Supplementary materials for "Properties of Inverse Probability of Adherence Weighted Estimator of the Per-protocol Effect for Sustained Treatment Strategies under Different Data-generating Mechanisms and Adherence Patterns"

A Performance Measures in Simulations

This section contains the statistical properties used to evaluate each effect estimator as part of this simulation. These statistical properties have been calculated for each estimator. Each property includes a brief description as well as the formula for the estimate and the standard error for the estimate (if applicable).

For this section let $\hat{\theta}_i$ denote the estimated treatment effect for iteration i of the simulation, where $i \in 1, ..., n_{sim}$. The true treatment effect is denoted θ, and the average of the estimated treatment effect across all iterations is $\bar{\theta}$. Additionally, let p_i denote the p-value associated with the coefficient for the treatment effect at iteration i, and $\hat{\theta}_{high,i}, \hat{\theta}_{low,i}$ denote the upper and lower 95% confidence intervals for the estimated treatment effect for iteration i. Finally, let $\widehat{\text{Var}}(\hat{\theta}_i)$ be the estimated variance associated with the effect estimate $\hat{\theta}_i$. In these definitions we also use $\mathbf{I}$ to denote the indicator function. The indicator function returns a value of 1 if the input condition is true, and 0 otherwise.

Convergence: this denotes what proportion of iterations of the simulation yielded valid treatment effect estimates.

$$\frac{1}{n_{sim}} \sum_{i=1}^{n_{sim}} \mathbf{I}(\theta_i \in \mathbf{R})$$

Bias: quantifies how far each treatment effect estimates is from the true treatment effect.

$$\frac{1}{n_{sim}} \sum_{i=1}^{n_{sim}} \hat{\theta}_i - \theta \qquad \sqrt{\frac{1}{n_{sim}(n_{sim}-1)} \sum_{i=1}^{n_{sim}} \left(\hat{\theta}_i - \bar{\theta}\right)^2}$$

Coverage Probability: quantifies what proportion of confidence intervals include the true treatment effect. If an estimate is unbiased, the coverage probability should be equal to the confidence level for the calculated intervals.

$$\frac{1}{n_{sim}} \sum_{i=1}^{n_{sim}} \mathbf{I}(\hat{\theta}_{low,i} \le \theta \le \hat{\theta}_{high,i}) \qquad \sqrt{\frac{\text{Cover.} \times (1 - \text{Cover.})}{n_{sim}}}$$

Bias-Adjusted Coverage Probability: quantifies what proportion of confidence intervals include the average estimated treatment effect. The coverage probability should be equal to the confidence level for the calculated intervals, regardless of whether the estimator is unbiased.

$$\frac{1}{n_{sim}} \sum_{i=1}^{n_{sim}} \mathbf{I}(\hat{\theta}_{low,i} \le \theta \le \hat{\theta}_{high,i}) \qquad \sqrt{\frac{\text{B.A. Cover.} \times (1 - \text{B.A. Cover.})}{n_{sim}}}$$

Empirical Standard Error (EmpSE): quantifies the standard error between the treatment effect estimates and

the average estimated treatment effect across iterations.

$$\sqrt{\frac{1}{n_{sim}-1}\sum_{i=1}^{n_{sim}}(\hat{\theta}_i-\bar{\theta})^2} \qquad\qquad \frac{\widehat{EmpSE}}{\sqrt{2(n_{sim}-1)}}$$

Average Model Standard Error (ModSE): averages the standard error estimated for each treatment effect estimate.

$$\sqrt{\frac{1}{n_{sim}}\sum_{i=1}^{n_{sim}}\widehat{\mathrm{Var}}(\hat{\theta}_i)} \qquad\qquad \sqrt{\frac{\widehat{\mathrm{Var}}[\widehat{\mathrm{Var}}(\hat{\theta}_i)]}{4n_{sim}\times\widehat{ModSE}^2}}$$

Mean Squared Error (MSE): quantifies the bias and variance of a given treatment effect simultaneously.

$$\frac{1}{n_{sim}}\sum_{i=1}^{n_{sim}}\left(\hat{\theta}_i-\theta\right)^2 \qquad\qquad \sqrt{\frac{\sum_{i=1}^{n_{sim}}\left[(\hat{\theta}_i-\theta)^2-\widehat{MSE}\right]^2}{n_{sim}\times(n_{sim}-1)}}$$

Power or Type I Error (Power): quantifies what proportion of iterations yielded a statistically significant treatment effect estimate for a given significance level α.

$$\frac{1}{n_{sim}}\sum_{i=1}^{n_{sim}}\mathbf{I}\left(p_i\leq\alpha\right) \qquad\qquad \sqrt{\frac{\mathrm{Power}\times(1-\mathrm{Power})}{n_{sim}}}$$

Confidence Interval Length: calculates the average confidence interval length for an estimator across all iterations of the simulation.

$$\frac{1}{n_{sim}}\sum_{i=1}^{n_{sim}}\hat{\theta}_{high,i}-\hat{\theta}_{low,i}$$

B   Cumulative survival type estimates

All cumulative survival type estimates calculated the log(OR) estimate of the treatment effect as:

$$\log\left(\frac{1-S_1(K)}{1-S_0(K)}\right).$$

Where $S_z(K)$ is the cumulative survival among arm z at time point K, which is defined as:

$$S_z(K)=\prod_{t=0}^{K}\frac{n_{tz}-\sum_{i=1}^{n_{tz}}y_{t,i}}{n_{tz}}.$$

C   Simulation Settings

Appendix Table C.1: Table of parameters used in the simulations. Settings that varied two parameters explored different combinations of the parameters simultaneously. The parameters in this table correspond to the data generation mechanism (DGM) as defined in equations (2.1)-(2.5): β parameters are from equations (2.1)-(2.2), α parameters are from equations (2.3)-(2.4), θ parameters are from equation (2.5).

Section	Setting	Parameters Varied	Range of Parameter Values Considered
Aim 1: Structure of DGM	Diagram 1 (i)	Vary θ_2	$\theta_2 \in \{$ -1.3, -1, -0.7, -0.3, 0, 0.3, 0.7, 1, 1.3$\}$
	Diagram 1 (ii)	Fix: $\beta_{12} = \beta_{13} = \beta_{24} = \beta_{25} = \beta_{26} = 0$, then vary θ_2	$\theta_2 \in \{$ -1.3, -1, -0.7, -0.3, 0, 0.3, 0.7, 1, 1.3$\}$
	Diagram 2 (i)	Fix: $\alpha_{41} = \alpha_{40} = 0$, then vary θ_2	$\theta_2 \in \{$ -1.3, -1, -0.7, -0.3, 0, 0.3, 0.7, 1, 1.3$\}$
	Diagram 2 (ii)	Fix: $\alpha_{41} = \alpha_{40} = 0$, $\beta_{12} = \beta_{13} = \beta_{24} = \beta_{25} = \beta_{26} = 0$, then vary θ_2	$\theta_2 \in \{$ -1.3, -1, -0.7, -0.3, 0, 0.3, 0.7, 1, 1.3$\}$
	Diagram 3 (i)	Fix: $\alpha_{41} = \alpha_{40} = 0$, $\beta_{1-1} = \beta_{2-1} = 0$, then vary θ_2	$\theta_2 \in \{$ -1.3, -1, -0.7, -0.3, 0, 0.3, 0.7, 1, 1.3$\}$
	Diagram 3 (ii)	Fix: $\alpha_{41} = \alpha_{40} = 0$, $\beta_{1-1} = \beta_{2-1} = 0$, $\beta_{12} = \beta_{13} = \beta_{24} = \beta_{25} = \beta_{26} = 0$, then vary θ_2	$\theta_2 \in \{$ -1.3, -1, -0.7, -0.3, 0, 0.3, 0.7, 1, 1.3, 1.5$\}$
	Diagram 4 (i)	Fix: $\alpha_{41} = \alpha_{40} = 0$, $\beta_{1-1} = \beta_{2-1} = 0$, $\theta_1 = 0$, then vary θ_2	$\theta_2 \in \{$ -1.3, -1, -0.7, -0.3, 0, 0.3, 0.7, 1, 1.3$\}$
	Diagram 4 (ii)	Fix: $\alpha_{41} = \alpha_{40} = 0$, $\beta_{1-1} = \beta_{2-1} = 0$, $\theta_1 = 0$, $\beta_{12} = \beta_{13} = \beta_{24} = \beta_{25} = \beta_{26} = 0$, then vary θ_2	$\theta_2 \in \{$ -1.3, -1, -0.7, -0.3, 0, 0.3, 0.7, 1, 1.3, 1.5$\}$
Aim 2: Non-Adherence	Non-differential	Vary α_{01}, α_{00}	to fix proportion of protocol deviation and event rate via grid search: Adherence rates in both arms $\in \{0\%, 20\%, 40\%, 60\%, 80\%\}$
	Differential	Vary α_{01}, α_{00}	For each adherence rates in treated arm $\in \{0\%, 20\%, 40\%, 60\%, 80\%\}$, adherence rates in control arm $\in \{0\%, 20\%, 40\%, 60\%, 80\%\}$
	Measurement Schedule,	For each m	$m \in \{$ 1, 6, 12, 18, 24 $\}$,

Section	Setting	Parameters Varied	Range of Parameter Values Considered
	different combinations of A and L	Vary α_{01} and α_{00}	adherence rates in both arms $\in \{0\%, 20\%, 40\%, 60\%, 80\%\}$
Additional Simulations	Varying Treatment Effect and Trial Size	For each n,	$n \in \{200, 1000, 2000\}$
		vary θ_2	$\theta_2 \in \{ $ -1.3, -1, -0.7, -0.5, -0.3, 0, 0.3, 0.5, 0.7, 1, 1.3 $ \}$
	Event Rate	Vary θ_0	$\theta_0 \in \{$-19, -17.9, -17.3, -16, -16.8,-15, -14.6, -14.2, -13.8,-13.4, -12.8, -12.3, -11.8, -11.3, -10.8 $\}$
	Effect of Time On Y	Vary θ_5	$\theta_5 \in \{ $ 0, 0.01, 0.02, 0.03, 0.04, 0.05,0.075, 0.1 $ \}$
	Measurement Schedule,	For each m,	
	different combinations of A and L	Vary m	$m \in \{ $ 1, 3, 6, 9, 12, 15, 18, 21, 24, 27, 30 $ \}$

D Simulation Results
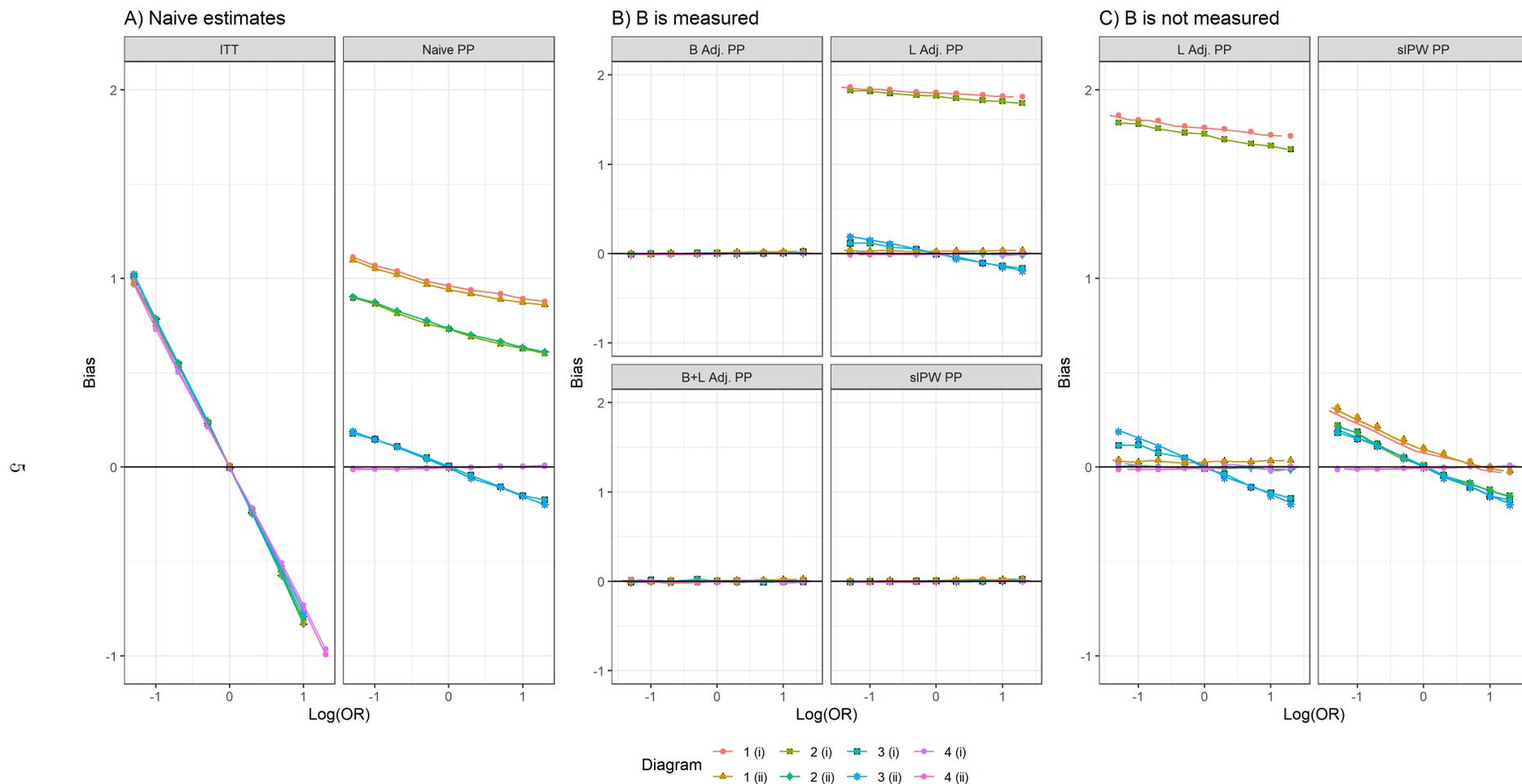
Appendix Figure D.1: A) Bias plotted against increasing treatment effect for two naive effect estimates: intention-to-treat and naive per-protocol. B) Five versions of per-protocol estimates (when B is measured, used during estimation). C) Two versions of per protocol estimates are presented (when B is not measured, not used during estimation). Different lines corresponds to different data generating mechanisms. A log(OR) of 0 corresponds to a null treatment effect. Abbreviations: ITT: intention to treat; Naive PP: Naive per-protocol; B Adj. PP: Baseline adjusted per-protocol; L Adj. PP: post-baseline prognostic factor adjusted-per-protocol; B+L Adj. PP: baseline and post-baseline prognostic factor adjusted-per-protocol; sIPW PP: stabilized inverse probability weighted per-protocol.
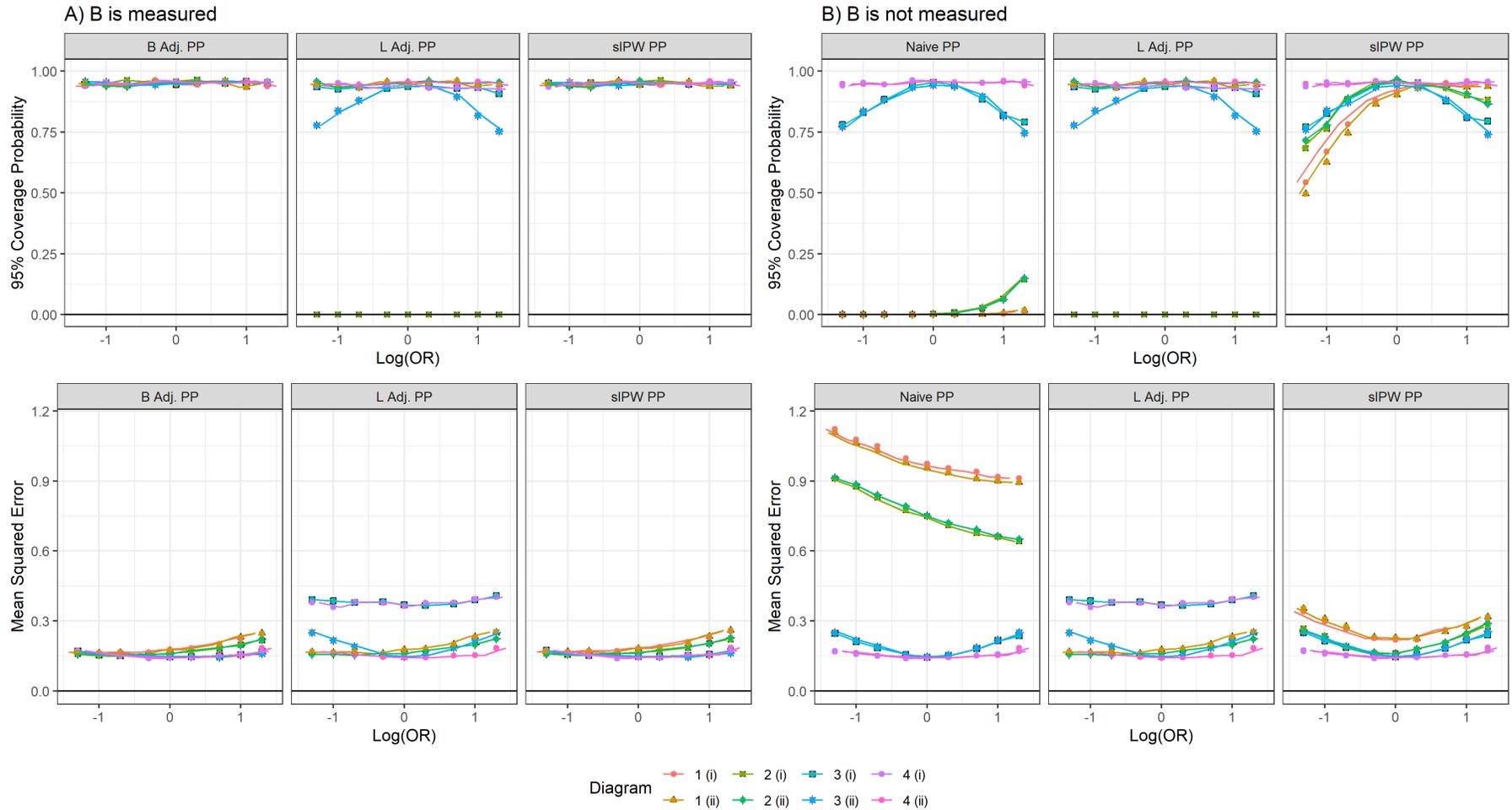
Appendix Figure D.2: A) 95 percent coverage probability (upper panel) and mean squared error (lower panel) plotted against increasing treatment effect for two versions of per-protocol estimates (when B is measured). B) Same estimates are presented (when B is not measured or adjusted). Different lines corresponds to different data generating mechanisms. A log(OR) of 0 correspond to a null treatment effect. Abbreviations: B Adj. PP: Baseline adjusted per-protocol; L Adj. PP: post-baseline prognostic factor adjusted-per-protocol; sIPW PP: stabilized inverse probability weighted per-protocol.
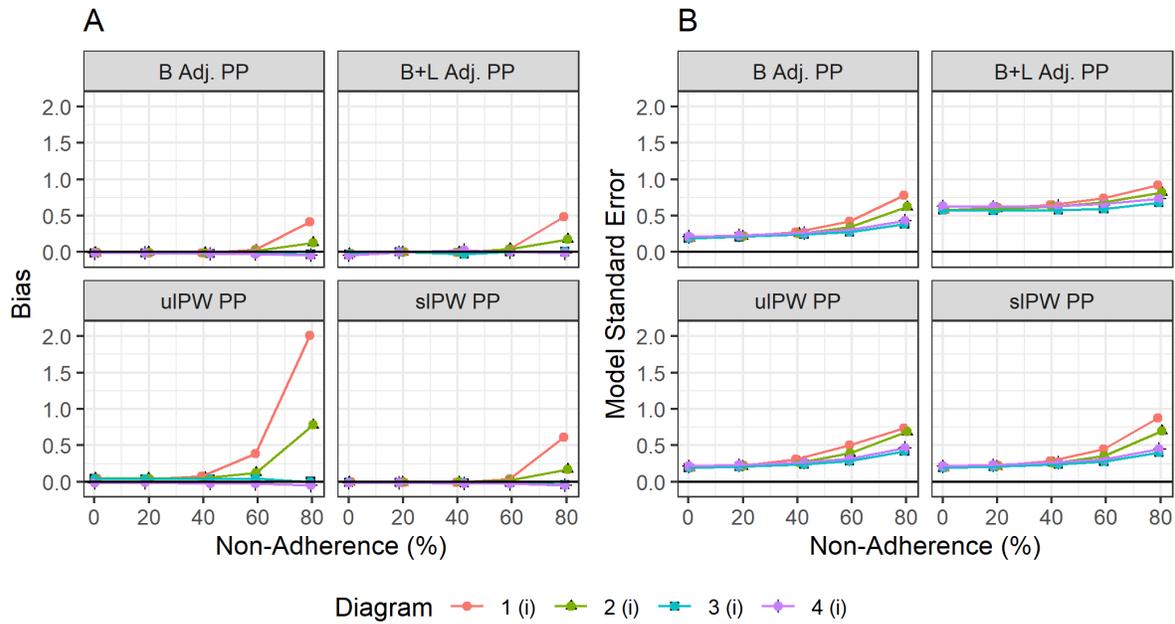
Appendix Figure D.3: Bias (plot A) and average model standard error (plot B) against non-adherence rates for each effect estimate. Non-adherence rates in both arms are the same in these simulations. We only considered situations where treatment received A impacts future post-baseline prognostic factors L. Abbreviations: B Adj. PP: Baseline adjusted per-protocol; B+L Adj. PP: baseline and post-baseline prognostic factor adjusted-per-protocol; uIPW PP: unstabilized inverse probability weighted per-protocol; sIPW PP: stabilized inverse probability weighted per-protocol.
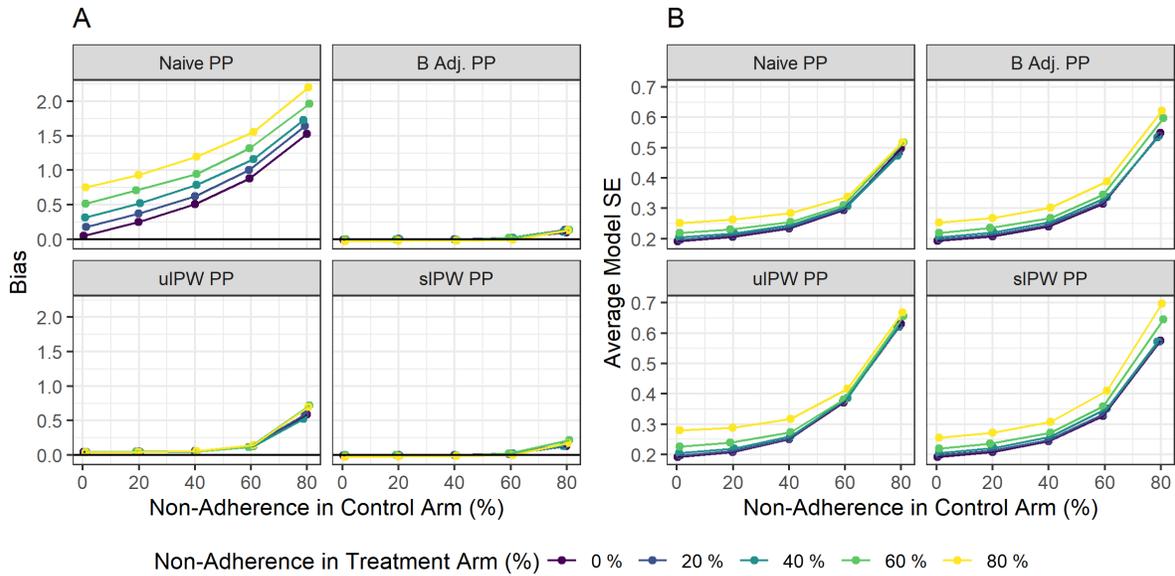
Appendix Figure D.4: Bias (plot A) and average model standard error (plot B) against non-adherence rates in the control arm for each effect estimate. Non-adherence rates in both arms can be different in these simulations. Data was generated based on Diagram (i) data generating process. B Adj. PP: Baseline adjusted per-protocol; B+L Adj. PP: baseline and post-baseline prognostic factor adjusted-per-protocol; uIPW PP: unstabilized inverse probability weighted per-protocol; sIPW PP: stabilized inverse probability weighted per-protocol.
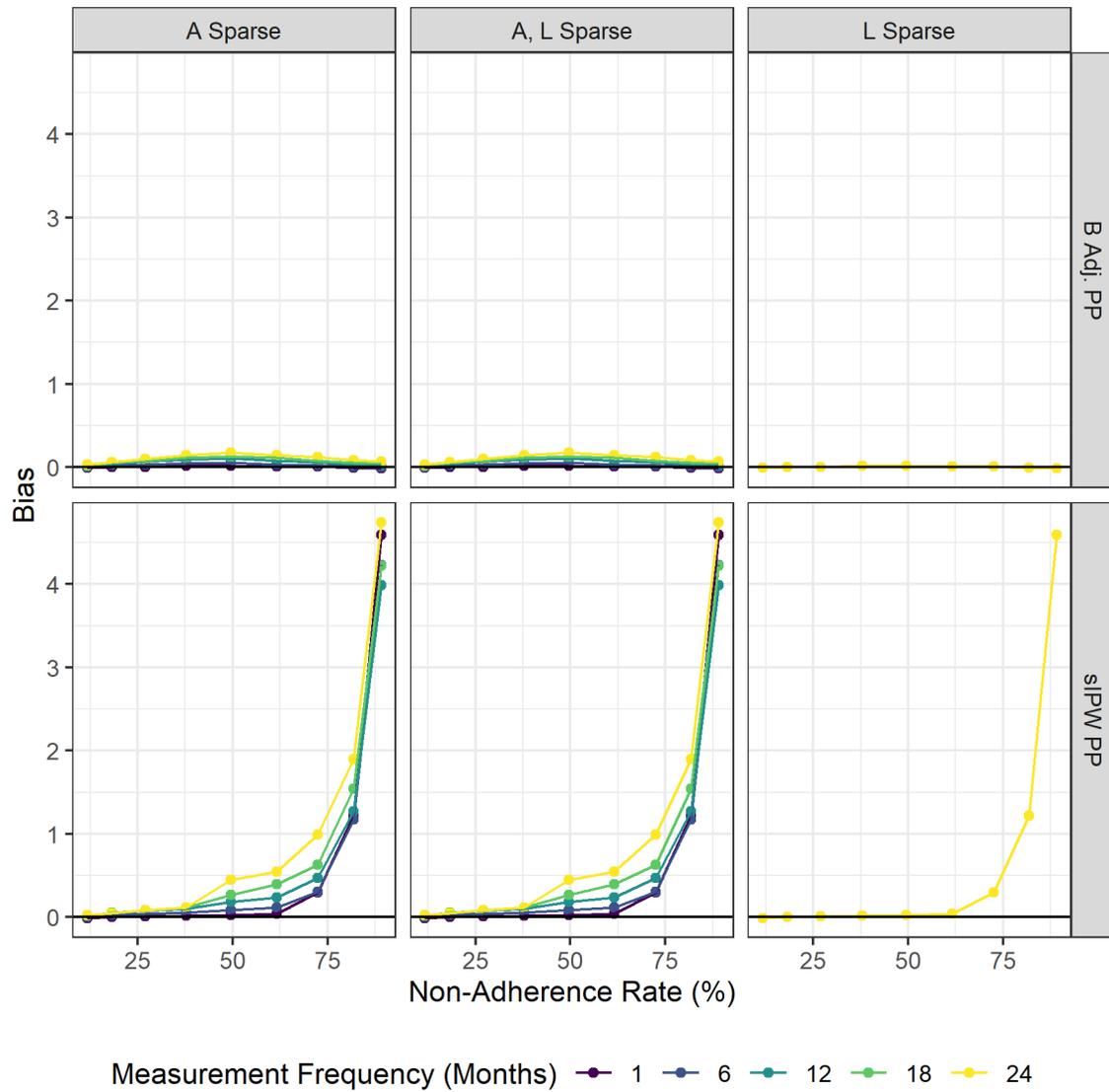
Appendix Figure D.5: Bias observed in each estimation method across 1,000 iterations of the simulation as the rate of non-adherence varies between 10 percent and 90 percent with equal non-adherence rates in each arm. Colour denotes the time between measurements for the sparse adherence and time-varying covariates. Both time-varying covariates and adherence were sparse during the follow-up period and were imputed using LOCF. B Adj. PP: Baseline adjusted per-protocol; sIPW PP: stabilized inverse probability weighted per-protocol.

E   Additional Simulation Settings and Results

Checking the validity of simulation algorithm

Varying Treatment Effect and Trial Size
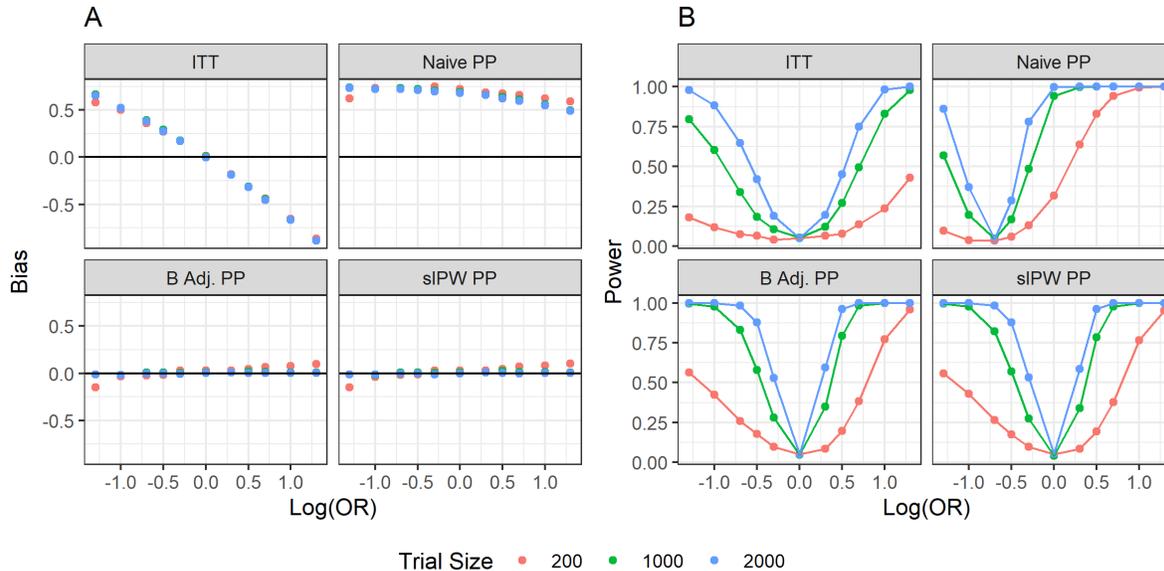
Appendix Figure E.1: A) Bias plotted against treatment effect for each effect estimate where the colour corresponds to the trial size in terms of number of participants per arm. B) Power or proportion of times the treatment effect was found to be statistically significant plotted against the treatment effect for each effect estimate where the colour corresponds to the trial size. Note that for both figures a log(OR) effect of treatment of 0 correspond to a null treatment effect.

For our base scenario Diagram 1 (i), Figure E.1 A illustrates the bias observed for each estimation method as the treatment effect varies. In this scenario, as a result to changing treatment effect, event rates varied between 5% (associated with θ_2 = -1.3) and 15% (associated with θ_2 = 1.3), with adherence rates fixed at ∼ 25%. As expected, stabilized IPW and baseline adjusted per-protocol estimates are approximately unbiased for all assessed treatment effects, with some observed bias when the trial size is small (e.g., n = 200 participants per arm). Figure E.1 B illustrates the power observed as the treatment effect changes, with a U-shaped pattern for all effect estimates. For the naive per-protocol estimate we see the U-shape is not centred at the null treatment effect indicating undesirable performance.

Considering the precision of the effect estimates, the empirical standard error and model standard error results can be seen in Figure E.2. For all effect estimates, the empirical standard error decreases as the $log(OR)$ treatment effect increases. This is likely due to strong positive treatment effects leading to increases in the event rate, allowing greater precision when estimating. Empirical standard error also decreases as the trial size increases, for all effect estimates. For all effect estimates presented, the empirical standard error and the model standard error align well for all treatment effects, if the trial size is n = 1,000 or larger. This indicates that our models accurately estimate the standard error of the treatment effect.
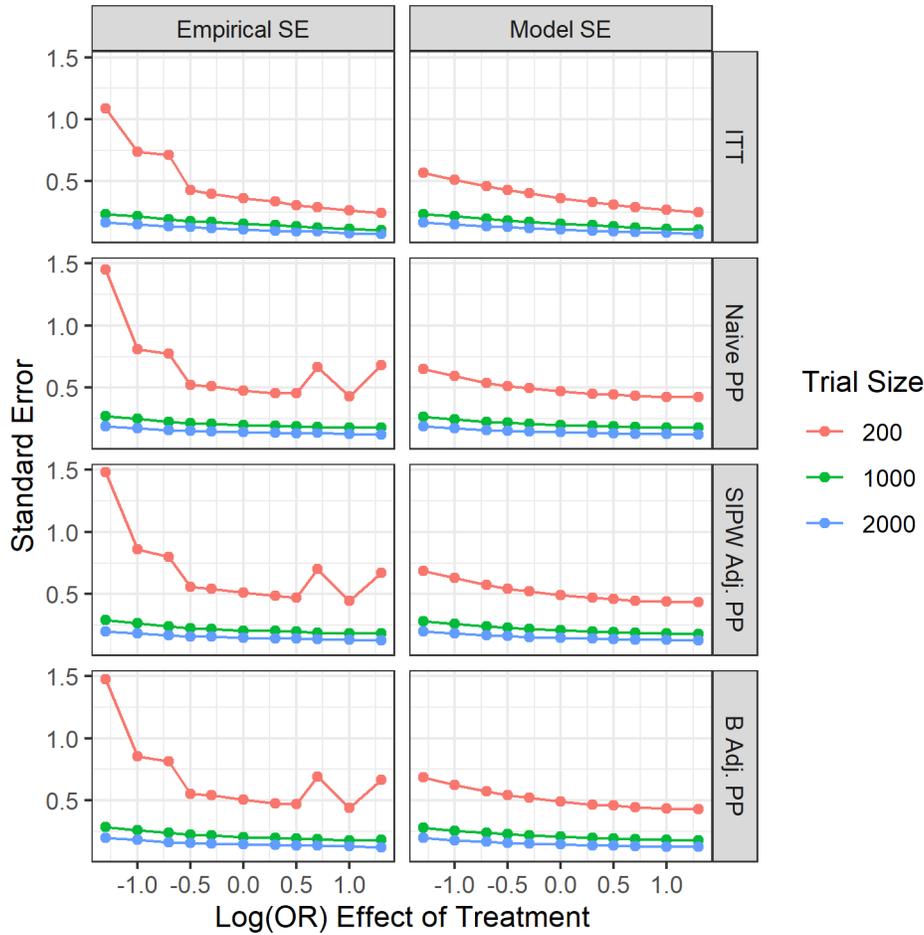
Appendix Figure E.2: Empirical standard error and model standard error (columns) for each effect estimate (rows) plotted against the $log(OR)$ effect of treatment, for various trial sizes (colour).

A final key statistical property to assess is the performance of the confidence intervals using coverage probability. The results for coverage probability and unbiased coverage probability as the treatment effect varies for different trial sizes can be seen in Figure E.3. Coverage probability is the proportion of confidence intervals that include the true treatment effect. For a traditional 95% confidence interval, the coverage probability should equal 95%. Under or over-coverage may be attributed to bias or incorrect interval width. To remove the effect of bias, Morris et al. (Morris et al., 2019) proposed unbiased coverage probability as a metric to assess the accuracy of the confidence interval length and thus standard error. The ITT and naive per-protocol estimate largely suffer from under-coverage, due to bias. The ITT effect has appropriate coverage probability for a null treatment effect (when it is also unbiased), and both the ITT and naive per-protocol estimate have unbiased coverage probability approximately equal to 95% for all treatment effects assessed. For the stabilized IPW and baseline adjusted per-protocol estimates we see coverage and unbiased coverage probabilities approximately equal to 95% for all treatment effects and trial sizes.
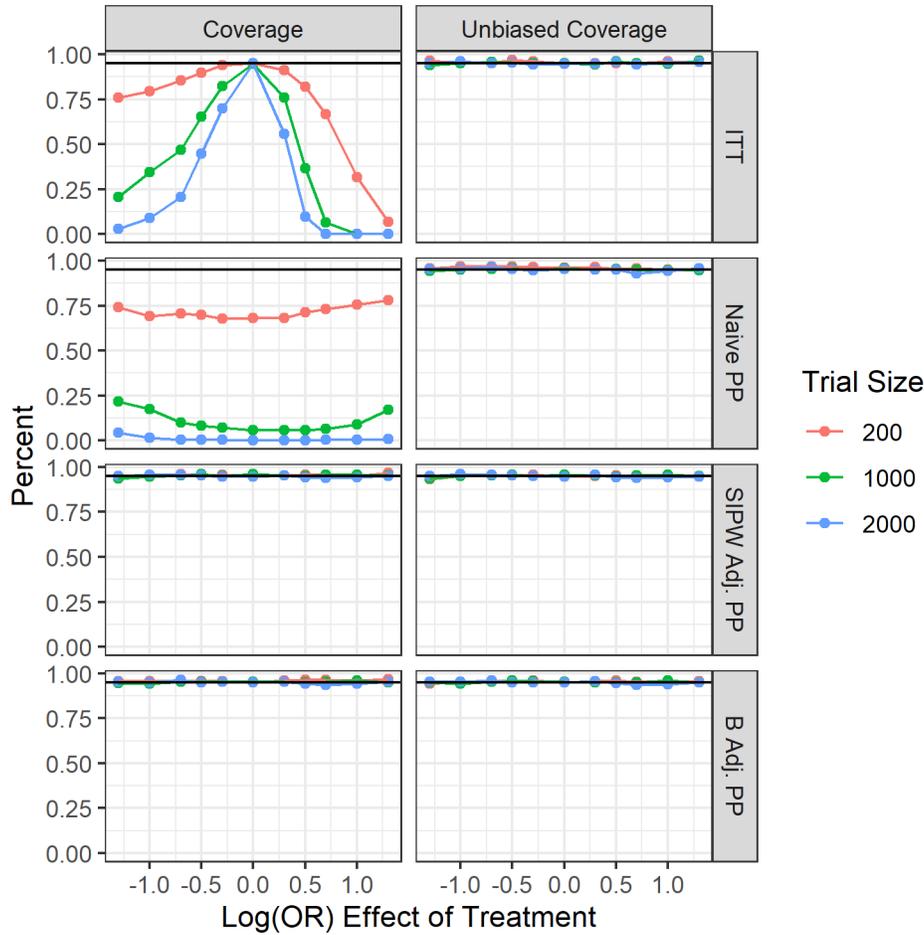
Appendix Figure E.3: Coverage probability and unbiased coverage probability (columns) for each effect estimate (rows) plotted against the $log(OR)$ effect of treatment, for various trial sizes (colour).

Event Rate

Model-based estimates   For this set of simulations, the event rates were varied between 0.1% and 75% among all participants. Figure E.4 plot A illustrates the bias observed as the event rate varies. Notably, we see that the stabilized IPW estimate as well as the baseline adjusted per-protocol estimate are unbiased for event rates between approximately 1% and 75%. The ITT estimate and the naive per-protocol estimate are biased for all event rates. We see that the estimates that are largely unbiased (stabilized IPW and baseline adjusted per-protocol estimate) have greater bias for small event rates (< 1%). In this scenario, adherence rates varied between 20% and 30%. The variability in bias between iterations and the average model standard error (see Figure E.4 plot B) also increases with small event rates (< 1%) for all estimation methods.
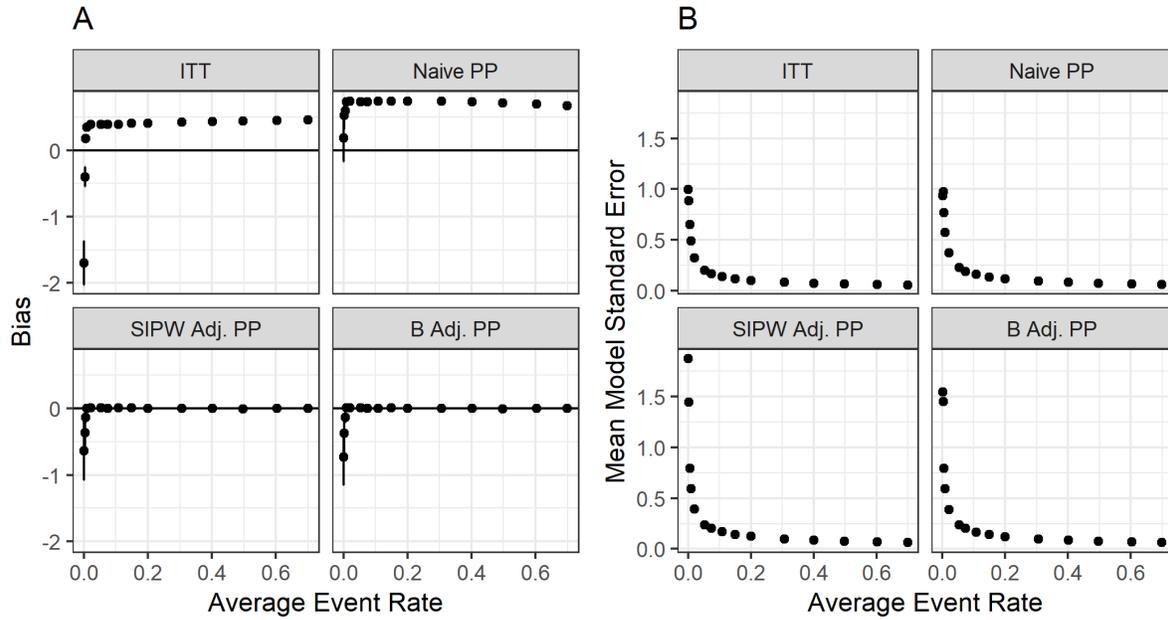
Appendix Figure E.4: Bias (plot A) and average model standard error (plot B) observed in each estimation method across 1,000 iterations of the simulation as the event rate varies between 0.001 and 0.75.

Cumulative survival methods The cumulative survival method was also used to calculate the ITT, naive per-protocol, and unstabilized IPW per-protocol effect estimates as the event rate varied. In Appendix E, Figure E.5 illustrates the bias and convergence rates for the cumulative survival estimates. We see the biases observed for the cumulative survival ITT and naive per-protocol estimates are similar to the model-based ITT and naive per-protocol estimates at approximately 0.5 and 0.7 respectively. For the cumulative survival unstabilized IPW per-protocol estimate we see greater bias than with the model-based stabilized IPW per-protocol estimate. For these simple estimates, low event rates did not increase the variability of the effect estimates, but rather decreased the proportion of iterations that converged. With the event rate below 1%, as few as 30% of iterations produced valid effect estimates.
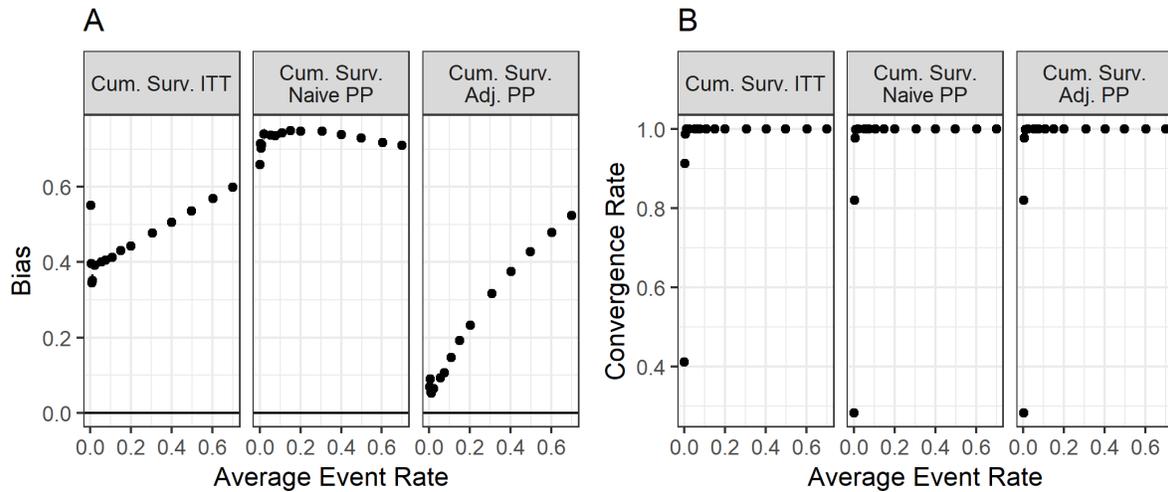
Appendix Figure E.5: Bias (plot A) and convergence rate (plot B) observed in each of the cumulative survival type estimates across 1,000 iterations of the simulation as the event rate varies between 0.001 and 0.75.

The Effect of Time on the Outcome

The impact of the impact of time on the outcome for these estimates was assessed for linear effects, with the effect of time varying from having no impact on the outcome to having a moderate to strong impact on the outcome. The bias (plot A) and average model standard error (plot B) for these simulations are presented in Figure E.6. For this simulation, time was set to have no effect on the outcome when the event rate was approximately 3%. The effect of time on the outcome was then increased until a total event rate of approximately 58% was observed, corresponding to a 55% increase in the event rate. In this scenario, adherence rates varied between $\sim 20\%$ and $\sim 30\%$. The stabilized IPW and baseline adjusted per-protocol estimates were unbiased for all effects of time assessed. In contrast, the ITT and naive per-protocol estimates were biased for all effects of time assessed, where time having a strong impact on the outcome led to the most biased effect estimates. When time had a strong impact on the outcome, more events were observed, resulting in lower average model standard error for all effect estimates.
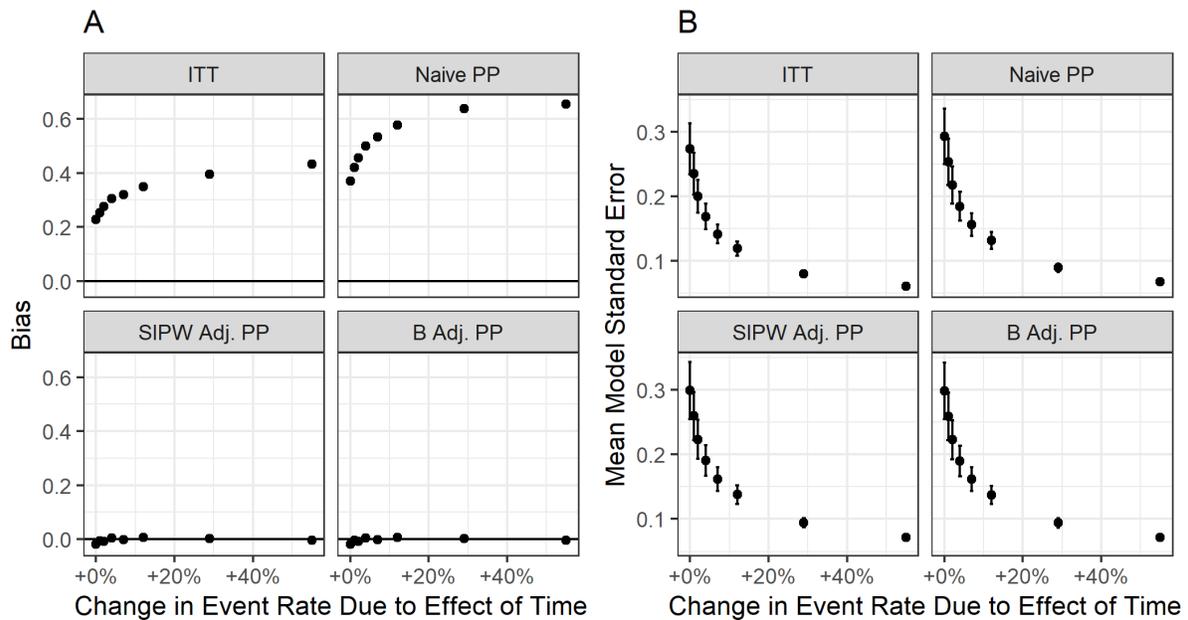
Appendix Figure E.6: Bias (plot A) and average model standard error (plot B) against the increase in event rate due to the effect of time on the outcome.

Sparse Follow-Up Measurements

Figure E.8 illustrates power for each estimate, for both missing data handling methods as the effect of treatment is varied. In this scenario, we had adherence rates fixed at $\sim 25\%$ and event rates fixed at $\sim 20\%$. No substantial difference was observed in power depending on which variables were sparse during follow-up, so this plot presents the findings for when both adherence and time-varying covariates were sparse. Complete-case (CC) analysis consistently has less power than LOCF imputation. The difference in power between CC and LOCF analyses is most pronounced for strong non-zero treatment effects for the ITT, stabilized IPW, and baseline adjusted per-protocol estimates. The ITT, stabilized IPW, and baseline adjusted per-protocol estimates all achieve the desired power of 0.05 for null treatment effects regardless of missing data method. The baseline adjusted per-protocol estimate combined with LOCF achieves 100% power sooner than the stabilized IPW per-protocol estimate combined with LOCF ($|log(OR)| = 0.7$ compared to $|log(OR)| = 1$).
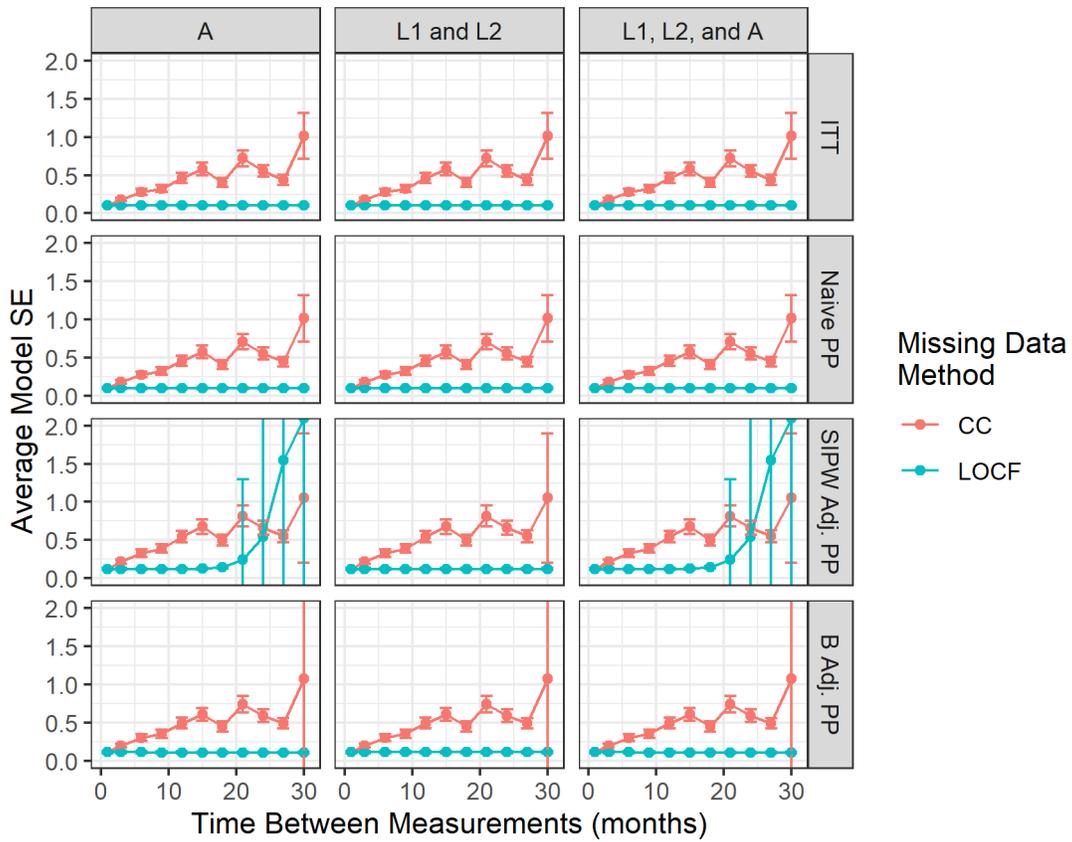
Appendix Figure E.7: Average model standard error observed in each estimation method as the measurement schedule during the follow-up period varies from monthly measures ($m = 1$) to measures every 2 years ($m = 24$). Columns correspond to whether the adherence (A), time-varying covariates (L_1 and L_2), or both (A, L_1 and L_2) were sparse during the follow-up period. Rows correspond to the 4 main estimation methods. Colour denotes whether the analysis was based on CC or LOCF imputation. Note that in this Figure some error bars have been truncated to due to outliers.
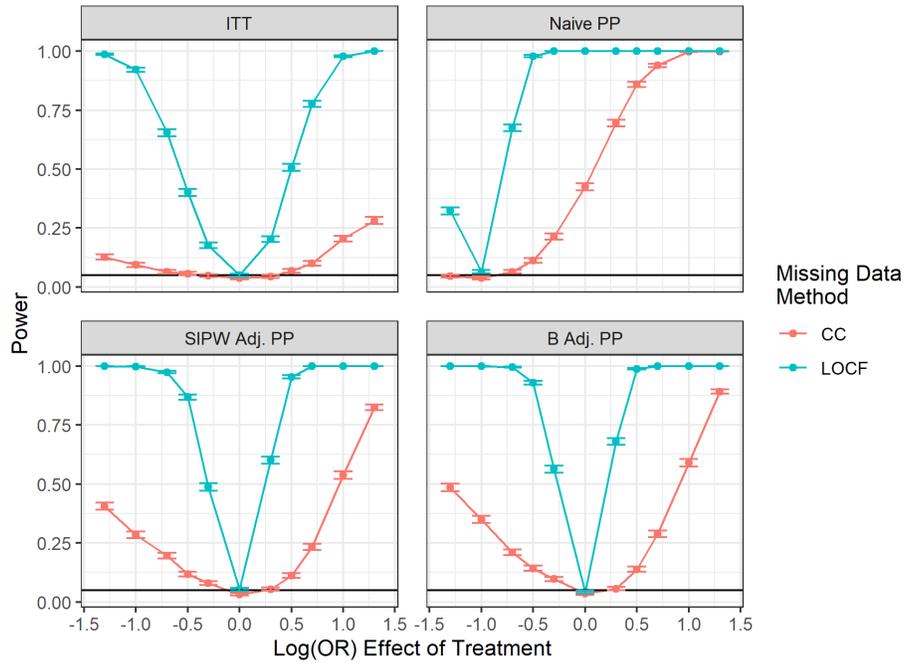
Appendix Figure E.8: Power observed in each estimation method across 1,000 iterations of the simulation as the treatment effect varied. Colour denotes whether the analysis was based on CC or LOCF imputation. Note that no difference was observed depending on whether the adherence (A), time-varying covariates (L_1 and L_2), or both (A, L_1 and L_2) were sparse during the follow-up period. This Figure corresponds to when both adherence and time-varying covariates were subject to sparse measurement during the follow-up period

Differential Non-Adherence Rates with Unmeasured Variable
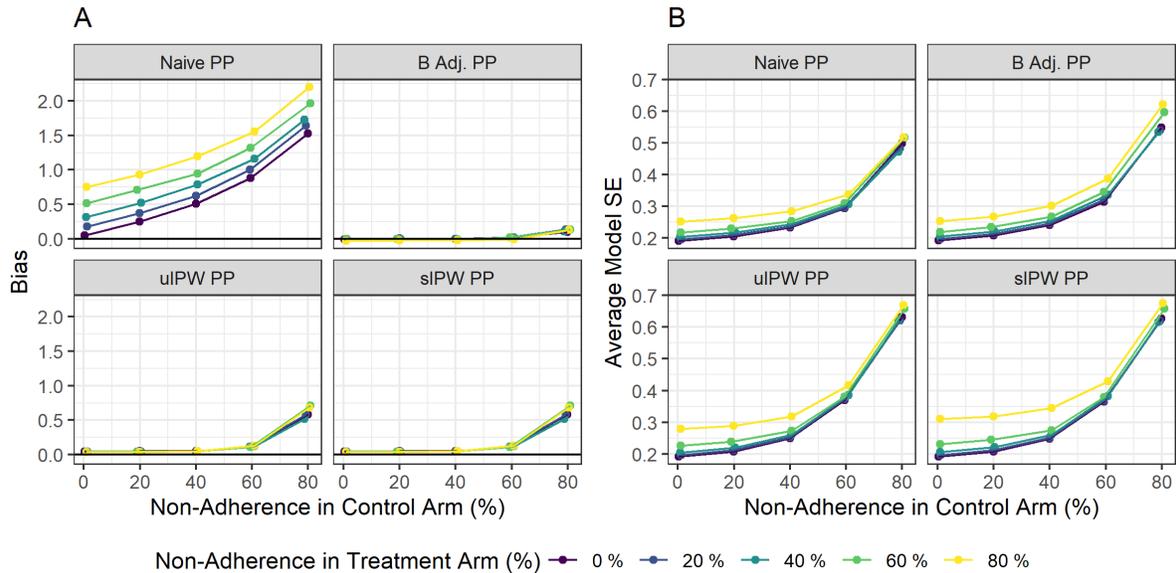
Appendix Figure E.9: Bias (plot A) and average model standard error (plot B) against non-adherence rates in the control arm for each effect estimate when treatment arm non-adherence rates can be different than that of the control arm and the baseline prognostic factor B is unmeasured. The data was generated from the process outlined in Diagram (i).

F Details about the Case Study

Measurement process

The first four visits were for screening and identifying eight prognostic risk strata. A total of $3,550$ eligible participants were randomized at the fifth visit and followed for a minimum of seven years. Visits six and seven occurred at two weeks intervals. All subsequent visits occurred at two months intervals. More than fifty types of forms were used for data collection, including outcome ascertainment, treatment assignment, screening, follow-up visit, hospitalization, and event record. However, not all forms were filled in each visit, and participants could have missed the occasional visit. Covariate and adherence values were carried forward from the most recent visit for up to five visits (Wanis et al., 2020). Participants were censored if they missed six consecutive visits.

Adherence

Participants received four packets per day of cholestyramine or a placebo at the fifth visit (considered as the baseline of the trial), which was increased to six packets per day after the sixth visit. Participants returned their unused medication packets and received a new supply of medication at each visit. A static treatment regime was considered, and all patients would be expected to continue the study medication. Since cholestyramine and placebo had no major side-effects, the static treatment strategies assumption was reasonable. Medication adherence was measured using counts of unused medication packets, where we defined satisfactory adherence as $\geq 80\%$. In other words, taking $< 80\%$ medication at a particular visit was considered as deviating from the protocol, i.e., nonadherent at that visit. For participants who were adherent in a particular visit, we censored them at the first visit they become non-adherent. The overall nonadherence rate was then quantified by calculating the proportion of participants who became nonadherent by their last observation.

Appendix Table F.1: Estimated effect of cholestyramine treatment on coronary heart disease death or non-fatal myocardial infarction.

Method	Weights		Coef. (log(OR))		OR	
	Mean	Min-Max	Estimate	SE	Estimate	95% CI
ITT			-0.16	0.13	0.85	0.66-1.09
Naïve PP			-0.22	0.29	0.80	0.45-1.41
B Adj. PP			-0.25	0.29	0.78	0.45-1.37
L Adj. PP			0.18	0.33	1.20	0.63-2.28
uIPW PP	1.34	1.00-172.49	-0.79	0.50	0.46	0.17-1.21
uIPW PP (5% truncated)	1.16	1.00-1.44	-0.27	0.29	0.76	0.43-1.34
sIPW PP	1.01	0.16-10.52	-0.31	0.29	0.74	0.42-1.29

OR: odds ratio; CI: confidence interval; SE: robust standard error; ITT: intention to treat; Naive PP: naive per-protocol; B Adj. PP: baseline adjusted per-protocol; L Adj. PP: time-varying covariates adjusted per-protocol; uIPW PP: unstabilized inverse probability weighted per-protocol; sIPW PP: stabilized inverse probability weighted per-protocol.

Covariates

Baseline Covariates (B)

The baseline covariates include baseline risk strata, age at randomization, physical activity level at work at baseline, educational status, and race. Baseline risk strata were created using the combination of three binary prognostic variables: i) positive or negative ECG test, ii) LDL cholesterol greater than 215 mg/dl, and iii) a multiple logistic risk function estimated using the covariates age, number of cigarettes smoked, and diastolic blood pressure at pre-randomization visits one and two. Baseline measurements of physical activity were categorized into five groups based on levels of activities.

Time-varying covariates (L)

The time-varying covariates were systolic blood pressure, diastolic blood pressure, total serum cholesterol, LDL cholesterol, HDL cholesterol, triglyceride, cigarettes smoked per day, physical activity outside of work, body mass index (grams/sq.cm), nausea, vomiting, diarrhea, calories, protein, total fat, total carbohydrates, angina, GXT, ischemia, total number of aspirin-cmpd tablets past week, P/S ratio, sodium, phosphorus, potassium, thyroxine, iron, total bilirubin, direct bilirubin, alkaline phosphatase, creatinine, fasting glucose, albumin, calcium, white blood cell count, vascular events, non-fatal/non myocardial infarction coronary events, vascular events history, and non-fatal/non myocardial infarction coronary events history. We considered these baseline and time-varying covariates from the previous studies (The Lipid Research Clinics Program, 1979; Lipid Research Clinics Program, 1984; Wanis et al., 2020). Time-varying measurements of physical activity were categorized into five groups based on levels of activities, blood pressure measurements were taken using the standard sphygmomanometer, and lipid data measurement were recorded using bloodwork. We considered flexible functions for cholesterol, LDL, HDL, and triglyceride since these four covariates were frequently measured variables in the trial. We considered restricted cubic splines with knots at the 5th, 35th, 65th, and 95th percentiles of the covariates as was done in previous research (Wanis et al., 2020). In other words, since blood lipids were frequently measured variables in the trial, our adherence or outcome models were adjusted for flexible functions of these covariates in the past six visits.

G Adjusting for Baseline Covariates in Pragmatic Trials

For RCTs, there remains much debate regarding whether baseline covariates should be adjusted in the analysis, as baseline confounding is supposed to be handled by design (Austin et al., 2010). Some argue that adjusting for known prognostic factors in an RCT has methodological advantages, i.e., bias reduction and increased precision (Ciolino et al., 2019; Kahan et al., 2014). However, the CONSORT statement on RCT advises against identifying such adjustment variables via empirical exploration (e.g., statistical significance) of the same data (Moher et al., 2010). In the context of pragmatic trials with more flexible designs, however, the existence of confounding may be more relevant. There are pragmatic trials where adjustment of baseline characteristics was necessary to correct for baseline imbalances (Rossouw et al., 2002; Holme et al., 2018; Carroll et al., 2018). A recent guideline draft also proposed adjusting for prognostic factors that meet a pre-specified threshold for imbalance and suggested to use IPW approaches for adjustment to preserve the marginal interpretations of the estimates (Murray et al., 2019).

H Future Works

In this study, adherence was measured as a binary factor. While it is common to dichotomize adherence (Murray and Hernán, 2016, 2018), it has been shown to introduce bias in effect estimation (Shrier et al., 2018). Additional work could compare similar effect estimates using partial adherence and more complex patterns of adherence over a trial (Wanis et al., 2020; Sanders, 2019).

Our work was limited to estimating OR rather than risk difference. In practice, both OR and RD are regularly reported means to quantify a treatment effect, each with their own merits (Toh et al., 2010). In our simulation work, calculation of the RD was difficult, requiring experimenting with various starting values during model fitting. OR estimation did not require such considerations but suffers from known limitations (Pang et al., 2013).

In this work, we showed results associated with baseline covariate generated from beta distribution. A future study could also consider studying the impact of choosing different distributions for the generation of the baseline variable(s). Future studies could consider the impact of model-misspecification, and whether double robust or flexible versions of the estimates can address the problem (Zhong et al., 2022). When the complexity of the data generating process increases, investigation of possibly introducing collider bias while adjusting covariates could be of future interest.

Sparse values are problematic in data analysis, but common when working with healthcare data. In pragmatic trials where researchers aim to utilize non-intrusive forms of follow-up, existing electronic health records and administrative databases may be the source of key follow-up information (Blaschke et al., 2012; Loudon et al., 2015; Thorpe et al., 2009; Ford and Norrie, 2016). However, few countries have a single comprehensive data source

for all forms of health data, requiring follow-up information to be derived by linking data from multiple sources leading to sparse values during follow-up (Blaschke et al., 2012; Ford and Norrie, 2016).

Future work for this active research area, would also benefit from exploring a greater range of treatment effects, such as where both the most recent treatment received as well as the cumulative dose received are able to affect outcomes (Young et al., 2019). Additional directions for future works include comparisons to additional estimation techniques such as g-estimation, the parametric g-computation, and instrumental variable methods (Hernán et al., 2013; Hernán and Robins, 2017).

References

Austin, P. C., Manca, A., Zwarenstein, M., Juurlink, D. N., and Stanbrook, M. B. A substantial and confusing variation exists in handling of baseline covariates in randomized controlled trials: a review of trials published in leading medical journals. *Journal of Clinical Epidemiology*, 63(2):142–153, Feb 2010.

Blaschke, T. F., Osterberg, L., Vrijens, B., and Urquhart, J. Adherence to medications: Insights arising from studies on the unreliable link between prescribed and actual drug dosing histories. *Annual Review of Pharmacology and Toxicology*, 52(1):275–301, 2012. doi: 10.1146/annurev-pharmtox-011711-113247. URL https://doi.org/10.1146/annurev-pharmtox-011711-113247. PMID: 21942628.

Carroll, J. K., Pulver, G., Dickinson, L. M., Pace, W. D., Vassalotti, J. A., Kimminau, K. S., Manning, B. K., Staton, E. W., and Fox, C. H. Effect of 2 Clinical Decision Support Strategies on Chronic Kidney Disease Outcomes in Primary Care: A Cluster Randomized Trial. *JAMA Netw Open*, 1(6):e183377, 10 2018.

Ciolino, J. D., Palac, H. L., Yang, A., Vaca, M., and Belli, H. M. Ideal vs. real: a systematic review on handling covariates in randomized controlled trials. *BMC Med Res Methodol*, 19(1):136, 07 2019.

Ford, I. and Norrie, J. Pragmatic trials. *New England Journal of Medicine*, 375(5):454–463, 2016.

Hernán, M. A. and Robins, J. M. Per-protocol analyses of pragmatic trials. *New England Journal of Medicine*, 377 (14):1391–1398, May 2017. doi: 10.1056/nejmsm1605385.

Hernán, M. A., Hernández-Díaz, S., and Robins, J. M. Randomized trials analyzed as observational studies. *Annals of Internal Medicine*, Oct 2013. doi: 10.7326/0003-4819-159-8-201310150-00709.

Holme, O., Loberg, M., Kalager, M., Bretthauer, M., Hernán, M. A., Aas, E., Eide, T. J., Skovlund, E., Lekven, J., Schneede, J., Tveit, K. M., Vatn, M., Ursin, G., and Hoff, G. Long-Term Effectiveness of Sigmoidoscopy Screening on Colorectal Cancer Incidence and Mortality in Women and Men: A Randomized Trial. *Ann. Intern. Med.*, 168(11):775–782, 06 2018.

Kahan, B. C., Jairath, V., Doré, C. J., and Morris, T. P. The risks and rewards of covariate adjustment in randomized trials: an assessment of 12 outcomes from 8 studies. *Trials*, 15:139, Apr 2014.

Lipid Research Clinics Program. The lipid research clinics coronary primary prevention trial results I. Reduction in Incidence of Coronary Heart Disease. *JAMA*, 251(3):351–364, 1984.

Loudon, K., Treweek, S., Sullivan, F., Donnan, P., Thorpe, K. E., and Zwarenstein, M. The PRECIS-2 tool: designing trials that are fit for purpose. *Bmj*, 350(h2147), Aug 2015. doi: 10.1136/bmj.h2147.

Moher, D., Hopewell, S., Schulz, K. F., Montori, V., Gøtzsche, P. C., Devereaux, P. J., Elbourne, D., Egger, M., and Altman, D. G. CONSORT 2010 explanation and elaboration: updated guidelines for reporting parallel group randomised trials. *BMJ (clinical research edition)*, 340(c869), Mar 2010. doi: 10.1136/bmj.c869.

Morris, T. P., White, I. R., and Crowther, M. J. Using simulation studies to evaluate statistical methods. *Statistics in Medicine*, 38(11):2074–2102, 2019. doi: 10.1002/sim.8086.

Murray, E. J. and Hernán, M. A. Adherence adjustment in the coronary drug project: A call for better per-protocol effect estimates in randomized trials. *Clinical Trials: Journal of the Society for Clinical Trials*, 13(4):372–378, Jul 2016. doi: 10.1177/1740774516634335.

Murray, E. J. and Hernán, M. A. Improved adherence adjustment in the coronary drug project. *Trials*, 19(158), May 2018. doi: 10.1186/s13063-018-2519-5.

Murray, E. J., Swanson, S. A., and Hernán, M. A. Guidelines for estimating causal effects in pragmatic randomized trials, 2019.

Pang, M., Kaufman, J. S., and Platt, R. W. Mixing of confounding and non-collapsibility: a notable deficiency of the odds ratio. *American Journal of Cardiology*, 111(2):302–303, 2013.

Rossouw, J. E., Anderson, G. L., Prentice, R. L., LaCroix, A. Z., Kooperberg, C., Stefanick, M. L., Jackson, R. D., Beresford, S. A., Howard, B. V., Johnson, K. C., Kotchen, J. M., and Ockene, J. Risks and benefits of estrogen plus progestin in healthy postmenopausal women: principal results From the Women's Health Initiative randomized controlled trial. *JAMA*, 288(3):321–333, Jul 2002.

Sanders, E. Incorporating partial adherence into the principal stratification analysis framework, 2019.

Shrier, I., Platt, R. W., Steele, R. J., and Schnitzer, M. Estimating causal effects of treatment in a randomized trial when some participants only partially adhere. *Epidemiology*, 29(1):78–86, 2018. doi: 10.1097/ede.0000000000000771.

The Lipid Research Clinics Program. The coronary primary prevention trial: design and implementation. *Journal of Chronic Diseases*, 32(9-10):609–631, 1979. ISSN 0021-9681.

Thorpe, K. E., Zwarenstein, M., Oxman, A. D., Treweek, S., Furberg, C. D., Altman, D. G., Tunis, S., Bergel, E., Harvey, I., Magid, D. J., and et al. A pragmatic-explanatory continuum indicator summary (PRECIS): a tool to help trial designers. *Canadian Medical Association Journal*, 180(10), 2009. doi: 10.1503/cmaj.090523.

Toh, S., Hernández-Díaz, S., Logan, R., Robins, J. M., and Hernán, M. A. Estimating absolute risks in the presence of nonadherence. *Epidemiology*, 21(4):528:539, 2010. doi: 10.1097/ede.0b013e3181df1b69.

Wanis, K. N., Madenci, A. L., Hernán, M. A., and Murray, E. J. Adjusting for adherence in randomized trials when adherence is measured as a continuous variable: An application to the lipid research clinics coronary primary prevention trial. *Clinical Trials*, page 1740774520920893, 2020.

Young, J. G., Vatsa, R., Murray, E. J., and Hernán, M. A. Interval-cohort designs and bias in the estimation of per-protocol effects: a simulation study. *Trials*, 20(1), May 2019. doi: 10.1186/s13063-019-3577-z.

Zhong, Y., Brooks, M. M., Kennedy, E. H., Bodnar, L. M., and Naimi, A. I. Use of machine learning to estimate the per-protocol effect of low-dose aspirin on pregnancy outcomes: a secondary analysis of a randomized clinical trial. *JAMA network open*, 5(3):e2143414–e2143414, 2022.