart catheterization data	Sensitivity 6

5. Primary Analysis — Calibration

5.1 Observed-to-expected ratio

The O/E ratio is computed as total observed deaths divided by the sum of individual UCSRS-predicted probabilities, with a 95% confidence interval by the Poisson method. Reported overall, within each of the four risk strata (<4%, 4–7.9%, 8–19.9%, ≥20%), and for each parent model for comparison.

5.2 Calibration slope and intercept

A logistic regression of observed outcome on the logit of predicted probability yields the calibration slope; the intercept is estimated with the slope fixed at 1. Perfect calibration corresponds to slope 1 and intercept 0. Slopes below 1 indicate overfitting or excessive spread of predictions; intercepts departing from 0 indicate systematic over- or under-prediction.

5.3 Brier score and calibration plot

The Brier score is reported with its scaled version relative to the null model. Calibration plots present observed against predicted probability across deciles of predicted risk with loess smoothing and 95% confidence bands, for UCSRS and both parent models on common axes.

5.4 Hosmer–Lemeshow test

Reported as a secondary calibration metric across 10 deciles. Interpreted with caution: the test is sensitive to sample size and provides no measure of the magnitude or direction of miscalibration. Calibration slope, intercept, and the calibration plot are the primary calibration outputs.

6. Primary Analysis — Discrimination

The area under the receiver operating characteristic curve is computed for UCSRS, STS-PROM where available, EuroSCORE II, and the Layer 1 baseline. Pairwise comparisons use the DeLong method for correlated ROC curves with 95% confidence intervals and two-sided p-values.

Pre-specified comparisons: UCSRS versus Layer 1 baseline; UCSRS versus EuroSCORE II alone; and, in the subset with an official value available, UCSRS versus STS-PROM alone.

A significance threshold of 0.05 is applied to the co-primary discrimination comparison. No multiplicity adjustment is applied across the two co-primary endpoints, as calibration and discrimination address distinct questions and both are reported regardless of the other's result.

7. Secondary Analyses

7.1 Decision curve analysis

Net benefit is plotted across threshold probabilities from 0 to 30%, with formal comparison at the pre-specified 4% and 8% thresholds corresponding to the low-to-intermediate and intermediate-to-high risk boundaries. Curves are presented for UCSRS, both parent models, treat-all, and treat-none.

7.2 Reclassification

Net reclassification improvement is computed across the three-stratum classification (<4%, 4–7.9%, ≥8%) with separate reporting for events and non-events, and with bootstrap 95% confidence intervals from 2,000 resamples. Integrated discrimination improvement is reported alongside. Reclassification tables are presented in full.

7.3 Morbidity and resource use

The morbidity composite comprises stroke, new renal failure requiring dialysis, prolonged ventilation exceeding 24 hours, deep sternal wound infection, reoperation for any cause, and major bleeding, defined per STS ACSD version 4.20. Association with UCSRS is examined by logistic regression. Length of stay is analyzed by quantile regression given expected right skew.

8. Sensitivity Analyses

Ten sensitivity analyzes are pre-specified. Each is reported regardless of result.

No.	Analysis	Method
1	Domain/coefficient-block weight optimization	Logistic regression to derive empirically optimized block multipliers across the ~40-coefficient v3.0 vector where pooled events permit. Internal validation by 10-fold cross-validation with 200 bootstrap resamples for optimism correction. Produces the candidate next numbered coefficient set; full per-coefficient re-estimation is not expected to be supported at the 5,000-patient target (see Section 3).

No.	Analysis	Method
2	MELD contribution	UCSRS computed with and without the Layer 2a MELD correction. Change in AUC, calibration, and net benefit reported.
3	Frailty domain — removal	UCSRS computed with the Layer 2b multiplier fixed at ×1.00 for all patients. Quantifies the discriminatory and clinical value contributed by the frailty domain.
4	Layer 1 baseline comparison	Comparison of the implemented v3.0 Layer 1 log-odds baseline against EuroSCORE II alone and, in the subset with an official value, STS-PROM alone. The prior 50/50 blended-ensemble comparison in earlier SAP versions no longer applies, since Layer 1 no longer blends with either comparator.
5	Procedure-specific comparison	UCSRS compared against the 2024 STS procedure-specific calculators in the applicable procedure categories, where site-supplied values are available.
6	Tier comparison	UCSRS at Tier 1 (clinic inputs only) versus Tier 3 (full catheterization laboratory data) in patients with complete Tier 3 data.
7	LV and SYNTAX	UCSRS with and without the LV correction; with and without the SYNTAX correction; and empirical optimization of both correction terms.
8	Optional domain impact	Performance in the complete-data set versus patients with one or more absent optional domains, quantifying prognostic information lost when optional inputs are not entered.
9	Partial versus full FT	Performance in patients with all FT components assessed versus those with laboratory components only, quantifying the effect of incomplete frailty assessment in urgent and emergency presentation.
10	Domain overlap assessment	See Section 8.1 below.

8.1 Sensitivity 10 — Domain overlap assessment

Serum albumin contributes to the FT score in Layer 2b, while MELD in Layer 2a is computed from bilirubin, INR, and creatinine. Albumin does not enter the MELD formula, so no direct mathematical double-counting occurs. In patients with hepatic congestion or cirrhosis, however, hypoalbuminemia and elevated MELD may reflect a shared underlying process.

MELD is an optional domain and is not computed in the majority of patients. This analysis is therefore restricted to the hepatic subset, in whom bilirubin, INR, and creatinine were all recorded — in routine practice, patients in whom hepatic dysfunction was clinically suspected.

- Pearson correlation between serum albumin and MELD score within the hepatic subset, with 95% CI.
- Variance inflation factor across the full UCSRS variable set. A VIF exceeding 5 for albumin is flagged.
- Logistic regression entering both albumin and MELD as predictors of 30-day mortality with the Layer 1 baseline as an adjustment term, testing whether albumin retains independent predictive value after MELD adjustment.
- Observed versus predicted mortality in patients with both MELD ≥9 and albumin <3.5 g/dL, compared against the remainder of the cohort.

Pre-specified interpretation: An adjusted albumin odds ratio with 95% CI excluding 1.0 indicates independence and no correction is warranted. A CI including 1.0 combined with an albumin–MELD correlation exceeding 0.40 and an O/E ratio in the both-abnormal subgroup significantly below 1.0 indicates material overlap. In that event a MELD-conditional albumin weighting is derived from the data

and implemented in a subsequent numbered version, with the derivation reported in full. No coefficient is altered in anticipation of this result.

This analysis is performed as a planned interim assessment at approximately 1,000 enrolled patients and repeated on the full cohort.

8.2 Declared overlap — MELD and the Layer 1 renal term

Serum creatinine enters UCSRS twice: once in the Layer 1 renal term, and again inside MELD (Layer 2a), which is itself a function of creatinine, bilirubin and INR. This is a structural, investigator-declared overlap, distinct from the albumin–MELD overlap examined in Sensitivity 10 above — it exists by construction wherever MELD is entered, not only in the hepatic subset, because creatinine sets a hard lower bound on the achievable MELD score. At a serum creatinine of 2.0 mg/dL the minimum representable MELD is approximately 15; at 4.0 mg/dL it is approximately 22 (holding INR and bilirubin at their lowest normal values). A patient combination implying both an elevated creatinine and a low MELD is not constructible in this engine.

This overlap was examined but not corrected, by investigator ruling of 20 September 2026, following a 1,000,000-patient high-risk validation run (complete MELD and complete FT data; no outcome simulated at any stage) that quantified its consequence rather than its existence: in the profile examined, UCSRS read 17–to–45× EuroSCORE II, which does not score albumin, INR, bilirubin, hemoglobin or frailty at all. The ruling stands that this reflects the score capturing physiology its comparators cannot see, not an artifact requiring correction, and no coefficient was altered in response. Two related architectural findings from the same run are recorded for attention when outcome data becomes available rather than acted on now: the 70% score cap was reached by 40.1% of that high-risk population, and the MELD and FT multipliers compound multiplicatively, so their combined effect is larger than either alone. Full results are archived in calibration/logs/1M_highrisk_eft_meld_20Sep.log and UCSRS_v3.1_Calibration_and_Architecture_Review.md, Section 7b. Update, v3.2 (24 October 2026): the compounding finding was addressed in part by lowering the MELD slope to 0.086 per point, counting aortic disease once, and applying the frailty ladder to the odds; in a fully populated simulated cohort the proportion at the 70% cap fell from 0.86% to 0.17%. The creatinine overlap itself is unchanged, and the 1,000,000-patient high-risk validation was re-run on v3.2 (mean predicted risk 52.3% to 20.8%; share at the 70% cap 40.1% to 0.003%; ratio to EuroSCORE II 23.5× to 10.1×). See Specification of Record v3.2, Sections 1 and 3.1.

9. Subgroup Analyses

Pre-specified subgroups. All are reported with AUC, O/E ratio, and calibration slope, and forest plots present the primary discrimination result across subgroups.

Category	Subgroups
Procedure	Isolated CABG; isolated AVR; isolated MVR or repair; isolated tricuspid; CABG plus valve; multiple valve; thoracic aortic surgery; redo sternotomy
Urgency	Elective; urgent; emergency; salvage
Frailty	FT 0-1; FT 2-3; FT 4-6; partial FT versus full FT
Hepatic	MELD not calculated; MELD <9; MELD 9–15; MELD 16–20; MELD >20
Etiology	Rheumatic valve disease versus degenerative valve disease
Region	Pre-specified geographic groupings with adequate enrollment

Category	Subgroups
Calibration failure	Sites in countries with documented parent-model O/E departure exceeding 1.5 or below 0.7, analyzed as a pre-specified primary stratum rather than post hoc
Center volume	Institutional annual adult cardiac surgery volume above versus below 500 cases
Special populations	Acute aortic dissection; active endocarditis; preoperative mechanical support

The high-calibration-failure stratum is designated a pre-specified primary stratum. Prior registry data document parent-model O/E ratios of 3.60 in Brazil, 1.89 in Japan, 1.74 in Korea, and 1.57 in Turkey, and the UCSRS is expected to demonstrate its largest calibration benefit in these populations. Designating this stratum in advance avoids the appearance of post hoc subgroup selection.

10. Machine Learning Comparison

Gradient boosting models are fitted to the same variable set as the UCSRS and compared against it, addressing the calibration reporting gap identified in the published meta-analytic literature. This is a secondary, conditional analysis. It is performed only once pooled data reach the threshold in Section 3 (approximately 5,000 patients and 150 events), and it uses only data from sites that have agreed to share de-identified patient-level data; sites contributing summary results only are not included in it.

- Algorithms: XGBoost, LightGBM, random forest, and elastic-net logistic regression.
- Validation: nested 5-fold cross-validation with hyperparameter tuning in inner folds. Performance estimates derive from held-out outer folds only.
- Comparison metrics: AUC by DeLong, calibration slope and intercept, Brier score, and net benefit at 4% and 8% — the same metrics applied to the UCSRS.
- Interpretability: SHAP values reported at cohort and individual level to identify which variables drive predictions.
- Fairness of comparison: because the UCSRS is pre-specified and the machine learning models are fitted to the data, a cross-validated UCSRS is also reported alongside the full-dataset UCSRS so that like is compared with like.

A published systematic review found that only 26% of artificial intelligence studies in cardiac surgery included external validation and 19% adhered to TRIPOD reporting standards; a meta-analysis of 26 studies covering 565,063 participants was unable to perform a quantitative calibration comparison because calibration was inconsistently reported. This analysis is designed to supply that comparison.

11. Missing Data

Extent of missingness	Handling
Under 5%	Complete-case analysis
5% to 20%	Multiple imputation by chained equations, 20 imputations, combined by Rubin’s rules
Over 20%	Excluded from the primary model; reported descriptively and examined in sensitivity analysis

Optional domains that are deliberately not entered are not treated as missing data. Absence of MELD, LV dimensions, SYNTAX, or RHC inputs is an intended feature of the tiered architecture, and the score is computed with those corrections set to zero as specified. Sensitivity 8 quantifies the consequence.

Mandatory variables — Layer 1 inputs, hemoglobin, albumin, and the 30-day outcome — cannot be missing; records lacking them are excluded from the full analysis set and reported in the participant flow diagram.

12. Center Effects and Recalibration

12.1 Center effects

Random-effects logistic regression with center as a random intercept is fitted for the pooled analysis. Center-specific calibration plots are produced for every site contributing 100 or more patients. Between-center heterogeneity is quantified by the intraclass correlation coefficient.

Site contribution cap: To prevent any single high-volume site from dominating the pooled coefficient derivation, Sensitivity 1 is additionally performed with site-weighted estimation in which no individual site contributes more than 25% of the effective sample. Both unweighted and site-weighted coefficient sets are reported.

12.2 Recalibration thresholds

Threshold	Update permitted	Parameters estimated	Output
~1,000 patients, single site	Local recalibration	All 8 domain parameters against ~150 events (v1.x/v2.x); under v3.0, coefficient-block multipliers against ~150 events -- full re-estimation of the ~40-coefficient vector is not achievable at this n (Section 3)	Site-specific recalibrated score. Publishable as a single-center validation. Does not create a global version.
~5,000 patients pooled	Full coefficient re-estimation	Coefficient-block multipliers against ~150 events; full re-estimation of the ~40-coefficient v3.0 vector is not achievable at this n (Section 3)	Candidate next numbered UCSRS version. Requires its own validation publication.
~10,000+ with regional depth	Region-specific coefficient sets	Coefficient-block multipliers per region, each requiring ~150 regional events; full regional re-estimation of all ~40 coefficients is likewise not supported at this n	Regional coefficient sets sharing the identical architecture.

Where the O/E ratio departs from 1.0 by more than 20%, logistic recalibration follows the Van Calster method. Uncalibrated and recalibrated performance are always reported together. A minimum of 1,000 local cases is required before local recalibration is attempted. Recalibration and re-estimation are secondary and conditional: they apply only at the thresholds above and use only patient-level data from sites that agreed to share it. The primary analysis fits no model, and the locked UCSRS is unchanged.

13. Interim Analysis

A single planned interim analysis is performed when pooled enrollment reaches approximately 1,000 patients, or when any individual site reaches 1,000 patients, whichever occurs first.

- Purpose: early assessment of discrimination and calibration, and execution of Sensitivity 10.
- No stopping rules apply. This is an observational study with no randomized allocation and no intervention; there is no basis for stopping for efficacy or futility.
- No alpha adjustment is applied, as the interim analysis does not test the primary hypothesis for a stopping decision.
- Interim results may be reported as a preliminary publication provided the interim nature is stated explicitly and the analysis is identified as pre-specified.

14. Software and Reproducibility

- The primary analysis is performed in Python 3 using numpy, pandas and scipy. The secondary and sensitivity analyses (Sections 10 and 12.2, conditional on the pooled threshold and on site agreement to share patient-level data) additionally use scikit-learn, statsmodels, XGBoost, LightGBM and lifelines; R may be used for decision curve analysis via the rmda package and for calibration via CalibrationCurves.
- The complete analysis script is version-controlled and published as a supplementary file with the primary manuscript.
- The algorithm specification file (ucsrs_spec_v3.0.json) is archived with a permanent identifier so that any published result can be reproduced against the exact coefficient set used.
- Sites electing local analysis under data sovereignty provisions run the identical script and transmit summary statistics only.
- Random seeds are fixed and reported for all resampling and cross-validation procedures.

15. Reporting Standards

The primary manuscript follows the TRIPOD statement for prediction model validation studies, with the TRIPOD checklist submitted as a supplementary file. Observational reporting follows STROBE where TRIPOD is silent.

All ten sensitivity analyzes are reported regardless of direction of result. Subgroup analyzes are presented with explicit acknowledgment of multiplicity and are interpreted as hypothesis-generating unless designated a pre-specified primary stratum. Any deviation from this plan is documented with its date and rationale.

16. Known Specification Divergences

The following discrepancies between the published manuscript and the implemented specification are documented in advance and will be addressed transparently in the validation report rather than discovered during analysis.

Item	Divergence	Resolution for this SAP
Layer 1 architecture	Published v1.0/v1.1 specify a 50/50 blend of a physiology-derived baseline and EuroSCORE II. UCSRS v3.0 retires this blend entirely.	v3.0 Layer 1 is a single log-odds baseline; EuroSCORE II is comparator-only and never enters the score. This SAP evaluates the current engine (SPEC VERSION 3.2.0).
Layer 1 independence claim	Earlier documents described Layer 1 as independent of EuroSCORE II.	Withdrawn. Layer 1's age bands are fitted to EuroSCORE II's own published log-odds age curve. UCSRS's claim to added value rests on Layers 2a, 2b, 2c and 3, not on Layer 1 independence.
Calibration status	Resolved by dated amendment, 20 September 2026. The registry-anchored calibration run has been executed twice at 30,000,000 simulated patients against SPEC_VERSION 3.1.0 (confirmatory, then a locking run re-solving the intercept from scratch to 0.431392 against the shipped 0.433039, within solver tolerance). See Section 1.5.1. Re-solved at v3.2 (24 October 2026) to 0.581 at 325,000 simulated patients and confirmed by a 30,000,000-patient locking run (median O/E 1.0011; hold-outs 0.9968 and 0.9936); see Section 1.5.1.	Median O/E 0.9965 across the thirteen registries at lock, reproduced on two hold-out realisations (0.9976, 0.9971). The v3.1 instruction that no coefficient or constant moves before ATLAS reaches 5,000 enrolled patients was set aside for v3.2, which is made before any enrolment; it applies to v3.2 from 24 October 2026. No performance or discrimination claim follows from this run — it is an anchoring against published mortality, not a fit, and no simulated patient is assigned an outcome.
Frailty instrument	Published specification uses CFS; v1.1 used the Essential Frailty Toolset (EFT, 0-5).	v3.0 uses Frailty Test (FT, 0-6), adding a second hemoglobin point below 8.0 g/dL (v3.1 added a second albumin point below 3.0 g/dL; the total is capped at 6). Multiplier ladder re-set against FRAILTY-AVR reference ranges. Documented in the registration and site package.
MELD correction	Earlier versions used a single-segment percentage-point MELD slope.	v3.0 rescales MELD to log-odds, two-segment. At v3.2 the slope is 0.086/point from MELD 9-20 and 0.038/point above 20 (v3.0 and v3.1: 0.18 and 0.08), capped at MELD 40.
Renal term	Earlier versions scored creatinine clearance (Cockcroft-Gault).	v3.0 scores serum creatinine directly; clearance is retained only inside the internal EuroSCORE II sub-computation. Dialysis patients are scored, in both Layer 1 and MELD, as if creatinine = 4.0 mg/dL; sites submit the patient's actual measured creatinine regardless.
Procedure increments	Earlier versions carried an inversion in which isolated AVR read below isolated CABG.	Corrected in v3.0; procedure increments re-anchored to each class's own reference operation.
Coefficients without a published anchor	Aortic arch involvement (0.75) and the three-tier SYNTAX correction (0/2.5/4.5) are investigator clinical judgment; no published coefficient exists to fit them to.	Disclosed as judgment-based, not derived, pending sufficient trial events to re-estimate empirically.
Overlapping-weight items	Immunosuppression and prior mediastinal radiation are both weighted 0.40 in Layer 1 and are not clinically equivalent risks.	Retained as identical pending investigator decision; flagged for

Item	Divergence	Resolution for this SAP
		empirical differentiation once event counts permit (see Sensitivity 1).
Primary endpoint	An earlier draft of this plan named 30-day all-cause mortality as the primary outcome, following Section 8.4 of the original publication.	Per the ATLAS Study Protocol v4.0 (as approved by WCG IRB), 30-day MACCE (all-cause mortality, permanent stroke, renal failure, prolonged ventilation >24h) is the primary endpoint; 30-day all-cause mortality, a component of MACCE, is the second primary outcome for calibration. Discrimination is evaluated on MACCE and calibration on all-cause mortality; Secondary 3 reports the morbidity components individually.

17. Approval

This Statistical Analysis Plan is finalized before the first patient is enrolled and is submitted with the ClinicalTrials.gov registration.

_____ _____

Atiq Rehman, MD Date

Coordinator, UCSRS ATLAS Consortium

Union Health System, Terre Haute, Indiana 47804

Email: atiqmd@aol.com

Published algorithm: J Cardiothorac Surg 2026. DOI: 10.1186/s13019-026-04433-x

Calculator: https://ucsrs-calculator.netlify.app — Source: https://github.com/atiqmdai/UCSRS