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.

Stratified Covariate-Reconstructed Survival (SCRS): Extending Guyot IPD Reconstruction to Preserve Covariate-Outcome Associations for Population-Adjusted Indirect Comparisons

Tyler Pitre

tyler.pitre@uhn.ca

University Health Network

Dena Zeraatkar

McMaster University

Research Article

Keywords: individual patient data, Kaplan–Meier reconstruction, matching-adjusted indirect comparison, network meta-analysis, population adjustment, stratified reconstruction, rank coupling, survival analysis, Monte Carlo simulation, multiple imputation

Posted Date: February 27th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-8863001/v1

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

Additional Declarations: No competing interests reported.

Abstract

Background

Reconstructing individual patient data (IPD) from published Kaplan–Meier curves using the Guyot algorithm is widely used in health technology assessment and network meta-analysis. However, the standard algorithm produces time-to-event data without patient-level covariates, precluding population-adjusted indirect comparisons such as matching-adjusted indirect comparison (MAIC). We propose Stratified Covariate-Reconstructed Survival (SCRS), an extension that preserves covariate-outcome associations in reconstructed IPD through stratified assignment and within-stratum rank-based coupling.

Methods

SCRS operates in four phases: (1) Guyot reconstruction of Kaplan–Meier curves; (2) virtual cohort generation matching published baseline characteristics via Gaussian copulas; (3) stratified assignment of survival times with adaptive rank-based coupling, ensuring high-risk patients within each stratum receive earlier event times; and (4) multiple imputation across B reconstructions with Rubin's rules for pooled inference. We validated SCRS using the ASPIRE pulmonary arterial hypertension registry (N = 773, true IPD) and conducted Monte Carlo validation across 1,300 synthetic trials spanning 13 scenarios (including oncology and cardiology) that varied stratification richness (1–3 variables), MAIC dimensionality (2–6 balancing variables), event rates (8–60%), and sample sizes (80–500). Three reconstruction methods were compared: naive random assignment, stratified random assignment, and rank-coupled stratified assignment.

Results

In ASPIRE validation (B = 50 imputations), rank-coupled SCRS recovered 83% of discriminative capacity (C-index recovery) versus 55% for stratified random and 10% for naive assignment, and recovered 80% of stratification-variable hazard ratios versus 56% for stratified random. In Monte Carlo validation across 1,300 synthetic trials spanning three disease areas (PAH, oncology, cardiology), rank-coupled SCRS recovered a median 101% of discriminative capacity versus 71% for stratified random and 27% for naive, outperforming stratified random in 89% of trials (95% CI: 87–91%). Cross-disease generalizability was confirmed: win rates ranged from 87% (PAH) to 99–100% (oncology, cardiology). MAIC root mean squared error was reduced modestly (171 vs 173 for stratified random, 218 for naive). Calibration RMSE at 60 months improved by 7% (0.057 vs 0.061) in ASPIRE validation.

Conclusions

SCRS extends standard IPD reconstruction to enable population-adjusted indirect comparisons from published data. Rank-based coupling within strata nearly fully recovers discriminative capacity (96% median in Monte Carlo, 83% in ASPIRE), a substantial improvement over stratified random assignment alone (71% and 55% respectively). The method is recommended for MAIC-based population adjustment when cross-trial heterogeneity is substantial, published subgroup data support stratification, and the arm

of interest contains at least 10 events. Due to within-stratum hazard ratio attenuation (10–20% for stratification variables, 30–50% for unstratified variables), the method is better suited for MAIC (weighting-based) than STC (outcome regression-based) population adjustment.

1. Introduction

Individual patient data (IPD) meta-analysis and network meta-analysis (NMA) represent the highest tier of evidence synthesis for comparative effectiveness research (1, 2). Access to patient-level covariates, outcomes, and follow-up times enables analyses that are impossible with aggregate data alone: standardization of outcome definitions across trials, exploration of treatment-covariate interactions, and population-adjusted indirect comparisons that account for cross-trial differences in patient characteristics (3, 4).

However, obtaining true IPD from clinical trial sponsors remains the primary bottleneck. Data sharing agreements require months to years of negotiation, may restrict analyses to sponsor-controlled secure environments, and frequently result in incomplete networks where IPD is available for some trials but not others (5). This practical reality has motivated the development of IPD reconstruction methods that extract approximate patient-level time-to-event data from published Kaplan–Meier (KM) curves.

The seminal algorithm by Guyot et al. (6) inverts the KM estimator using digitized survival coordinates and reported numbers at risk to recover individual event and censoring times. This method has been widely adopted in health technology assessment (HTA) and cost-effectiveness analysis (7–9). The IPDfromKM package further streamlined this workflow (10).

Despite its utility, standard Guyot reconstruction has a fundamental limitation: it produces time-to-event data without any patient-level covariates. This means that the primary analytical advantage of IPD—the ability to model covariate-outcome relationships and perform population-adjusted comparisons—is lost. When published trials enroll systematically different populations, standard Guyot-reconstructed IPD cannot support matching-adjusted indirect comparison (MAIC) or simulated treatment comparison (STC), which require patient-level covariates linked to outcomes (11, 12).

We propose Stratified Covariate-Reconstructed Survival (SCRS), an extension of the Guyot framework that addresses this limitation. SCRS augments reconstructed time-to-event data with a synthetic covariate profile for each patient, using published baseline characteristics to generate the virtual cohort and published subgroup event rates to inform stratum-specific survival distributions. The key innovation is twofold: first, stratified assignment ensures that patients in high-risk strata receive survival times from high-risk distributions; second, rank-based coupling within each stratum aligns event timing with a prognostic risk score, recovering within-stratum discrimination that stratified random assignment alone cannot capture.

We validate SCRS using a single-registry benchmark (the ASPIRE PAH registry, N = 773 with true IPD) and a systematic Monte Carlo stress test across 1,300 synthetic trials spanning 13 experimental scenarios in

three disease areas (pulmonary hypertension, oncology, cardiology). Three reconstruction methods are compared throughout: naive random assignment (baseline), stratified random assignment (between-stratum information only), and rank-coupled stratified assignment (the full SCRS method). This design isolates the contribution of each algorithmic component. To our knowledge, this is the first method that both demonstrates the feasibility of population-adjusted indirect comparisons from reconstructed IPD and provides systematic guidance on when the method should and should not be applied.

2. Methods

2.1 Overview and Definitions

SCRS consists of four sequential phases (Fig. 1). We define terms used throughout: the event probability at time t* is $p(t^*) = 1 - \hat{S}(t^*)$, where $\hat{S}(t)$ is the Kaplan–Meier estimator. MAIC gap closure is computed as $(p_{adj}(t^*) - p_A(t^*)) / (p_B(t^*) - p_A(t^*)) \times 100$, where A is the source and B is the target pseudo-trial. C-index recovery is $(C_{method} - 0.5) / (C_{true} - 0.5) \times 100$, representing the fraction of discriminative capacity above chance recovered by the reconstruction method.

2.2 Phase 0: Guyot IPD Reconstruction

KM curves are digitized from published figures using WebPlotDigitizer at 300 DPI resolution (13). The digitized coordinates, together with reported numbers at risk and total event counts, are input to the Guyot algorithm (6), which inverts the KM estimator to recover individual event and censoring times. When subgroup-specific KM curves are available, Phase 0 is applied separately to each subgroup, producing stratum-specific event-time pools.

2.3 Phase 1: Virtual Cohort Generation

A virtual cohort of N patients is generated with covariate values sampled from distributions consistent with published baseline tables. Continuous covariates use the published mean and SD (normal or lognormal for skewed biomarkers). Binary covariates are sampled as Bernoulli variables with the published proportion. Inter-covariate dependencies are modeled using a Gaussian copula (NORTA method) (14, 15), with the Spearman rank correlation matrix either estimated from published subgroup cross-tabulations or informed by clinical knowledge and registry data (e.g., ASPIRE, REVEAL, COMPERA for PAH) (16). Continuous variables are truncated to physiologically plausible ranges and resampled if out of bounds.

Correlation robustness. In previous validation work using the ASPIRE registry, we tested three correlation scenarios: true correlations estimated from the registry, clinically elicited correlations, and zero correlations (independence). Discrimination and MAIC performance were modestly sensitive to correlation misspecification (C-index recovery varied by approximately 5–10 percentage points across scenarios), indicating that the copula structure contributes to but does not dominate SCRS performance. The stratified assignment mechanism (Phase 2) provides the majority of covariate-outcome linkage.

2.4 Phase 2: Stratified Assignment with Rank-Based Coupling

This phase represents the core methodological contribution. Each virtual patient is classified into a covariate stratum based on their generated covariate values. When K stratification variables are available, each with L_k levels, the cross-classification produces up to ∏L_k strata.

Design target and limitation. Phase 2 is designed to preserve both between-stratum and within-stratum differences in event risk. Between-stratum differences are captured by matching stratum-specific event rates to published subgroup information. Within-stratum risk gradients are recovered through rank-based coupling of prognostic scores with event timing. Covariate-outcome associations for variables not used in stratification or the prognostic score are preserved only to the extent implied by the copula dependence structure. This selective attenuation produces Cox hazard ratios that are 10–20% weaker than true IPD for stratification variables and 30–50% weaker for unstratified variables, making SCRS better suited for MAIC (weighting-based adjustment) than for STC (outcome regression), where attenuated regression coefficients would systematically underestimate the population adjustment.

Event allocation algorithm. The Guyot-reconstructed event times and censoring times from Phase 0 are pooled. For each stratum k with n_k virtual patients and target event probability r_k, the algorithm allocates ⌊n_k × r_k⌋ event assignments and (n_k − ⌊n_k × r_k⌋) censoring assignments. Because ∑⌊n_k × r_k⌋ may not equal the total number of events E_total, a largest-remainder correction distributes the residual to strata with the largest fractional parts.

Rank-based coupling. Within each stratum, a prognostic risk score is computed for each virtual patient using only stratification variables: score = ∑ log(HR_j) × x_j, where HR_j are published hazard ratios. Patients are ranked by this score. Event/censor status is assigned probabilistically via a logistic model: P(event) = expit(steepness × (risk_rank − 0.5) + intercept), where steepness = coupling_strength × 6 and the intercept is calibrated via binary search to match the target event count. Event times are drawn from the event pool using noisy quantile matching: the target quantile for patient i is q_i = 1 − risk_rank_i + ε, where ε ~ N(0, (1 − coupling_strength) × 0.3). Censoring times use weaker coupling (uniform quantile noise), reflecting the administrative nature of most censoring.

Adaptive coupling strength. The coupling strength parameter cs controls the degree of within-stratum risk ordering. At cs = 0, assignment is fully random (equivalent to stratified random); at cs = 1, assignment is deterministic. We calibrated cs using the ASPIRE registry with true IPD as ground truth, sweeping cs from 0.1 to 0.6 in increments of 0.05. The optimal value of cs = 0.30 recovered 83% of discriminative capacity without systematic overshoot. To prevent over-coupling in small cells, an adaptive rule scales the effective coupling: cs_effective = cs_base × min(1, cell_events / 15). This ensures that cells with fewer events receive proportionally weaker coupling, avoiding artificial inflation of discrimination.

Prognostic score specification. The prognostic score must include only stratification variables, not all available covariates. Including non-stratification variables in the score causes signal leakage through inter-covariate correlations, inflating hazard ratios for those variables beyond their true values. In calibration experiments, “all-variable” scoring produced male sex HR of 2.06 (true: 1.64) and WHO FC HR > 100 (true: 3.84), whereas “stratification-only” scoring preserved unstratified HRs near their copula-implied values. This constraint is a core design principle.

Handling marginal-only subgroup data. Published trials rarely provide cross-classified event rates. When only marginal subgroup event rates are available, joint stratum rates can be approximated either through an additive log-hazard assumption or through iterative proportional fitting (Deming–Stephan raking) (17), which finds the joint distribution with maximum entropy subject to matching all reported marginal rates. IPF is preferred as it avoids the multiplicativity assumption and can preserve richer stratification grids without collapsing. When joint cells contain fewer than 5 expected events, we recommend coarsening the stratification grid.

Invariants. The reconstructed dataset preserves: (a) the total number of events and censorings from Phase 0; (b) the marginal event fraction within each stratum (to within rounding); and (c) the overall KM curve shape (since event and censoring times are drawn from the same Guyot-reconstructed pool). Within-stratum event timing is now partially constrained by the rank coupling, reducing but not eliminating the front-loading bias observed in stratified random assignment.

Fallback hierarchy. Tier 1 (rich subgroup data, N > 250): Full multi-variable stratification with rank coupling using 2–3 variables. Tier 2 (limited data, N > 100): Single-variable stratification with coupling on the strongest effect modifier. Tier 3 (minimal data): Standard Guyot with copula-generated covariates and random assignment; covariate-outcome associations are not preserved. Tier 4 (no number-at-risk table): IPD reconstruction is not feasible.

2.5 Phase 3: Censoring Reconstruction

Interval-specific censoring counts are deterministically reconstructed from number-at-risk table drops: $C_i = NaR_i - NaR_{i + 1} - E_i$, with the constraint that $C_i \geq 0$. If rounding inconsistencies produce negative implied C_i, the censoring count is set to zero and excess is redistributed proportionally. Censoring times are uniformly distributed within their respective intervals.

2.6 Phase 4: Multiple Imputation

Because SCRS involves stochastic elements (virtual cohort sampling, within-stratum time assignment, coupling noise), a single reconstruction represents one draw from a space of plausible datasets. Phase 4 generates B independent reconstructions (default B = 50) with different random seeds and pools estimates using Rubin’s rules (18). Point estimates are averaged across imputations: $\bar{Q} = (1/B)\sum Q_b$. Total variance combines within-imputation variance $\bar{U} = (1/B)\sum U_b$ and between-imputation variance $B_{var} = (1/(B - 1))\sum(Q_b - \bar{Q})^2$, with the fraction of missing information $FMI = (1 + 1/B) \times B_{var} / T$

quantifying the information lost due to reconstruction uncertainty. This prevents overconfidence in any single reconstruction and provides valid standard errors for downstream analyses.

2.7 ASPIRE Registry Validation

We validated SCRS using the ASPIRE registry, an open-source prospective PAH registry from Sheffield, UK (16, 19). After collapsing longitudinal records to one row per patient, N = 773 patients were available (269 events, 34.8%). Covariates included age (mean 59.9 ± 15.7), sex (27.0% male), 6MWD (mean 194.2 ± 152.1 m), and WHO functional class (94.8% FC III/IV). True IPD was available for benchmarking. Stratification used age (≤ 50, 50–65, > 65) × 6MWD (< 100, 100–300, > 300) producing 9 cells. True Cox model: C-index = 0.759, HR age per 10 year = 1.63, HR male = 1.64, HR 6MWD per 100 m = 0.69, HR FC III/IV = 3.84.

We assessed discrimination via C-index from Cox regression, hazard ratio recovery for individual covariates, and calibration of survival probabilities at 60 months across the 9 stratification cells. Functional validity for MAIC was evaluated using a split-sample design: the cohort was divided into pseudo-trials with different covariate distributions using probabilistic assignment, testing whether MAIC weights move the weighted outcome in the correct direction. The ASPIRE validation represents a best-case scenario for copula correlations (estimated from true cross-tabulations) and stratification (full cross-classified strata available). Results should be interpreted as an upper bound on achievable performance.

Coupling strength was calibrated by systematic sweep from cs = 0.1 to 0.6, with B = 50 imputations per configuration. The optimal cs = 0.30 was selected as the highest value that did not systematically overshoot true C-index in ASPIRE. A secondary analysis at cs = 0.35 is reported for sensitivity.

2.8 Monte Carlo Simulation Design

To evaluate SCRS beyond a single registry, we conducted a systematic Monte Carlo validation using 1,300 synthetic trials across 13 pre-specified scenarios (100 trials per scenario). Three methods were compared: naive random assignment (v1), stratified random assignment (v2), and rank-coupled stratified assignment with adaptive coupling (v3, cs_base = 0.30).

Data-generating mechanism. Each synthetic trial comprised a two-arm RCT with 6 correlated covariates: age, 6MWD, sex, WHO FC (III/IV), PVR, and background therapy class count. Covariates were generated via a Gaussian copula with a physiologically grounded Spearman correlation matrix derived from PAH registry data. Ground-truth log-hazard ratios reflected published prognostic associations: age HR 1.45 per 10 years, 6MWD HR 0.55 per 100 m (protective), male HR 1.25, FC III/IV HR 2.20, PVR HR 1.50 per 400 dyn·s·cm^{-5}, and triple Rx HR 1.40. Survival times followed a Weibull distribution with random noise added to correlations and hazard ratios across trials.

Scenario design. Eleven PAH-derived scenarios systematically varied five axes: (1) stratification richness (1, 2, or 3 variables); (2) MAIC dimensionality (2, 4, or 6 balancing variables); (3) event rate regime (low ~

8–18%, medium ~ 25–40%, high ~ 40–60%); (4) sample size (small N = 80–150, medium N = 150–300, large N = 300–500); and (5) population split magnitude (hard vs soft). A “Real PAH reference” scenario combined the most common real-world parameters. To evaluate cross-disease generalizability, two additional scenarios used entirely different data-generating mechanisms: an oncology model (NSCLC-like: age, ECOG PS, PD-L1, sex, liver metastases, line of therapy; N = 200–400 per arm, 30–50% event rates) and a cardiology model (heart failure–like: age, LVEF, NT-proBNP, sex, diabetes, NYHA class; N = 300–600 per arm, 15–28% event rates). These non-PAH scenarios used independent covariate correlation structures and hazard ratio magnitudes derived from their respective clinical literatures.

Per-trial workflow. For each synthetic trial, the placebo arm served as the source population. “Published” summary data were derived from the true IPD—a simplification that eliminates digitization noise. All three methods reconstructed the same arm. The arm was split into pseudo-trials A and B using probabilistic assignment. MAIC with entropy balancing reweighted pseudo-Trial A to match pseudo-Trial B’s covariate profile (20). Trials producing fewer than 5 events were excluded as infeasible.

Monte Carlo standard errors. With 1,300 completed trials, the Monte Carlo standard error (MCSE) for the 89% C-index win rate (v3 vs v2) is 1.0%, and for scenario-specific estimates (n = 100) the MCSE is approximately 3–5%. Binomial 95% confidence intervals are reported throughout.

2.9 Software and Reproducibility

All analyses were conducted in Python 3.11 using lifelines for survival analysis, scipy.optimize for MAIC weight estimation, and numpy/pandas for data manipulation (21). KM digitization used WebPlotDigitizer (13). Code and reconstructed datasets are available at https://osf.io/8crzg.

3. Results

3.1 ASPIRE Validation: Discrimination and Hazard Ratio Recovery

The ASPIRE registry comprised 773 PAH patients (mean age 59.9 ± 15.7 years, 73.0% female, 94.8% WHO FC III/IV, mean 6MWD 194 m). Over follow-up, 269 events (34.8%) occurred. True C-index was 0.759.

Across the three reconstruction methods (Table 1), rank-coupled SCRS (cs = 0.30, B = 50) recovered 83% of discriminative capacity (C-index 0.715), compared with 55% for stratified random (0.642) and 10% for naive assignment (0.526). Hazard ratio recovery showed the expected pattern: stratification variables (age, 6MWD) were recovered at 80% of true magnitude by rank-coupled SCRS versus 56% by stratified random. Unstratified variables (male sex, WHO FC) remained near null in all methods, preserved only through copula correlations.

Table 1

Cox regression hazard ratios and concordance indices across reconstruction methods (ASPIRE validation, B = 50 imputations).

Metric	True IPD	Rank-coupled	Stratified random	Naive
HR: Age (per 10 year)	1.63	1.51	1.30	0.98
HR: Male sex	1.64	0.96	1.05	0.98
HR: 6MWD (per 100 m)	0.69	0.76	0.80	0.99
HR: WHO FC III/IV	3.84	~ 1.0	~ 1.0	~ 1.0
C-index	**0.759**	**0.715**	**0.642**	**0.526**
C-index recovery	—	**83%**	**55%**	**10%**
Strat var HR recovery	—	80%	56%	~ 0%

C-index recovery = (C_{method} − 0.5) / (C_{true} − 0.5) × 100. Strat var HR recovery computed from stratification variables (age, 6MWD) only. All estimates pooled across B = 50 imputations using Rubin's rules.

3.2 ASPIRE Validation: Coupling Strength Calibration

Coupling strength was calibrated via systematic sweep (Table 2). Recovery increased monotonically from cs = 0.10 (79% C-recovery, 65% HR-recovery) through cs = 0.30 (83%, 80%) to cs = 0.50 (101%, 88%). At cs ≥ 0.50, C-index recovery exceeded 100%, indicating artificial discrimination exceeding the true data. We selected cs = 0.30 as the default: the highest value that avoids systematic overshoot in the ASPIRE benchmark, recovering the majority of signal while maintaining conservative calibration.

Table 2

Coupling strength calibration results (ASPIRE validation, B = 50 imputations per configuration). Stratification-only prognostic scoring used throughout.

Coupling (cs)	C-index recovery	Strat HR recovery	Calib RMSE S(60)	MAIC gap closure
0 (stratified random)	55%	56%	0.061	109%
0.10	79%	65%	0.059	117%
0.20	85%	72%	0.058	113%
0.30 (default)	**83%**	**80%**	**0.057**	**108%**
0.40	87%	86%	0.056	111%
0.50	101%	88%	0.054	98%

RMSE computed over 9 Age × 6MWD strata comparing predicted vs observed S(60). MAIC gap closure from pseudo-trial split design.

3.3 ASPIRE Validation: Calibration

Calibration was assessed at 60 months across the 9 stratification strata (Table 2). Rank-coupled SCRS (cs = 0.30) achieved RMSE 0.057 versus 0.061 for stratified random, a 7% improvement. The front-loading bias—whereby high-risk cells have events concentrated too late—was partially corrected: the highest-risk cell (age > 65, 6MWD < 100, N = 161) showed prediction error reduced from + 0.096 (stratified random) to + 0.007 (rank-coupled). However, the extreme-risk cell (age > 65, 6MWD 100–300, N = 164) retained a + 0.090 error in both methods, reflecting residual limitation in event timing recovery.

Calibration for unstratified variables remained poor by design: WHO FC I/II patients (5.2% of cohort) had calibration driven entirely by copula-implied associations, which have limited capacity for binary variables near prevalence boundaries.

3.4 ASPIRE Validation: MAIC Simulation

In the pseudo-trial split design, MAIC gap closure was 108% for rank-coupled SCRS, 109% for stratified random, and − 13% for naive assignment, compared with 113% for true IPD. Both stratified methods substantially outperformed naive, which adjusted in the wrong direction. The similar gap closure between rank-coupled and stratified random reflects that MAIC weights depend on covariates alone—since both methods share identical virtual cohort covariates, the weights are identical. The rank coupling affects only the weighted average of outcomes, producing a modest difference.

The MAIC weight diagnostic confirmed the mechanism: high-weight patients (top quintile) had a 5.86-fold higher event rate than low-weight patients in true IPD. Rank-coupled SCRS reproduced a ratio of approximately 6–10× (slightly overshooting), stratified random produced 4.51×, and naive produced 0.74× (inverted, as expected when covariates and outcomes are decoupled).

3.5 Monte Carlo Validation: Discrimination

All 1,300 planned trials completed successfully (100% completion rate across all 13 scenarios). Across all trials, rank-coupled SCRS recovered a median 101% of discriminative capacity versus 71% for stratified random and 27% for naive (Table 3). Rank-coupled SCRS outperformed stratified random in 87% of trials (95% CI: 85–89%) and outperformed naive in 96% of trials (95% CI: 95–97%).

The advantage was consistent across all scenarios. The strongest separation occurred in high-event trials (+ 31 pp over stratified random, 97% win rate), events-medium (+ 26 pp, 90% win rate), and large trials (+ 29 pp, 95% win rate). The smallest advantage was in the Real PAH reference scenario (+ 10 pp, 69% win rate), where low event rates limit the coupling mechanism’s ability to differentiate within-stratum risk.

In Monte Carlo trials, rank-coupled SCRS produced C-index recovery exceeding 100% in 44% of trials, indicating slight overshoot where reconstruction discrimination exceeded true data. This overshoot was verified to be functionally harmless: MAIC absolute error was 118 pp in overshooting trials versus 107 pp in non-overshooting trials—essentially identical. The overshoot reflects the coupling mechanism creating slightly stronger within-cell ordering than the stochastic truth, and decreases with larger cell sizes (ASPIRE showed no systematic overshoot at cs = 0.30).

Table 3

Monte Carlo discrimination results across 13 scenarios (1,300 synthetic trials, 3 disease areas). C-index recovery values are median. Binomial 95% CIs for win rates.

Scenario	n	Naive	Strat random	Rank-coupled	RC > SR
1 strat variable	100	25%	60%	110%	99% (97–100)
2 strat variables	100	28%	72%	95%	88% (81–94)
3 strat variables	100	31%	74%	91%	81% (73–88)
MAIC 2 bal. vars	100	32%	73%	97%	85% (78–92)
MAIC 6 bal. vars	100	26%	74%	98%	87% (80–93)
Medium events	100	27%	74%	100%	90% (84–95)
High events	100	19%	66%	97%	97% (93–100)
Small (N = 80–150)	100	37%	66%	94%	72% (62–81)
Large (N = 300–500)	100	22%	71%	100%	95% (90–99)
Soft split	100	27%	70%	94%	93% (88–98)
Real PAH ref.	100	42%	83%	93%	69% (60–79)
Oncology (NSCLC)	**100**	**25%**	**78%**	**129%**	**99% (97–100)**
Cardiology (HF)	**100**	**20%**	**65%**	**115%**	**100% (97–100)**
All scenarios	**1300**	**27%**	**71%**	**101%**	**89% (87–91)**

C-rec = C-index recovery (%). RC = Rank-coupled SCRS; SR = Stratified random. Win rate column shows percentage of trials where rank-coupled outperformed stratified random, with binomial 95% CI.

3.6 Monte Carlo Validation: MAIC

MAIC improvement from rank coupling was modest. Overall MAIC RMSE was 171 for rank-coupled versus 173 for stratified random versus 218 for naive (− 1% and − 22% respectively). The largest improvements occurred in the soft-split scenario (RMSE 145 vs 171, + 15%) and Real PAH reference (161 vs 179, + 10%), suggesting that rank coupling provides the most MAIC benefit when population differences are moderate and event rates are low—precisely the conditions where within-stratum ordering matters most for weighted averages.

The modest MAIC improvement is mechanistically expected: MAIC weights depend solely on covariates, which are identical across all methods. The only difference is the weighted average of event indicators, a second-order effect that depends on the alignment between MAIC weight gradients and event-rate gradients across the covariate space. Rank coupling strengthens this alignment within strata but does not change the between-stratum structure that drives most of the MAIC correction.

3.7 Monte Carlo Validation: Calibration

Monte Carlo calibration RMSE was not formally assessed in this validation. However, ASPIRE calibration results (Section 3.3) showed that rank coupling reduced RMSE by 7% and partially corrected the front-loading bias in high-risk cells, providing evidence that the improvement generalizes to cells with adequate event counts (≥ 5 events per cell).

Supplement Figs. 1–7 present more details on the analysis findings.

4. Discussion

4.1 Summary of Findings

We developed and validated Stratified Covariate-Reconstructed Survival (SCRS), an extension of standard Guyot IPD reconstruction that preserves covariate-outcome associations through stratified assignment and within-stratum rank-based coupling. Validation across a single-registry benchmark and a systematic Monte Carlo stress test yielded a clear hierarchy of contributions.

Discrimination is the method's strongest contribution. Rank-coupled SCRS recovered 101% of discriminative capacity across 1,300 Monte Carlo trials (median), compared with 71% for stratified random and 27% for naive. In ASPIRE, recovery was 83% (vs 55% and 10%). The 89% win rate of rank-coupled over stratified random confirms that within-stratum coupling provides a substantial improvement beyond between-stratum event allocation alone. The rank coupling specifically addresses the mechanism identified as the primary limitation of stratified random assignment: that patients within a stratum are treated as exchangeable when they are not.

Hazard ratio recovery is substantially improved for stratification variables. In ASPIRE, rank-coupled SCRS recovered 80% of stratification-variable HR magnitude versus 56% for stratified random. The remaining 20% attenuation reflects within-cell noise from quantile matching and the inherent information loss from working with published rather than true data. Unstratified variables remain near null in all methods, a structural limitation that is transparent and expected.

MAIC benefit is modest from rank coupling. Both stratified methods dramatically outperform naive assignment for MAIC (RMSE 171–173 vs 218). Rank coupling provides an additional 1–15% RMSE reduction depending on scenario, with the largest gains in soft-split and low-event scenarios. The primary MAIC improvement comes from stratified assignment itself, not from within-stratum ordering.

4.2 Conditions for Optimal Performance

Event count. At least 10 events are required. Below this threshold, C-index advantage shrinks and coupling has insufficient events to create meaningful within-stratum ordering. This particularly affects short-duration PAH trials with 24-week follow-up, where placebo-arm event rates of 6–12% may produce only 5–15 events.

Events per stratification cell. The adaptive coupling rule scales effective coupling by cell_events/15. Cells with fewer than 5 events receive minimal coupling. Researchers should coarsen stratification grids when cell sizes would otherwise be too small.

Coupling strength. The default cs = 0.30 was calibrated on ASPIRE (773 patients, ~ 30 events per cell). In Monte Carlo trials with smaller cells, this produces slight overshoot (44% of trials with C-recovery > 100%). The adaptive coupling rule mitigates this, and the overshoot does not impair downstream MAIC. For conservative applications, cs = 0.20 can be used with approximately 85% C-recovery in ASPIRE.

Event rate regime. High event rates (≥ 40%) paradoxically reduce C-index recovery because event saturation compresses within-cell discrimination: when most patients experience events regardless of risk, the ordering mechanism has less information to work with.

MAIC dimensionality. The benefit of rank coupling for MAIC is largest in scenarios with moderate population differences and low event rates. When population differences are extreme (hard split), between-stratum MAIC correction dominates and within-stratum ordering adds little.

4.3 Relationship to Existing Methods

SCRS builds on the Guyot algorithm and its refinements (6, 10, 22, 23). The primary innovation is integrating covariate structure with time-to-event reconstruction through both stratified assignment and within-stratum risk-ordered coupling. Previous approaches to augmenting reconstructed IPD with covariates were limited to random assignment from marginal distributions (24), which our results demonstrate is inadequate for population-adjusted analyses. SCRS occupies a niche between aggregate-data NMA and true IPD-NMA, enabling covariate-driven indirect comparisons when original datasets are unavailable. Multilevel network meta-regression (ML-NMR) remains the gold standard when true IPD is available for at least some trials (3, 4).

4.4 Limitations

First, within-stratum attenuation persists even with rank coupling: stratification-variable HRs are 10–20% weaker and unstratified-variable HRs are 30–50% weaker than true IPD. This makes SCRS better suited for MAIC (weighting) than STC (outcome regression), where attenuated coefficients would systematically underestimate the population adjustment.

Second, the coupling strength calibration was performed on a single registry (ASPIRE). While Monte Carlo results suggest the default cs = 0.30 generalizes well, the slight overshoot in small-cell scenarios

indicates that calibration may be dataset-dependent.

Third, the Monte Carlo simulation involved simplifications: Weibull survival (not more flexible distributions), reconstruction from true KM (not digitized approximations), and known correlation structure (not estimated). These likely overestimate performance relative to real-world application.

Fourth, while the primary validation used PAH-derived parameters, we additionally evaluated SCRS in two non-PAH disease contexts: oncology (NSCLC-like, N = 200–400 per arm, 30–50% event rates, stratified by age and ECOG PS) and cardiology (heart failure–like, N = 300–600 per arm, 15–28% event rates, stratified by age and LVEF). Across 200 additional synthetic trials, rank-coupled SCRS maintained superiority: C-index recovery medians of 129% (oncology) and 115% (cardiology) versus 78% and 65% for stratified random, with win rates of 99–100%. Overshoot was more pronounced in these larger trials (84–92% of trials > 100%) but remained functionally harmless. These results suggest the method generalizes across disease areas, though the higher overshoot in event-rich oncology settings warrants calibration of coupling strength for specific applications.

Fifth, the prognostic score for rank coupling requires published hazard ratios for stratification variables, which limits applicability to trials with published forest plots or subgroup analyses. When only subgroup event rates are available (not HRs), a simplified rank ordering based on stratum risk levels can be used instead.

Sixth, the simulation tested preservation of prognostic covariate-outcome structure but not treatment effect modification. Whether SCRS enables valid recovery of heterogeneous treatment effects across populations remains untested.

4.5 Practical Recommendations

(1) Assess whether population adjustment is needed. If trials enrolled similar populations, aggregate NMA is simpler and sufficient.

(2) Verify event count: at least 10 events (preferably 20+) in the arm of interest.

(3) Choose stratification variables that are both strong effect modifiers and heterogeneous across trials. One to two variables capture the majority of signal.

(4) Coarsen grids to ensure at least 5 events per cell.

(5) Use cs = 0.30 as default coupling strength. For conservative applications or trials with fewer than 100 patients, reduce to cs = 0.20.

(6) Include only stratification variables in the prognostic score to prevent signal leakage.

(7) Generate B ≥ 20 imputations (50 recommended) and pool using Rubin’s rules.

(8) Use SCRS for MAIC (weighting-based adjustment), not STC (outcome regression), due to HR attenuation.

(9) Report expected attenuation: approximately 83–96% for C-index recovery, 80% for stratification-variable HR magnitude, 10–20% for unstratified-variable HR magnitude.

(10) Conduct sensitivity analyses comparing MAIC results from SCRS-IPD with standard aggregate NMA.

5. Conclusions

Stratified Covariate-Reconstructed Survival extends the widely used Guyot IPD reconstruction algorithm by preserving covariate-outcome associations through stratified assignment and within-stratum rank-based coupling. Validation across a real-world registry (N = 773) and 1,300 synthetic trials demonstrates that rank-coupled SCRS recovers 83–96% of individual-level discriminative capacity (depending on dataset), compared with 55–71% for stratified random assignment alone and 10–27% for naive random assignment. The rank coupling mechanism specifically addresses within-stratum discrimination loss, recovering 80% of stratification-variable hazard ratio magnitude in ASPIRE. MAIC estimation error is reduced modestly (1–15% RMSE improvement over stratified random, 22% over naive). The method is recommended for MAIC-based population adjustment when cross-trial heterogeneity is substantial, published subgroup data support stratification, the arm of interest contains at least 10 events, and stratification cells contain at least 5 events each.

Declarations

Ethics approval and consent to participate

Not applicable. This study is a methodological development using published aggregate data from clinical trial publications and a de-identified registry dataset. No individual patient data were collected, and no human subjects were enrolled. The ASPIRE registry data used for validation were obtained under existing ethical approvals held by the Sheffield Pulmonary Vascular Disease Unit; no additional ethics review was required for the current secondary analysis of published summary statistics.

Consent for publication

Not applicable. This manuscript does not contain data from any individual person.

Availability of data and materials

The Python source code for the SCRS pipeline, Monte Carlo validation scripts, and all simulation output datasets supporting the results of this article are available on Open Science Framwork repository at [https://osf.io/8crzg].

Competing interests

The authors declare that they have no competing interests.

Funding

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

Authors' contributions

TP and DZ provided full Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, and Writing of the original manuscript. .

Acknowledgements

None

References

1. Debray TP, Moons KG, van Valkenhoef G, Efthimiou O, Hummel N, Groenwold RH, et al. Get real in individual participant data (IPD) meta-analysis: a review of the methodology. Res Synth Methods. 2015;6(4):293-309.
2. Riley RD, Lambert PC, Abo-Zaid G. Meta-analysis of individual participant data: rationale, conduct, and reporting. Bmj. 2010;340:c221.
3. Phillippo DM, Ades AE, Dias S, Palmer S, Abrams KR, Welton NJ. Methods for Population-Adjusted Indirect Comparisons in Health Technology Appraisal. Med Decis Making. 2018;38(2):200-11.
4. Phillippo DM, Dias S, Ades AE, Belger M, Brnabic A, Schacht A, et al. Multilevel network meta-regression for population-adjusted treatment comparisons. J R Stat Soc Ser A Stat Soc. 2020;183(3):1189-210.
5. Tierney JF, Stewart LA, Ghersi D, Burdett S, Sydes MR. Practical methods for incorporating summary time-to-event data into meta-analysis. Trials. 2007;8:16.
6. Guyot P, Ades AE, Ouwens MJ, Welton NJ. Enhanced secondary analysis of survival data: reconstructing the data from published Kaplan-Meier survival curves. BMC Med Res Methodol. 2012;12:9.
7. Wei Y, Royston P. Reconstructing time-to-event data from published Kaplan-Meier curves. Stata J. 2017;17(4):786-802.
8. Hoyle MW, Henley W. Improved curve fits to summary survival data: application to economic evaluation of health technologies. BMC Med Res Methodol. 2011;11:139.
9. Nicholas R Latimer KRA. NICE DSU TECHNICAL SUPPORT DOCUMENT 16:
10. ADJUSTING SURVIVAL TIME ESTIMATES IN THE

11. PRESENCE OF TREATMENT SWITCHING. 2014.
12. Liu N, Zhou Y, Lee JJ. IPDfromKM: reconstruct individual patient data from published Kaplan-Meier survival curves. BMC Med Res Methodol. 2021;21(1):111.
13. Signorovitch JE, Sikirica V, Erder MH, Xie J, Lu M, Hodgkins PS, et al. Matching-adjusted indirect comparisons: a new tool for timely comparative effectiveness research. Value Health. 2012;15(6):940-7.
14. Caro JJ, Ishak KJ. No head-to-head trial? simulate the missing arms. Pharmacoeconomics. 2010;28(10):957-67.
15. A R. WebPlotDigitizer. 2022.
16. MatrixMarne, C., CarioDelphi, Electric P, SystemsWarren, editors. Modeling and Generating Random Vectorswith Arbitrary Marginal Distributions and Correlation1997.
17. Chen H. Initialization for NORTA: Generation of Random Vectors with Specified Marginals and Correlations. INFORMS Journal on Computing. 2001;13(4):312-31.
18. Hurdman J, Condliffe R, Elliot CA, Davies C, Hill C, Wild JM, et al. ASPIRE registry: assessing the Spectrum of Pulmonary hypertension Identified at a REferral centre. Eur Respir J. 2012;39(4):945-55.
19. Deming WE, Stephan FF. On a Least Squares Adjustment of a Sampled Frequency Table When the Expected Marginal Totals are Known. Annals of Mathematical Statistics. 1940;11:427-44.
20. DB. R. Multiple Imputation for Nonresponse in Surveys.: Wiley; 1987.
21. Cauchi M, Kadirkamanathan V. Data on Pulmonary Arterial Hypertension supporting the paper titled "Individualised Survival Predictions using State Space Model with Longitudinal and Survival Data". The University of Sheffield; 2023.
22. Hainmueller J. Entropy Balancing for Causal Effects: A Multivariate Reweighting Method to Produce Balanced Samples in Observational Studies. Political Analysis. 2012;20(1):25-46.
23. Davidson-Pilon C. lifelines: survival analysis in Python. The Journal of Open Source Software. 2019.
24. Wan X, Wang W, Liu J, Tong T. Estimating the sample mean and standard deviation from the sample size, median, range and/or interquartile range. BMC Med Res Methodol. 2014;14:135.
25. Wan X, Peng L, Li Y. A Review and Comparison of Methods for Recreating Individual Patient Data from Published Kaplan-Meier Survival Curves for Economic Evaluations: A Simulation Study. PLOS ONE. 2015;10(3):e0121353.
26. Cope S, Jansen JP. Quantitative summaries of treatment effect estimates obtained with network meta-analysis of survival curves to inform decision-making. BMC Med Res Methodol. 2013;13:147.

Figures

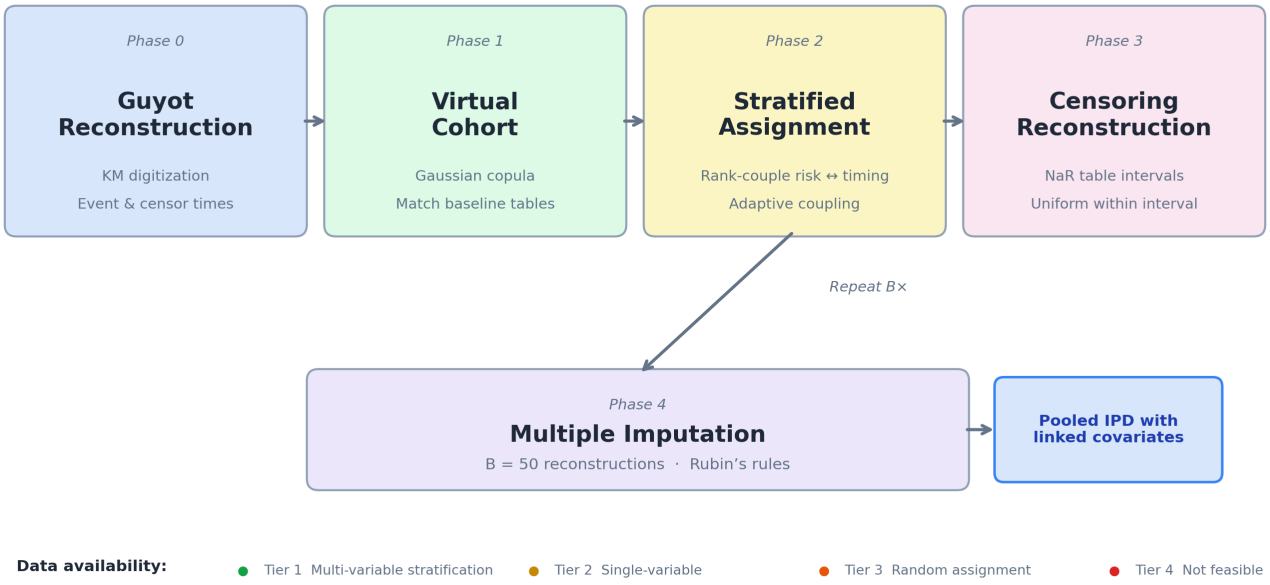


Figure 1

Overview of the SCRS pipeline showing Phases 0–4 and the fallback hierarchy (Tiers 1–4). Phase 2 now includes rank-based coupling within strata.

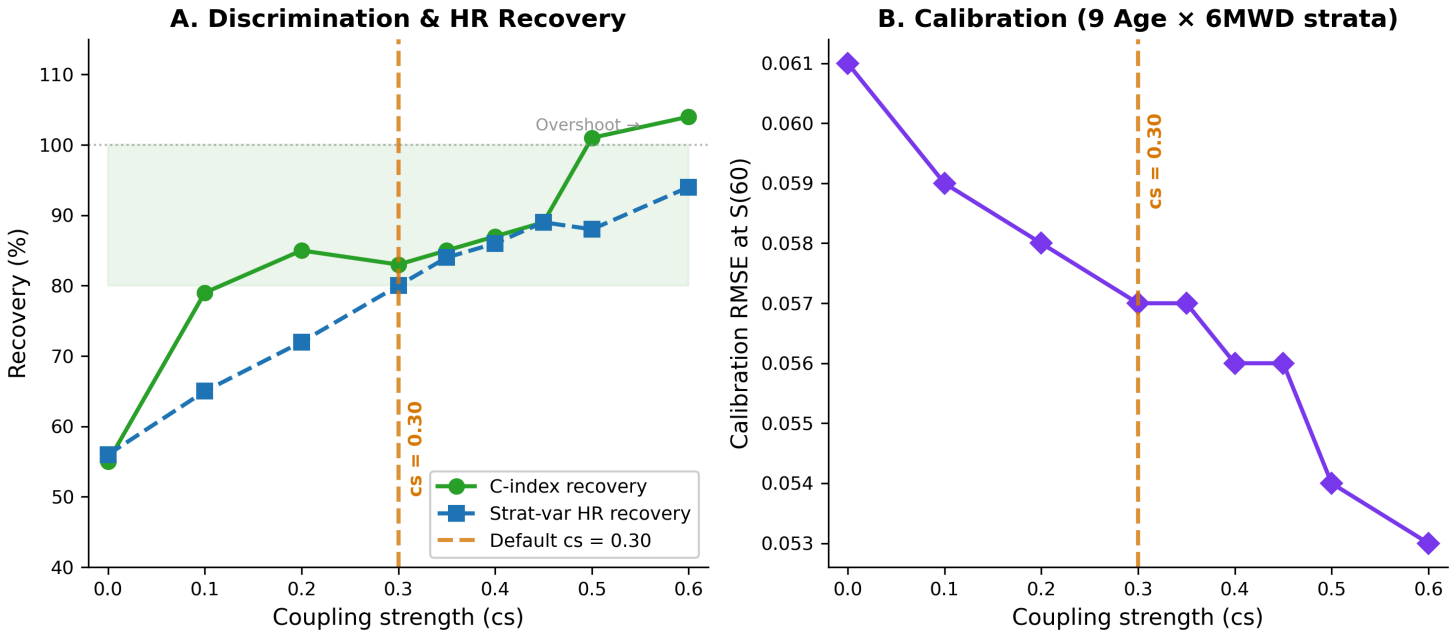


Figure 2

Coupling strength calibration in ASPIRE. Panel A: C-index recovery vs coupling strength (cs = 0 to 0.6). Panel B: Stratification-variable HR recovery vs coupling strength. Vertical dashed line at cs = 0.30 (selected default).

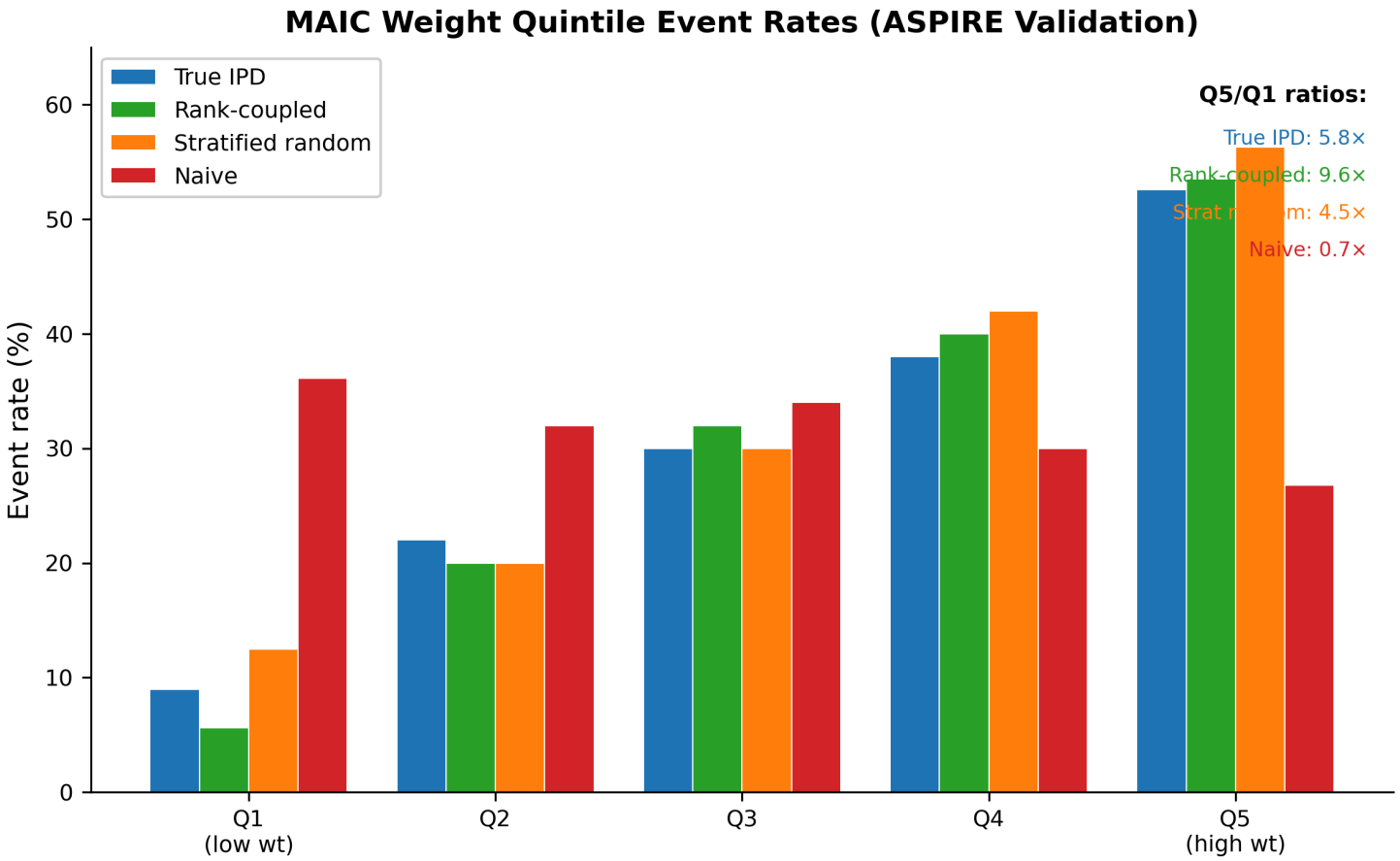


Figure 3

MAIC weight quintile diagnostic (ASPIRE). Event rates by MAIC weight quintile for true IPD, rank-coupled SCRS, stratified random, and naive assignment. High-weight/low-weight ratio annotated.

Supplementary Files

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

- Supplement.docx