Methods for external control borrowing in hybrid control arm designs

Mingyang Shan

Real World Analytics
Eli Lilly and Company

August 20, 2022

Outline

- Background
- Methods
  - Causal Inference Framework and Assumptions
  - Propensity Score Integrated Bayesian Methods
- Simulation
  - Type I error calibration
- Additional Data Integration Framework

Background and Motivation

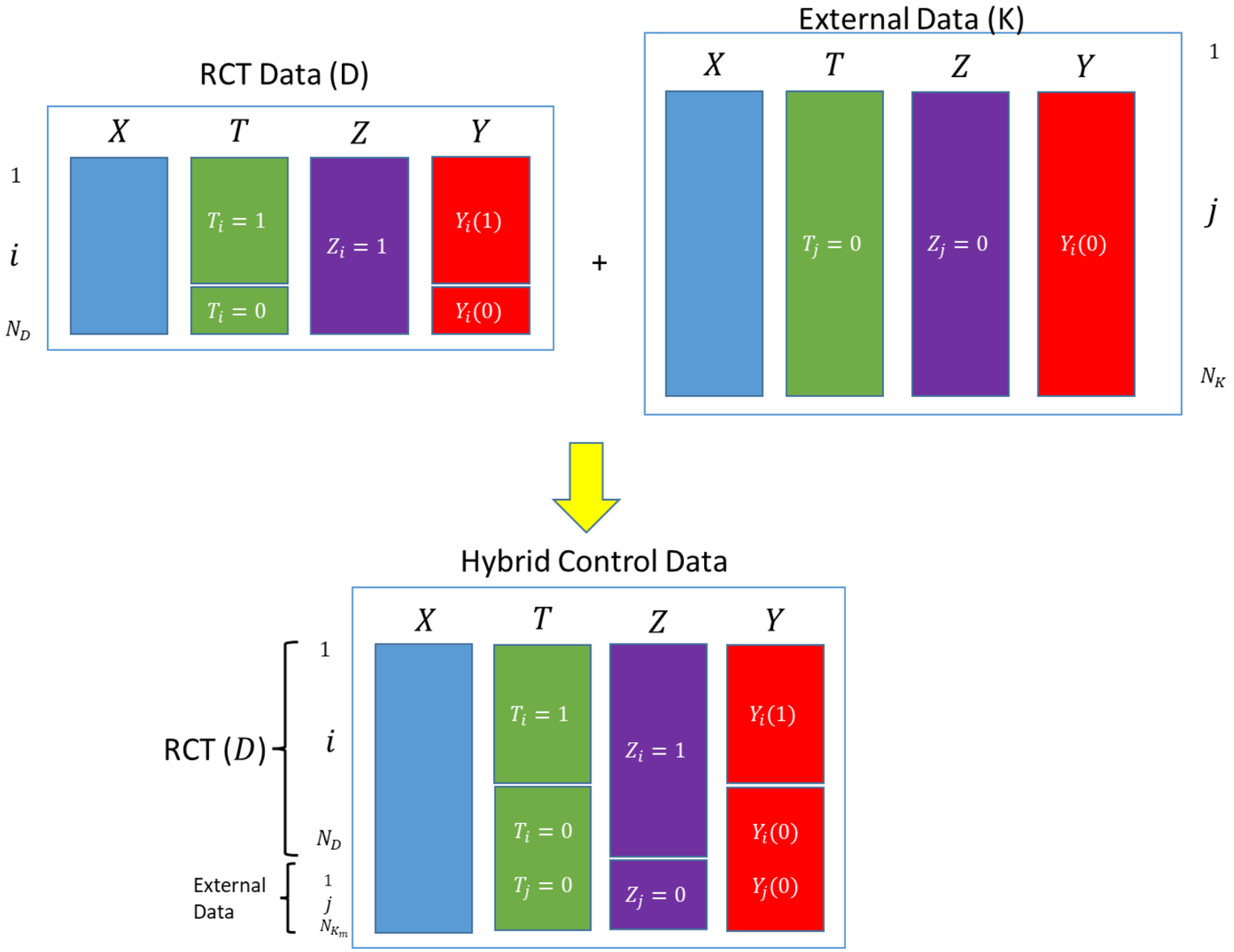
- Why consider external controls
  - Randomization to a control group is unethical or unfeasible (e.g., limited or no effective treatments are available, rare diseases)
  - External controls can:
    - Accelerate drug development by eliminating or reducing number of patients on placebo, thereby reducing the duration and cost of the trial
    - Provide strong evidence for the efficacy and safety of a treatment

- Types of External Control Arms
  - Single Arm Study – trial enrolls patients only the active treatment, external subjects are selected to serve as a comparator control arm
  - Hybrid Control Arm – trial enrolls patients to a small concurrent control arm, which is then augmented using external subjects

Regulatory Concerns with External Controls Arms

- "Food and Drug Administration (2019) Rare Diseases: Natural History Studies for Drug Development," notes concerns with the use of external control groups:
  - Selection Bias: external data can differ from the RCT in baseline disease characteristics
  - Unmeasured Confounding: critical patient disease characteristics may not have been assessed or may have been assessed different.
  - Lack of Concurrency: RCT and external may have been conducted in different time frames or care settings
  - Measurement Error: data collection intervals and quality may lack consistency
  - Outcome Validity: outcomes in external data sources may be measured differently than in RCTs or they may not be well defined or reliable

Hybrid Control Data Structure



Notation:

D: current trial data (RCT)
K: external data

X: a vector of P covariates
T: treatment assignment (1 if treated, 0 if not treated. Note $T_j = 0 \; \forall \; j$)
Z: data source (1 if current trial, 0 if external data)
Y: observed outcome
$Y(0), Y(1)$: potential outcomes

Causal Inference Framework

- Target Causal Estimand: $E(Y(1) - Y(0)|D)$
- We focus on continuous outcomes generated from the linear model

$$Y = \beta X + \beta_T T + \varepsilon, \varepsilon \sim N(0, \sigma^2)$$

- Joint Distribution of Information

$$f(X, Y(0), Y(1), T, Z) = f(X) f(Y(0), Y(1)|X) f(T|X, Y(0), Y(1)) f(Z|T, Y(0), Y(1), X)$$

- Key Assumptions:
  - Stable Unit Treatment Value Assumption (SUTVA)

$$Y_i^{obs} = Y_i(1)T_i + Y_i(0)(1 - T_i)$$

  - Strong Ignorability of Treatment Assignment (Unconfoundedness)

$$f(T|X, Y(0), Y(1)) = f(T|X)$$

  - Strong Data Source Ignorability

$$f(Z|X, Y(0), Y(1), T) = f(Z|X, T)$$

Selection of External Data Source

- Researchers should make significant efforts to select high quality, "fit for purpose" data that minimizes the risk for potential bias when selecting external control cohorts.
  - External control group should be similar to target trial population in all aspects that can impact outcomes.
  - Similar inclusion and exclusion criteria should be applied to the external controls to satisfy the positivity assumption.
  - Select data sources with similar index dates, duration and frequency of follow-up, and definition of outcomes.
  - Ensure important confounding variables are available and recorded consistently, or link to other data sources to obtain this information.
- Extensive guidance is provided in draft guidance from FDA: Considerations for the Use of Real-World Data and Real World Evidence to Support Regulatory Decision-Making for Drug and Biological Products (Dec. 2021).
- Industry guidance for statistical methods have not yet been released.

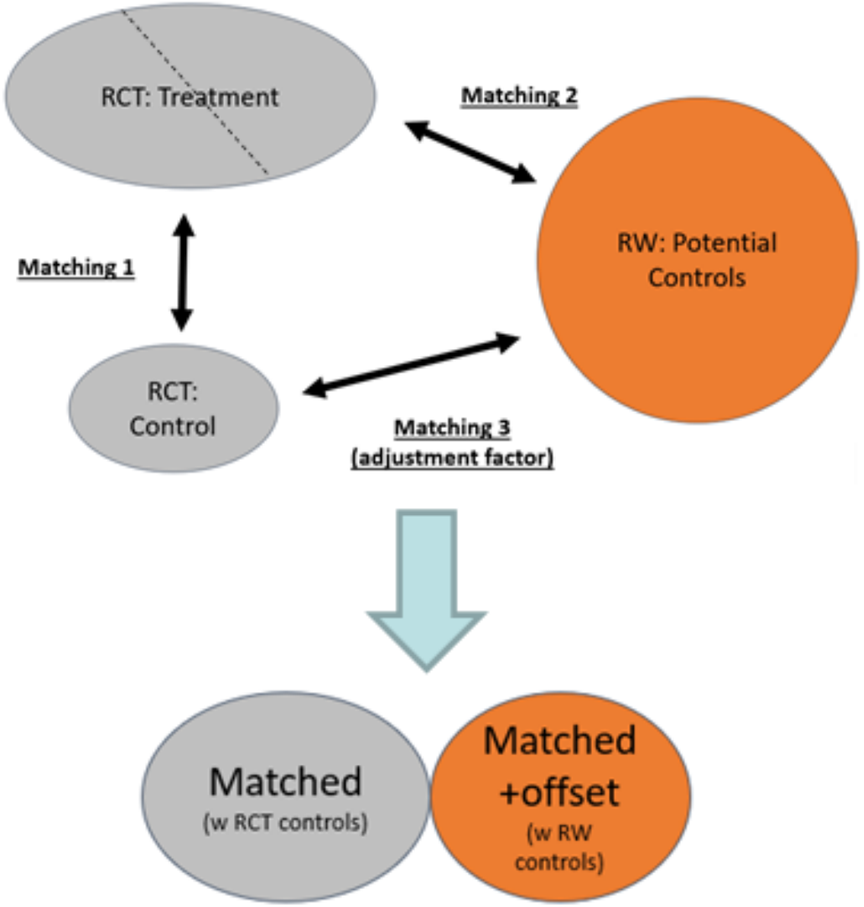
Methods for Fixed External Borrowing

- Frequentist
  - Matching, weighting (IPTW or entropy balancing)
  - Test-then-pool (Viele et. al. 2012)
  - Matching and Bias Adjustment (Stuart and Rubin 2008)
- Bayesian –most methods can incorporate matching prior to posterior estimation
  - Pocock's Method (Pocock (1976)
  - Power Prior
  - Meta-analytic Predictive prior
  - Commensurate prior
  - Mixture Prior

Test then Pool (Viele et. al. (2012))

- A frequentist hypothesis test is conducted to determine whether there is a difference in distribution between concurrent and external controls.
- For instance, under the model $Y(0)|\beta, \delta_0, \sigma^2 \sim N(\beta X + \delta_0 Z, \sigma^2)$, a test of $H_0: \delta_0 = 0$ vs $H_1: \delta_0 \neq 0$ can be performed.
- If the null hypothesis of equality is not rejected, the external control subjects are all pooled with the current trial subjects for analysis.
- If the null hypothesis is rejected, the treatment effect is estimated using the current trial data only.
- Viele et. al. (2012) notes the size of the test can be increased to control the potential type I error inflation from a pooled analysis.

Matching and Bias Adjustment



- Three matched sets are constructed:
  1. RCT treated : RCT control
  2. Unmatched RCT treated : RW control
  3. RCT control : RW control
- A data-source level bias adjustment term δ_0 is estimated from the Bayesian model $Y(0)|\beta, \delta_0, \sigma^2 \sim N(\beta X + \delta_0 Z, \sigma^2)$
- Posterior estimates of δ_0 are drawn and added to the observed outcome for each external control subject
- Rubin's rules for multiple imputation are used to construct point and interval estimates for β_T

Pocock's Method

- Assumes external controls are potentially biased from concurrent controls, with bias represented as $\delta = \theta_{CD} - \theta_{CK}$.
- The parameter(s) for K are subsequently modeled by $\theta_{CD} - \delta$.
- δ is typically given a $N(0, \sigma_\delta^2)$ prior.
- Pocock's posterior takes the form

$$p\left(\theta_{TD}, \theta_{CD}, \theta_{CK}, \delta \middle| D, K, \sigma_\delta^2\right) \propto p\left(\delta \middle| \sigma_\delta^2\right) L(\theta_{TD}, \theta_{CD} | D) L(\theta_{CD} - \delta | K)$$

Power Prior

Power Prior: **Discounts** each RW control patient by a factor of $\alpha, 0 \leq \alpha \leq 1$. Increasing α increases the influence of RWD in the estimation of the treatment effect.

RWD Likelihood

$$\underbrace{p(\theta|RCT, RW, \alpha)}_{\text{Power prior}} \propto p(\theta) \underbrace{L(\theta|RCT)}_{\text{RCT Likelihood}} L(\theta|RW)^{\alpha}$$

Power prior
Posterior for
treatment effect

Meta-analytic Predictive Prior

- The outcome parameters between the RCT and external data are assumed to be exchangeable and can be represented by the following hierarchical models:

$$\mu_D = \beta X + \beta_T T + \eta_D$$
$$\mu_K = \beta X + \beta_T T + \eta_K$$
$$\eta_D, \eta_K \sim N(0, \tau^2)$$

- η_D and η_K denote trial-level error terms that adjust for heterogeneity between data sources. τ^2 represents the between data source variance and controls the amount of borrowing from external data.

Simulation Evaluation for Fixed Borrowing Methods (Shan et. al. 2022)

- When the strong data source ignorability assumption is valid, the single external control arm and all hybrid control arm methods provide unbiased estimates of the treatment effect. Single arm trials are the most efficient design and attain power and type I error rates similar to 1:1 RCTs.
- When the strong data source ignorability assumption is violated, a single external control arm and hybrid control arm methods with fixed amount of borrowing (power prior, test-then pool) that do not incorporate a bias adjustment term are biased and may produce incorrect inference.
- Matching and bias adjustment (Stuart and Rubin 2008) produces unbiased estimates under different violations of the strong data source ignorability assumption and maintains close to nominal type I error rates. However, it produces larger standard errors compared to partial RCT only analysis, which results in higher MSE and lower power.
- MAP priors also produce unbiased estimates under different violations of the strong data source ignorability assumption. MAP priors have lower standard error and higher power compared to the partial RCT analysis, but have inflated type I error rates.

Propensity Score Integrated Bayesian Methods

Several Bayesian methods to incorporate subject-level external information with the propensity score used to assess or mitigate data heterogeneity have been proposed under two popular classes of priors: power prior and meta-analytic predictive prior

1. Power Prior: Discounts the contribution to the likelihood of each external control subject by a factor of α (power parameter), $0 \leq \alpha \leq 1$.
   - Subject-specific propensity score weights as power parameter (Lin et. al. 2018)
   - Propensity score stratification with stratum-specific power parameter (Wang et.al. 2019)
   - Propensity score matching with subject-specific posterior predictive p-value as power parameter (Kwiatkowski et. al 2022, forthcoming)

2. Meta-Analytic Predictive Prior: A hierarchical random effects meta-analysis model is used to account for between-data heterogeneity and discount the external data
   - Propensity score stratification with stratum-specific MAP-priors (Liu et. al. 2021)

Power Prior with Subject-specific PS weights

- Propensity Score Weighting Power Prior Parameter (Lin et.al. 2018)
  - The propensity score is estimated as probability of trial inclusion, given covariates ($\hat{e}(X) = \Pr(Z = 1|X = x)$). Matching is performed using $\hat{e}(X)$ to select a set of K_m external control subjects such that 1:1 randomization is achieved with hybrid control arm
  - The estimated $\hat{e}(X)$ is used as subject specific power parameter for external subjects such that $\alpha_j = \hat{e}(X_j)$
  - The posterior distribution takes the form:

$$p(\theta|\boldsymbol{\alpha}, D, K_m) \propto p(\theta)L(\theta|D) \prod_{j \in Km} f\left(Y_j \middle| T_j, X_j\right)^{\alpha_j}$$

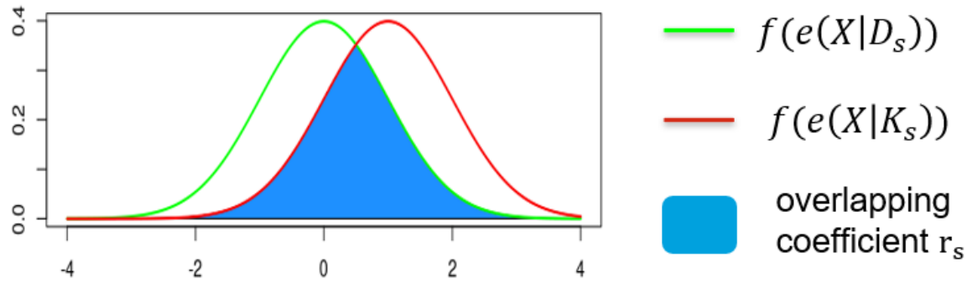
- Normalized Propensity Score Weighting Power Prior Parameter
  - The use of $\hat{e}(X_j)$ may be unstable and may not always lead to higher borrowing when covariate distributions are homogeneous.
  - We propose a normalized power prior weight

$$\alpha_j = \frac{\hat{e}(X_j)/(1 - \hat{e}(X_j))}{\max_{i \in D}(\hat{e}(X_i)/(1 - \hat{e}(X_i)))}$$

where $\hat{e}(X_i)$ is the estimated propensity score among RCT treated subjects.

Power Prior with PS stratum-specific weights (Wang et.al. 2019)

1. Estimate the propensity score ($\hat{e}(X)$) and use $\hat{e}(X|D)$ to create $s = 1, \ldots, S$ strata.
2. An overlapping coefficient $\mathrm{r_s}$ is calculated between the PS score densities for $\mathrm{D_s}$ and K_s partitioned to stratum s.



3. The stratum-level power prior weight is $\alpha_s = \min\left(\frac{Av_s}{n_{K,s}}, 1\right)$ where $\mathrm{v_s}=\frac{\mathrm{r_s}}{\sum_s \mathrm{r_s}}$, $A = \mathrm{N_{D_T}} - \mathrm{N_{D_C}}$, and $n_{K,s}$ is the number of subjects in K_s
4. The PS stratification integrated power prior posterior is expressed as

$$p(\theta|\boldsymbol{\alpha}, D, K) \propto \prod_{s=1}^{S} p(\theta_s) L(\theta_s|D_s) L(\theta_s|K_s)^{\alpha_s}$$

5. Stratum-specific parameters are estimated and a pooled estimate of θ is obtained for inference.

Power Prior with Posterior Predictive P-value weights (Kwiatowski et. al. (2022, forthcoming))

- Subject-specific weights will be assigned to external controls based on heterogeneity between outcome distributions rather than covariate distributions. The posterior predictive p-value captures the probability of obtaining a value of $Y_K^{pred}|D$ at least as extreme as Y_K^{obs}.
- PS matching (similar procedure as in Lin et. al. (2018) can be performed to select subjects with balanced covariate distributions
- The Posterior Predictive P-value weights are then calculated as follows:
  1. Fit the posterior distribution of $p(\theta|D) \propto p(\theta)L(\theta|D)$ and sample values of θ
  2. Given each value of θ, sample Y_K^{pred} from the posterior predictive distribution $p\left(Y_K^{pred}\middle|X_K,\theta\right)$ and calculate $p\left(Y_K^{pred}\middle|X_K,\theta\right)$ and $p\left(Y_K^{obs}\middle|X_K,\theta\right)$
  3. Calculate Posterior Predictive p-value: $PPP = \int \Pr\left(p\left(Y_K^{pred}\middle|X_K,\theta\right) < p\left(Y_K^{obs}\middle|X_K,\theta\right)\right) p(\theta|D)d\theta$
  4. Apply subject specific Posterior Predictive p-value as power parameter for external subjects: $\alpha_j = PPP_j$

Meta-analytic predictive prior with PS stratification (Liu et. al. 2021)

1. Estimate the propensity score and use $\hat{e}(X|D)$ to create $s = 1, \ldots, S$ strata. Overlapping coefficients $R = (r_1, .., r_S)$ are calculated between the PS score densities for D_s and K_s in stratum s.
2. Mean exchangeability is assumed within strata and heterogeneity between strata is controlled by random effect terms η :

$$Y_{s,D} \sim N\left(\boldsymbol{X_{D_s}\beta_s} + \beta_{T_s} T_{D_s} + \eta_{s,D}, \sigma^2_{s,D}\right)$$
$$Y_{s,K} \sim N\left(\boldsymbol{X_{K_s}\beta_s} + \eta_{s,K}, \sigma^2_{s,K}\right)$$
$$\eta_{s,D}, \eta_{s,K} \sim N(0, \tau_s^2)$$

   where $\tau_s^2 \sim$ half−normal($k_s *$ t), $k_s = \frac{med(R)}{r_s}$, and t reflects a baseline variance. A calibration procedure for t based on target ESS was proposed to obtain the desired amount of borrowing.
3. The PS-MAP posterior distribution:

$$p(\theta|D, K, R) \propto \prod_{s=1}^{S} p\left(\boldsymbol{\beta_s}, \beta_{T_s}\right) p(\sigma_s^2) p(\eta_s, \tau_s) L(\theta_s|D_s) L(\theta_s|K_s)$$

4. Strata specific parameters are estimated and pooled together

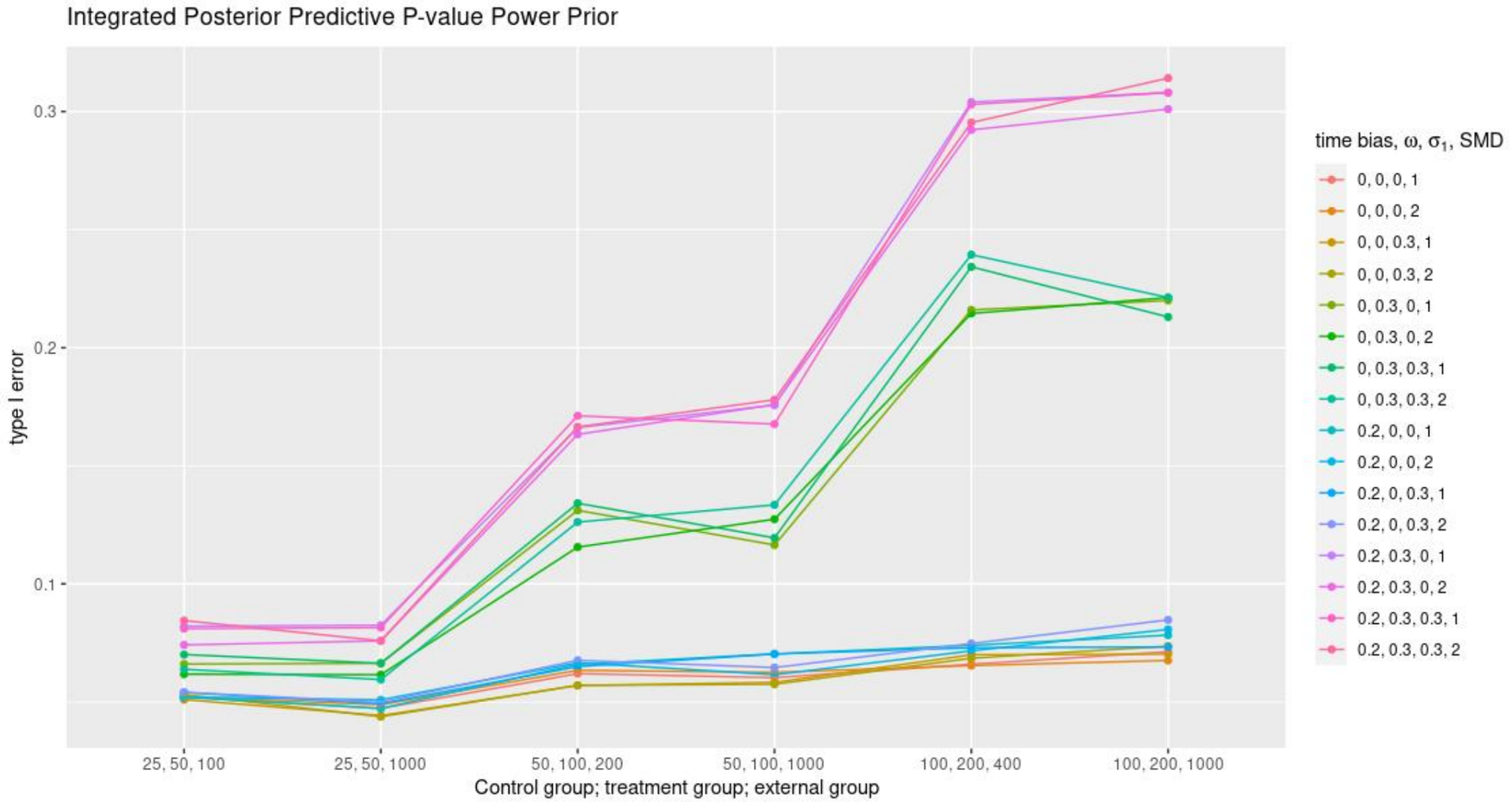
Simulation parameters

$Y_i = \boldsymbol{X_i}\beta + \beta_T T_i + \omega U_i + \varepsilon_i$
$Y_j = \boldsymbol{V}(\boldsymbol{X_j}\beta + \omega U_j) + \delta_0 + \delta_1 Y_{baseline_j} + \varepsilon_j$
Treatment effect: $\beta_T = (0, 0.3)$

RCT control sample size	$N_{D_C} = (25, 50, 100)$
RCT treated sample size	$N_{D_T} = 2 * N_{D_C}$
External control sample size	$N_{K_C} = (2 * N_{D_T}, 1000)$
Selection bias	$SMD = (obs, 2 * obs)$
Concurrency bias	$\delta_0 = (0, 0.2)$
Measurement Error	$\delta_1 = (0, 0.3)$
Unmeasured Confounding	$U_i \sim N(0,1), U_j \sim N(0.5, 1), cor(U_j, \boldsymbol{X_j}) = 0.3,$ $\omega = (0, 0.3)$
Outcome Validity	$\boldsymbol{V}(\boldsymbol{X_j}\beta + \omega U_j) = \{\boldsymbol{X_j}\beta + \omega U_j, (\boldsymbol{X_j}\beta + \omega U_j)^2\}$

* $\boldsymbol{X_i}$ were simulated according to oncology clinical trial data
* $\boldsymbol{X_j}$ were simulated according to the Flatiron Health Spotlight data

Type I Error Inflation

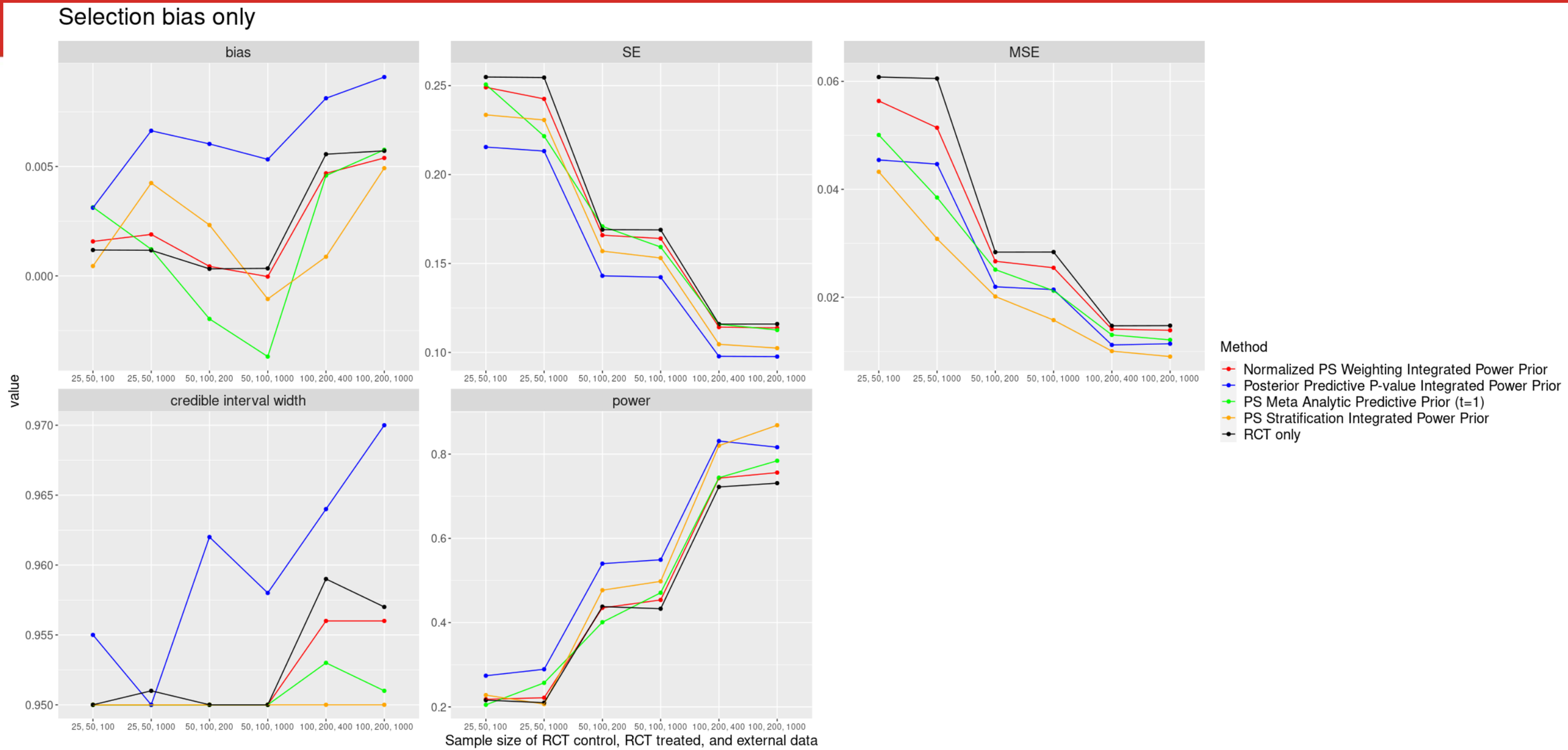


- In the presence of certain types of bias, type I error is not controlled at 0.05
- The width of the credible interval for each method under each bias scenario and sample size configuration was calibrated to control type I error to enable valid power comparisons between methods.

Type I Error Calibration Procedure

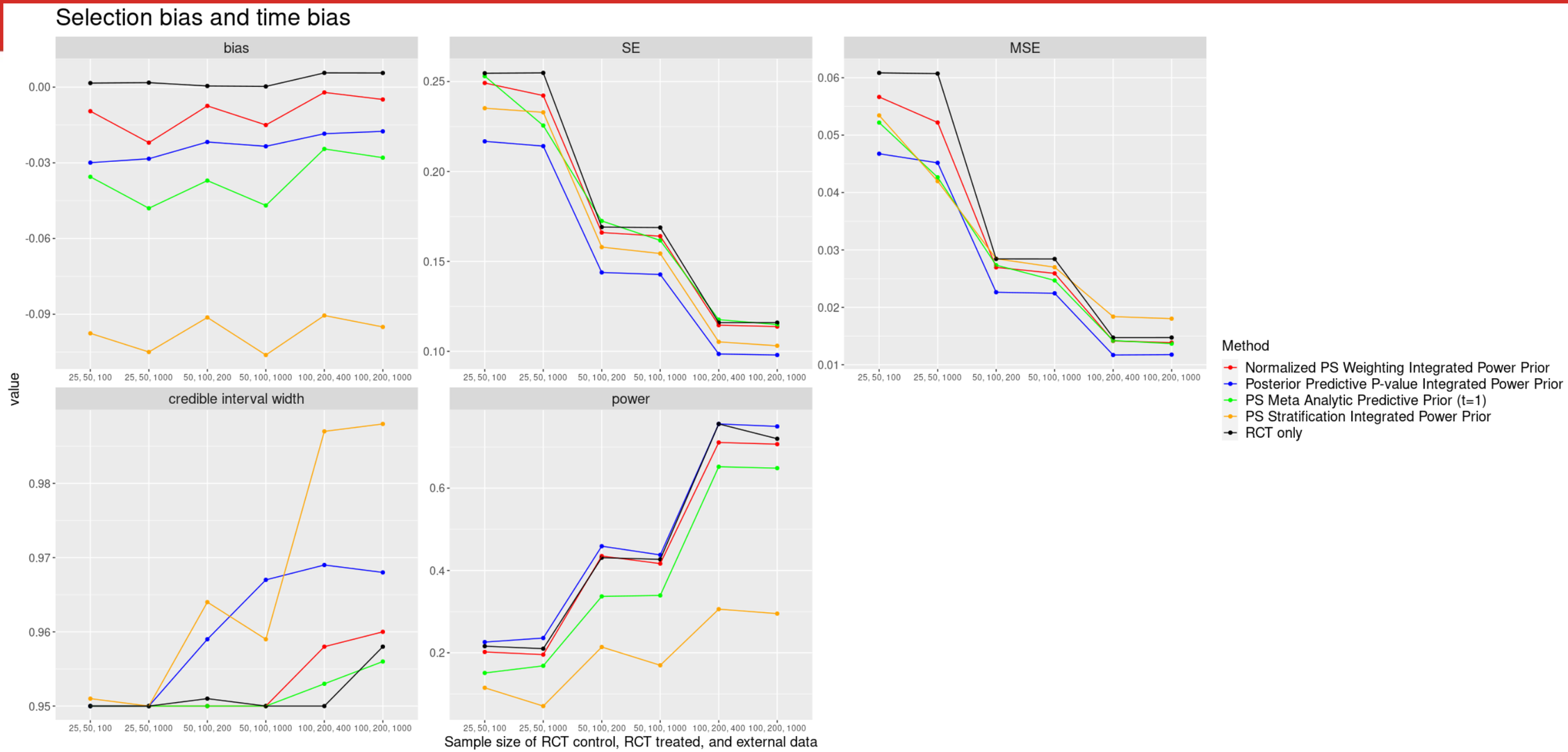
1. Simulate data under Null hypothesis, i.e. no treatment effect
2. Run Bayesian model and record the empirical 95% credible interval for β_T
3. Check whether null ($\mathrm{H}_0: \beta_T = 0$) is included in the credible interval for β_T
4. Repeat above steps to get the empirical type I error rate.
5. If this empirical type I error is **larger** than 5% (Type I error inflation), then increase width of credible interval by 0.01% and repeat above steps until type I error is controlled at 5%. If this empirical type I error is **smaller** than 5% (Type I error deflation), then keep using 95% credible interval

Simulation Results: Scenario with Selection Bias Only



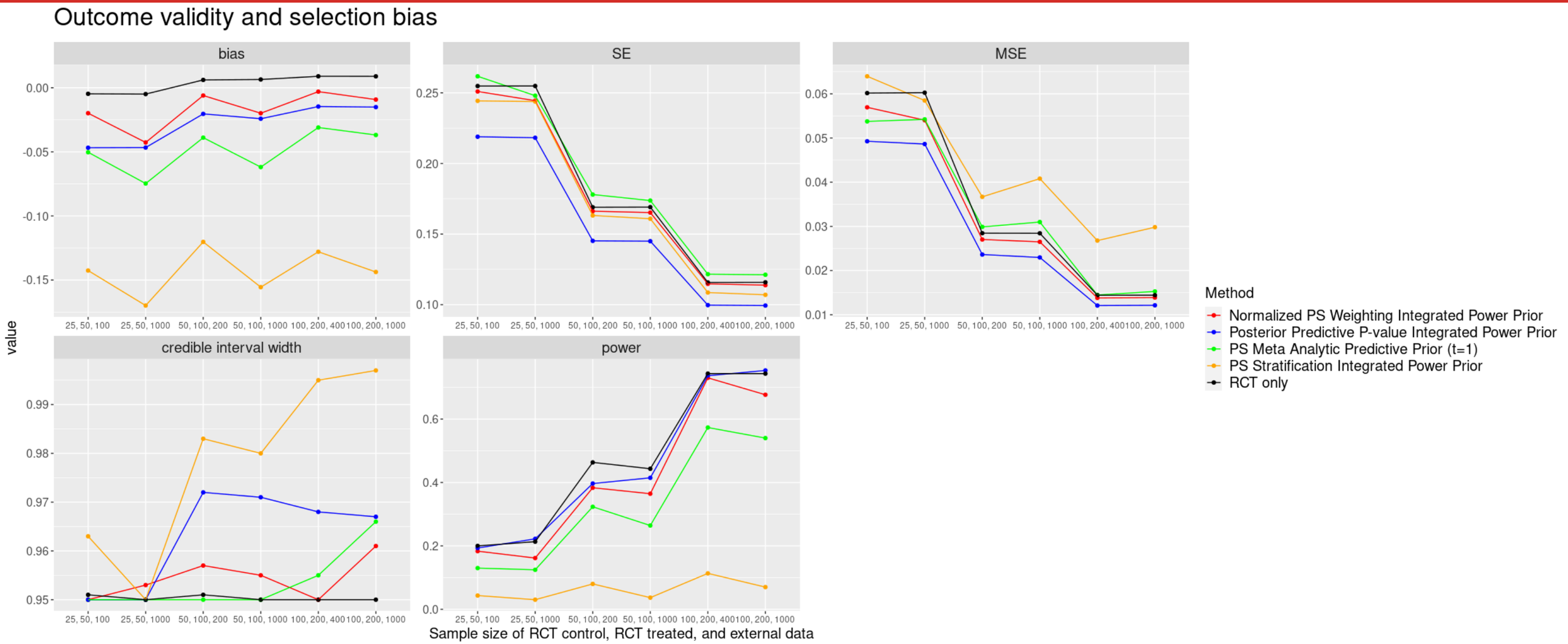
- Small bias across all methods and all methods have greater power than partial RCT only
- PPP (blue) had type I error inflation, but still resulted in highest power with adjusted credible intervals
- Normalized PS weighting (red) had slightly better power than partial RCT only (black)

Simulation Results: Scenario with Concurrency Bias and Selection Bias



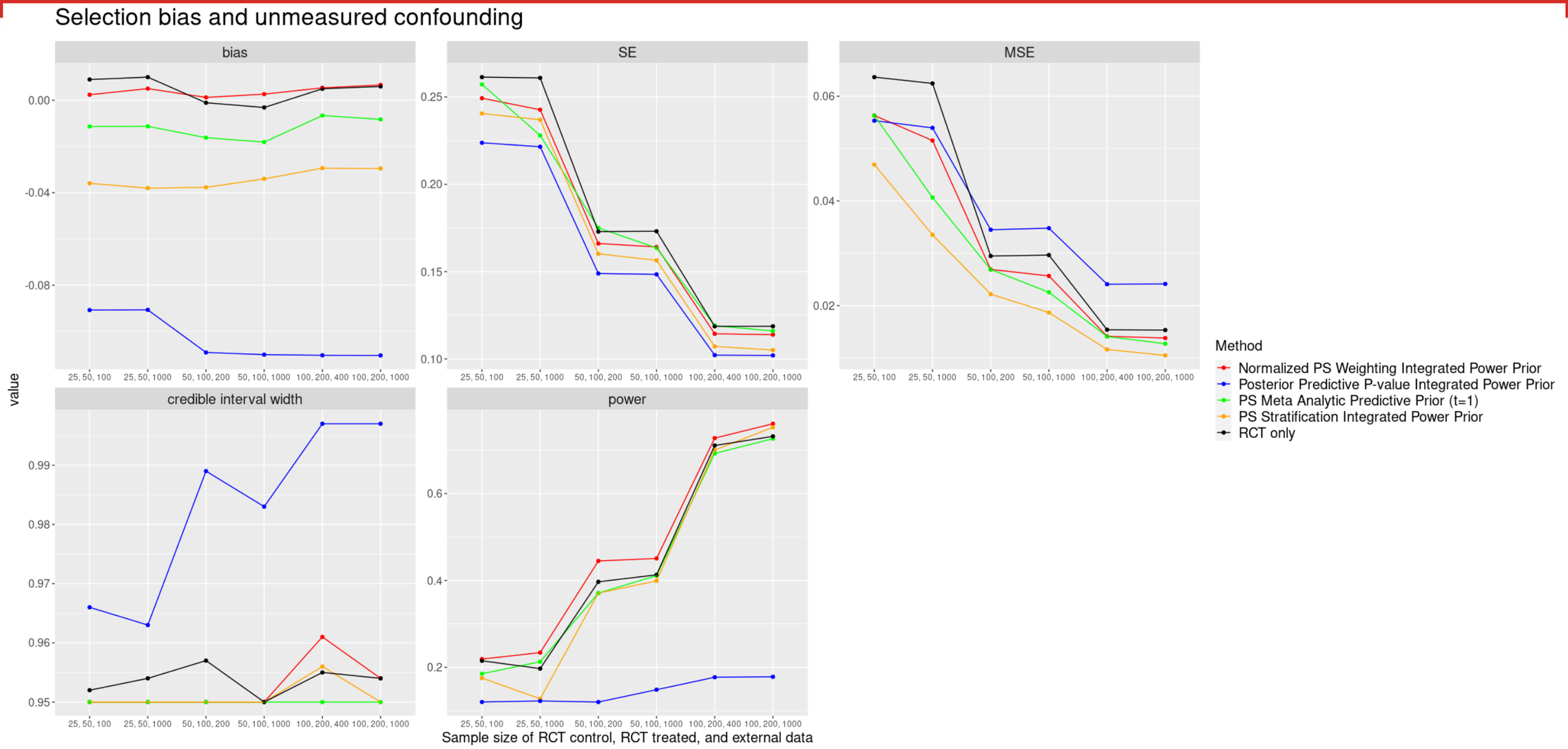
- PPP (blue) and Normalized PS weighting (red) have small bias but significantly lower SE and MSE than partial RCT only.
- Type I error inflation was seen for PPP, but still resulted in the highest power after type I error calibration
- Normalized PS weighting had slight type I error inflation at higher sample sizes,

Simulation Results: Scenario with Outcome Validity and Selection Bias



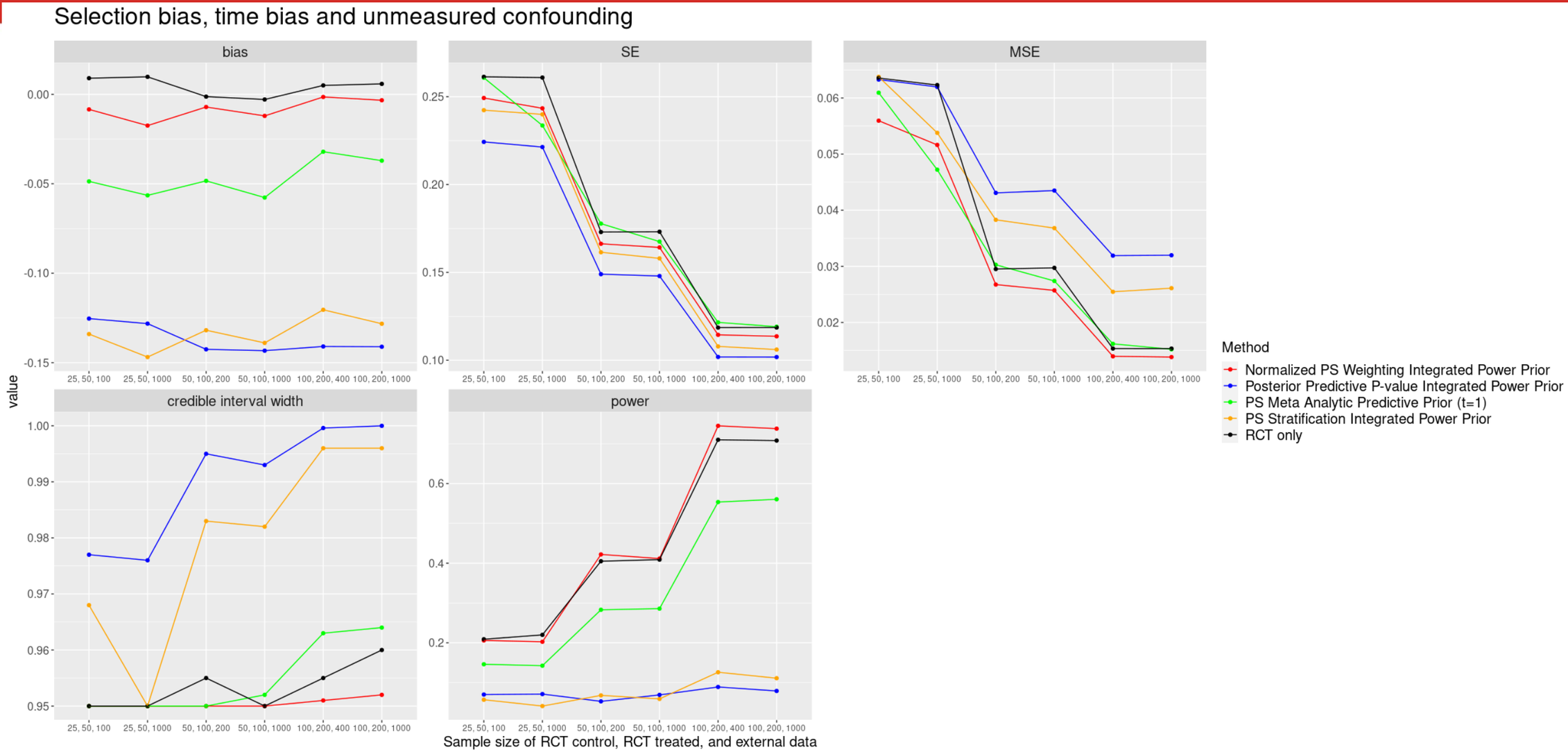
- PPP (blue) and Normalized PS weighting (red) have small bias but significant lower SE and MSE than partial RCT only.
- PPP has slight type I error inflation, which resulted in similar calibrated power as partial RCT only.
- Normalized PS weighting power prior had lower calibrated power than hybrid RCT only.

Simulation Results: Scenario with Unmeasured Confounding and Selection Bias



- PPP (blue) is biased, has severe type I error inflation, and resulted in poor calibrated power.
- Normalized PS weighting was unbiased and had higher power over partial RCT only.

Simulation Results: Scenario with Time bias, Unmeasured Confounding, and Selection Bias



- Normalized PS weighting (red) is the only method with small bias (conservative borrowing). Has similar or slightly improved calibrated power over partial RCT only.
- PPP (blue) has large bias and type I error inflation.

Conclusions

- In general, borrowing from external controls offer the potential to increase power and trial efficiency while still maintaining robust inference on the treatment effect.
- Normalized PS weighting power prior is a conservative borrowing method that minimizes potential bias in most scenarios compared to other methods but will also have the least potential gain in power.
- PPP performs well (low bias, improvements in power) in scenarios where there is no unmeasured confounding. In the presence of unmeasured confounders, PPP is unable to identify the violation of exchangeability and can lead to large bias.
- Stratification (PS integrated MAP, PS stratification power prior) should be used cautiously in small sample sizes. Operating characteristics seem to be sensitive to the number of strata and parameters being estimated.

References

- Ibrahim JG and Chen MH. Power prior distributions for regression models. Stat Sci 2000; 15: 46–60.
- Lin J, Gamalo-Siebers M, Tiwari R. Propensity-score-based priors for Bayesian augmented control design. Pharm Stat. 2019 Mar;18(2):223-238. doi: 10.1002/pst.1918. Epub 2018 Dec 9. PMID: 30537087.
- Liu, M, Bunn, V, Hupf, B, Lin, J, Lin, J. Propensity-score-based meta-analytic predictive prior for incorporating real-world and historical data. *Statistics in Medicine*. 2021; 40: 4794– 4808. https://doi.org/10.1002/sim.9095
- Pocock SJ. The combination of randomized and historical controls in clinical trials. J Chron Dis 1976; 29: 175–188
- Schmidli H, Gsteiger S, Roychoudhury S, et al. Robust meta-analytic-predictive priors in clinical trials with historical control information. Biometrics 2014; 70: 1023–1032.
- Shan, M., Faries, D., Dang, A. *et al.* A Simulation-Based Evaluation of Statistical Methods for Hybrid Real-World Control Arms in Clinical Trials. *Stat Biosci* **14,** 259–284 (2022). https://doi.org/10.1007/s12561-022-09334-w
- Viele K, Berry S, Neuenschwander B, et al. Use of historical control data for assessing treatment effects in clinical trials. Pharm Stat 2014; 13: 41–54
- Wang C, Li H, Chen W-C, et al. Propensity score-integrated power prior approach for incorporating real-world evidence in single-arm clinical studies. *J Biopharm Stat*. 2019;**29**(5):731-748.

Alternative Data Integration Frameworks

- Record Linkage – Covariates, treatment assignment, and outcomes for each subject are separated across two or more data sources. Causal Inference requires first linking the two data sources together by identifying subjects that overlap in both data sources.

- Data Generalizability – The observed study sample is a sample from a larger/broader target population. Interest is to generalize treatment effect estimates from the observed sample to the larger target population.

- Data Transportability – The target population is only partially overlapping or potentially non-overlapping with the observed sample.

Causal Inference and Record Linkage

- Assume the observed data are split across two files A and B, with n_A and n_B records, respectively. Let C represent a $n_A \times n_B$ matrix indicating the linkage structure between records.
- An additional necessary assumption for causal inference is that the linkage decisions are strongly non-informative. That is, the linkage decisions are conditionally independent of the potential outcomes.

$$f(C|X, Z, Y(0), Y(1)) = f(C|X, Z).$$

- This allows Bayesian inference of treatment effect to be expressed as

$$f\left(\tau\middle|X, Z, Y^{obs}\right) = \int f\left(\tau\middle|Y^{obs}, Y^{mis}, C\right)p(C|X, Z, \theta_C)f\left(Y(0)^{mis}\middle|X, \theta_{Y0}\right)p\left(\theta_{Y0}\middle|X, Y(0)^{obs}\right)$$
$$f\left(Y(1)^{mis}\middle|X, \theta_{Y1}\right)p\left(\theta_{Y1}\middle|X, Y(1)^{obs}\right)d\theta_C d\theta_{Y0} d\theta_{Y1} dC dY(0)^{mis} dY(1)^{mis}$$

Bayesian Record Linkage and Causal Inference Methods

1. Scenario 1: File A contains covariate information and outcome information. File B contains treatment assignment. Linkage informs treatment for the records in file A.
   - Mingyang Shan, Kali S. Thomas, Roee Gutman. "A multiple imputation procedure for record linkage and causal inference to estimate the effects of home-delivered meals." The Annals of Applied Statistics, 15(1) 412-436 March 2021.
2. Scenario 2: File A contains covariate and treatment assignment information. File B contains outcome information. Linkage informs outcomes for the records in file A.
   - Wortman JH, Reiter JP. Simultaneous record linkage and causal inference with propensity score subclassification. Stat Med. 2018 Oct 30;37(24):3533-3546. doi: 10.1002/sim.7911. Epub 2018 Aug 1.
   - Sharmistha Guha, Jerome P. Reiter, Andrea Mercatanti. "Bayesian Causal Inference with Bipartite Record Linkage." Bayesian Analysis, Advance Publication 1-25 2022. https://doi.org/10.1214/21-BA1297
3. Scenario 3: File A contains a subset of covariates, the treatment assignment, and outcome information. File B contains additional covariate information. Linkage would reduce the risk of potential unmeasured confounding for records in file A.
   - N/A ☹

Two Stage Multiple Imputation Procedure for Causal Inference on Record Linked Data