Performance of Different Propensity Score Methods in Simulated Cohort Studies with Time-to-Event Outcomes

Sentinel Methods

Susan Gruber[1,2], Zilu Zhang[1], Robert Wellman[3], Yueqin Zhao[4], Rima Izem[4], Jennifer Clark Nelson[3], Jessica Franklin[5], Christian Hampp[4], Mark Levenson[4], Katherine Freitas[1], Judith Maro[1], Catherine Rogers[1], Darren Toh[1], Joshua Gagne[5], Sebastian Schneeweiss[5], Richard Wyss[5], Laura Amsden[6], Bruce Fireman[6]

1. Department of Population Medicine, Harvard Pilgrim Health Care Institute and Harvard Medical School, Boston, MA; 2. Putnam Data Sciences, LLC, Cambridge, MA; 3. Kaiser Permanente Washington Health Research Institute, Seattle, WA; 4. Office of Biostatistics, Center for Drug Evaluation and Research, FDA, Silver Spring, MD; 5. Division of Pharmacoepidemiology and Pharmacoeconomics, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA; 6. Kaiser Permanente Northern California, Oakland, CA

Version 1.0

May 1, 2019

The Sentinel System is sponsored by the U.S. Food and Drug Administration (FDA) to proactively monitor the safety of FDA-regulated medical products and complements other existing FDA safety surveillance capabilities. The Sentinel System is one piece of FDA's Sentinel Initiative, a long-term, multi-faceted effort to develop a national electronic system. Sentinel Collaborators include Data and Academic Partners that provide access to healthcare data and ongoing scientific, technical, methodological, and organizational expertise. The Sentinel Coordinating Center is funded by the FDA through the Department of Health and Human Services (HHS) Contract number HHSF223201400030I.

Performance of Different Propensity Score Methods in Simulated Cohort Studies with Time-to-Event Outcomes

Sentinel Methods

Table of Contents

History of Modifications

Version	Date	Modification	Author
1.0	5/1/2019	Original Version	Sentinel Operations Center

Executive Summary

We compared different propensity score (PS)-based methods that are used to control for confounding in a cohort study to estimate the effect of a treatment on a time-to-event outcome. We simulated treatment and outcomes based on the covariate profiles of real people resampled from data contributed by two FDA Sentinel data partners. We varied the strength of the treatment effect, strength of confounding, incidence of outcome events, prevalence of treatment, heterogeneity of the treatment effect, and censoring. The PS methods we considered estimate either a conditional or marginal hazard ratio (HR) among the treated (ATT) or among the study population (ATE) using matching, stratification, inverse weighting, or adjusting for PS-based covariates in a Cox regression model. In scenarios where treatment affected risk, methods that condition on the PS yielded estimates of the conditional HR that were biased towards 1.0 by amounts that increased over time, as outcome events differentially depleted treatment and comparator groups of high-risk individuals. Methods that estimate a marginal HR overestimated the treatment effect when there was much early, albeit uninformative, censoring. PS stratification performed better when strata were fine rather than coarse. Fine stratification and weighting methods performed well with respect to bias and precision.

Introduction

We assessed the performance of different ways that propensity scores (PS) can be used to control for confounding in a cohort study to evaluate the effect of a treatment on a time-to-event outcome. A PS is a person's probability of receiving the treatment, conditional on covariates that are potential confounders. PSs can be used to balance comparator groups in cohort studies by 1) matching on the PS, 2) stratifying on quantiles of the PS, 3) weighting by the inverse of the PS, or 4) adjusting for PS-based covariates. We evaluated multiple variants of each of these ways of using a PS in simulated scenarios that varied the strength of the treatment effect, strength of confounding, outcome incidence, treatment prevalence, heterogeneity of the treatment effect, and censoring.

Our work was motivated by the US Food and Drug Administration's Sentinel System (1), which monitors medical product safety with data from partners covering nearly 200 million people. PS-based methods are especially useful to Sentinel because they help preserve privacy by letting individual-level data remain behind the firewalls of Sentinel's partners (2).

PS-based methods have additional advantages: they can be parsimonious ways to adjust for many confounders, and they can provide intuitive ways to emulate randomized trials. Our aim was to inform Sentinel about the performance of its PS-based analytic tools and potential enhancements.

Rosenbaum and Rubin (3) showed that the PS can balance comparisons of treated with untreated individuals, and thereby permit a cohort study to emulate a randomized trial (if causal assumptions are met). They showed that a regression analysis of observational data, conditioned on the PS, can consistently estimate an additive treatment effect, such as a difference in means of a continuous outcome. However, drug safety studies often focus on binary outcomes in censored follow-up (4). They often target an effect that is multiplicative rather than additive, either estimating an odds ratio by logistic regression or a hazard ratio (HR) by Cox regression.

There are concerns that PS-based methods are biased in such studies (5,6). The bias arises as the cohort is depleted by outcome events in higher risk individuals, the marginal HR (HR_m, averaged over the cohort) diverges from the conditional HR (HR_c, conditional on baseline covariates), and PS-based estimators of the HR_c tend to land between the HR_c and the HR_m. PS-based estimators of the HR_m encounter a related problem when follow-up is heavily censored: analyses of the uncensored survivors

become biased as the treatment groups are differentially depleted of higher-risk individuals. The HR_m is a moving target over time that can be hard to hit (7). We address these concerns as we compare the different ways of using the PS. In all, we compared 47 methods across 21 scenarios. Evaluation of such a broad set of methods in a broad range of scenarios is relatively rare in the epidemiology literature.

Methods

We conducted plasmode simulations (8) based on Sentinel's surveillance of the oral anticoagulants rivaroxaban and warfarin (9). Using de-identified data from two Sentinel data partners, we sampled with replacement to simulate cohorts with the same characteristics and size as the Sentinel cohorts: 31,791 users of either rivaroxaban or warfarin from one Sentinel partner and 7,681 from the other. By sampling real data, we simulated cohorts in which covariates are realistically distributed and correlated.

Treatments and outcomes were not sampled; they were allocated by mechanisms tailored to generate the scenarios in Table 1. For our baseline scenario, 25% of patients received treatment drug A while 75% received comparator drug B. The outcome was observed in 5% of the cohort. Drug A doubled each individual's risk at every timepoint. Covariates positively associated with the outcome were less prevalent in users of A, resulting in strong confounding in the negative direction. Specifically, we tailored the baseline scenario so that an unadjusted analysis yields an HR_c estimate of 1.0 when the truth is 2.0. Thus, an unadjusted analysis is falsely consistent with the null hypothesis.

The other scenarios varied features of the baseline scenario to challenge our methods. Details are in Appendix 1. Related work on similar scenarios with binary outcomes is summarized elsewhere (10).

Data-Generating Mechanisms

We used 15 covariates: age, sex, and 13 binary covariates with the most confounding potential, as measured by Bross's formula (11).

Treatment was allocated according to a PS based on a scenario-specific function of the 15 baseline covariates:

$$\text{logit of probability of treatment with Drug A} = \alpha_0 + \alpha_1 \times X_1 + \ldots + \alpha_{15} \times X_{15}$$

In our baseline scenario the C-statistic averaged 0.64, indicating good overlap between the treated and comparator populations (see Appendix 2 for more on the PS).

Outcomes follow a Weibull distribution based on another scenario-specific function of the baseline covariates:

$$\text{time\_to\_event} = \{-ln(u) \,/\, exp(\beta_0 + \beta_1 \times X_1 + \ldots + \beta_{15} \times X_{15} + ln(HR_c) \times TX)\}^{1/k},$$

where TX is the treatment, u is random uniform, and k is 1.5. The coefficients ($\alpha_0 - \alpha_{15}$, $\beta_0 - \beta_{15}$) were tailored for the baseline scenario and then modified for the other scenarios (see (12) for a discussion of simulating from a Weibull distribution).

Follow-up was censored by a mechanism inspired by the distribution of censoring times in Sentinel's rivaroxaban surveillance. Follow-up ended at the earliest of: an outcome event, a lapse in treatment, or completion of two years of treatment (see Appendix 1). Outcome incidence and censoring are described in Figure 1.

Data Analyses

For each scenario 1,000 datasets were generated and analyzed using SAS 9.3 (13). In each analysis, the PS was estimated using a logistic regression model that was consistent with our treatment-generating mechanism (except in scenario 19, which features a residual confounder). The model was fit at each site separately as is done in routine Sentinel analysis.

All analyses used Cox regression to estimate an HR but they estimated different types of HRs, either conditional (HR_c) or marginal (HR_m). An HR_c multiplies each individual's hazard and is conditional on the individual's covariates. A marginal, or population-averaged effect, HR_m, multiplies the average risk and is standardized according to the distribution of covariates in either the entire cohort (ATE) or in the treated group (ATT). For each scenario studied, the HR_c was fixed by our data generating mechanism. We calculated the HR_m (both ATE and ATT) by Cox regression using simulated counterfactual cohorts followed for two years without censoring.

PS-Based Estimators of the Treatment Effect

The estimators we studied are described in Table 2 (see Appendix 3 for technical definitions). The estimators in the top ten rows of Table 2 target an HR_c. The top two rows describe benchmark estimators that adjust for individual covariates rather than the PS. Recent literature demonstrated that PS-based estimators can diverge from this true HR_c due to conditioning on the PS instead of on covariates (5,6). One of our major aims was to better understand how this bias arises, even if the PS is estimated from a correct model. The estimators in the bottom 12 rows of Table 2 describe PS-based estimators of the HR_m.

The estimators (Table 2) that adjust for confounding by (a) covariate adjustment (b) matching, (c) stratification, or (d) inverse probability of treatment weighting (IPTW) were implemented as follows:

Estimators that are adjusted for PS-based covariates (2.1 – 2.3)

These used either polynomial terms (PS, PS^2, PS^3) (method 2.1), dummy variables defined by site-specific deciles of the PS (2.2), or cubic B-splines with knots at site-specific quintiles of the PS among the treated (2.3).

Estimators that match on the PS (3.1 – 3.3, 4.1 – 4.5)

These used "greedy" nearest neighbor matching. Three such methods (3.1 – 3.3) used 1:1 fixed ratio matching, and five (4.1 - 4.5) used 1:M variable ratio matching with up to 10 comparators per treated subject. These methods estimate an ATT.

The matched data analyses in 3.1 and 4.1 stratify on matched set; they estimate a conditional HR. The other matching estimators ignore the matching and instead estimate a marginal HR (ATT) (2). The marginal estimators with 1:M matching either conditioned the Cox regression analysis on M (4.2), or else weighted the Cox regression (4.3-4.5). No weighting or conditioning was used for the marginal estimator that matched 1:1 (3.3).

The matching estimators handled potential heterogeneity by site in one of four ways: conditioning the Cox model by site (3.2, 4.2), adding site indicators as covariates (4.3), ignoring site (4.4), or using meta-analysis to average results from separate site-specific analyses (4.5).

Estimators that stratify on the PS (5.1 - 5.4)

These either used 10 strata (deciles), 20 strata, or fine stratification (for the ATT: five treated subjects per stratum plus all comparators within the PS range of the five treated subjects; for the ATE: five subjects per stratum) (14). Method 5.1 conditions the Cox regression on the strata. Methods 5.2 and 5.3

weight the Cox regression to estimate either an ATT or ATE. Method 5.4 uses meta-analysis to combine site-specific Cox regression analyses as done by Method 5.3.

Estimators that use IPTW (6.1-6.3)

These estimators of an HR_m (15, 16), were implemented using stabilized or unstabilized weights, either with or without truncation.

We assessed bias, precision, mean squared error (MSE), and coverage. Bias was calculated on the HR scale relative to either HR_c or HR_m (ATE or ATT). Standard errors (SE) and MSE were evaluated on the log(HR) scale. Our SE estimates were taken directly from the SAS model output for unweighted Cox regressions. We used robust sandwich estimators when the regression was weighted (17).

Results

The Baseline Scenario

The performance measures for each estimator in our baseline scenario are reported in Table 3. The top two rows show the benchmark estimators that adjust for covariates individually. As expected, they were very near the true HR_c.

Figure 2A shows HR estimates for the baseline scenario in relation to reference lines at the marginal ATE, the marginal ATT and the HR_c. The HR_c was fixed at 2.0 for all individuals (the conditional ATE and ATT are identical). The true marginal ATE and ATT diverged from 2.0, and were 1.57 and 1.65, respectively. The PS-based estimators of the HR_c all landed below their 2.0 target. The 1:1 matched conditional estimator was 0.1 (5%) below the HR_c; the 1:M matched conditional estimator was 0.3 (16%) below, while the three estimators using PS regression were 0.26 to 0.34 (13% to 18%) below the HR_c. PS regression performed better with polynomial terms or splines than with PS deciles. The six stratified estimators of the HR_c ranged from 0.21 to 0.41 (11% to 21%) below the true HR_c. The finer the stratification, the smaller the bias. The three stratified estimators of the ATE (shown in black) performed better than their ATT counterparts (in gray); performance improved when we bounded strata by PS cut-points from the entire cohort rather than the treated group.

Most HR_m estimators landed closer to their targets than did the HR_c estimators. The six IPTW estimators of the ATT were about 0.02 (1%) above their 1.65 target; the six IPTW estimators of the ATE were about 0.02 (3%) above their 1.57 target. The 1:1 matched estimators of the ATT performed similarly to the IPTW estimators of the ATT with respect to bias, but were much less precise. 1:M matching estimates were near the ATT target when the analysis was stratified on M, but fell below the ATT when the analysis was weighted. Stratified estimators generally performed better with finer stratification.

Scenarios 2-13

In contrast with the baseline scenario, HR_m diverged less from HR_c when the outcome incidence was lower (scenarios 2 and 3), when the treatment effect was smaller (scenarios 6-7, 10-12), and when the covariates were less predictive of the outcome but highly predictive of treatment (scenario 9) (Appendix 1). When the treatment had no effect on the outcome (scenarios 8, 11), the marginal and conditional targets are 1.0.

The two benchmark estimators landed very near HR_c in scenarios 2-13, as expected. The 1:1 matched conditional estimator was less biased than the other PS-based conditional estimators, except in scenario 3 (rarest event) where some analyses failed to converge due to the rarity of outcomes, and scenario 10 (protective treatment effect) where 1:M matching performed as well as 1:1 matching.

As in the baseline scenario, finer stratification yielded less bias. The IPTW estimators were near their HR_m target and so were the unconditional 1:1 matched estimators. Weighted stratified estimators that used coarser stratification or matching were more biased.

Scenarios with heterogeneous effects (Scenarios 14-18)

The HR_c is undefined in heterogeneous scenarios (only subgroup-specific conditional HRs are defined). Nevertheless, our conditional estimators may still be compared with each other. In scenario 15, where many observed outcomes occurred in the high-risk subgroup, all 10 PS-based conditional estimators yielded HR_c estimates below the benchmarks. PS-stratification using 10 or 20 strata were farther below the benchmarks than were the other estimators.

HR_m targets are well-defined in heterogeneous scenarios. The IPTW and 1:1 matched estimators landed close to their targets despite the heterogeneity. In scenario 15, where a relatively high proportion of the censored outcomes were in the high-risk subgroup, the IPTW estimators were somewhat higher than their targets. Unlike the other scenarios, finer stratification performed slightly worse than coarser stratification.

Scenarios with residual confounding, informative censoring or heavier censoring (scenarios 19-21)

All estimators were substantially above their targets in scenario 19 where an unmeasured covariate was a confounder in the positive direction. Here the more coarsely stratified estimators, which performed poorly in other scenarios, were closer to their targets, as the negative residual confounding within the coarse strata offset some of the positive residual confounding from the unmeasured covariate.

Similarly, when informative censoring depleted the cohort differentially in scenario 20, the conditional PS-based estimators landed below their target and the marginal estimators landed above their targets. In this scenario, the 1:1 matched conditional estimator landed near the HR_c, as the bias toward the null (that was always observed in PS-based conditional estimators) was offset by the bias from informative censoring.

Uninformative but heavy censoring in scenario 21 had more impact on the marginal estimators than on the conditional estimators. It tended to bias the marginal estimates away from 1.0 toward the HR_c.

Whereas Figure 2 shows the average estimate for each method in a single scenario, Figure 3 presents a box plot for a variant of each type of method showing the spread of the bias over the thousand replications for each scenario. The better-performing methods in our baseline scenario were usually the better-performing methods in our other scenarios. However, some important differences emerged in the more challenging scenarios. In our rare event scenario (scenario 3), estimators of the HR_c produced biased and variable point estimates, while some estimators of an HR_m were unbiased (1:1 matching, fine stratification, IPTW). Also, when exposure greatly increased the hazard (scenario 13) the HR_c was underestimated by HR_c estimators while the HR_m estimators were less biased.

Precision

Plots of SEs are shown in Figure 2B. Each color-coordinated pair of symbols shows the mean of our estimated SEs in comparison with the Monte Carlo standard deviation (MCSD) of the HRs. When these two statistics are similar, the estimated SEs are accurate. With few exceptions, this was found to be the case. In rare event and poor overlap scenarios 3 and 9 the estimated SEs for PS-stratified methods were lower than the Monte Carlo standard deviations when estimating a marginal ATE. This would result in confidence intervals that are too narrow.

The conditional 1:1 matched estimators were much less precise than the other estimators in all scenarios except 9, where several of the stratified estimators were even less precise. In all scenarios, PS regression yielded HR_c estimates that were similar in precision to the benchmark estimators but with much higher MSE. In all scenarios, 1:1 matching was less precise than weighting methods in estimating the marginal ATT.

In the baseline scenario, the IPTW estimators of ATE were more precise than the benchmarks or any of the other conditional estimators. Only in scenario 9 (poor overlap) was IPTW less precise than the benchmarks. PS regression was also relatively precise, especially if SEs are considered relative to the size of the estimands.

Coverage for methods that estimated a marginal effect was usually near the nominal 95% rate. An exception was low coverage for PS-stratification using 10 or 20 strata. Coverage for methods that estimated a conditional effect was low, since these methods tended to yield an effect estimate far from the HR_c (toward the HR_m).

Hypothesis Testing

In our baseline scenario where the HRc is 2.0, the marginal ATE and ATT are 1.57 and 1.65, respectively. If drug safety signals are evaluated by Wald tests (based on the ratio of logHR to its SE), then tests of the HRc would tend to be more powerful than tests of HRm when the estimates are equally precise. However, at our large sample size and effect sizes, IPTW estimators of HRm performed equally well at rejecting false null hypotheses. In all scenarios where treatment affected risk rejection rates were especially favorable for the IPTW estimators of the ATT (Appendix 1, Tables A1.24, A1.25).

Discussion

We compared 47 ways of using a PS across 21 scenarios in cohort studies with time-to-event data. These methods use matching, stratification, weighting, or adjustment for PS-based covariates. All methods use Cox regression, but they estimate different kinds of HRs. The methods that adjust for PS-based covariates target a HR_c; the IPTW estimators target a HR_m. Methods that match or stratify on the PS can target either an HR_c or an HR_m depending on the kind of Cox model that is used (see Appendices 4 and 6).

Our scenarios were designed to pose various challenges to the PS-based methods for estimating a target HR_c or HR_m. Crafting scenarios where the challenges were well understood allowed us to gain insight into drivers of bias and precision in these PS-based methods. However, as with all simulation studies, our findings may not generalize to scenarios that are markedly different from those we examined. In particular, how the methods perform in more complex settings where many of these challenges are simultaneously present merits further in-depth study.

To assess the relative performance of these methods in scenarios challenged by low exposure and rare outcomes we first used straightforward simulation settings where the HR_c was homogeneous and constant over time, PSs were estimated from a correct model, and there was good overlap between treatment and comparator populations. All methods had similar performance in several of these scenarios, but differences in finite sample bias emerged as either exposure or the outcome grew increasingly rare. We subsequently introduced additional complexity to better understand how methods might be differentially affected by challenges of poor overlap, treatment effect heterogeneity, informative censoring, and PS misspecification (18, 19).

In many drug safety studies the HR_m is nearly the same as the HR_c, and there is little cause for concern about distinctions between them. The HR_m and HR_c do not diverge at all if treatment has no effect on

risk. Even if treatment has a big effect on risk, the HR_c and HR_m are nearly equal when the outcome incidence is very low, as often happens in drug safety surveillance. When the HR_m does diverge from the HR_c the HR_m is always closer to 1.

In our baseline scenario treatment increased risk. Over time the treated group became more depleted than the untreated group of high-risk individuals; thus, the instantaneous HR attenuated towards 1. Because the HR_m amounts to the average of instantaneous HRs over all event times, the HR_m moves away from the HR_c and toward 1 as follow-up time increases. When follow-up time is extended until outcome events have occurred in all individuals (as in the simulation studies in reference (5)) this divergence is more marked and consequential than in our scenarios, designed to mimic realistic safety studies (see Appendix 4).

The three PS regression estimators (2.1-2.3) that included functions of the PS in the Cox model yielded biased estimates of the HR_c that were between the HR_c and the HR_m. This bias was reduced, but remained substantial, when we used splines or polynomial PS terms as covariates rather than a single main term. Our findings demonstrate that whether PS regression estimators land closer to the HR_c or the HR_m is related to how closely the PS is correlated with the disease risk score (20), a function of the covariates that best predicts the HR_c under no exposure (this was proven in a linear context in (6)). The higher the absolute value of the correlation of the PS with the diseases risk score, the closer the PS-regression estimator approximates the HR_c. Conversely, the lower the absolute value of this correlation, the closer the PS-regression estimator approximates the HR_m (Appendix 4).

The 1:1 matched conditional estimator (3.1) was least biased in our baseline scenario. This is because whenever a subject experienced an outcome or was censored the matched conditional analysis censored the subject's entire matched set. This minimized the extent to which the depletion of susceptibles was differential. 1:M matching stratifying on matched set (4.1) was more precise than 1:1 matching but more vulnerable to depletion of susceptibles.

If it is feasible to adjust for the individual covariates, doing so can adjust for the differential depletion of susceptibles and estimate the true HR_c consistently. Because this is not always feasible we are investigating methods for reducing the bias of the conditional PS-based estimators that periodically update the PS as the cohort is depleted by outcomes and censoring (Appendix 4). Marginal structural modeling is one of several approaches that successfully incorporates time-varying propensity scores to adjust for selection bias, and also bias due to informative censoring and time-dependent confounding (16, 21).

The IPTW estimators and our 1:1 matched estimator of the marginal ATT estimated their marginal targets with little bias (although the IPTW ATE estimate was biased when overlap was poor). The small amounts of bias observed tended to be away from 1 in the direction of the conditional HR_c because censoring -- despite being unassociated with treatment and covariates – biases these marginal estimators toward the HR_c. When follow-up was heavily censored this bias was substantial. Censoring – even when unassociated with treatment and covariates – leaves the estimated HR_m disproportionately influenced by the earlier events which are more likely than later events to be observed (i.e. uncensored) and which occur before there has been as much time for the HR_m to diverge from the HR_c.

Several findings can be discussed in terms of bias-precision tradeoffs that often arise in evaluations of methods. Among our PS-based estimators of the HR_c, the 1:1 matched estimator was the least biased, but since it was also the least precise the result of any single data analysis may be far from the truth. Among the PS-stratified conditional estimators, bias decreased as the number of strata increased, but overly fine stratification decreased precision. The use of finer rather than coarser strata also reduced

bias in the stratified estimators of the HR_m. The 1:1 matched estimator of the marginal ATT using un-stratified Cox regression (3.3), and the IPTW estimators of the ATT (6.1, 6.2), performed similarly well with respect to bias, but the 1:1 matched estimator was less precise because potentially informative comparators are not used in the analysis.

Privacy preserving PS-based tools that implement 1:1 matching or 1:M matching or PS stratification are already in use for Sentinel surveillance. These simulations extend our understanding of their strengths and limitations. The weighted estimators that used fine stratification or IPTW performed as well or better than the currently available estimators; it seems worthwhile for Sentinel and others users of PS-based methods to consider these weighted estimators.

Tables and Figures

Table 1. Features of Scenarios 1 – 21. The Conditional Treatment Effect is Homogeneous Across Patient-Level Characteristics and Site, Except Where Noted. Unmeasured Confounding and Informative Censoring Were Built into Scenarios 19 and 20, Respectively.

Scenario	Outcome Incidence	Treatment Prevalence	Unadj HR	Cond HR	Notes / Changes from baseline scenario
1	0.05	0.25	1	2	Baseline scenario
2	0.01	0.25	1	2	Low incidence
3	<0.01	0.25	1	2	Lower incidence
4	0.05	0.05	1	2	Lower treatment prevalence
5	0.05	0.5	1	2	Higher treatment prevalence
6	0.05	0.25	1	1.25	Smaller treatment effect
7	0.05	0.25	1	1.5	Small treatment effect
8	0.05	0.25	2	1	No treatment effect
9	0.05	0.25	1	2	Poor overlap in PS distributions for treated and comparator groups
10	0.05	0.25	0.4	0.8	Treatment reduces outcome risk
11	0.05	0.25	0.5	1	No effect of treatment, confounding in negative direction
12	0.05	0.25	0.625	1.25	Smaller treatment effect, confounding in negative direction
13	0.05	0.25	2.5	5	Larger treatment effect
14	0.05	0.25	1		Heterogeneity by site: HR=2 at smaller site, HR=1 at larger site
15	0.05	0.25	1		Heterogeneity by quintile of DRS: highest HR=2, others have HR=1
16	0.05	0.25	1		Heterogeneity by quintile of PS: lowest HR=2, others have HR=1
17	0.05	0.25	1		Heterogeneity by outcome subtype: common HR=0.8, rare HR=2
18	0.05	0.25	1		Heterogeneity by time-to-drug: 1st 30 days HR=2, then HR=1
19	0.05	0.25	1	2	Residual confounding*
20	0.05	0.25	1	2	Informative censoring**
21	0.015	0.25	1	2	Low incidence, early uninformative censoring

PS: Propensity Score, HR: Hazard Ratio, DRS: Disease Risk Score

* Residual confounding by a strong binary residual confounder correlated with the true PS.

** Informative censoring according to a censoring mechanism driven by a measured covariate and by a covariate not predictive of treatment.

Table 2. Variants of the Methods Examined. Each Method Estimates a Conditional or Marginal Hazard Ratio Corresponding to the Average Treatment Effect (ATE) or the Average Treatment Effect Among the Treated (ATT).

	Method	ID	Description	ATE	ATT
Conditional	Covariate Regression	1.1	Regress (T, Δ) on Z and X	✓	
		1.2	Fixed effects meta-analysis	✓	
	PS Regression	2.1	Adjust for polynomial PS terms (PS, PS^2, PS^3)	✓	
		2.2	Adjust for site-specific PS decile, $PS_{1,k,d}$	✓	
		2.3	Adjust for cubic B-spline PS deciles, $PS_{1,k,q}$	✓	
	1:1 Matching[a]	3.1	Stratify on site + matched set		✓
	1:M variable ratio matching[a]	4.1	Stratify on site + matched set		✓
	PS-stratification (10,20, fine) [a]	5.1	Stratify on site + PS strata	✓	✓
Marginal	1:1 Matching	3.2	Stratify on site only		✓
		3.3	No weighting or stratification		✓
	1:M variable ratio matching	4.2	Stratify on site + matching ratio		✓
		4.3	Weight, adjusting for site		✓ [b]
		4.4	Weight		✓ [b]
		4.5	Fixed effects meta-analysis		✓
	PS-stratification (10,20, fine)	5.2	Weight, adjust for site	✓ [c]	✓ [d]
		5.3	Weight	✓ [c]	✓ [d]
		5.4	Fixed effects meta-analysis	✓	✓
	IPTW (unstabilized or stabilized, with and without truncation)	6.1	Weight, adjust for site	✓ [e]	✓ [f]
		6.2	Weight	✓ [e]	✓ [f]
		6.3	Fixed effects meta-analysis	✓	✓

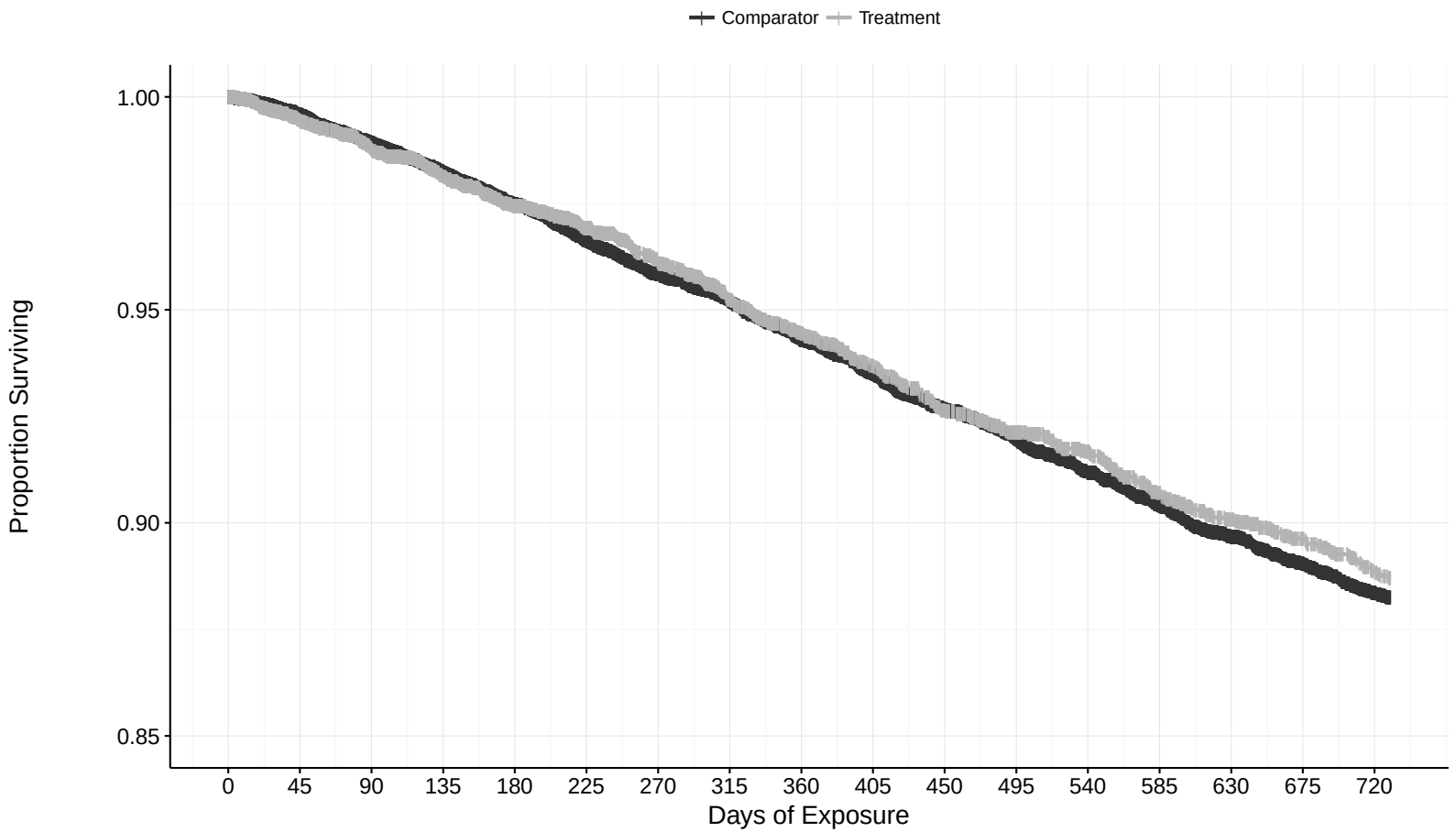
[a]Method is in production use for Sentinel surveillance. [b]ATT weights equal 1 for treated and $1/m_j$ for comparator subjects, where m_j denotes the number of comparators in matched set j. [c]*ATE weights equal* $1/n_{1,s,k}$ *for treated and* $1/n_{0,s,k}$ *for comparator subjects.* [d]*ATT weights equal 1 for treated and* $n_{1,s,k}/n_{0,s,k}$ *for comparator subjects, where* $(n_{1,s,k}, n_{0,s,k})$ *are the numbers of treated and comparator subjects in stratum s at site k.* [e]Unstabilized ATE weights equal 1/PS for treated subjects and 1/(1-PS) for comparator subjects. Stabilized IPTW weights for treated subjects = $\bar{Z}$ * $wt_{unstabilized}$, where $\bar{Z}$ is the proportion treated. Stabilized IPTW weights for comparator subjects = $(1 - \bar{Z})$ * $wt_{unstabilized}$. [f]Unstabilized ATT weights equal 1 for treated subjects and PS/(1-PS) for comparator subjects. Stabilized ATT weights = 1 for treated subjects and equal $(1 - \bar{Z})/\bar{Z}$ * $wt_{unstabilized}$ for untreated subjects. IPTW weights were truncated by setting any weight > 50 to 50.

Table 3. Scenario 1 Results. Mean Estimated Hazard Ratio ((HR)ˆ), Standard Error ((SE)ˆ), Monte Carlo Standard Deviation (MC SD), Bias, Mean Squared Error (MSE), and Coverage Of 95% Confidence Intervals for ATT And ATE Parameters. SE, SD, MSE Reported on the Log(HR) Scale. Bias (HR Scale), was Calculated With Respect to Conditional HR = 2 for Methods 1.1-5.1 (Top), and Marginal ATT= 1.647, Marginal ATE = 1.574 for Methods 3.2–6.3 (Bottom).

Method	ATT						ATE					
	$\widehat{HR}$	$\widehat{SE}$	MC SD	Bias	MSE	Coverage	$\widehat{HR}$	$\widehat{SE}$	MC SD	Bias	MSE	Coverage
Cov Reg 1.1	*						2.006	0.054	0.054	0.006	0.012	0.953
1.2	*						2.009	0.054	0.054	0.009	0.012	0.948
PS Reg 2.1	*						1.741	0.053	0.054	-0.259	0.076	0.249
2.2	*						1.659	0.053	0.053	-0.341	0.124	0.052
2.3	*						1.743	0.053	0.054	-0.257	0.075	0.251
1:1 Match 3.1	1.906	0.115	0.116	-0.094	0.059	0.915	*					
1:M Match 4.1	1.677	0.066	0.065	-0.323	0.116	0.215	*					
Strat-10 5.1	1.587	0.053	0.052	-0.413	0.177	0.004	1.656	0.053	0.053	-0.344	0.126	0.057
Strat-20 5.1	1.659	0.053	0.052	-0.341	0.124	0.058	1.702	0.053	0.053	-0.298	0.097	0.141
Strat-fine 5.1	1.742	0.057	0.057	-0.258	0.076	0.300	1.790	0.076	0.111	-0.210	0.078	0.710
1:1 Match 3.2	1.667	0.073	0.069	0.020	0.014	0.967	*					
3.3	1.667	0.073	0.069	0.020	0.014	0.966	*					
1:M Match 4.2	1.638	0.053	0.052	-0.010	0.007	0.949	*					
4.3	1.590	0.053	0.050	-0.057	0.010	0.916	*					
4.4	1.589	0.053	0.050	-0.058	0.010	0.917	*					
4.5	1.590	0.053	0.050	-0.057	0.009	0.917	*					
Strat-10 5.2	1.563	0.053	0.050	-0.084	0.013	0.857	1.558	0.065	0.061	-0.016	0.009	0.951
5.3	1.563	0.053	0.050	-0.084	0.013	0.856	1.557	0.065	0.062	-0.017	0.009	0.951
5.4	1.564	0.053	0.050	-0.083	0.013	0.861	1.560	0.053	0.049	-0.014	0.006	0.955
Strat-20 5.2	1.618	0.053	0.051	-0.029	0.008	0.942	1.558	0.066	0.062	0.014	0.010	0.958
5.3	1.618	0.053	0.051	-0.029	0.008	0.942	1.558	0.066	0.062	0.014	0.010	0.960
5.4	1.619	0.053	0.051	-0.028	0.007	0.942	1.591	0.053	0.049	0.017	0.006	0.963
Strat-fine 5.2	1.665	0.056	0.053	0.018	0.008	0.959	1.659	0.065	0.067	0.085	0.020	0.877
5.3	1.665	0.056	0.053	0.018	0.008	0.958	1.658	0.065	0.066	0.084	0.019	0.880
5.4	1.667	0.056	0.053	0.020	0.008	0.958	1.659	0.065	0.067	0.085	0.020	0.877
IPTW 6.1	1.666	0.054	0.050	0.019	0.007	0.955	1.609	0.054	0.050	0.035	0.008	0.951
stab-IPTW 6.1	1.667	0.054	0.050	0.020	0.007	0.954	1.610	0.054	0.050	0.036	0.008	0.951
IPTW 6.2	1.666	0.054	0.050	0.019	0.007	0.955	1.609	0.054	0.050	0.035	0.008	0.950
stab-IPTW 6.2	1.667	0.054	0.050	0.020	0.007	0.955	1.610	0.054	0.050	0.036	0.008	0.949
IPTW 6.3	1.667	0.054	0.050	0.020	0.007	0.954	1.610	0.054	0.050	0.036	0.008	0.949
stab-IPTW 6.3	1.668	0.054	0.050	0.021	0.007	0.954	1.611	0.054	0.050	0.037	0.008	0.949

*Methods to estimate these parameters were not included in this study.

Figure 1. Baseline Scenario 1 Distribution of Event and Censoring Times by Treatment Group (One Replicate Dataset, No Covariate Adjustment).



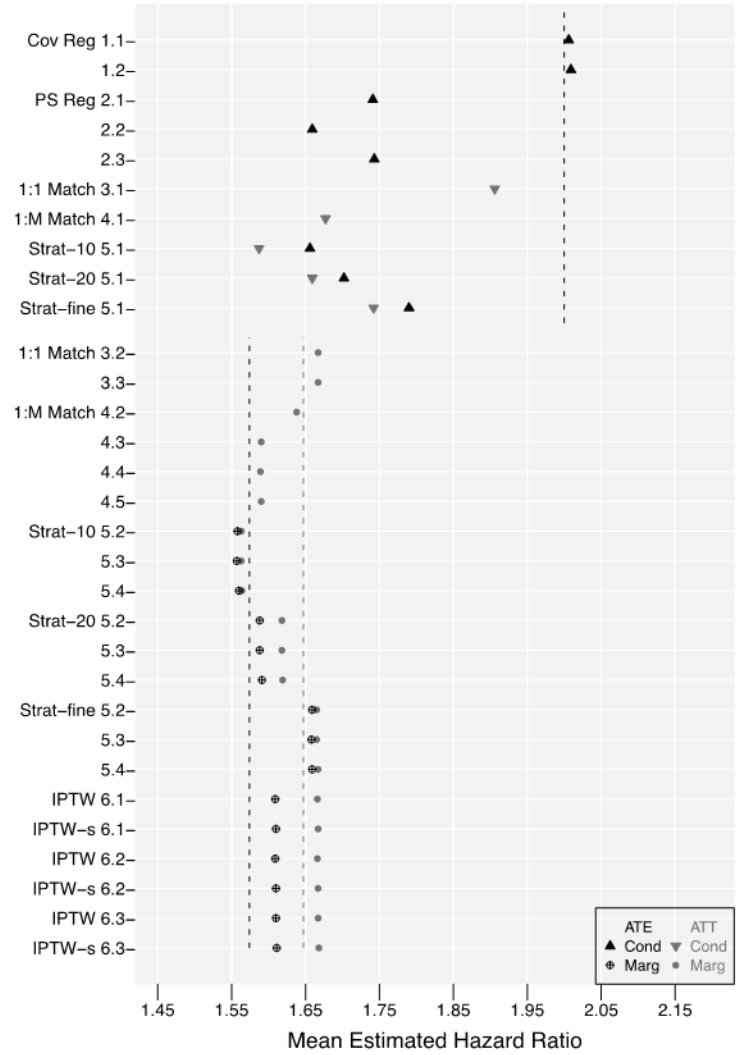
Number at risk (number of events)

Comparator	29944 (0)	22589 (114)	19860 (254)	16488 (374)	15289 (497)	14134 (629)	13084 (738)	12128 (823)	11265 (933)	10474 (1026)	9683 (1115)	8983 (1188)	8355 (1261)	7755 (1330)	7127 (1390)	6608 (1439)	6142 (1488)	
Treatment	9836 (0)	7388 (49)	6546 (95)	5458 (134)	5064 (171)	4704 (196)	4361 (234)	4037 (272)	3727 (305)	3472 (333)	3237 (371)	3009 (389)	2803 (403)	2608 (432)	2421 (449)	2231 (461)	2080 (480)	

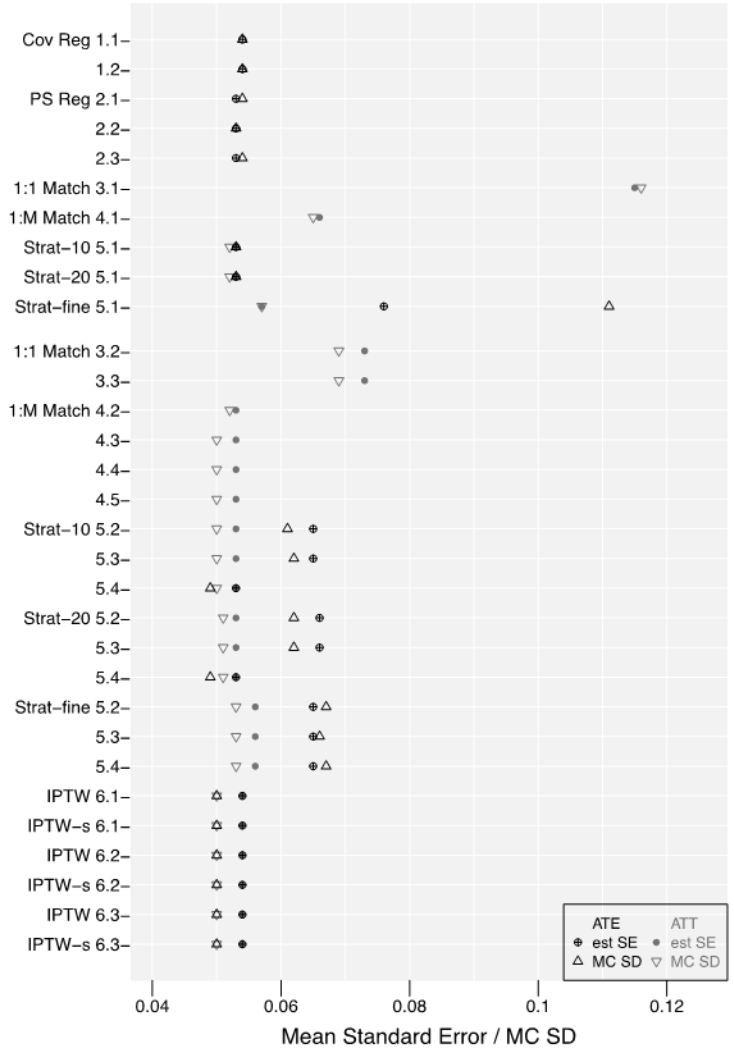
Cumulative number of censoring

Comparator	0	7241	9830	13082	14158	15181	16122	16993	17746	18444	19146	19773	20328	20859	21427	21897	22314	
Treatment	0	2399	3195	4244	4601	4936	5241	5527	5804	6031	6228	6438	6630	6796	6966	7144	7276	

Figure 2. Scenario 1 Results: Mean Estimated Hazard Ratios (A) Standard Errors (SE) and Monte Carlo Standard Deviations (MC SD) (B). HR c = 2, HRm ATE = 1.57, HRm ATT = 1.65.

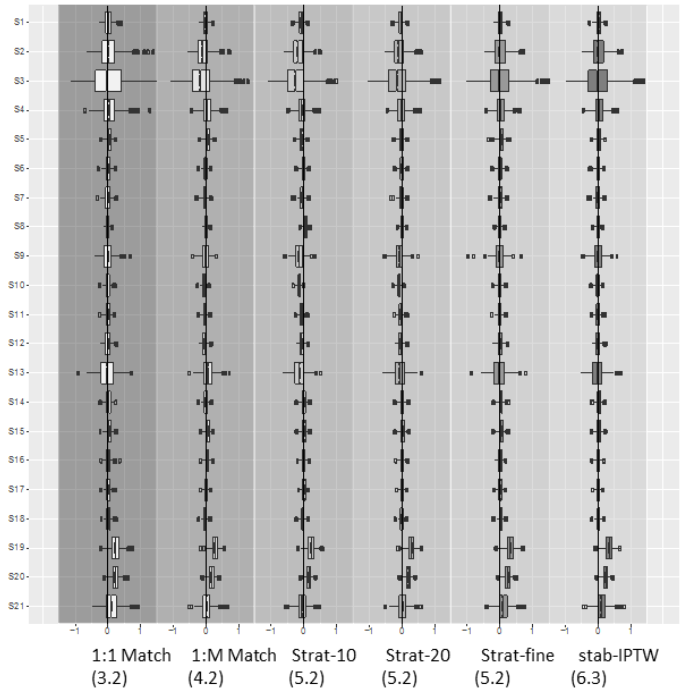


(A)

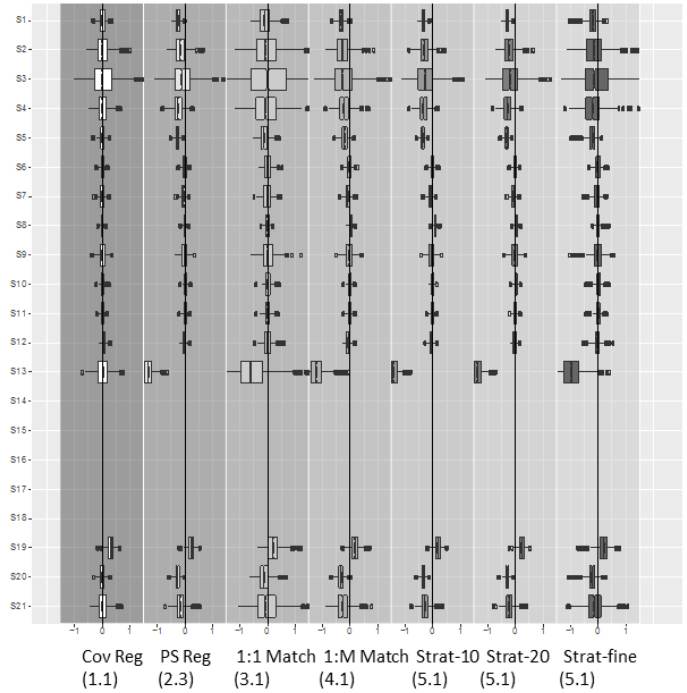


(B)

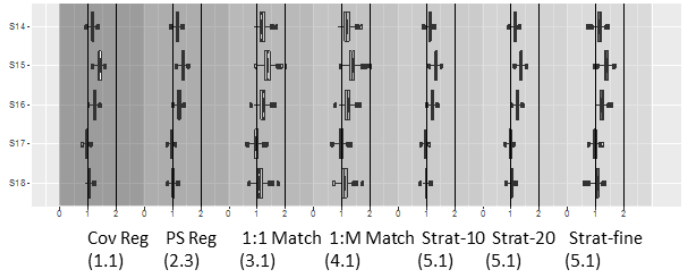
Figure 3. Bias in Marginal ATT Hazard Ratio Estimates for Matching, PS Stratification and IPTW Methods (A), Bias in Conditional Hazard Ratio Estimates for Covariate Adjusted Regression, PS Regression, Matching, PS Stratification Scenarios With a Homogeneous Treatment Effect (B), Estimated Conditional Effects for Covariate Adjusted Regression, PS Regression, Matching and Stratification Methods in Scenarios With a Heterogeneous Treatment Effect (C).



(A)



(B)



(C)

Acknowledgements

The authors would like to thank Lingling Li and the Sentinel Data Partners who provided data used in the analysis: Aetna (Aetna Informatics, Blue Bell, PA) and Humana (Humana, Inc., Comprehensive Health Insights, Miramar, FL).

This project was supported by Task Order HHSF22301012T under Master Agreement HHSF223200910006I from the US Food and Drug Administration (FDA). The FDA approved the study protocol including statistical analysis plan, and reviewed and approved this manuscript. Coauthors from the FDA participated in the results interpretation and in the preparation and decision to submit the manuscript for publication. The FDA had no role in data collection, management, or analysis.

References

1. Berman RE, Benner JS, Brown JS, et al. Developing the Sentinel System — A National Resource for Evidence Development. *N Engl J Med.* 2011;364(6):498-499.
2. Toh S, Reihman ME, Houstoun M, et al. Multivariable confounding adjustment in distributed data networks without sharing of patient-level data. *Pharmacoepidemiol Drug Saf*. 2013;22(11):1171-7. doi: 10.1002/pds.3483. Epub 2013 Jul 23.
3. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika.* 1983;70:41-55.
4. Schneeweiss S, Suissa S. Advanced Approaches to Controlling Confounding in Pharmacoepidemiologic Studies. In: Brian L. Strom, Stephen E. Kimmel, Sean Hennessy*, eds. Pharmacoepidemiology*, Fifth Edition. Hoboken, NJ: John Wiley & Sons, Ltd.; 2012: Chapter 47.
5. Austin PC. The performance of different propensity score methods for estimating marginal hazard ratios. *Stat Med.* 2013;32(16):2837-2849.
6. Hade EM, Lu B. Bias associated with using the estimated propensity score as a regression covariate. *Stat Med.* 2014;33(1):74–87. doi:10.1002/sim.5884.
7. Hernán MA. The hazards of hazard ratios. *Epidemiology.* 2010;21(1):13-15.
8. Franklin JM, Schneeweiss S, Polinski JM, et al. Plasmode simulation for the evaluation of pharmacoepidemiologic methods in complex healthcare databases. *Comput Stat Data Anal* 2014;72:219-226.
9. Chrischilles EA, Gagne JJ, Fireman B, et al. Prospective surveillance pilot of rivaroxaban safety within the US Food and Drug Administration Sentinel System. *Pharmacoepidemiol Drug Saf*. 2018;27(3): 263-271.
10. Evaluation of Propensity Score Based Methods in Sentinel Study Settings Using Simulation Experiments. *Sentinel Coordinating Center.* June 15, 2017. https://www.sentinelinitiative.org/sentinel/methods/evaluation-propensity-score-based-methods-sentinel-study-settings-using-simulation. Accessed March 28, 2018.
11. Bross IDJ. Spurious effects from an extraneous variable. *J. Chron. Dis.* 1966;19:637-647.
12. Cook AJ, Wellman RD, Izem R, et al. Comparison of Safety Signaling Methods for Survival Outcomes to Control for Confounding in the MSDD. The Sentinel Initiative. https://www.sentinelinitiative.org/sentinel/methods/safety-signaling-methods-survival-outcomes-control-confounding-msdd. Published October 13, 2015. Accessed March 30, 2018.
13. SAS version 9.3, SAS Institute, Cary NC.
14. Rosenbaum PR, Rubin DB. Reducing Bias in Observational Studies Using Subclassification on the Propensity Score. *J Am Stat Assoc.* 1984;79:516-524.
15. Hernán MA, Brumback B, Robins JM. Marginal structural models to estimate the causal effect of Zidovudine on the survival of HIV-positive men. *Epidemiology.* 2000;11:561-570.

16. Robins JM, Hernán MA, Brumback B. Marginal structural models and causal inference in epidemiology. *Epidemiology.* 2000;11:550-560.
17. Lin, D. Y. and Wei, L. J. The Robust Inference for the Proportional Hazards Model. *J Am Stat Assoc.* 1989;84:1074–1078.
18. Pirracchio R, Petersen ML, van der Laan M. Improving Propensity Score Estimators' Robustness to Model Misspecification Using Super Learner, *Am J Epidemiol.* 2015; 181(2):108–119. doi:10.1093/aje/kwu253.
19. Lee BK, Lessler J, Stuart EA. Improving propensity score weighting using machine learning. *Stat Med.* 2010;29(3):337-346. doi:10.1002/sim.3782.
20. Hansen BB. The prognostic analogue of the propensity score. *Biometrika.* 2008;95(2):481-488.
21. M Petersen, J Schwab, S Gruber, et al. Targeted maximum likelihood estimation for dynamic and static longitudinal marginal structural working models. *J Causal Inference*. 2014; 2(2):147-185.

Appendix 1. Simulation Studies with Time to Event Outcomes

Data Generation

Coefficients in the propensity score (PS) models were based on covariate-treatment relationships in the claims data. We fit a logistic model regressing treatment Z on 15 pre-selected covariates. Coefficient values were then modified to produce the desired scenario-specific level of exposure (Table A1.1). Coefficients for PS Models 1-4 were set to the estimated coefficients in the model fit to the claims data, except for the interept. PSs in all scenarios were designed to provide reasonable overlap except for PS Model 7 used in Scenario 9 (see Appendix 3 for sample PS diagnostics).

Table A 1. 1 Propensity Score Model Coefficients Used in Scenarios 1 – 21. Coefficients in PS Models 1-4 Equal the Coefficient Estimates Obtained When Fitting the PS Logistic Regression Model to the Claims Data, With a Modified Intercept.

	Claims	Propensity Score Model*								
	Data	1	2	3	4	5	6	7	8	9
Intercept	1.08	0.682	-	1.825	0.6815	2.28	5.62	2.551	1.78	0.647
Age	-0.02					-	-	-0.042		-0.02
Ambul-Visits	-0.01					-	-	-0.021		-0.01
Outpt-Visits	-0.01					-	-	-0.021		-0.01
Inpt-Visits	0.06					0.093	0.24	0.126		0.059
ComorbidScore	-0.07					-	-	-0.147		-
Atrial flutter	0.33					0.217	1.32	0.693		0.323
Diabetes	-0.19					-	-	-0.399		-
Prior GI Bleed	-0.14					-	-	-0.294		-
Prior MI	-0.13					-	-	-0.273		-
Renal Disease	-0.08					-	-	-0.168		-
Diuretic	-0.22					-	-	-0.462		-
DrugClasses	0.03					0.093	0.12	0.063		0.029
Prior Ischemic	-0.12					-	-	-0.252		-
Prior Bleed	-0.21					-	-	-0.441		-
Sex	-0.07					-	-	-0.147		-

* Model 1: Scenarios 1-3, 20, 21; Model 2: Scenario 4; Model 3: Scenario 5; Model 4: Scenarios 6, 7, 13-15, 17, 18; Model 5: Scenario 8; Model 6: Scenario 9; Model 7: Scenarios 10-12; Model 8: Scenario 19; Model 9: Scenario 16.

Coefficients in the Weibull model used to generate outcome event times were also inspired by outcome-covariate relationships in the claims data. We fit a logistic model regressing outcome Y on 15 pre-determined covariates. These covariates were modified to produce the conditional scale parameters to yield the desired event times and level of confounding for each scenario (Table A1.2). The shape parameter for event time generation was set to 1.5 (increasing risk over time).

Censoring patterns were also motivated by the distribution of censoring times in the claims data, in which spikes at 30-day intervals were observed. Aside from uninformative censoring at a scenario-specific background rate subjects could be censored at 36, 66, or 96 days.

Table A 1. 2 Estimated Coefficients in Claims Data Regression Model and Outcome Event Time Model Coefficients Used In the Weibull Model for Each Scenario: A) Scenarios 1 – 10, and B) Scenarios 11 – 21.

	Claims	Scenario									
	Data	1	2	3	4	5	6	7	8	9	10
Intercept	-6.22	-20.90	-20.10	-20.70	-20.45	-19.50	-13.58	-15.61	-6.75	-14.56	-17.54
Treatment		0.693	0.693	0.693	0.693	0.693	0.223	0.405	0.000	0.693	-0.223
Age	0.02	0.071	0.033	0.024	0.072	0.051	0.012	0.020	-0.080	0.014	0.016
Ambul-Visits	-0.01	-0.016	-0.016	-0.016	-0.016	-0.016	-0.006	-0.010	-0.015	-0.007	0.025
Outpt-Visits	0.01	0.016	0.016	0.016	0.016	0.016	0.006	0.010	-0.008	0.007	0.015
Inpt-Visits	0.08	0.124	0.124	0.124	0.124	0.124	0.050	0.081	-0.300	0.057	0.120
ComorbScore	0.09	0.430	0.430	0.430	0.430	0.430	0.172	0.280	-0.338	0.198	0.405
Atrial flutter	0.16	0.248	0.248	0.248	0.248	0.248	0.099	0.161	0.008	0.114	0.240
Diabetes	0.03	0.046	0.046	0.046	0.047	0.047	0.019	0.300	0.008	0.021	0.045
Prior GI Bleed	0.81	1.700	1.700	1.700	1.700	1.700	0.680	1.105	-0.638	0.782	1.215
Prior MI	0.26	0.403	0.403	0.403	0.403	0.403	0.161	0.262	0.005	0.185	0.390
Renal Disease	0.14	0.217	0.217	0.217	0.217	0.217	0.087	0.141	0.015	0.100	0.210
Diuretic	0.20	0.310	0.310	0.310	0.310	0.310	0.124	0.202	0.030	0.143	0.300
DrugClasses	0.04	0.062	0.062	0.062	0.062	0.062	0.025	0.040	0.150	0.029	0.060
Prior Ischemic	0.08	0.124	0.124	0.124	0.124	0.124	0.050	0.081	-0.008	0.057	0.120
Prior Bleed	0.07	0.109	0.109	0.109	0.109	0.109	0.043	0.071	-0.030	0.050	0.105
Sex	-0.05	-0.078	-0.078	-0.078	-0.078	-0.078	-0.031	-0.050	0.075	-0.036	-0.075

(A)

	Scenario										
	11	12	13	14	15	16	17	18	19	20	21
Intercept	-17.54	-17.16	-18.63	-15.48	-18.82	-13.37	-11.80	-17.63	-16.03	-21.94	-20.52
Treatment	0.000	0.223	1.609	(heterogeneous by subgroup: 0 or 0.693)					0.693	0.693	0.693
Age	0.040	0.040	0.047	0.020	0.015	0.009	0.000	0.031	0.016	0.090	0.066
Ambul-Visits	-0.020	-0.020	-0.024	-0.010	-0.007	-0.005	0.000	-0.016	-0.008	-0.020	-0.014
Outpt-Visits	0.020	0.020	0.024	0.010	0.007	0.005	0.000	0.016	0.008	0.020	0.014
Inpt-Visits	0.160	0.160	0.188	0.079	0.059	0.037	-0.001	0.124	0.064	0.156	0.116
ComorbScore	0.180	0.180	0.212	0.275	0.205	0.128	-0.004	0.430	0.224	0.542	0.401
Atrial flutter	0.320	0.320	0.376	0.159	0.118	0.074	-0.002	0.248	0.129	0.313	0.231
Diabetes	0.060	0.060	0.071	0.298	0.022	0.014	0.000	0.047	0.024	0.059	0.043
Prior GI Bleed	1.620	1.620	1.904	1.088	0.811	0.506	-0.017	1.700	0.884	2.144	1.586
Prior MI	0.520	0.520	0.611	0.258	0.192	0.120	-0.004	0.403	0.210	0.508	0.376
Renal Disease	0.280	0.280	0.329	0.139	0.104	0.065	-0.002	0.217	0.113	0.274	0.202
Diuretic	0.400	0.400	0.470	0.198	0.148	0.092	-0.003	0.310	0.161	0.391	0.289
DrugClasses	0.080	0.080	0.094	0.040	0.030	0.018	-0.001	0.062	0.032	0.078	0.058
Prior Ischemic	0.160	0.160	0.188	0.079	0.059	0.037	-0.001	0.124	0.064	0.156	0.116
Prior Bleed	0.140	0.140	0.165	0.069	0.052	0.032	-0.001	0.109	0.056	0.137	0.101
Sex	-0.100	-0.100	-0.118	-0.050	-0.037	-0.023	0.001	-0.078	-0.040	-0.098	-0.072

(B)

Probabilities of being censored were drawn from a multinomial distribution: each subject has a 68.73% chance of being censored according to some background mechanism, a 19.67% chance of being censored at 36 days, a 4.47% chance of being censored at 66 days, and a 7.13% chance of being censored at 96 days. Background censoring followed an exponential distribution with the rate set to provide the desired event incidence (scenarios 1-19, 21). Informative censoring for scenario #20 was a function of *ComorbidityScore* and *Diuretic.*

Simulation Studies

Data Analysis

Each method was applied in Monte Carlo simulation studies (1000 replicates per scenario). PSs were estimated separately at each site using correctly specified models. These estimated PSs were incorporated into each PS-based method to estimate a HR as described in Appendix 3. Robust sandwich estimators of the variance of the estimated HRs were used for PS-regression analyses 2.1-2.3 and for all methods that estimate a marginal (ATE or ATT) effect. Covariate adjusted regression models provide a performance benchmark.

Results

Mean HR estimates, mean estimated standard errors (SE), and Monte Carlo standard deviation (MCSD) for each scenario are plotted in Figs. A1.1- A1.7. These values, mean bias, mean squared error (MSE), and coverage of 95% confidence intervals are provided in Tables A1.3- A1.23. Blank rows in the tables correspond to methods that were not included in our study.

Box plots illustrate the relative performance of methods across scenarios 1- 21 with time to event outcomes (Figs. A1.8- A1.14). Where possible bias was plotted, and ideal boxplots would be centered at 0. In scenarios 14-18 with heterogeneous treatment effect it is not possible to calculate bias with respect to a conditional HR, so estimates are plotted instead of bias.

In a hypothesis testing framework the relevant question is whether treatment effects the outcome of interest. Rejection rates for a two-sided test of the null hypothesis of no treatment effect ($\alpha = 0.05$) are plotted for each method (Fig. A1.15, Tables A1.24, A1.25). These rates demonstrate the relative power of each method to reject the null hypothesis of no treatment effect, and also reveal incorrect rejection of the null hypothesis in scenarios 8 and 11 where the null does not hold. Rejection rates < 0.80 are bolded for methods that had less than 80% power to detect an effect in scenarios where there is a non-null treatment effect.

The null holds for scenarios 8 and 11, so we'd expect a rejection rate of approximately 5%. In heterogeneous scenarios 17 and 18 although 20% of the population is at increased risk, the marginal HR is quite close to 1, so it is difficult to reject the null hypothesis at this sample size.

Discussion

With some exceptions, performance within each type of method was fairly robust with respect to implementation choices.

- Variants of 1:1 matching had similar performance whether estimating a conditional effect or a marginal ATT (Figs. A1.8, A1.10).
- Variants of 1:M variable ratio matching had similar performance whether estimating a conditional effect or a marginal ATT (Figs. A1.8, A1.10).
- PS regression method 2.2 (using categorical PS terms) was slightly more biased than the other PS regression variants (adjusting for polynomial of B-spline PS terms) when estimating a marginal ATT (Fig. A1.9). More flexibly modeling the PS term in the Cox regression improved performance.
- Variants of PS stratification using a fixed number of strata had similar performance, with meta-analysis 5.4 having less bias in Scenarios 2 and 3 than the other variants for estimating a marginal ATT or marginal ATE (Fig. A1.10 - A1.12). Finer stratification typically improved performance, except when overly fine strata were created.
- All variants of IPTW to estimate a marginal ATT had similar performance (Fig. A1.13). There were some differences when estimating a marginal ATE (Fig. A1.14). In rare exposure scenario 4 and poor overlap scenario 9 weight stabilization decreased finite sample bias.

Relative Performance in Response to Challenges

Low Incidence

In scenarios 1-3 as the event rate decreases (5%, 1%, .01%) the variance of the estimates grew. In the most challenging rare event scenario 3, 1:1 Matching remains an unbiased estimator of both the conditional and marginal HR. PS stratification using fine strata and IPTW remain unbiased estimators of the marginal ATT. As the box plots illustrate, when the event rate is low the result from any single data analysis is likely to be further away from the truth even when the estimator is unbiased.

Rare Exposure

In scenario 4 only 5% of the population were treated with the study drug. 1:1 Matching is successful at estimating a marginal ATT, but has large bias when estimating a conditional effect. Fine stratification and IPTW to estimate a marginal ATT have bias centered at 0 and narrow box plots. Unstabilized IP weights ranged from 1 to 93.34 (Appendix 2), thus in the analyses some weights were truncated at 50, which helped reduce finite sample bias.

Common Exposure

In scenario 5 50% of the population were treated with the study drug. Box plots of bias for methods that estimate a conditional effect were narrow, but centered to the left of zero. Those for methods that estimate a marginal ATT were narrow and centered on zero.

Small Effect Size

In scenarios 6 and 7 the true conditional HRs were 1.25 and 1.5, with marginal HR even closer to the null. All box plots are narrow and centered at 0.

Null Treatment Effect

In scenarios 8 and 11 the true conditional and marginal HRs are 1, and all methods had similar performance.

Poor Overlap

In scenario 9 there is poor overlap in the PS distributions in treatment and comparator populations. Despite this, there are comparator subjects with PSs like those observed in the treated group. All methods were fairly successful, however 1:1 matching produced more outlying HR estimates than other methods. PS-stratification using 10 or 20 strata was biased for the marginal ATT, while fine stratification was unbiased. However, fine stratification produced more outlying HR estimates when estimating a conditional effect.

Negative Confounding

In scenario 10 treatment is protective, while in scenario 12 treatment leads to a mild risk increase but a crude analysis would produce a misleading HR indicating that treatment was protective. The direction of confounding had no impact on the relative performance of methods.

Large Treatment Effect

In scenario 13 although the true conditional HR is 5, methods that estimate a conditional effect are highly biased, with box plots centered well to the left of 0. Methods that estimate a marginal ATT are more successful, particularly 1:1 matching (though the box plot is a little wide), fine stratification, and IPTW.

Heterogeneous Treatment Effect

In scenarios 14-18 methods that estimate a conditional effect appear to be estimating a weighted average of the conditional HRs in each subgroup or time-period. PS stratification methods had the narrowest box plots. All methods that estimate a marginal ATT performed well in these heterogeneous scenarios.

Residual Bias

All methods were biased in scenario 19 due to unmeasured confounding. They were also biased in scenario 20 due to informative censoring.

Strong Early Censoring

A great deal of early censoring caused most PS-based methods to miss their targets in scenario 21, despite it being uninformative.

Rejection Rates

In rejection rates were quite similar for all methods except in scenarios 3, 8, 10, 11, 12, 14, 16. Important findings include low power for all methods when the outcome incidence is low (scenario 3), or when the treatment effect is close to the null (scenarios 14, 17, 18). In scenarios 8 and 11 the true HR is 1. Ideally methods would reject at the nominal rate of 5%, however PS regression using deciles erroneously rejected the null hypothesis 30% of the time.

Overall, 1:1 Matching and fine stratification were the least biased methods for estimating a conditional effect. 1:1 Matching had higher variability, however 1:M matching was more biased. In our scenarios fine stratification and IPTW had the best overall performance for estimating a marginal ATT. This may partly be due to the fact that there was good overlap in all scenarios except 9. In this scenario estimating the marginal ATE was challenging for IPTW. However, there was sufficient support in the data for estimating the ATT.

Figure A 1. 1 Box Plots of Bias in Covariate Adjusted Regression (A) and PS Regression (B) Conditional HR Estimates for Scenarios With a Homogeneous Treatment Effect. Box Plots of Conditional HR Estimates for Scenarios With a Heterogeneous Treatment Effect are Shown for Covariate Adjusted Regression (C) and PS Regression (D), Bias In Marginal ATT HR Estimates From All Variants of 1:1 Matching and 1:M Variable Ratio Matching (E).

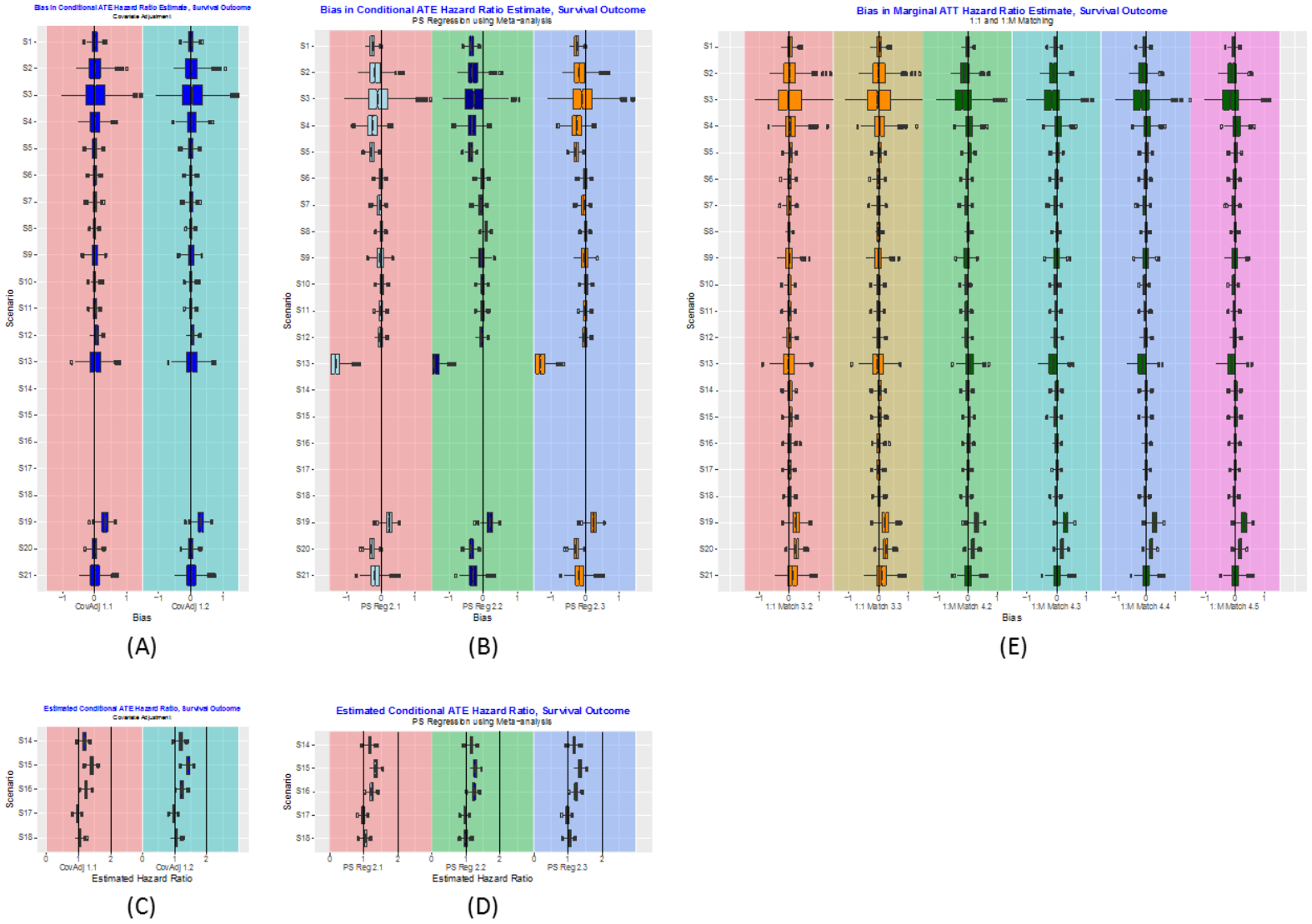


Figure A 1. 2 Box Plots of Bias in Marginal HR Estimates From All Variants of PS-Stratification Using 10 Strata, ATT (A) and ATE (B), and Using 20 Strata ATT(C) ATE (D).

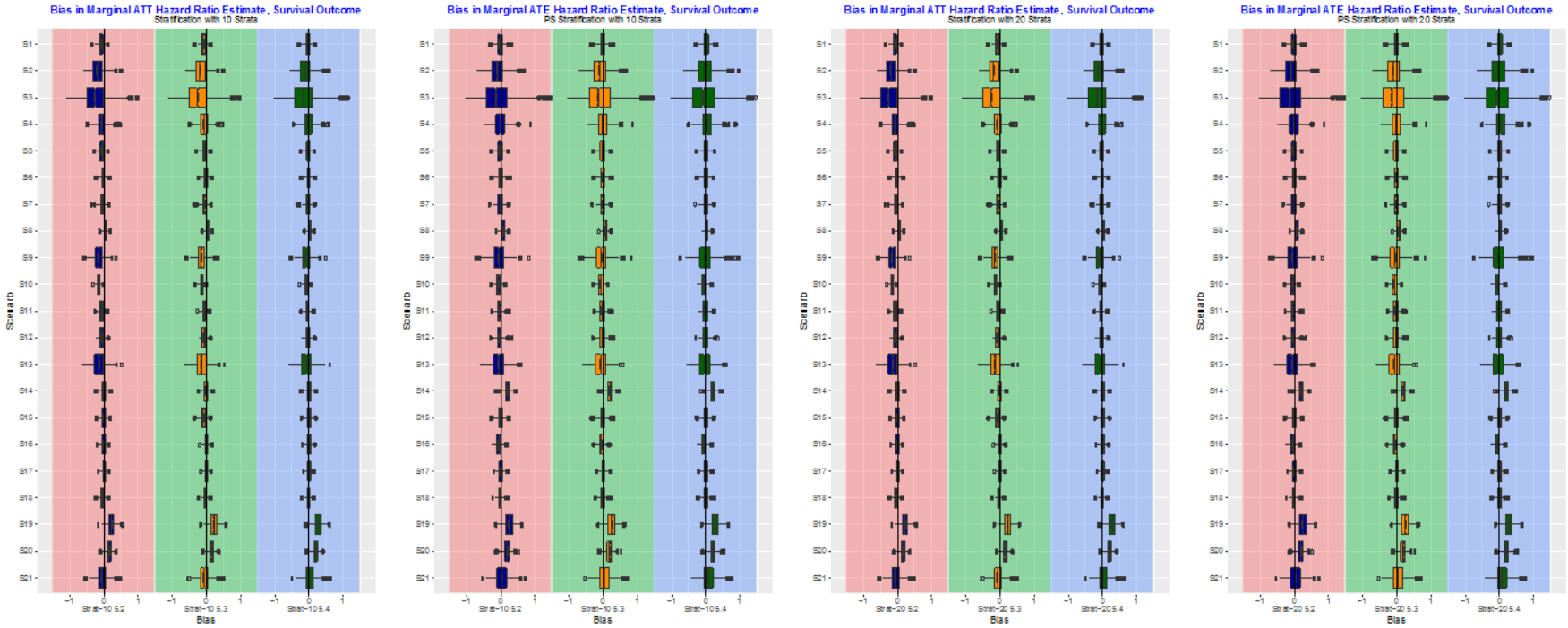


Figure A 1. 3 Box Plots of Bias in Marginal HR Estimates From All Variants of PS-Stratification Using Fine Strata, ATT (A) and ATE (B).

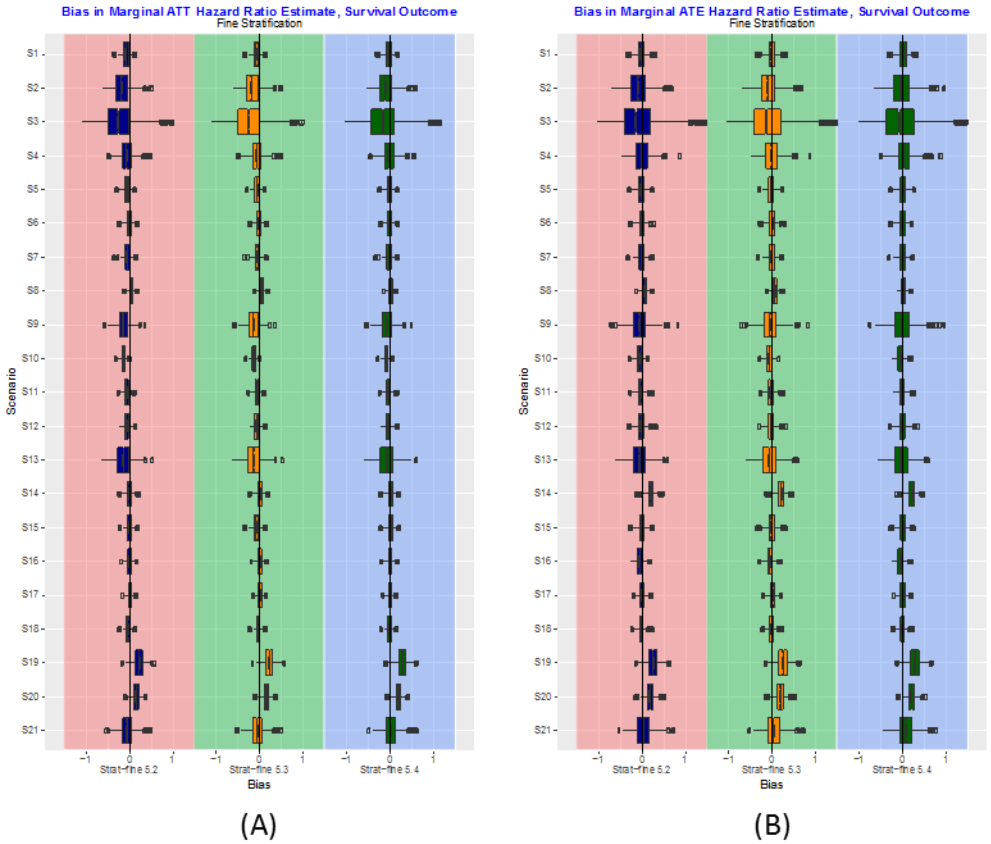


Figure A 1. 4 Box Plots of Bias in Marginal HR Estimates From All Variants of Stabilized and Unstabilized IPTW Methods, ATT (A), ATE (B).

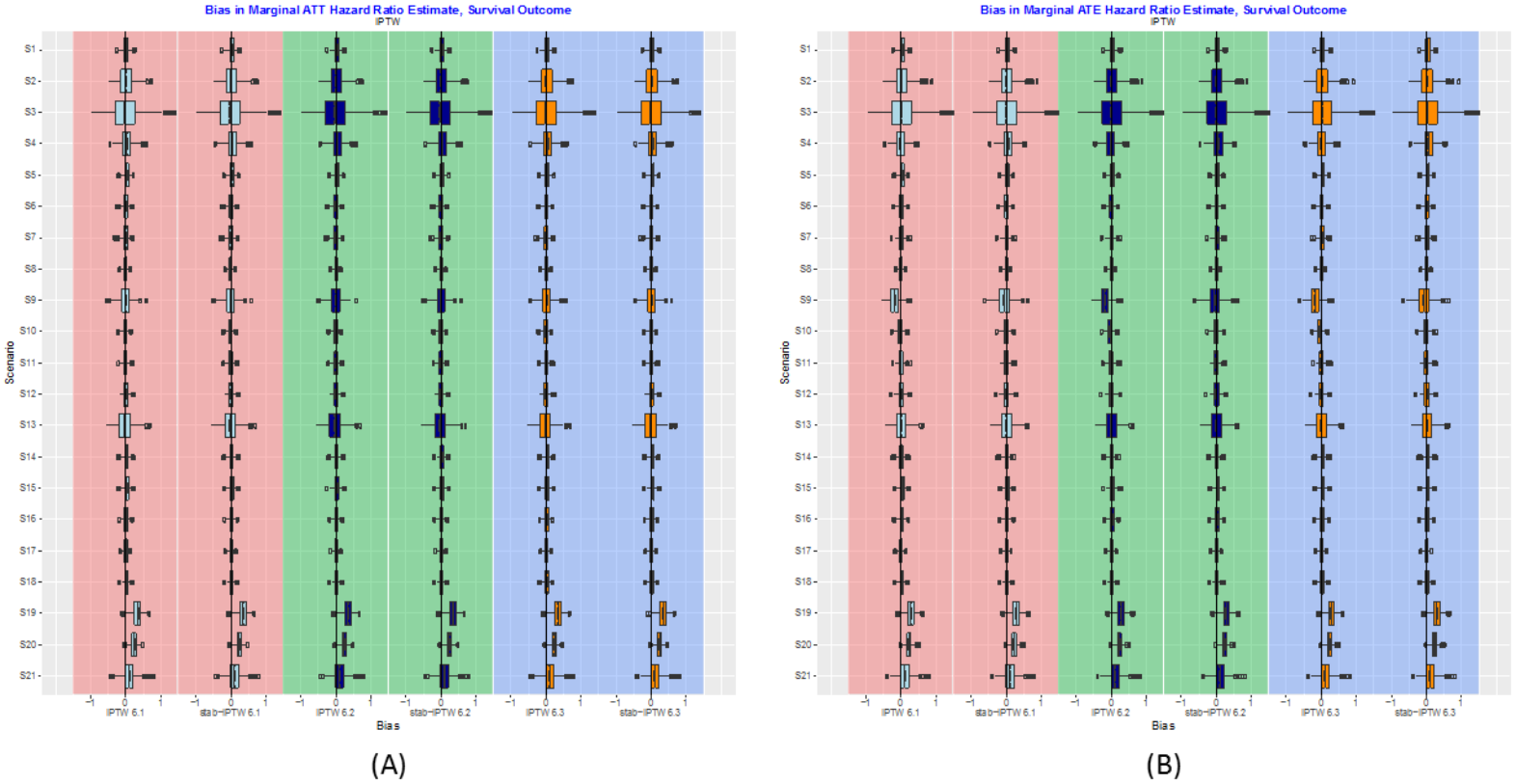


Table A 1. 3 Scenario 1 Results. Mean Estimated Hazard Ratio ($\widehat{HR}$), Standard Error ($\widehat{SE}$), Monte Carlo Standard Deviation (MC SD), Bias, Mean Squared Error (MSE), And Coverage of 95% Confidence Intervals for ATT and ATE Parameters. SE and SD are Reported on the Log HR Scale. Bias and MSE on the HR Scale were Calculated With Respect to the True Conditional HR = 2 for Methods in Top Section of the Table, and With Respect to Marginal ATT= 1.647, Marginal ATE = 1.574 for Remaining Methods.

Method	ATT						ATE					
	$\widehat{HR}$	$\widehat{SE}$	MC SD	Bias	MSE	Coverage	$\widehat{HR}$	$\widehat{SE}$	MC SD	Bias	MSE	Coverage
Cov Reg 1.1							2.006	0.054	0.054	0.006	0.012	0.953
1.2							2.009	0.054	0.054	0.009	0.012	0.948
PS Reg 2.1							1.741	0.053	0.054	-0.259	0.076	0.249
2.2							1.659	0.053	0.053	-0.341	0.124	0.052
2.3							1.743	0.053	0.054	-0.257	0.075	0.251
1:1 Match 3.1	1.906	0.115	0.116	-0.094	0.059	0.915						
1:M Match 4.1	1.677	0.066	0.065	-0.323	0.116	0.215						
Strat-10 5.1	1.587	0.053	0.052	-0.413	0.177	0.004	1.656	0.053	0.053	-0.344	0.126	0.057
Strat-20 5.1	1.659	0.053	0.052	-0.341	0.124	0.058	1.702	0.053	0.053	-0.298	0.097	0.141
Strat-fine 5.1	1.742	0.057	0.057	-0.258	0.076	0.300	1.790	0.076	0.111	-0.210	0.078	0.710
1:1 Match 3.2	1.667	0.073	0.069	0.020	0.014	0.967						
3.3	1.667	0.073	0.069	0.020	0.014	0.966						
1:M Match 4.2	1.638	0.053	0.052	-0.010	0.007	0.949						
4.3	1.590	0.053	0.050	-0.057	0.010	0.916						
4.4	1.589	0.053	0.050	-0.058	0.010	0.917						
4.5	1.590	0.053	0.050	-0.057	0.009	0.917						
Strat-10 5.2	1.563	0.053	0.050	-0.084	0.013	0.857	1.558	0.065	0.061	-0.016	0.009	0.951
5.3	1.563	0.053	0.050	-0.084	0.013	0.856	1.557	0.065	0.062	-0.017	0.009	0.951
5.4	1.564	0.053	0.050	-0.083	0.013	0.861	1.560	0.053	0.049	-0.014	0.006	0.955
Strat-20 5.2	1.618	0.053	0.051	-0.029	0.008	0.942	1.558	0.066	0.062	0.014	0.010	0.958
5.3	1.618	0.053	0.051	-0.029	0.008	0.942	1.558	0.066	0.062	0.014	0.010	0.960
5.4	1.619	0.053	0.051	-0.028	0.007	0.942	1.591	0.053	0.049	0.017	0.006	0.963
Strat-fine 5.2	1.665	0.056	0.053	0.018	0.008	0.959	1.659	0.065	0.067	0.085	0.020	0.877
5.3	1.665	0.056	0.053	0.018	0.008	0.958	1.658	0.065	0.066	0.084	0.019	0.880
5.4	1.667	0.056	0.053	0.020	0.008	0.958	1.659	0.065	0.067	0.085	0.020	0.877
IPTW 6.1	1.666	0.054	0.050	0.019	0.007	0.955	1.609	0.054	0.050	0.035	0.008	0.951
stab-IPTW 6.1	1.667	0.054	0.050	0.020	0.007	0.954	1.610	0.054	0.050	0.036	0.008	0.951
IPTW 6.2	1.666	0.054	0.050	0.019	0.007	0.955	1.609	0.054	0.050	0.035	0.008	0.950
stab-IPTW 6.2	1.667	0.054	0.050	0.020	0.007	0.955	1.610	0.054	0.050	0.036	0.008	0.949
IPTW 6.3	1.667	0.054	0.050	0.020	0.007	0.954	1.610	0.054	0.050	0.036	0.008	0.949
stab-IPTW 6.3	1.668	0.054	0.050	0.021	0.007	0.954	1.611	0.054	0.050	0.037	0.008	0.949

Table A 1. 4 Scenario 2 Results. Mean Estimated Hazard Ratio ($\widehat{HR}$), Standard Error ($\widehat{SE}$), Monte Carlo Standard Deviation (MC SD), Bias, Mean Squared Error (MSE), And Coverage of 95% Confidence Intervals for ATT and ATE Parameters. SE and SD are Reported on the Log HR Scale. Bias and MSE on the HR Scale were Calculated With Respect to the True Conditional HR = 2 For Methods In Top Section of The Table, and With Respect to Marginal ATT= 1.798, Marginal ATE = 1.734 for Remaining Methods.

Method	ATT						ATE					
	$\widehat{HR}$	$\widehat{SE}$	MC SD	Bias	MSE	Coverage	$\widehat{HR}$	$\widehat{SE}$	MC SD	Bias	MSE	Coverage
Cov Reg 1.1							2.020	0.122	0.124	0.020	0.065	0.951
1.2							2.031	0.123	0.125	0.031	0.067	0.952
PS Reg 2.1							1.842	0.120	0.121	-0.158	0.075	0.897
2.2							1.705	0.119	0.119	-0.295	0.129	0.736
2.3							1.843	0.120	0.121	-0.157	0.075	0.898
1:1 Match 3.1	1.993	0.267	0.269	-0.007	0.305	0.950						
1:M Match 4.1	1.723	0.147	0.152	-0.277	0.146	0.799						
Strat-10 5.1	1.613	0.119	0.119	-0.387	0.187	0.536	1.694	0.119	0.114	-0.306	0.131	0.722
Strat-20 5.1	1.704	0.119	0.119	-0.296	0.129	0.739	1.769	0.120	0.120	-0.231	0.099	0.824
Strat-fine 5.1	1.834	0.129	0.128	-0.166	0.083	0.895	1.865	0.195	0.216	-0.135	0.179	0.885
1:1 Match 3.2	1.828	0.170	0.165	0.030	0.094	0.964						
3.3	1.828	0.170	0.165	0.030	0.094	0.964						
1:M Match 4.2	1.698	0.119	0.120	-0.100	0.052	0.916						
4.3	1.690	0.120	0.119	-0.108	0.052	0.919						
4.4	1.690	0.120	0.119	-0.108	0.052	0.919						
4.5	1.694	0.120	0.118	-0.104	0.051	0.921						
Strat-10 5.2	1.622	0.119	0.119	-0.176	0.068	0.862	1.643	0.146	0.138	-0.092	0.060	0.947
5.3	1.622	0.119	0.119	-0.176	0.068	0.862	1.642	0.146	0.139	-0.092	0.060	0.948
5.4	1.626	0.120	0.119	-0.172	0.067	0.866	1.636	0.120	0.114	-0.098	0.044	0.935
Strat-20 5.2	1.706	0.120	0.119	-0.092	0.050	0.925	1.716	0.148	0.146	-0.019	0.063	0.950
5.3	1.706	0.120	0.119	-0.092	0.050	0.925	1.715	0.148	0.146	-0.019	0.063	0.951
5.4	1.711	0.120	0.119	-0.087	0.049	0.929	1.705	0.122	0.120	-0.029	0.043	0.950
Strat-fine 5.2	1.809	0.128	0.124	0.011	0.051	0.959	1.778	0.149	0.146	0.044	0.068	0.960
5.3	1.809	0.128	0.124	0.011	0.051	0.958	1.778	0.148	0.146	0.044	0.068	0.958
5.4	1.816	0.127	0.124	0.018	0.052	0.959	1.779	0.148	0.146	0.044	0.069	0.954
IPTW 6.1	1.812	0.121	0.121	0.014	0.048	0.958	1.754	0.125	0.125	0.020	0.049	0.944
stab-IPTW 6.1	1.813	0.121	0.121	0.015	0.048	0.958	1.755	0.125	0.125	0.021	0.049	0.944
IPTW 6.2	1.812	0.121	0.121	0.014	0.048	0.957	1.754	0.125	0.125	0.020	0.049	0.946
stab-IPTW 6.2	1.813	0.121	0.121	0.015	0.048	0.957	1.755	0.125	0.125	0.021	0.049	0.946
IPTW 6.3	1.817	0.121	0.120	0.019	0.048	0.956	1.758	0.124	0.124	0.024	0.048	0.947
stab-IPTW 6.3	1.818	0.121	0.120	0.020	0.048	0.956	1.759	0.125	0.124	0.025	0.049	0.947

Table A 1. 5 Scenario 3 results. Mean Estimated Hazard Ratio ($\widehat{HR}$), Standard Error ($\widehat{SE}$), Monte Carlo Standard Deviation (MC SD), Bias, Mean Squared Error (MSE), And Coverage of 95% Confidence Intervals for ATT and ATE Parameters. SE and SD are Reported on the Log HR Scale. Bias and MSE on the HR Scale were Calculated With Respect to the True Conditional HR = 2 for Methods in Top Section of the Table, and With Respect to Marginal ATT = 1.894, marginal ATE = 1.815 for Remaining Methods.

Method	ATT						ATE					
	$\widehat{HR}$	$\widehat{SE}$	MC SD	Bias	MSE	Coverage	$\widehat{HR}$	$\widehat{SE}$	MC SD	Bias	MSE	Coverage
Cov Reg 1.1							2.043	0.215	0.221	0.0430	0.203	0.949
1.2							2.071	0.219	0.224	0.0710	0.216	0.952
PS Reg 2.1							1.926	0.212	0.219	-0.0740	0.180	0.944
2.2							1.751	0.210	0.217	-0.2490	0.202	0.909
2.3							1.927	0.212	0.219	-0.0730	0.181	0.944
1:1 Match 3.1	2.4e+8	119.245	0.959	2.4e+8	5.8e+19	0.954						
1:M Match 4.1	1.773	0.258	0.267	-0.227	0.279	0.909						
Strat-10 5.1	1.641	0.208	0.218	-0.359	0.254	0.823	1.741	0.209	0.219	-0.2589	0.209	0.894
Strat-20 5.1	1.742	0.209	0.219	-0.258	0.209	0.895	1.818	0.210	0.219	-0.1823	0.189	0.916
Strat-fine 5.1	1.914	0.225	0.233	-0.086	0.207	0.938	1.975	0.307	0.328	-0.0252	0.430	0.926
1:1 Match 3.2	1.974	0.305	0.312	0.080	0.439	0.955						
3.3	1.974	0.305	0.312	0.079	0.439	0.954						
1:M Match 4.2	1.746	0.210	0.219	-0.150	0.165	0.931						
4.3	1.746	0.211	0.220	-0.148	0.166	0.935						
4.4	1.746	0.211	0.220	-0.148	0.166	0.934						
4.5	1.754	0.211	0.487	-0.140	0.175	0.937						
Strat-10 5.2	1.657	0.209	0.219	-0.237	0.185	0.899	1.731	0.254	0.263	-0.084	0.217	0.930
5.3	1.657	0.209	0.219	-0.237	0.185	0.900	1.731	0.254	0.263	-0.084	0.217	0.931
5.4	1.664	0.210	0.481	-0.230	0.191	0.904	1.703	0.213	0.547	-0.112	0.161	0.936
Strat-20 5.2	1.755	0.210	0.220	-0.140	0.164	0.936	1.799	0.256	0.265	-0.016	0.230	0.947
5.3	1.754	0.210	0.220	--0.140	0.164	0.934	1.799	0.256	0.265	-0.016	0.230	0.946
5.4	1.763	0.211	0.482	--0.132	0.173	0.936	1.773	0.215	0.553	-0.042	0.166	0.941
Strat-fine 5.2	1.901	0.223	0.229	0.006	0.191	0.946	2.011	0.264	0.303	0.195	9.095	0.952
5.3	1.901	0.223	0.229	0.006	0.191	0.946	2.009	0.264	0.303	0.194	8.949	0.952
5.4	1.917	0.223	0.455	0.023	0.207	0.940	1.946	0.261	0.643	0.131	0.885	0.943
IPTW 6.1	1.897	0.213	0.222	0.003	0.174	0.947	1.852	0.220	0.229	0.037	0.179	0.938
stab-IPTW 6.1	1.897	0.213	0.222	0.003	0.174	0.947	1.853	0.220	0.230	0.038	0.179	0.938
IPTW 6.2	1.897	0.213	0.222	0.003	0.174	0.947	1.852	0.220	0.229	0.037	0.179	0.938
stab-IPTW 6.2	1.897	0.213	0.222	0.003	0.174	0.947	1.853	0.220	0.230	0.038	0.180	0.938
IPTW 6.3	1.906	0.213	0.481	0.012	0.188	0.945	1.857	0.220	0.549	0.042	0.193	0.936
stab-IPTW 6.3	1.907	0.213	0.444	0.012	0.188	0.946	1.858	0.220	0.505	0.043	0.193	0.936

Table A 1. 6 Scenario 4 results. Mean Estimated Hazard Ratio ($\widehat{HR}$), Standard Error ($\widehat{SE}$), Monte Carlo Standard Deviation (MC SD), Bias, Mean Squared Error (MSE), And Coverage of 95% Confidence Intervals for ATT and ATE Parameters. SE and SD are Reported on the Log HR Scale. Bias and MSE on the HR Scale were Calculated With Respect to the True Conditional HR = 2 for Methods in Top Section of the Table, and With Respect to Marginal ATT = 1.621, marginal ATE = 1.540 for Remaining Methods.

Method	ATT						ATE					
	$\widehat{HR}$	$\widehat{SE}$	MC SD	Bias	MSE	Coverage	$\widehat{HR}$	$\widehat{SE}$	MC SD	Bias	MSE	Coverage
Cov Reg 1.1							2.013	0.105	0.104	0.013	0.044	0.949
1.2							2.024	0.105	0.104	0.024	0.045	0.944
PS Reg 2.1							1.750	0.104	0.105	-0.250	0.097	0.751
2.2							1.681	0.104	0.105	-0.319	0.133	0.600
2.3							1.751	0.104	0.106	-0.249	0.096	0.753
1:1 Match 3.1	1.991	0.274	0.277	-0.009	0.334	0.947						
1:M Match 4.1	1.779	0.129	0.131	-0.221	0.103	0.831						
Strat-10 5.1	1.591	0.104	0.103	-0.409	0.193	0.375	1.662	0.104	0.105	-0.338	0.145	0.553
Strat-20 5.1	1.660	0.104	0.103	-0.340	0.145	0.548	1.713	0.104	0.106	-0.287	0.115	0.677
Strat-fine 5.1	1.724	0.107	0.107	-0.276	0.110	0.719	1.807	0.176	0.193	-0.193	0.162	0.882
1:1 Match 3.2	1.678	0.164	0.150	0.057	0.068	0.972						
3.3	1.677	0.164	0.150	0.056	0.068	0.971						
1:M Match 4.2	1.666	0.109	0.105	0.045	0.033	0.945						
4.3	1.665	0.109	0.105	0.044	0.032	0.948						
4.4	1.664	0.109	0.105	0.043	0.032	0.946						
4.5	1.670	0.109	0.104	0.049	0.033	0.946						
Strat-10 5.2	1.556	0.104	0.097	-0.065	0.027	0.954	1.530	0.132	0.125	-0.010	0.036	0.961
5.3	1.555	0.104	0.097	-0.066	0.027	0.953	1.529	0.133	0.125	-0.011	0.036	0.962
5.4	1.561	0.104	0.097	-0.060	0.026	0.953	1.537	0.108	0.101	-0.002	0.024	0.962
Strat-20 5.2	1.608	0.104	0.098	-0.013	0.025	0.964	1.580	0.136	0.127	0.040	0.042	0.952
5.3	1.607	0.104	0.098	-0.014	0.025	0.963	1.578	0.137	0.128	0.038	0.042	0.952
5.4	1.614	0.104	0.097	-0.007	0.025	0.963	1.583	0.111	0.102	0.043	0.028	0.955
Strat-fine 5.2	1.651	0.105	0.098	0.030	0.027	0.959	1.637	0.127	0.122	0.097	0.050	0.928
5.3	1.650	0.105	0.098	0.029	0.027	0.961	1.636	0.127	0.122	0.096	0.049	0.931
5.4	1.657	0.105	0.097	0.036	0.027	0.959	1.639	0.127	0.128	0.100	0.052	0.926
IPTW 6.1	1.661	0.104	0.098	0.040	0.028	0.956	1.532	0.108	0.102	-0.008	0.024	0.960
stab-IPTW 6.1	1.662	0.104	0.098	0.041	0.028	0.953	1.600	0.112	0.106	0.060	0.032	0.946
IPTW 6.2	1.660	0.104	0.098	0.039	0.028	0.955	1.531	0.108	0.102	-0.009	0.024	0.960
stab-IPTW 6.2	1.661	0.104	0.098	0.040	0.028	0.953	1.599	0.112	0.106	0.059	0.032	0.947
IPTW 6.3	1.667	0.104	0.098	0.046	0.029	0.954	1.537	0.108	0.101	-0.003	0.024	0.960
stab-IPTW 6.3	1.669	0.104	0.098	0.048	0.029	0.953	1.603	0.111	0.105	0.063	0.032	0.942

Table A 1. 7 Scenario 5 Results. Mean Estimated Hazard Ratio ($\widehat{HR}$), Standard Error ($\widehat{SE}$), Monte Carlo Standard Deviation (MC SD), Bias, Mean Squared Error (MSE), And Coverage of 95% Confidence Intervals for ATT and ATE Parameters. SE and SD are Reported on the Log HR Scale. Bias and MSE on the HR Scale were Calculated With Respect to the True Conditional HR = 2 for Methods in Top Section of the Table, and With Respect to Marginal ATT = 1.639, Marginal ATE = 1.593 for Remaining Methods.

Method	ATT						ATE					
	$\widehat{HR}$	$\widehat{SE}$	MC SD	Bias	MSE	Coverage	$\widehat{HR}$	$\widehat{SE}$	MC SD	Bias	MSE	Coverage
Cov Reg 1.1							2.005	0.047	0.047	0.005	0.009	0.956
1.2							2.007	0.047	0.048	0.007	0.009	0.957
PS Reg 2.1							1.731	0.046	0.047	-0.269	0.079	0.116
2.2							1.630	0.046	0.046	-0.370	0.143	0.005
2.3							1.731	0.046	0.047	-0.269	0.079	0.116
1:1 Match 3.1	1.869	0.082	0.085	-0.131	0.042	0.858						
1:M Match 4.1	1.797	0.068	0.070	-0.203	0.057	0.633						
Strat-10 5.1	1.579	0.046	0.045	-0.421	0.183	0.000	1.629	0.046	0.045	-0.371	0.143	0.004
Strat-20 5.1	1.647	0.046	0.046	-0.353	0.130	0.008	1.680	0.046	0.045	-0.320	0.108	0.031
Strat-fine 5.1	1.744	0.052	0.054	-0.256	0.075	0.259	1.777	0.064	0.093	-0.223	0.073	0.590
1:1 Match 3.2	1.666	0.053	0.050	0.028	0.008	0.945						
3.3	1.665	0.053	0.050	0.027	0.008	0.945						
1:M Match 4.2	1.687	0.047	0.046	0.049	0.009	0.904						
4.3	1.659	0.048	0.046	0.022	0.006	0.946						
4.4	1.659	0.048	0.046	0.021	0.006	0.944						
4.5	1.660	0.048	0.046	0.022	0.006	0.946						
Strat-10 5.2	1.574	0.047	0.045	-0.064	0.009	0.880	1.570	0.057	0.054	-0.024	0.008	0.947
5.3	1.573	0.047	0.045	-0.064	0.009	0.879	1.569	0.058	0.054	-0.025	0.008	0.947
5.4	1.574	0.047	0.044	-0.064	0.009	0.880	1.570	0.046	0.043	-0.024	0.005	0.951
Strat-20 5.2	1.626	0.047	0.045	-0.012	0.005	0.959	1.604	0.058	0.054	0.010	0.007	0.969
5.3	1.626	0.047	0.045	-0.012	0.005	0.960	1.603	0.058	0.054	0.010	0.007	0.967
5.4	1.626	0.047	0.045	-0.012	0.005	0.958	1.604	0.046	0.043	0.011	0.005	0.959
Strat-fine 5.2	1.672	0.054	0.052	0.034	0.009	0.948	1.651	0.056	0.054	0.057	0.011	0.924
5.3	1.671	0.054	0.052	0.034	0.009	0.950	1.650	0.056	0.053	0.056	0.011	0.925
5.4	1.673	0.054	0.052	0.035	0.009	0.948	1.651	0.056	0.054	0.057	0.011	0.923
IPTW 6.1	1.672	0.048	0.045	0.034	0.007	0.938	1.628	0.046	0.043	0.035	0.006	0.937
stab-IPTW 6.1	1.672	0.048	0.045	0.034	0.007	0.938	1.628	0.046	0.043	0.035	0.006	0.937
IPTW 6.2	1.671	0.048	0.045	0.033	0.007	0.939	1.628	0.046	0.043	0.034	0.006	0.937
stab-IPTW 6.2	1.671	0.048	0.045	0.033	0.007	0.939	1.628	0.046	0.043	0.034	0.006	0.937
IPTW 6.3	1.672	0.048	0.045	0.034	0.007	0.938	1.628	0.046	0.043	0.035	0.006	0.938
stab-IPTW 6.3	1.672	0.048	0.045	0.034	0.007	0.938	1.628	0.046	0.043	0.035	0.006	0.938

Table A 1. 8 Scenario 6 Results. Mean Estimated Hazard Ratio ($\widehat{HR}$), Standard Error ($\widehat{SE}$), Monte Carlo Standard Deviation (MC SD), Bias, Mean Squared Error (MSE), And Coverage of 95% Confidence Intervals for ATT and ATE Parameters. SE and SD are Reported on the Log HR Scale. Bias and MSE on the HR Scale were Calculated With Respect to the True Conditional HR = 1.25 for Methods in Top Section of the Table, and With Respect To Marginal ATT = 1.236, Marginal ATE = 1.227 for Remaining Methods.

Method	ATT						ATE					
	$\widehat{HR}$	$\widehat{SE}$	MC SD	Bias	MSE	Coverage	$\widehat{HR}$	$\widehat{SE}$	MC SD	Bias	MSE	Coverage
Cov Reg 1.1							1.252	0.053	0.053	0.002	0.004	0.959
1.2							1.253	0.053	0.053	0.003	0.004	0.961
PS Reg 2.1							1.245	0.053	0.053	-0.005	0.004	0.951
2.2							1.232	0.053	0.053	-0.018	0.005	0.946
2.3							1.245	0.053	0.053	-0.005	0.004	0.949
1:1 Match 3.1	1.248	0.113	0.114	-0.002	0.020	0.943						
1:M Match 4.1	1.223	0.073	0.072	-0.027	0.008	0.940						
Strat-10 5.1	1.218	0.053	0.053	-0.032	0.005	0.920	1.229	0.053	0.053	-0.021	0.005	0.935
Strat-20 5.1	1.231	0.053	0.053	-0.019	0.005	0.943	1.238	0.053	0.053	-0.012	0.004	0.948
Strat-fine 5.1	1.244	0.059	0.060	-0.006	0.006	0.949	1.245	0.086	0.089	-0.005	0.012	0.939
1:1 Match 3.2	1.238	0.067	0.066	0.002	0.007	0.956						
3.3	1.238	0.067	0.066	0.002	0.007	0.955						
1:M Match 4.2	1.224	0.053	0.053	-0.012	0.004	0.954						
4.3	1.223	0.053	0.053	-0.013	0.004	0.949						
4.4	1.222	0.053	0.053	-0.014	0.004	0.948						
4.5	1.223	0.053	0.053	-0.013	0.004	0.949						
Strat-10 5.2	1.220	0.053	0.053	-0.016	0.004	0.948	1.221	0.066	0.065	-0.007	0.006	0.956
5.3	1.219	0.053	0.053	-0.017	0.004	0.948	1.220	0.066	0.065	-0.007	0.006	0.957
5.4	1.220	0.053	0.053	-0.016	0.004	0.950	1.219	0.054	0.054	-0.008	0.004	0.954
Strat-20 5.2	1.231	0.053	0.053	-0.005	0.004	0.952	1.230	0.066	0.066	0.003	0.007	0.957
5.3	1.230	0.053	0.053	-0.006	0.004	0.951	1.230	0.067	0.066	0.003	0.007	0.955
5.4	1.231	0.053	0.053	-0.005	0.004	0.951	1.229	0.054	0.054	0.002	0.004	0.960
Strat-fine 5.2	1.240	0.057	0.057	0.004	0.005	0.950	1.235	0.063	0.064	0.007	0.006	0.938
5.3	1.239	0.057	0.057	0.003	0.005	0.952	1.235	0.063	0.064	0.007	0.006	0.937
5.4	1.240	0.057	0.057	0.004	0.005	0.949	1.235	0.063	0.064	0.008	0.006	0.942
IPTW 6.1	1.238	0.053	0.053	0.002	0.004	0.961	1.235	0.054	0.054	0.008	0.004	0.960
stab-IPTW 6.1	1.238	0.053	0.053	0.002	0.004	0.961	1.235	0.054	0.054	0.008	0.005	0.960
IPTW 6.2	1.238	0.053	0.053	0.002	0.004	0.960	1.235	0.054	0.054	0.008	0.004	0.960
stab-IPTW 6.2	1.238	0.053	0.053	0.002	0.004	0.960	1.235	0.054	0.054	0.008	0.004	0.960
IPTW 6.3	1.239	0.053	0.053	0.003	0.004	0.958	1.236	0.054	0.054	0.008	0.004	0.962
stab-IPTW 6.3	1.239	0.053	0.053	0.003	0.004	0.956	1.236	0.054	0.054	0.009	0.004	0.961

Table A 1. 9 Scenario 7 Results. Mean Estimated Hazard Ratio ($\widehat{HR}$), Standard Error ($\widehat{SE}$), Monte Carlo Standard Deviation (MC SD), Bias, Mean Squared Error (MSE), And Coverage of 95% Confidence Intervals for ATT and ATE Parameters. SE and SD are Reported on the Log HR Scale. Bias and MSE on the HR Scale were Calculated With Respect to the True Conditional HR = 1.5 For Methods in Top Section of the Table, and With Respect to Marginal ATT = 1.429, Marginal ATE = 1.394 for Remaining Methods.

Method	ATT						ATE					
	$\widehat{HR}$	$\widehat{SE}$	MC SD	Bias	MSE	Coverage	$\widehat{HR}$	$\widehat{SE}$	MC SD	Bias	MSE	Coverage
Cov Reg 1.1							1.504	0.053	0.053	0.004	0.006	0.959
1.2							1.505	0.053	0.053	0.005	0.006	0.961
PS Reg 2.1							1.451	0.053	0.053	-0.049	0.008	0.896
2.2							1.414	0.053	0.052	-0.086	0.013	0.807
2.3							1.451	0.053	0.053	-0.049	0.008	0.899
1:1 Match 3.1	1.485	0.110	0.111	-0.015	0.028	0.946						
1:M Match 4.1	1.411	0.067	0.067	-0.089	0.017	0.851						
Strat-10 5.1	1.383	0.053	0.052	-0.117	0.019	0.666	1.414	0.053	0.052	-0.086	0.013	0.799
Strat-20 5.1	1.416	0.053	0.052	-0.084	0.012	0.808	1.435	0.053	0.052	-0.065	0.010	0.863
Strat-fine 5.1	1.450	0.057	0.058	-0.050	0.009	0.894	1.457	0.075	0.089	-0.043	0.018	0.906
1:1 Match 3.2	1.430	0.070	0.067	0.001	0.009	0.961						
3.3	1.430	0.070	0.067	0.001	0.009	0.960						
1:M Match 4.2	1.402	0.053	0.052	-0.028	0.006	0.934						
4.3	1.391	0.053	0.051	-0.039	0.007	0.928						
4.4	1.391	0.053	0.051	-0.039	0.007	0.926						
4.5	1.391	0.053	0.051	-0.038	0.006	0.928						
Strat-10 5.2	1.380	0.053	0.051	-0.050	0.007	0.906	1.380	0.066	0.063	-0.014	0.008	0.952
5.3	1.379	0.053	0.051	-0.050	0.007	0.905	1.379	0.066	0.063	-0.014	0.008	0.951
5.4	1.380	0.053	0.051	-0.049	0.007	0.905	1.379	0.054	0.052	-0.015	0.005	0.945
Strat-20 5.2	1.407	0.053	0.051	-0.023	0.006	0.943	1.398	0.066	0.063	0.004	0.008	0.962
5.3	1.406	0.053	0.051	-0.023	0.006	0.943	1.398	0.066	0.063	0.004	0.008	0.960
5.4	1.407	0.053	0.051	-0.022	0.006	0.946	1.397	0.054	0.052	0.004	0.005	0.964
Strat-fine 5.2	1.430	0.057	0.056	0.000	0.006	0.963	1.424	0.064	0.062	0.030	0.009	0.943
5.3	1.430	0.057	0.056	0.000	0.006	0.961	1.424	0.064	0.062	0.030	0.009	0.942
5.4	1.431	0.057	0.056	0.002	0.006	0.963	1.424	0.064	0.062	0.031	0.009	0.943
IPTW 6.1	1.431	0.054	0.051	0.001	0.005	0.967	1.410	0.054	0.053	0.016	0.006	0.957
stab-IPTW 6.1	1.431	0.054	0.051	0.002	0.005	0.966	1.410	0.054	0.053	0.016	0.006	0.957
IPTW 6.2	1.430	0.054	0.051	0.001	0.005	0.967	1.409	0.054	0.053	0.016	0.006	0.957
stab-IPTW 6.2	1.431	0.054	0.051	0.001	0.005	0.967	1.410	0.054	0.053	0.016	0.006	0.957
IPTW 6.3	1.431	0.054	0.051	0.002	0.005	0.968	1.410	0.054	0.052	0.016	0.006	0.958
stab-IPTW 6.3	1.432	0.054	0.051	0.002	0.005	0.968	1.411	0.054	0.053	0.017	0.006	0.957

Table A 1. 10 Scenario 8 Results. Mean Estimated Hazard Ratio ($\widehat{HR}$), Standard Error ($\widehat{SE}$), Monte Carlo Standard Deviation (MC SD), Bias, Mean Squared Error (MSE), And Coverage of 95% Confidence Intervals for ATT and ATE Parameters. SE and SD are Reported on the Log HR Scale. Bias and MSE on the HR Scale were Calculated With Respect to the True Conditional HR = 1 for Methods in Top Section Of The Table, and With Respect to Marginal ATT = 1, Marginal ATE = 1 for Remaining Methods.

Method	ATT						ATE					
	$\widehat{HR}$	$\widehat{SE}$	MC SD	Bias	MSE	Coverage	$\widehat{HR}$	$\widehat{SE}$	MC SD	Bias	MSE	Coverage
Cov Reg 1.1							1.003	0.048	0.050	0.003	0.003	0.939
1.2							1.003	0.048	0.050	0.003	0.003	0.938
PS Reg 2.1							1.003	0.048	0.049	0.003	0.002	0.946
2.2							1.084	0.047	0.048	0.084	0.010	0.605
2.3							1.003	0.048	0.049	0.003	0.002	0.946
1:1 Match 3.1	1.002	0.076	0.076	0.002	0.006	0.952						
1:M Match 4.1	1.011	0.063	0.064	0.011	0.004	0.944						
Strat-10 5.1	1.046	0.048	0.048	0.046	0.005	0.847	1.082	0.047	0.047	0.082	0.009	0.631
Strat-20 5.1	1.024	0.048	0.049	0.024	0.003	0.918	1.046	0.048	0.048	0.046	0.005	0.846
Strat-fine 5.1	1.005	0.053	0.055	0.005	0.003	0.938	1.008	0.061	0.065	0.008	0.005	0.945
1:1 Match 3.2	1.004	0.054	0.054	0.001	0.003	0.941						
3.3	1.004	0.054	0.054	0.001	0.003	0.944						
1:M Match 4.2	1.009	0.049	0.050	0.007	0.003	0.946						
4.3	1.007	0.050	0.049	0.005	0.002	0.949						
4.4	1.007	0.050	0.049	0.005	0.002	0.948						
4.5	1.007	0.050	0.049	0.005	0.002	0.951						
Strat-10 5.2	1.046	0.048	0.048	0.044	0.004	0.871	1.060	0.061	0.060	0.059	0.007	0.852
5.3	1.046	0.048	0.048	0.044	0.004	0.871	1.060	0.061	0.059	0.059	0.007	0.851
5.4	1.046	0.048	0.048	0.044	0.004	0.865	1.064	0.049	0.048	0.063	0.007	0.771
Strat-20 5.2	1.024	0.049	0.047	0.021	0.003	0.940	1.030	0.062	0.060	0.029	0.005	0.940
5.3	1.024	0.049	0.047	0.021	0.003	0.941	1.030	0.061	0.060	0.029	0.005	0.937
5.4	1.024	0.049	0.047	0.022	0.003	0.939	1.033	0.050	0.049	0.032	0.004	0.912
Strat-fine 5.2	1.003	0.055	0.052	0.001	0.003	0.960	1.004	0.056	0.055	0.003	0.003	0.951
5.3	1.003	0.055	0.052	0.001	0.003	0.959	1.004	0.056	0.055	0.003	0.003	0.954
5.4	1.004	0.055	0.052	0.001	0.003	0.962	1.004	0.056	0.055	0.003	0.003	0.954
IPTW 6.1	1.002	0.050	0.050	-0.001	0.002	0.954	1.003	0.050	0.049	0.002	0.002	0.952
stab-IPTW 6.1	1.002	0.050	0.050	-0.001	0.002	0.954	1.002	0.050	0.049	0.002	0.002	0.951
IPTW 6.2	1.002	0.050	0.050	-0.001	0.002	0.954	1.003	0.050	0.049	0.002	0.002	0.950
stab-IPTW 6.2	1.002	0.050	0.050	0.000	0.002	0.953	1.003	0.050	0.049	0.002	0.002	0.950
IPTW 6.3	1.002	0.050	0.050	0.000	0.002	0.953	1.003	0.050	0.049	0.002	0.002	0.951
stab-IPTW 6.3	1.003	0.050	0.050	0.000	0.002	0.954	1.003	0.050	0.049	0.002	0.002	0.952

Table A 1. 11 Scenario 9 Results. Mean Estimated Hazard Ratio ($\widehat{HR}$), Standard Error ($\widehat{SE}$), Monte Carlo Standard Deviation (MC SD), Bias, Mean Squared Error (MSE), And Coverage of 95% Confidence Intervals for ATT and ATE Parameters. SE and SD are Reported on the Log HR Scale. Bias and MSE on the HR Scale were Calculated With Respect to the True Conditional HR = 2 for Methods in Top Section of the Table, and With Respect to Marginal ATT = 1.984, Marginal ATE = 1.897 for Remaining Methods.

Method	ATT						ATE					
	$\widehat{HR}$	$\widehat{SE}$	MC SD	Bias	MSE	Coverage	$\widehat{HR}$	$\widehat{SE}$	MC SD	Bias	MSE	Coverage
Cov Reg 1.1							2.009	0.062	0.062	0.009	0.015	0.957
1.2							2.010	0.062	0.062	0.010	0.016	0.956
PS Reg 2.1							1.977	0.064	0.063	-0.023	0.016	0.951
2.2							1.950	0.064	0.063	-0.050	0.017	0.934
2.3							1.978	0.064	0.063	-0.022	0.016	0.951
1:1 Match 3.1	2.014	0.122	0.123	0.014	0.062	0.953						
1:M Match 4.1	1.969	0.080	0.079	-0.031	0.025	0.952						
Strat-10 5.1	1.738	0.061	0.058	-0.262	0.079	0.379	1.951	0.064	0.062	-0.049	0.017	0.935
Strat-20 5.1	1.845	0.063	0.060	-0.155	0.036	0.750	1.971	0.064	0.063	-0.029	0.016	0.949
Strat-fine 5.1	1.973	0.069	0.072	-0.027	0.019	0.958	1.979	0.087	0.106	-0.021	0.040	0.930
1:1 Match 3.2	1.980	0.084	0.084	-0.003	0.028	0.949						
3.3	1.980	0.084	0.084	-0.003	0.028	0.949						
1:M Match 4.2	1.960	0.065	0.064	-0.024	0.016	0.953						
4.3	1.965	0.069	0.068	-0.019	0.018	0.954						
4.4	1.965	0.069	0.068	-0.019	0.018	0.954						
4.5	1.966	0.069	0.068	-0.018	0.018	0.953						
Strat-10 5.2	1.837	0.077	0.076	-0.147	0.041	0.830	1.843	0.109	0.110	-0.053	0.043	0.926
5.3	1.837	0.077	0.076	-0.147	0.041	0.829	1.842	0.110	0.110	-0.054	0.044	0.924
5.4	1.846	0.075	0.074	-0.137	0.038	0.845	1.829	0.100	0.105	-0.067	0.041	0.909
Strat-20 5.2	1.907	0.081	0.080	-0.076	0.029	0.933	1.891	0.123	0.126	-0.005	0.058	0.935
5.3	1.907	0.081	0.080	-0.077	0.029	0.933	1.890	0.124	0.127	-0.007	0.058	0.933
5.4	1.921	0.078	0.077	-0.063	0.026	0.945	1.850	0.109	0.120	-0.047	0.052	0.893
Strat-fine 5.2	1.978	0.082	0.083	-0.006	0.026	0.959	1.970	0.077	0.078	0.074	0.029	0.927
5.3	1.977	0.082	0.083	-0.006	0.026	0.959	1.970	0.077	0.078	0.074	0.029	0.926
5.4	1.985	0.081	0.084	0.002	0.026	0.956	1.970	0.077	0.080	0.074	0.029	0.926
IPTW 6.1	1.981	0.083	0.083	-0.003	0.027	0.952	1.715	0.082	0.083	-0.182	0.053	0.753
stab-IPTW 6.1	1.979	0.082	0.081	-0.004	0.025	0.951	1.848	0.105	0.109	-0.048	0.043	0.909
IPTW 6.2	1.981	0.083	0.083	-0.003	0.027	0.952	1.713	0.082	0.083	-0.184	0.054	0.748
stab-IPTW 6.2	1.979	0.082	0.081	-0.005	0.025	0.951	1.847	0.105	0.109	-0.049	0.043	0.909
IPTW 6.3	1.996	0.080	0.080	0.012	0.026	0.948	1.716	0.081	0.082	-0.181	0.053	0.755
stab-IPTW 6.3	1.993	0.080	0.079	0.009	0.025	0.949	1.830	0.100	0.107	-0.067	0.043	0.891

Table A 1. 12 Scenario 10 results. Mean Estimated Hazard Ratio ($\widehat{HR}$), Standard Error ($\widehat{SE}$), Monte Carlo Standard Deviation (MC SD), Bias, Mean Squared Error (MSE), And Coverage of 95% Confidence Intervals for ATT and ATE Parameters. SE and SD are Reported on the Log HR Scale. Bias and MSE on the HR Scale were Calculated With Respect to the True Conditional HR = 0.8 for Methods in Top Section of the Table, and With Respect to Marginal ATT =0.850, marginal ATE = 0.857 for Remaining Methods.

Method	ATT						ATE					
	$\widehat{HR}$	$\widehat{SE}$	MC SD	Bias	MSE	Coverage	$\widehat{HR}$	$\widehat{SE}$	MC SD	Bias	MSE	Coverage
Cov Reg 1.1							0.801	0.079	0.079	0.001	0.004	0.944
1.2							0.802	0.079	0.080	0.002	0.004	0.946
PS Reg 2.1							0.837	0.080	0.078	0.037	0.006	0.922
2.2							0.783	0.079	0.077	-0.017	0.004	0.944
2.3							0.837	0.080	0.077	0.037	0.006	0.926
1:1 Match 3.1	0.819	0.137	0.138	0.019	0.013	0.954						
1:M Match 4.1	0.788	0.087	0.086	-0.012	0.005	0.944						
Strat-10 5.1	0.683	0.078	0.075	-0.117	0.016	0.465	0.801	0.075	0.074	0.001	0.003	0.957
Strat-20 5.1	0.743	0.079	0.075	-0.057	0.006	0.859	0.815	0.079	0.077	0.015	0.004	0.950
Strat-fine 5.1	0.830	0.082	0.080	0.030	0.005	0.930	0.824	0.109	0.135	0.024	0.012	0.919
1:1 Match 3.2	0.841	0.099	0.092	-0.009	0.006	0.960						
3.3	0.841	0.099	0.092	-0.009	0.006	0.962						
1:M Match 4.2	0.792	0.080	0.077	-0.058	0.007	0.869						
4.3	0.801	0.081	0.076	-0.049	0.006	0.899						
4.4	0.801	0.081	0.077	-0.049	0.006	0.899						
4.5	0.802	0.081	0.077	-0.047	0.006	0.901						
Strat-10 5.2	0.709	0.079	0.075	-0.141	0.023	0.336	0.800	0.098	0.092	-0.067	0.010	0.884
5.3	0.709	0.079	0.075	-0.140	0.023	0.333	0.801	0.098	0.093	-0.066	0.010	0.892
5.4	0.710	0.079	0.075	-0.139	0.022	0.342	0.803	0.086	0.082	-0.064	0.008	0.876
Strat-20 5.2	0.766	0.080	0.076	-0.084	0.010	0.768	0.817	0.107	0.098	-0.051	0.009	0.929
5.3	0.766	0.080	0.076	-0.083	0.010	0.770	0.817	0.108	0.099	-0.050	0.009	0.932
5.4	0.768	0.080	0.076	-0.082	0.010	0.775	0.819	0.096	0.086	-0.048	0.007	0.940
Strat-fine 5.2	0.837	0.085	0.081	-0.013	0.005	0.956	0.855	0.088	0.085	-0.013	0.006	0.958
5.3	0.837	0.085	0.081	-0.013	0.005	0.957	0.855	0.088	0.085	-0.012	0.005	0.964
5.4	0.839	0.084	0.081	-0.010	0.005	0.958	0.856	0.088	0.085	-0.012	0.005	0.959
IPTW 6.1	0.839	0.081	0.077	-0.010	0.004	0.953	0.835	0.099	0.092	-0.033	0.007	0.945
stab-IPTW 6.1	0.839	0.081	0.077	-0.010	0.004	0.953	0.858	0.107	0.102	-0.009	0.008	0.957
IPTW 6.2	0.840	0.081	0.077	-0.010	0.004	0.955	0.835	0.099	0.092	-0.032	0.007	0.944
stab-IPTW 6.2	0.840	0.081	0.077	-0.010	0.004	0.955	0.858	0.107	0.102	-0.009	0.008	0.958
IPTW 6.3	0.841	0.081	0.077	-0.008	0.004	0.956	0.834	0.098	0.092	-0.034	0.007	0.945
stab-IPTW 6.3	0.841	0.081	0.077	-0.008	0.004	0.956	0.853	0.104	0.099	-0.014	0.007	0.958

Table A 1. 13 Scenario 11 Results. Mean Estimated Hazard Ratio ($\widehat{HR}$), Standard Error ($\widehat{SE}$), Monte Carlo Standard Deviation (MC SD), Bias, Mean Squared Error (MSE), And Coverage of 95% Confidence Intervals for ATT and ATE Parameters. SE and SD are Reported on the Log HR Scale. Bias and MSE on the HR Scale were Calculated With Respect to the True Conditional HR = 1 for Methods in Top Section of the Table, and With Respect to Marginal ATT =1, Marginal ATE = 1 for Remaining Methods.

Method	ATT						ATE					
	$\widehat{HR}$	$\widehat{SE}$	MC SD	Bias	MSE	Coverage	$\widehat{HR}$	$\widehat{SE}$	MC SD	Bias	MSE	Coverage
Cov Reg 1.1							1.001	0.065	0.065	0.001	0.004	0.952
1.2							1.002	0.066	0.065	0.002	0.004	0.952
PS Reg 2.1							1.002	0.066	0.064	0.002	0.004	0.963
2.2							0.979	0.066	0.063	-0.021	0.004	0.953
2.3							1.002	0.066	0.064	0.002	0.004	0.961
1:1 Match 3.1	1.005	0.112	0.114	0.005	0.013	0.953						
1:M Match 4.1	0.972	0.074	0.072	-0.028	0.006	0.940						
Strat-10 5.1	0.919	0.065	0.062	-0.081	0.010	0.746	0.979	0.066	0.063	-0.021	0.004	0.949
Strat-20 5.1	0.959	0.065	0.063	-0.041	0.005	0.914	0.994	0.066	0.064	-0.006	0.004	0.958
Strat-fine 5.1	1.000	0.069	0.067	0.000	0.005	0.958	0.994	0.083	0.097	-0.006	0.009	0.934
1:1 Match 3.2	1.001	0.082	0.078	-0.004	0.006	0.959						
3.3	1.001	0.082	0.078	-0.004	0.006	0.959						
1:M Match 4.2	0.971	0.066	0.063	-0.034	0.005	0.927						
4.3	0.977	0.067	0.065	-0.028	0.005	0.937						
4.4	0.977	0.067	0.065	-0.028	0.005	0.938						
4.5	0.978	0.067	0.065	-0.027	0.005	0.942						
Strat-10 5.2	0.935	0.067	0.064	-0.070	0.008	0.817	0.965	0.088	0.082	-0.035	0.007	0.957
5.3	0.935	0.067	0.064	-0.070	0.008	0.820	0.965	0.088	0.082	-0.035	0.007	0.959
5.4	0.937	0.067	0.064	-0.069	0.008	0.827	0.965	0.076	0.069	-0.035	0.006	0.947
Strat-20 5.2	0.970	0.067	0.065	-0.036	0.005	0.925	0.989	0.091	0.084	-0.011	0.007	0.966
5.3	0.970	0.067	0.065	-0.036	0.005	0.927	0.989	0.091	0.085	-0.011	0.007	0.966
5.4	0.971	0.067	0.065	-0.035	0.005	0.934	0.987	0.078	0.072	-0.013	0.005	0.967
Strat-fine 5.2	1.001	0.072	0.070	0.001	0.005	0.954	0.999	0.077	0.074	0.000	0.005	0.962
5.3	1.001	0.072	0.070	0.001	0.005	0.955	0.999	0.077	0.074	0.000	0.005	0.963
5.4	1.002	0.072	0.070	0.002	0.005	0.956	1.000	0.077	0.074	0.000	0.005	0.964
IPTW 6.1	1.002	0.068	0.065	0.002	0.004	0.959	0.990	0.080	0.075	-0.010	0.006	0.964
stab-IPTW 6.1	1.002	0.068	0.065	0.002	0.004	0.959	1.000	0.083	0.079	0.00	0.006	0.968
IPTW 6.2	1.002	0.068	0.065	0.002	0.004	0.959	0.990	0.080	0.074	-0.010	0.006	0.964
stab-IPTW 6.2	1.002	0.068	0.065	0.002	0.004	0.959	1.000	0.083	0.079	0.000	0.006	0.968
IPTW 6.3	1.003	0.068	0.065	0.003	0.004	0.960	0.990	0.079	0.074	-0.010	0.006	0.964
stab-IPTW 6.3	1.003	0.068	0.065	0.003	0.004	0.959	0.998	0.082	0.077	-0.002	0.006	0.963

Table A 1. 14 Scenario 12 Results. Mean Estimated Hazard Ratio ($\widehat{HR}$), Standard Error ($\widehat{SE}$), Monte Carlo Standard Deviation (MC SD), Bias, Mean Squared Error (MSE), And Coverage of 95% Confidence Intervals for ATT and ATE Parameters. SE and SD are Reported on the Log HR Scale. Bias and MSE on the HR Scale were Calculated With Respect to the True Conditional HR = 1.25 for Methods in Top Section of the Table, and With Respect to Marginal ATT =1.220, Marginal ATE = 1.195 for Remaining Methods.

Method	ATT						ATE					
	$\widehat{HR}$	$\widehat{SE}$	MC SD	Bias	MSE	Coverage	$\widehat{HR}$	$\widehat{SE}$	MC SD	Bias	MSE	Coverage
Cov Reg 1.1							1.253	0.062	0.062	0.058	0.009	0.875
1.2							1.254	0.062	0.062	0.060	0.010	0.874
PS Reg 2.1							1.226	0.062	0.060	-0.024	0.006	0.945
2.2							1.199	0.062	0.060	-0.051	0.008	0.894
2.3							1.226	0.062	0.060	-0.024	0.006	0.947
1:1 Match 3.1	1.244	0.123	0.123	-0.006	0.024	0.952						
1:M Match 4.1	1.190	0.076	0.076	-0.060	0.012	0.905						
Strat-10 5.1	1.126	0.061	0.059	-0.124	0.020	0.583	1.195	0.062	0.061	-0.055	0.008	0.891
Strat-20 5.1	1.174	0.061	0.060	-0.076	0.011	0.824	1.216	0.062	0.061	-0.034	0.007	0.932
Strat-fine 5.1	1.224	0.066	0.065	-0.026	0.007	0.943	1.224	0.097	0.112	-0.026	0.019	0.923
1:1 Match 3.2	1.220	0.079	0.076	0.000	0.009	0.962						
3.3	1.220	0.079	0.076	0.000	0.009	0.962						
1:M Match 4.2	1.188	0.062	0.060	-0.032	0.006	0.938						
4.3	1.192	0.063	0.061	-0.028	0.006	0.946						
4.4	1.192	0.063	0.061	-0.028	0.006	0.946						
4.5	1.193	0.063	0.061	-0.027	0.006	0.947						
Strat-10 5.2	1.145	0.063	0.060	-0.075	0.010	0.832	1.165	0.081	0.078	-0.030	0.009	0.950
5.3	1.145	0.063	0.060	-0.076	0.010	0.833	1.165	0.081	0.078	-0.030	0.009	0.948
5.4	1.146	0.063	0.060	-0.074	0.010	0.839	1.165	0.069	0.066	-0.030	0.007	0.948
Strat-20 5.2	1.185	0.064	0.061	-0.036	0.007	0.933	1.191	0.084	0.081	-0.004	0.009	0.953
5.3	1.184	0.064	0.061	-0.036	0.007	0.932	1.190	0.085	0.081	-0.004	0.009	0.954
5.4	1.186	0.064	0.061	-0.034	0.006	0.935	1.190	0.072	0.068	-0.005	0.007	0.965
Strat-fine 5.2	1.220	0.068	0.067	-0.001	0.007	0.968	1.211	0.072	0.072	0.016	0.008	0.950
5.3	1.219	0.068	0.067	-0.001	0.007	0.968	1.211	0.072	0.072	0.016	0.008	0.951
5.4	1.222	0.068	0.066	0.001	0.007	0.969	1.212	0.072	0.072	0.017	0.008	0.949
IPTW 6.1	1.222	0.064	0.061	0.001	0.006	0.964	1.194	0.073	0.069	-0.001	0.007	0.967
stab-IPTW 6.1	1.222	0.064	0.061	0.001	0.006	0.963	1.204	0.076	0.073	0.009	0.008	0.962
IPTW 6.2	1.221	0.064	0.062	0.001	0.006	0.963	1.194	0.073	0.069	-0.001	0.007	0.966
stab-IPTW 6.2	1.222	0.064	0.061	0.001	0.006	0.963	1.204	0.076	0.073	0.009	0.008	0.962
IPTW 6.3	1.223	0.064	0.061	0.003	0.006	0.966	1.193	0.073	0.069	-0.002	0.007	0.967
stab-IPTW 6.3	1.223	0.064	0.061	0.003	0.006	0.966	1.202	0.075	0.072	0.007	0.008	0.965

Table A 1. 15 Scenario 13 Results. Mean Estimated Hazard Ratio ($\widehat{HR}$), Standard Error ($\widehat{SE}$), Monte Carlo Standard Deviation (MC SD), Bias, Mean Squared Error (MSE), And Coverage of 95% Confidence Intervals for ATT and ATE Parameters. SE and SD are Reported on the Log HR Scale. Bias and MSE on the HR Scale were Calculated With Respect to the True Conditional HR = 5 for Methods in Top Section of the Table, and With Respect to Marginal ATT =3.468, Marginal ATE = 3.248 for Remaining Methods.

Method	ATT						ATE					
	$\widehat{HR}$	$\widehat{SE}$	MC SD	Bias	MSE	Coverage	$\widehat{HR}$	$\widehat{SE}$	MC SD	Bias	MSE	Coverage
Cov Reg 1.1							5.023	0.049	0.048	0.023	0.059	0.949
1.2							5.033	0.049	0.049	0.033	0.061	0.946
PS Reg 2.1							3.665	0.047	0.053	-1.335	1.820	0.000
2.2							3.581	0.047	0.050	-1.419	2.045	0.000
2.3							3.673	0.047	0.053	-1.327	1.798	0.000
1:1 Match 3.1	4.451	0.140	0.141	-0.549	0.712	0.838						
1:M Match 4.1	3.775	0.075	0.076	-1.225	1.583	0.041						
Strat-10 5.1	3.471	0.047	0.049	-1.529	2.366	0.000	3.563	0.047	0.050	-1.437	2.096	0.000
Strat-20 5.1	3.569	0.047	0.050	-1.431	2.079	0.000	3.621	0.047	0.051	-1.379	1.937	0.000
Strat-fine 5.1	3.728	0.055	0.060	-1.272	1.667	0.001	4.006	0.085	0.115	-0.994	1.177	0.298
1:1 Match 3.2	3.457	0.070	0.079	-0.011	0.074	0.915						
3.3	3.455	0.070	0.079	-0.014	0.075	0.914						
1:M Match 4.2	3.529	0.047	0.051	0.061	0.036	0.915						
4.3	3.355	0.048	0.057	-0.113	0.049	0.825						
4.4	3.352	0.048	0.057	-0.116	0.050	0.819						
4.5	3.357	0.048	0.057	-0.111	0.049	0.824						
Strat-10 5.2	3.335	0.047	0.055	-0.134	0.051	0.808	3.185	0.057	0.065	-0.063	0.046	0.898
5.3	3.332	0.047	0.055	-0.136	0.052	0.799	3.181	0.058	0.065	-0.067	0.048	0.898
5.4	3.337	0.047	0.055	-0.132	0.051	0.810	3.214	0.046	0.057	-0.035	0.034	0.883
Strat-20 5.2	3.399	0.048	0.057	-0.069	0.042	0.874	3.213	0.058	0.065	-0.035	0.045	0.910
5.3	3.397	0.048	0.057	-0.072	0.043	0.864	3.209	0.058	0.066	-0.040	0.046	0.909
5.4	3.402	0.048	0.057	-0.067	0.042	0.872	3.240	0.046	0.057	-0.008	0.035	0.890
Strat-fine 5.2	3.451	0.052	0.064	-0.017	0.050	0.899	3.400	0.059	0.070	0.152	0.082	0.845
5.3	3.448	0.052	0.064	-0.020	0.050	0.894	3.396	0.059	0.070	0.148	0.080	0.845
5.4	3.456	0.052	0.064	-0.012	0.050	0.901	3.399	0.059	0.070	0.151	0.081	0.843
IPTW 6.1	3.449	0.048	0.059	-0.019	0.042	0.879	3.254	0.046	0.059	0.006	0.037	0.889
stab-IPTW 6.1	3.455	0.048	0.059	-0.013	0.042	0.886	3.263	0.047	0.058	0.015	0.036	0.895
IPTW 6.2	3.446	0.048	0.059	-0.022	0.043	0.880	3.251	0.046	0.059	0.003	0.037	0.884
stab-IPTW 6.2	3.453	0.048	0.059	-0.016	0.042	0.883	3.260	0.047	0.058	0.012	0.037	0.891
IPTW 6.3	3.451	0.048	0.059	-0.017	0.042	0.882	3.253	0.046	0.059	0.005	0.037	0.889
stab-IPTW 6.3	3.458	0.048	0.059	-0.010	0.042	0.880	3.262	0.047	0.058	0.014	0.036	0.895

Table A 1. 16 Scenario 14 Results. Mean Estimated Hazard Ratio ($\widehat{HR}$), Standard Error ($\widehat{SE}$), Monte Carlo Standard Deviation (MC SD), Bias, Mean Squared Error (MSE), and Coverage of 95% Confidence Intervals for ATT And ATE Parameters. SE and SD Reported on the LogHR Scale. Bias and MSE are not Calculated for Methods in Top Section of the Table Because the Conditional Effect Is not Defined Under Heterogeneity. Other Calculations are With Respect to Marginal ATT = 1.140 and Marginal ATE = 1.115.

Method	ATT						ATE					
	$\widehat{HR}$	$\widehat{SE}$	MC SD	Bias	MSE	Coverage	$\widehat{HR}$	$\widehat{SE}$	MC SD	Bias	MSE	Coverage
Cov Reg 1.1							1.170	0.054	0.055			
1.2							1.187	0.055	0.054			
PS Reg 2.1							1.176	0.055	0.055			
2.2							1.158	0.055	0.057			
2.3							1.176	0.055	0.055			
1:1 Match 3.1	1.178	0.107	0.110									
1:M Match 4.1	1.130	0.069	0.071									
Strat-10 5.1	1.126	0.054	0.060				1.117	0.056	0.056			
Strat-20 5.1	1.142	0.054	0.058				1.132	0.056	0.056			
Strat-fine 5.1	1.156	0.059	0.061				1.143	0.077	0.094			
1:1 Match 3.2	1.164	0.069	0.070	0.024	0.007	0.928						
3.3	1.162	0.069	0.070	0.023	0.007	0.931						
1:M Match 4.2	1.133	0.054	0.059	-0.007	0.004	0.926						
4.3	1.143	0.055	0.059	0.003	0.004	0.933						
4.4	1.141	0.055	0.058	0.001	0.004	0.933						
4.5	1.148	0.055	0.057	0.008	0.004	0.940						
Strat-10 5.2	1.137	0.055	0.060	-0.003	0.005	0.927	1.314	0.069	0.066	0.199	0.047	0.323
5.3	1.136	0.055	0.059	-0.004	0.005	0.928	1.314	0.069	0.065	0.198	0.047	0.325
5.4	1.143	0.055	0.059	0.004	0.004	0.938	1.125	0.057	0.055	0.010	0.004	0.958
Strat-20 5.2	1.152	0.055	0.058	0.012	0.005	0.935	1.331	0.070	0.066	0.216	0.054	0.262
5.3	1.150	0.055	0.057	0.011	0.004	0.939	1.331	0.069	0.066	0.215	0.054	0.257
5.4	1.158	0.055	0.056	0.018	0.005	0.941	1.141	0.058	0.055	0.026	0.005	0.946
Strat-fine 5.2	1.164	0.058	0.060	0.024	0.005	0.921	1.149	0.066	0.083	0.033	0.012	0.920
5.3	1.162	0.058	0.060	0.022	0.005	0.929	1.148	0.066	0.083	0.032	0.011	0.922
5.4	1.162	0.059	0.060	0.022	0.005	0.926	1.159	0.067	0.082	0.043	0.012	0.914
IPTW 6.1	1.164	0.055	0.056	0.025	0.005	0.927	1.122	0.058	0.057	0.006	0.004	0.950
stab-IPTW 6.1	1.163	0.055	0.056	0.023	0.005	0.930	1.121	0.058	0.057	0.006	0.004	0.950
IPTW 6.2	1.163	0.055	0.056	0.023	0.005	0.932	1.121	0.058	0.056	0.006	0.004	0.951
stab-IPTW 6.2	1.163	0.055	0.056	0.023	0.005	0.932	1.122	0.058	0.057	0.006	0.004	0.950
IPTW 6.3	1.170	0.055	0.055	0.030	0.005	0.930	1.154	0.058	0.056	0.039	0.006	0.917
stab-IPTW 6.3	1.171	0.055	0.055	0.031	0.005	0.930	1.154	0.058	0.056	0.039	0.006	0.916

Table A 1. 17 Scenario 15 results. Mean Estimated Hazard Ratio ($\widehat{HR}$), Standard Error ($\widehat{SE}$), Monte Carlo Standard Deviation (MC SD), Bias, Mean Squared Error (MSE), and Coverage of 95% Confidence Intervals for ATT And ATE Parameters. SE and SD Reported on the LogHR Scale. Bias and MSE are not Calculated for Methods in Top Section of the Table Because the Conditional Effect Is not Defined Under Heterogeneity. Other Calculations are With Respect to Marginal ATT = 1.279 and Marginal ATE = 1.344.

	ATT						ATE					
Method	$\widehat{HR}$	$\widehat{SE}$	**MC SD**	**Bias**	**MSE**	**Coverage**	$\widehat{HR}$	$\widehat{SE}$	**MC SD**	**Bias**	**MSE**	**Coverage**
Cov Reg 1.1							1.409	0.053	0.053			
1.2							1.409	0.053	0.053			
PS Reg 2.1							1.361	0.053	0.053			
2.2							1.339	0.053	0.052			
2.3							1.362	0.053	0.053			
1:1 Match 3.1	1.361	0.106	0.106									
1:M Match 4.1	1.374	0.067	0.067									
Strat-10 5.1	1.318	0.053	0.052				1.338	0.053	0.053			
Strat-20 5.1	1.340	0.053	0.052				1.352	0.053	0.053			
Strat-fine 5.1	1.373	0.057	0.057				1.367	0.074	0.082			
1:1 Match 3.2	1.309	0.068	0.066	0.030	0.008	0.952						
3.3	1.309	0.068	0.066	0.030	0.008	0.950						
1:M Match 4.2	1.330	0.053	0.052	0.051	0.007	0.898						
4.3	1.286	0.053	0.052	0.007	0.004	0.959						
4.4	1.286	0.053	0.052	0.007	0.004	0.959						
4.5	1.287	0.053	0.052	0.008	0.004	0.961						
Strat-10 5.2	1.281	0.053	0.052	0.001	0.004	0.960	1.339	0.065	0.063	-0.005	0.007	0.951
5.3	1.281	0.053	0.052	0.001	0.004	0.957	1.338	0.066	0.063	-0.006	0.007	0.950
5.4	1.282	0.053	0.052	0.002	0.004	0.960	1.358	0.053	0.052	0.013	0.005	0.946
Strat-20 5.2	1.296	0.053	0.052	0.017	0.005	0.960	1.354	0.066	0.063	0.010	0.007	0.950
5.3	1.296	0.053	0.052	0.016	0.005	0.957	1.354	0.066	0.064	0.009	0.007	0.950
5.4	1.297	0.053	0.052	0.018	0.005	0.958	1.373	0.054	0.052	0.029	0.006	0.939
Strat-fine 5.2	1.309	0.057	0.056	0.030	0.006	0.946	1.333	0.063	0.066	-0.011	0.008	0.947
5.3	1.309	0.057	0.056	0.029	0.006	0.946	1.333	0.063	0.066	-0.011	0.008	0.946
5.4	1.311	0.057	0.056	0.031	0.006	0.942	1.334	0.063	0.066	-0.011	0.008	0.946
IPTW 6.1	1.310	0.053	0.052	0.030	0.005	0.944	1.383	0.054	0.052	0.039	0.007	0.925
stab-IPTW 6.1	1.310	0.053	0.052	0.031	0.006	0.944	1.384	0.054	0.052	0.040	0.007	0.923
IPTW 6.2	1.309	0.053	0.052	0.030	0.005	0.943	1.383	0.054	0.052	0.039	0.007	0.924
stab-IPTW 6.2	1.310	0.053	0.052	0.030	0.006	0.943	1.384	0.054	0.052	0.039	0.007	0.924
IPTW 6.3	1.311	0.053	0.052	0.031	0.006	0.940	1.383	0.054	0.052	0.038	0.007	0.927
stab-IPTW 6.3	1.311	0.053	0.052	0.032	0.006	0.940	1.384	0.054	0.052	0.039	0.007	0.923

Table A 1. 18 Scenario 16 results. Mean Estimated Hazard Ratio ($\widehat{HR}$), Standard Error ($\widehat{SE}$), Monte Carlo Standard Deviation (MC SD), Bias, Mean Squared Error (MSE), and Coverage of 95% Confidence Intervals for ATT And ATE Parameters. SE and SD Reported on the LogHR Scale. Bias and MSE are not Calculated for Methods in Top Section of the Table Because the Conditional Effect is not Defined Under Heterogeneity. Other Calculations are With Respect to Marginal ATT = 1.174 and Marginal ATE = 1.265.

Method	ATT						ATE					
	$\widehat{HR}$	$\widehat{SE}$	MC SD	Bias	MSE	Coverage	$\widehat{HR}$	$\widehat{SE}$	MC SD	Bias	MSE	Coverage
Cov Reg 1.1							1.233	0.053	0.054			
1.2							1.234	0.053	0.054			
PS Reg 2.1							1.230	0.053	0.053			
2.2							1.221	0.053	0.053			
2.3							1.230	0.053	0.053			
1:1 Match 3.1	1.203	0.103	0.106									
1:M Match 4.1	1.253	0.067	0.068									
Strat-10 5.1	1.212	0.053	0.053				1.220	0.053	0.053			
Strat-20 5.1	1.222	0.053	0.053				1.226	0.053	0.053			
Strat-fine 5.1	1.241	0.057	0.057				1.227	0.073	0.074			
1:1 Match 3.2	1.188	0.066	0.065	0.014	0.006	0.951						
3.3	1.188	0.066	0.065	0.014	0.006	0.953						
1:M Match 4.2	1.214	0.053	0.053	0.041	0.006	0.900						
4.3	1.175	0.053	0.053	0.002	0.004	0.956						
4.4	1.175	0.053	0.053	0.001	0.004	0.955						
4.5	1.177	0.053	0.053	0.003	0.004	0.956						
Strat-10 5.2	1.174	0.053	0.053	0.000	0.004	0.956	1.211	0.059	0.065	-0.055	0.009	0.828
5.3	1.174	0.053	0.052	0.000	0.004	0.956	1.210	0.059	0.065	-0.055	0.009	0.827
5.4	1.175	0.053	0.052	0.002	0.004	0.956	1.228	0.053	0.065	-0.037	0.008	0.839
Strat-20 5.2	1.181	0.053	0.053	0.007	0.004	0.959	1.217	0.060	0.066	-0.048	0.009	0.838
5.3	1.180	0.053	0.052	0.007	0.004	0.959	1.217	0.060	0.066	-0.049	0.009	0.840
5.4	1.182	0.053	0.052	0.008	0.004	0.957	1.235	0.053	0.066	-0.031	0.008	0.844
Strat-fine 5.2	1.186	0.056	0.055	0.013	0.004	0.956	1.202	0.063	0.066	-0.063	0.010	0.861
5.3	1.186	0.056	0.055	0.012	0.004	0.955	1.202	0.063	0.066	-0.063	0.010	0.863
5.4	1.188	0.056	0.055	0.014	0.005	0.952	1.203	0.063	0.066	-0.062	0.010	0.865
IPTW 6.1	1.187	0.053	0.053	0.013	0.004	0.958	1.288	0.054	0.053	0.023	0.005	0.943
stab-IPTW 6.1	1.187	0.053	0.053	0.013	0.004	0.958	1.289	0.054	0.053	0.023	0.005	0.942
IPTW 6.2	1.186	0.053	0.053	0.012	0.004	0.958	1.288	0.054	0.053	0.023	0.005	0.945
stab-IPTW 6.2	1.187	0.053	0.053	0.013	0.004	0.958	1.288	0.054	0.053	0.023	0.005	0.944
IPTW 6.3	1.188	0.053	0.053	0.014	0.004	0.955	1.287	0.054	0.053	0.022	0.005	0.948
stab-IPTW 6.3	1.188	0.053	0.053	0.015	0.004	0.955	1.288	0.054	0.053	0.022	0.005	0.947

Table A 1. 19 Scenario 17 results. Mean Estimated Hazard Ratio ($\widehat{HR}$), Standard Error ($\widehat{SE}$), Monte Carlo Standard Deviation (MC SD), Bias, Mean Squared Error (MSE), and Coverage of 95% Confidence Intervals for ATT And ATE Parameters. SE and SD Reported on the LogHR Scale. Bias and MSE are not Calculated for Methods in Top Section of the Table Because the Conditional Effect is not Defined Under Heterogeneity. Other Calculations are With Respect to Marginal ATT = 0.963 and Marginal ATE = 0.965.

	ATT						ATE					
Method	$\widehat{HR}$	$\widehat{SE}$	**MC SD**	**Bias**	**MSE**	**Coverage**	$\widehat{HR}$	$\widehat{SE}$	**MC SD**	**Bias**	**MSE**	**Coverage**
Cov Reg 1.1							0.971	0.053	0.053			
1.2							0.971	0.053	0.053			
PS Reg 2.1							0.971	0.053	0.053			
2.2							0.971	0.053	0.053			
2.3							0.971	0.053	0.053			
1:1 Match 3.1	0.977	0.099	0.103									
1:M Match 4.1	0.974	0.069	0.069									
Strat-10 5.1	0.971	0.053	0.053				0.970	0.053	0.053			
Strat-20 5.1	0.971	0.053	0.053				0.970	0.053	0.053			
Strat-fine 5.1	0.971	0.058	0.058				0.973	0.072	0.073			
1:1 Match 3.2	0.971	0.064	0.065	0.005	0.004	0.941						
3.3	0.971	0.064	0.065	0.005	0.004	0.941						
1:M Match 4.2	0.971	0.053	0.053	0.005	0.003	0.950						
4.3	0.971	0.053	0.054	0.005	0.003	0.943						
4.4	0.971	0.053	0.054	0.005	0.003	0.943						
4.5	0.972	0.053	0.054	0.006	0.003	0.946						
Strat-10 5.2	0.971	0.053	0.053	0.005	0.003	0.946	0.970	0.067	0.068	0.005	0.004	0.944
5.3	0.971	0.053	0.053	0.005	0.003	0.946	0.970	0.067	0.068	0.005	0.004	0.944
5.4	0.972	0.053	0.053	0.005	0.003	0.949	0.971	0.054	0.054	0.006	0.003	0.952
Strat-20 5.2	0.971	0.053	0.053	0.005	0.003	0.947	0.970	0.067	0.068	0.005	0.004	0.947
5.3	0.971	0.053	0.053	0.005	0.003	0.948	0.970	0.067	0.068	0.005	0.004	0.947
5.4	0.971	0.053	0.053	0.005	0.003	0.950	0.971	0.054	0.054	0.006	0.003	0.949
Strat-fine 5.2	0.971	0.056	0.056	0.005	0.003	0.952	0.970	0.062	0.063	0.005	0.004	0.952
5.3	0.971	0.056	0.056	0.005	0.003	0.952	0.970	0.062	0.063	0.005	0.004	0.952
5.4	0.972	0.056	0.056	0.006	0.003	0.952	0.971	0.062	0.063	0.005	0.004	0.952
IPTW 6.1	0.971	0.053	0.054	0.005	0.003	0.946	0.970	0.054	0.054	0.005	0.003	0.952
stab-IPTW 6.1	0.971	0.053	0.054	0.005	0.003	0.946	0.971	0.054	0.054	0.005	0.003	0.951
IPTW 6.2	0.971	0.053	0.054	0.005	0.003	0.946	0.970	0.054	0.054	0.005	0.003	0.952
stab-IPTW 6.2	0.971	0.053	0.054	0.005	0.003	0.946	0.971	0.054	0.054	0.005	0.003	0.950
IPTW 6.3	0.972	0.053	0.053	0.005	0.003	0.949	0.971	0.054	0.054	0.005	0.003	0.955
stab-IPTW 6.3	0.972	0.053	0.054	0.010	0.003	0.952	0.971	0.054	0.054	0.006	0.003	0.958

Table A 1. 20 Scenario 18 results. Mean Estimated Hazard Ratio ($\widehat{HR}$), Standard Error ($\widehat{SE}$), Monte Carlo Standard Deviation (MC SD), Bias, Mean Squared Error (MSE), and Coverage of 95% Confidence Intervals for ATT And ATE Parameters. SE and SD Reported on the LogHR Scale. Bias and MSE are not Calculated for Methods in Top Section of the Table Because the Conditional Effect is not Defined Under Heterogeneity. Other Calculations are With Respect to Marginal ATT = 1.005 and Marginal ATE = 1.007.

	ATT						ATE					
Method	$\widehat{HR}$	$\widehat{SE}$	**MC SD**	**Bias**	**MSE**	**Coverage**	$\widehat{HR}$	$\widehat{SE}$	**MC SD**	**Bias**	**MSE**	**Coverage**
Cov Reg 1.1							1.040	0.060	0.062			
1.2							1.040	0.060	0.062			
PS Reg 2.1							1.028	0.060	0.059			
2.2							0.989	0.060	0.058			
2.3							1.028	0.060	0.059			
1:1 Match 3.1	1.093	0.117	0.119									
1:M Match 4.1	0.994	0.072	0.072									
Strat-10 5.1	0.959	0.059	0.058				0.990	0.060	0.058			
Strat-20 5.1	0.992	0.060	0.058				1.011	0.060	0.058			
Strat-fine 5.1	1.030	0.064	0.061				1.040	0.081	0.099			
1:1 Match 3.2	1.025	0.076	0.071	0.020	0.006	0.956						
3.3	1.025	0.076	0.071	0.020	0.006	0.957						
1:M Match 4.2	0.977	0.060	0.058	-0.028	0.004	0.933						
4.3	0.982	0.060	0.057	-0.023	0.004	0.944						
4.4	0.982	0.060	0.057	-0.023	0.004	0.943						
4.5	0.983	0.060	0.057	-0.022	0.004	0.947						
Strat-10 5.2	0.969	0.060	0.057	-0.036	0.004	0.915	0.988	0.074	0.068	-0.019	0.005	0.953
5.3	0.969	0.060	0.057	-0.036	0.004	0.914	0.988	0.074	0.068	-0.019	0.005	0.954
5.4	0.969	0.060	0.057	-0.036	0.004	0.917	0.988	0.061	0.056	-0.019	0.003	0.940
Strat-20 5.2	0.998	0.060	0.057	-0.007	0.003	0.961	1.009	0.074	0.068	0.002	0.005	0.968
5.3	0.998	0.060	0.057	-0.007	0.003	0.960	1.009	0.075	0.068	0.002	0.005	0.970
5.4	0.999	0.060	0.057	-0.006	0.003	0.964	1.009	0.061	0.056	0.002	0.003	0.964
Strat-fine 5.2	1.024	0.063	0.059	0.019	0.004	0.960	1.025	0.071	0.068	0.018	0.005	0.960
5.3	1.024	0.063	0.059	0.019	0.004	0.958	1.025	0.071	0.067	0.018	0.005	0.960
5.4	1.025	0.063	0.059	0.020	0.004	0.959	1.025	0.071	0.067	0.018	0.005	0.960
IPTW 6.1	1.024	0.060	0.057	0.019	0.004	0.956	1.026	0.062	0.058	0.019	0.004	0.956
stab-IPTW 6.1	1.024	0.060	0.057	0.019	0.004	0.956	1.026	0.062	0.058	0.019	0.004	0.956
IPTW 6.2	1.024	0.060	0.057	0.019	0.004	0.956	1.026	0.062	0.058	0.019	0.004	0.957
stab-IPTW 6.2	1.024	0.060	0.057	0.019	0.004	0.956	1.026	0.062	0.058	0.019	0.004	0.957
IPTW 6.3	1.024	0.060	0.057	0.019	0.004	0.956	1.026	0.062	0.058	0.019	0.004	0.955
stab-IPTW 6.3	1.025	0.060	0.057	0.020	0.004	0.955	1.027	0.062	0.058	0.020	0.004	0.954

Table A 1. 21 Scenario 19 results. Mean Estimated Hazard Ratio ($\widehat{HR}$), Standard Error ($\widehat{SE}$), Monte Carlo Standard Deviation (MC SD), Bias, Mean Squared Error (MSE), and Coverage Of 95% Confidence Intervals for ATT and ATE Parameters. SE and SD Reported on the Log HR Scale. Bias and MSE on the HR Scale are Calculated With Respect to the True Conditional HR = 2 for Methods In Top Section of the Table, and With Respect to Marginal ATT = 1.889, Marginal ATE = 1.819 for Remaining Methods.

	ATT						ATE					
Method	$\widehat{HR}$	$\widehat{SE}$	**MC SD**	**Bias**	**MSE**	**Coverage**	$\widehat{HR}$	$\widehat{SE}$	**MC SD**	**Bias**	**MSE**	**Coverage**
Cov Reg 1.1							2.319	0.053	0.054	0.319	0.117	0.222
1.2							2.322	0.053	0.054	0.322	0.119	0.215
PS Reg 2.1							2.241	0.053	0.053	0.241	0.072	0.435
2.2							2.190	0.053	0.053	0.190	0.049	0.599
2.3							2.241	0.053	0.053	0.241	0.072	0.438
1:1 Match 3.1	2.206	0.113	0.115	0.206	0.108	0.885						
1:M Match 4.1	2.175	0.069	0.066	0.175	0.051	0.786						
Strat-10 5.1	2.069	0.052	0.051	0.069	0.016	0.908	2.190	0.053	0.053	0.190	0.050	0.594
Strat-20 5.1	2.148	0.053	0.052	0.148	0.034	0.738	2.224	0.053	0.053	0.224	0.064	0.489
Strat-fine 5.1	2.221	0.058	0.059	0.221	0.066	0.558	2.186	0.079	0.093	0.186	0.073	0.781
1:1 Match 3.2	2.121	0.072	0.073	0.229	0.077	0.647						
3.3	2.121	0.072	0.073	0.229	0.076	0.648						
1:M Match 4.2	2.175	0.053	0.053	0.284	0.093	0.248						
4.3	2.181	0.055	0.055	0.290	0.098	0.284						
4.4	2.180	0.055	0.055	0.289	0.098	0.283						
4.5	2.182	0.055	0.055	0.290	0.098	0.279						
Strat-10 5.2	2.112	0.055	0.054	0.221	0.061	0.469	2.069	0.067	0.066	0.249	0.080	0.523
5.3	2.112	0.055	0.054	0.220	0.061	0.472	2.068	0.067	0.066	0.248	0.080	0.528
5.4	2.113	0.055	0.054	0.221	0.062	0.465	2.069	0.057	0.056	0.249	0.075	0.380
Strat-20 5.2	2.173	0.055	0.054	0.282	0.093	0.299	2.103	0.069	0.067	0.283	0.100	0.440
5.3	2.173	0.055	0.054	0.282	0.093	0.302	2.102	0.069	0.067	0.282	0.099	0.448
5.4	2.174	0.055	0.054	0.283	0.094	0.292	2.103	0.058	0.057	0.283	0.094	0.300
Strat-fine 5.2	2.204	0.062	0.062	0.313	0.116	0.313	2.132	0.064	0.066	0.312	0.117	0.307
5.3	2.204	0.062	0.062	0.313	0.116	0.314	2.131	0.064	0.065	0.311	0.117	0.308
5.4	2.208	0.061	0.062	0.317	0.119	0.305	2.132	0.064	0.066	0.312	0.117	0.307
IPTW 6.1	2.231	0.056	0.055	0.340	0.131	0.176	2.098	0.059	0.058	0.277	0.092	0.324
stab-IPTW 6.1	2.232	0.056	0.055	0.341	0.131	0.173	2.110	0.060	0.059	0.289	0.099	0.315
IPTW 6.2	2.231	0.056	0.055	0.339	0.130	0.175	2.097	0.059	0.058	0.277	0.091	0.327
stab-IPTW 6.2	2.232	0.056	0.055	0.340	0.131	0.174	2.109	0.060	0.059	0.289	0.099	0.311
IPTW 6.3	2.233	0.056	0.055	0.341	0.131	0.166	2.097	0.058	0.058	0.277	0.091	0.314
stab-IPTW 6.3	2.234	0.056	0.055	0.342	0.132	0.165	2.107	0.060	0.059	0.287	0.098	0.301

Table A 1. 22 Scenario 20 Results. Mean Estimated Hazard Ratio ($\widehat{HR}$), Standard Error ($\widehat{SE}$), Monte Carlo Standard Deviation (MC SD), Bias, Mean Squared Error (MSE), and Coverage Of 95% Confidence Intervals for ATT and ATE Parameters. SE and SD Reported on the Log HR Scale. Bias and MSE on the HR Scale are Calculated With Respect to the True Conditional HR = 2 for Methods In Top Section of the Table, and With Respect to Marginal ATT =1.45, Marginal ATE = 1.406 for Remaining Methods.

	ATT						ATE					
Method	$\widehat{HR}$	$\widehat{SE}$	**MC SD**	**Bias**	**MSE**	**Coverage**	$\widehat{HR}$	$\widehat{SE}$	**MC SD**	**Bias**	**MSE**	**Coverage**
Cov Reg 1.1							2.003	0.052	0.052	0.003	0.011	0.949
1.2							2.005	0.053	0.052	0.005	0.011	0.948
PS Reg 2.1							1.732	0.052	0.051	-0.268	0.080	0.202
2.2							1.655	0.052	0.050	-0.345	0.126	0.036
2.3							1.733	0.052	0.051	-0.267	0.079	0.201
1:1 Match 3.1	1.879	0.110	0.112	-0.121	0.059	0.892						
1:M Match 4.1	1.664	0.066	0.064	-0.336	0.124	0.173						
Strat-10 5.1	1.595	0.051	0.049	-0.405	0.170	0.003	1.654	0.052	0.050	-0.346	0.126	0.034
Strat-20 5.1	1.658	0.052	0.050	-0.342	0.124	0.025	1.696	0.052	0.050	-0.304	0.099	0.100
Strat-fine 5.1	1.732	0.057	0.054	-0.268	0.081	0.174	1.768	0.075	0.105	-0.232	0.084	0.636
1:1 Match 3.2	1.685	0.071	0.066	0.234	0.067	0.452						
3.3	1.685	0.071	0.066	0.234	0.067	0.450						
1:M Match 4.2	1.630	0.052	0.049	0.179	0.039	0.394						
4.3	1.622	0.052	0.048	0.171	0.035	0.439						
4.4	1.622	0.052	0.048	0.171	0.035	0.437						
4.5	1.623	0.052	0.048	0.172	0.036	0.431						
Strat-10 5.2	1.606	0.052	0.048	0.155	0.030	0.507	1.590	0.064	0.058	0.184	0.042	0.520
5.3	1.606	0.052	0.048	0.155	0.030	0.510	1.589	0.064	0.058	0.183	0.042	0.525
5.4	1.607	0.052	0.048	0.156	0.030	0.509	1.593	0.052	0.048	0.187	0.041	0.342
Strat-20 5.2	1.650	0.052	0.049	0.199	0.046	0.308	1.617	0.065	0.058	0.211	0.053	0.405
5.3	1.650	0.052	0.049	0.199	0.046	0.305	1.616	0.065	0.058	0.210	0.053	0.409
5.4	1.651	0.052	0.049	0.200	0.046	0.310	1.620	0.053	0.047	0.214	0.052	0.229
Strat-fine 5.2	1.685	0.055	0.050	0.234	0.062	0.232	1.673	0.063	0.059	0.267	0.081	0.198
5.3	1.684	0.056	0.050	0.234	0.062	0.231	1.673	0.063	0.059	0.267	0.081	0.200
5.4	1.686	0.055	0.050	0.235	0.063	0.224	1.673	0.063	0.060	0.267	0.081	0.196
IPTW 6.1	1.687	0.052	0.048	0.236	0.062	0.180	1.635	0.053	0.048	0.229	0.059	0.180
stab-IPTW 6.1	1.688	0.052	0.048	0.237	0.063	0.178	1.636	0.053	0.048	0.230	0.059	0.179
IPTW 6.2	1.687	0.052	0.048	0.236	0.062	0.183	1.635	0.053	0.048	0.229	0.059	0.179
stab-IPTW 6.2	1.687	0.052	0.048	0.237	0.063	0.178	1.636	0.053	0.048	0.230	0.059	0.179
IPTW 6.3	1.688	0.052	0.048	0.237	0.063	0.178	1.635	0.053	0.048	0.229	0.059	0.181
stab-IPTW 6.3	1.688	0.052	0.048	0.238	0.063	0.176	1.636	0.053	0.048	0.230	0.059	0.178

Table A 1. 23 Scenario 21 results. Mean Estimated Hazard Ratio ($\widehat{HR}$), Standard Error ($\widehat{SE}$), Monte Carlo Standard Deviation (MC SD), Bias, Mean Squared Error (MSE), and Coverage Of 95% Confidence Intervals for ATT and ATE Parameters. SE and SD Reported on the Log HR Scale. Bias and MSE on the HR Scale are Calculated With Respect to the True Conditional HR = 2 for Methods In Top Section of the Table, and With Respect to Marginal ATT =1.682 and marginal ATE = 1.618 for Remaining Methods.

	ATT						ATE					
Method	$\widehat{HR}$	$\widehat{SE}$	**MC SD**	**Bias**	**MSE**	**Coverage**	$\widehat{HR}$	$\widehat{SE}$	**MC SD**	**Bias**	**MSE**	**Coverage**
Cov Reg 1.1							2.013	0.098	0.100	0.013	0.041	0.947
1.2							2.020	0.099	0.101	0.020	0.043	0.943
PS Reg 2.1							1.837	0.097	0.100	-0.163	0.061	0.855
2.2							1.718	0.096	0.097	-0.282	0.108	0.637
2.3							1.839	0.097	0.099	-0.161	0.060	0.856
1:1 Match 3.1	1.996	0.232	0.237	-0.004	0.237	0.951						
1:M Match 4.1	1.741	0.132	0.138	-0.259	0.125	0.779						
Strat-10 5.1	1.631	0.096	0.097	-0.369	0.161	0.407	1.715	0.096	0.098	-0.285	0.110	0.621
Strat-20 5.1	1.718	0.096	0.098	-0.282	0.108	0.632	1.777	0.097	0.099	-0.223	0.081	0.775
Strat-fine 5.1	1.845	0.111	0.113	-0.155	0.069	0.878	1.878	0.154	0.185	-0.122	0.131	0.879
1:1 Match 3.2	1.792	0.136	0.130	0.110	0.067	0.941						
3.3	1.792	0.136	0.130	0.110	0.067	0.941						
1:M Match 4.2	1.707	0.097	0.098	0.025	0.029	0.940						
4.3	1.681	0.097	0.097	-0.001	0.027	0.950						
4.4	1.681	0.097	0.097	-0.001	0.027	0.951						
4.5	1.683	0.097	0.097	0.002	0.027	0.950						
Strat-10 5.2	1.628	0.096	0.097	-0.054	0.028	0.937	1.654	0.119	0.118	0.035	0.039	0.950
5.3	1.628	0.096	0.097	-0.054	0.028	0.938	1.653	0.119	0.118	0.035	0.039	0.951
5.4	1.630	0.096	0.097	-0.052	0.028	0.937	1.648	0.097	0.097	0.029	0.027	0.936
Strat-20 5.2	1.703	0.097	0.097	0.021	0.028	0.937	1.702	0.120	0.118	0.083	0.047	0.937
5.3	1.703	0.097	0.097	0.021	0.028	0.938	1.701	0.120	0.118	0.083	0.047	0.935
5.4	1.705	0.097	0.097	0.024	0.028	0.936	1.696	0.098	0.097	0.078	0.033	0.919
Strat-fine 5.2	1.785	0.103	0.102	0.103	0.044	0.913	1.779	0.121	0.120	0.161	0.071	0.882
5.3	1.785	0.103	0.102	0.103	0.044	0.914	1.778	0.121	0.120	0.160	0.071	0.881
5.4	1.789	0.103	0.102	0.107	0.045	0.909	1.783	0.120	0.125	0.165	0.079	0.880
IPTW 6.1	1.788	0.098	0.098	0.106	0.042	0.901	1.734	0.100	0.101	0.116	0.044	0.893
stab-IPTW 6.1	1.788	0.098	0.098	0.106	0.043	0.898	1.735	0.100	0.101	0.117	0.045	0.892
IPTW 6.2	1.788	0.098	0.098	0.106	0.042	0.901	1.734	0.100	0.101	0.116	0.044	0.892
stab-IPTW 6.2	1.788	0.098	0.098	0.106	0.043	0.900	1.735	0.100	0.101	0.117	0.044	0.891
IPTW 6.3	1.791	0.098	0.098	0.109	0.043	0.898	1.736	0.100	0.100	0.118	0.045	0.889
stab-IPTW 6.3	1.791	0.098	0.098	0.110	0.043	0.898	1.738	0.100	0.100	0.120	0.045	0.889

Table A 1. 24 Rejection Rates by Scenario for Methods That Estimate Conditional or Marginal HR ATT. Rejection Rates < 0.8 for Scenarios With a Non-Null Treatment Effect are shown in bold.

		S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11	S12	S13	S14	S15	S16	S17	S18	S19	S20	S21
1:1 Match	3.1	1	**0.702**	**0.278**	**0.676**	**1**	**0.457**	0.939	0.048	1	**0.334**	0.047	**0.402**	1	**0.321**	0.815	**0.431**	**0.079**	**0.105**	1	1	0.819
1:M Match	4.1	1	0.939	**0.547**	0.986	1	**0.791**	0.999	0.056	1	0.806	0.06	**0.617**	1	**0.423**	0.995	0.902	**0.073**	**0.051**	1	1	0.982
Strat-10	5.1	1	0.968	**0.617**	0.990	1	0.960	0.999	0.153	1	1	0.254	**0.483**	1	**0.574**	0.999	0.943	**0.079**	**0.095**	1	1	0.998
Strat-20	5.1	1	0.993	**0.702**	0.994	1	0.971	1	0.082	1	0.977	0.086	**0.731**	1	**0.671**	1	0.959	**0.079**	**0.046**	1	1	0.999
Strat-fine	5.1	1	0.994	**0.772**	0.995	1	0.951	0.999	0.062	0.999	**0.633**	0.042	0.851	1	**0.687**	1	0.957	**0.080**	**0.063**	1	1	1
1:1 Match	3.2	1	0.953	**0.546**	0.891	1	0.884	0.999	0.053	1	**0.444**	0.040	**0.702**	1	**0.582**	0.980	**0.752**	**0.073**	**0.044**	1	1	0.992
	3.3	1	0.953	**0.548**	0.888	1	0.885	0.999	0.052	1	**0.450**	0.041	**0.704**	1	**0.581**	0.980	**0.751**	**0.074**	**0.046**	1	1	0.992
	4.2	1	0.991	**0.696**	0.996	1	0.968	1	0.054	1	0.875	0.059	**0.780**	1	**0.607**	1	0.950	**0.079**	**0.064**	1	1	0.999
1:M Match	4.3	1	0.990	**0.694**	0.996	1	0.960	0.999	0.049	1	0.821	0.054	**0.775**	1	**0.658**	0.998	0.842	**0.080**	**0.047**	1	1	0.999
	4.4	1	0.990	**0.694**	0.996	1	0.960	0.999	0.048	1	0.818	0.052	**0.775**	1	**0.65**	0.998	0.841	**0.080**	**0.045**	1	1	0.999
	4.5	1	0.990	**0.709**	0.996	1	0.960	0.999	0.047	1	0.816	0.053	**0.777**	1	**0.692**	0.998	0.846	**0.076**	**0.046**	1	1	0.999
Strat-10	5.2	1	0.972	**0.633**	0.981	1	0.961	0.999	0.142	1	0.997	0.156	**0.560**	1	**0.626**	0.998	0.833	**0.078**	**0.071**	1	1	0.996
	5.3	1	0.971	**0.632**	0.980	1	0.961	0.999	0.144	1	0.997	0.155	**0.561**	1	**0.626**	0.998	0.834	**0.077**	**0.074**	1	1	0.996
	5.4	1	0.974	**0.640**	0.982	1	0.960	0.999	0.151	1	0.996	0.150	**0.564**	1	**0.656**	0.998	0.838	**0.075**	**0.068**	1	1	0.996
Strat-20	5.2	1	0.99	**0.699**	0.993	1	0.970	0.999	0.069	1	0.944	0.068	**0.751**	1	**0.712**	0.998	0.859	**0.081**	**0.033**	1	1	0.999
	5.3	1	0.991	**0.699**	0.993	1	0.970	0.999	0.067	1	0.943	0.067	**0.748**	1	**0.707**	0.998	0.855	**0.080**	**0.034**	1	1	0.999
	5.4	1	0.991	**0.712**	0.995	1	0.972	0.999	0.070	1	0.939	0.062	**0.756**	1	**0.739**	0.998	0.865	**0.075**	**0.033**	1	1	0.999
Strat-fine	5.2	1	0.999	**0.766**	0.996	1	0.954	1	0.040	0.999	**0.593**	0.043	0.801	1	**0.729**	0.996	0.844	**0.070**	**0.048**	1	1	1
	5.3	1	0.999	**0.766**	0.996	1	0.954	1	0.039	0.999	**0.593**	0.044	0.801	1	**0.723**	0.996	0.844	**0.069**	**0.049**	1	1	1
	5.4	1	1	**0.790**	0.997	1	0.955	1	0.041	0.999	**0.583**	0.041	0.809	1	**0.714**	0.996	0.849	**0.067**	**0.05**	1	1	1
IPTW	6.1	1	0.999	**0.796**	0.997	1	0.984	1	0.046	1	**0.609**	0.038	0.864	1	**0.768**	0.998	0.880	**0.081**	**0.051**	1	1	1
IPTW-s	6.1	1	0.999	**0.796**	0.997	1	0.984	1	0.047	1	**0.609**	0.039	0.864	1	**0.765**	0.998	0.881	**0.081**	**0.05**	1	1	1
IPTW	6.2	1	0.999	**0.796**	0.997	1	0.986	1	0.046	1	**0.603**	0.038	0.865	1	**0.765**	0.998	0.881	**0.081**	**0.048**	1	1	1
IPTW-s	6.2	1	0.999	**0.796**	0.998	1	0.985	1	0.045	1	**0.605**	0.040	0.865	1	**0.764**	0.998	0.882	**0.081**	**0.048**	1	1	1
IPTW	6.3	1	0.999	0.806	0.998	1	0.984	1	0.045	1	**0.591**	0.039	0.868	1	0.798	0.998	0.886	**0.077**	**0.052**	**1**	1	1
IPTW-s	6.3	1	0.999	0.807	0.998	1	0.984	1	0.046	1	**0.591**	0.040	0.868	1	0.799	0.998	0.886	**0.077**	**0.053**	**1**	1	1

Table A 1. 25 Rejection Rates by Scenario for Methods That Estimate Conditional or Marginal HR ATE. Rejection Rates < 0.8 for Scenarios With a Non-Null Treatment Effect are Shown In Bold.

		S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11	S12	S13	S14	S15	S16	S17	S18	S19	S20	S21
Cov Reg	1.1	1	1	0.863	1	1	0.985	1	0.061	1	0.824	0.048	0.947	1	0.816	1	0.972	**0.081**	**0.092**	1	1	1
	1.2	1	1	0.865	1	1	0.985	1	0.062	1	0.826	0.048	0.946	1	0.865	1	0.974	**0.075**	**0.098**	1	1	1
PS Reg	2.1	1	1	0.821	0.996	1	0.985	1	0.054	1	**0.624**	0.037	0.900	1	0.827	1	0.970	**0.079**	**0.067**	1	1	1
	2.2	1	0.994	**0.709**	0.995	1	0.975	1	0.395	1	0.909	0.047	0.812	1	**0.755**	1	0.960	**0.082**	**0.05**	1	1	0.999
	2.3	1	1	0.819	0.996	1	0.985	1	0.054	1	**0.619**	0.039	0.898	1	0.823	1	0.969	**0.081**	**0.067**	1	1	1
Strat-10	5.1	1	0.992	**0.700**	0.992	1	0.968	1	0.369	1	0.859	0.051	0.801	1	**0.503**	1	0.958	**0.082**	**0.049**	1	1	0.999
Strat-20	5.1	1	0.997	**0.751**	0.996	1	0.980	1	0.154	1	**0.757**	0.042	0.880	1	**0.605**	1	0.964	**0.082**	**0.049**	1	1	0.999
Strat-fine	5.1	0.992	0.851	**0.532**	0.890	0.997	**0.710**	0.982	0.055	0.998	**0.440**	0.066	**0.557**	1	**0.451**	0.972	**0.788**	**0.077**	**0.114**	1	0.993	0.962
Strat-10	5.2	1	0.918	**0.524**	0.878	1	0.842	0.998	0.152	0.999	**0.634**	0.045	**0.459**	1	0.969	0.994	0.868	**0.084**	**0.041**	1	1	0.988
	5.3	1	0.916	**0.524**	0.873	1	0.837	0.998	0.151	0.999	**0.628**	0.044	**0.460**	1	0.970	0.994	0.866	**0.084**	**0.039**	1	1	0.985
	5.4	1	0.974	**0.663**	0.969	1	0.952	0.999	0.238	1	**0.753**	0.055	**0.582**	1	**0.552**	1	0.914	**0.082**	**0.053**	1	1	0.998
Strat-20	5.2	1	0.947	**0.567**	0.908	1	0.870	0.998	0.060	0.999	**0.481**	0.034	**0.528**	1	0.976	0.995	0.881	**0.085**	**0.031**	1	1	0.994
	5.3	1	0.946	**0.567**	0.903	1	0.870	0.998	0.065	0.999	**0.470**	0.034	**0.525**	1	0.976	0.995	0.881	**0.085**	**0.031**	1	1	0.994
Strat-20	5.4	1	0.990	**0.704**	0.978	1	0.966	1	0.089	1	**0.569**	0.034	**0.658**	1	**0.638**	1	0.929	**0.082**	**0.035**	1	1	0.999
Strat-fine	5.2	0.997	0.959	**0.664**	0.964	1	0.910	0.995	0.049	0.999	**0.457**	0.037	**0.752**	1	**0.525**	0.976	0.819	**0.081**	**0.046**	1	0.999	0.989
	5.3	0.997	0.959	**0.663**	0.963	1	0.909	0.995	0.047	0.999	**0.450**	0.037	**0.751**	1	**0.525**	0.975	0.818	**0.080**	**0.047**	1	0.999	0.989
	5.4	0.997	0.959	**0.669**	0.965	1	0.912	0.995	0.048	0.999	**0.448**	0.038	**0.754**	1	**0.580**	0.976	0.821	**0.077**	**0.047**	1	0.999	0.989
IPTW	6.1	1	0.991	**0.746**	0.963	1	0.970	1	0.049	1	**0.447**	0.037	**0.654**	1	**0.516**	1	0.995	**0.082**	**0.053**	1	1	1
IPTW-s	6.1	1	0.991	**0.746**	0.979	1	0.970	1	0.050	1	**0.327**	0.032	**0.669**	1	**0.511**	1	0.995	**0.082**	**0.054**	1	1	1
IPTW	6.2	1	0.991	**0.744**	0.963	1	0.971	1	0.050	1	**0.443**	0.037	**0.652**	1	**0.512**	1	0.995	**0.082**	**0.051**	1	1	1
IPTW-s	6.2	1	0.991	**0.744**	0.979	1	0.971	1	0.050	1	**0.324**	0.032	**0.670**	1	**0.512**	1	0.995	**0.082**	**0.051**	1	1	1
IPTW	6.3	1	0.990	**0.751**	0.965	1	0.971	1	0.048	1	**0.460**	0.038	**0.668**	1	**0.705**	1	0.995	**0.082**	**0.051**	1	1	1
IPTW-s	6.3	1	0.990	**0.751**	0.985	1	0.971	1	0.048	1	**0.351**	0.038	**0.671**	1	**0.705**	1	0.995	**0.082**	**0.051**	1	1	1

Appendix 2. Propensity Score Models and Diagnostics

Propensity Score Models for Data Generation

In keeping with the plasmode simulation approach, propensity scores (PS) for baseline scenario 1 were drawn from covariate-treatment associations in the real world data. A logistic regression of Z on X was fit to the data. The intercept was adjusted to yield a data generating model that preserved real-world associations while assigning 25% of the population to the treated group. This PS model was used in baseline scenario 1. These coefficients were modified for other scenarios to achieve the desired level and direction of confounding, and the designated proportion of treated subjects. PSs at both sites were generated from the same model (Table A1.1).

PS models were designed to assure reasonable overlap between treated and comparator subjects (except PS Model 6, designed for poor overlap). The C-statistic offers a summary of the degree of overlap (1). A C-statistic of 0.5 indicates the distribution of PSs in treated and comparator subjects are indistinguishable. A C-statistic of 1 indicates complete separability (no overlap), and no support in the data for estimating a treatment effect. Higher values reflect a decreasing amount of information in the data for estimating the causal contrast, and likely a higher degree of variability in estimates produced by PS-based methods. C-statistics for our PSs ranged from 0.61 to 0.73 for all scenarios except Scenario 9, where the C-statistic = 0.83.

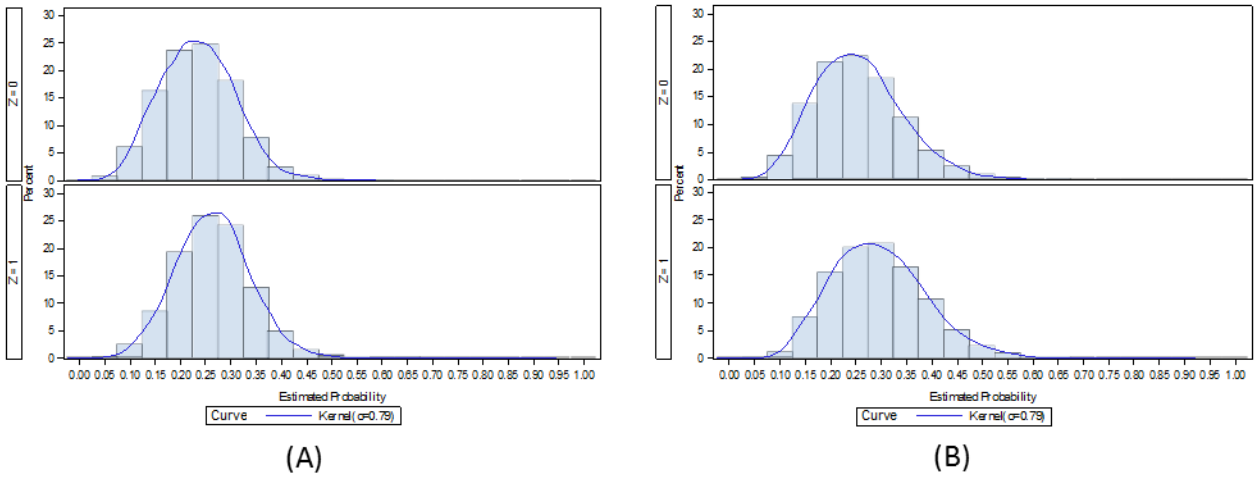
Propensity Score Model Diagnostics

C-statistics associated with PSs generated by each model were calculated as an average over 100 datasets. PS distributions, covariate balance established by 1:1 and 1:M variable ratio matching and distribution of unstabilized and stabilized IPTW weights are shown for a single representative dataset for two PS models that were used in nine of our scenarios (diagnostics for additional scenarios are available online (2)). Although these diagnostics provide useful information, imbalances in weak confounders (or instrumental variables) might not introduce noticeable bias, and perfect balance on measured covariates does not imply that unmeasured confounders are balanced (3-6).

PS Model 1: Scenarios 1, 2, 3, 6, 7, 13, 14, 17

PS Diagnostics: The C-statistic of 0.62 indicates good overlap in the distributions of PSs in treated and comparator groups at both sites (Fig. A3.1). 9,942 out of 9,950 treated subjects were retained in the post-match dataset. Under variable ratio matching the number of comparator subjects increased to 29,830. Covariate balance in the pooled data from both sites as measured by the standardized mean difference improved markedly after 1:1 or 1:M matching (Table A2.1).

Figure A 2. 1 Distribution of PS Model 1 Propensity Scores in Treated (Bottom) and Comparator (Top) Populations at Site 1 (A) and Site 2 (B).



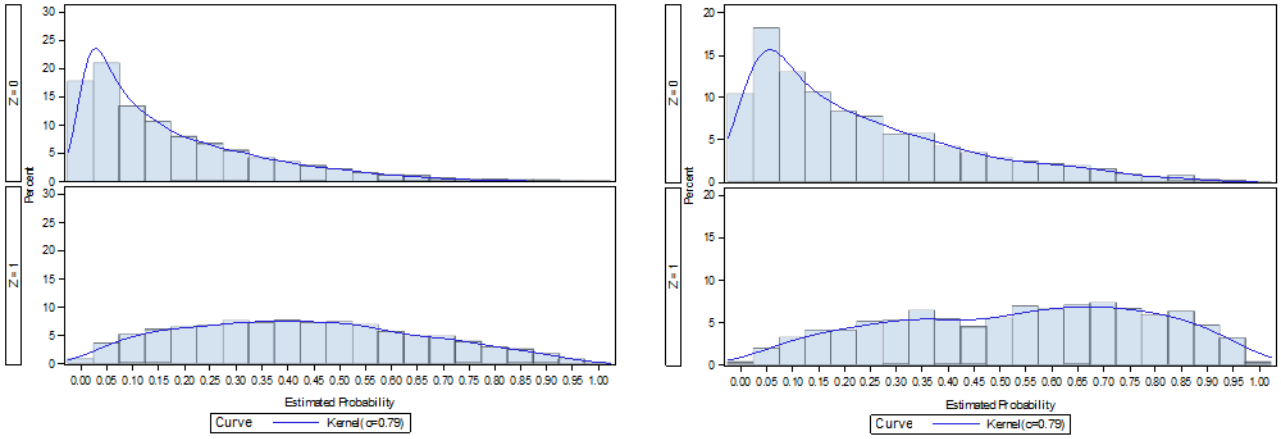
(A) (B)

Consistent with theory, stabilized IP weights had a mean of 1.00 for ATT (min = 0.04, max = 5.44) and for ATE (min = 0.38, max = 5.22). Unstabilized weights had a mean equal to 0.49 for ATT (min = 0.013, max = 1.77), and 2 for ATE (min = 1.01, max = 21.25). As expected, unstabilized ATE weights sum to $2n$, while unstabilized ATT weights sum to $2n_{\text{treated}}$. Since all weights were less than 50, none were truncated. (A maximum weight of 50 was chosen to allow us to see effects of large weights on the analysis while still implementing the estimator in a way that is consistent with recommended practice, analogous to setting a caliper of 0.2 in the nearest neighbor matching algorithm.)

PS Model 6: Scenario 9

PS Diagnostics: The C-statistic of 0.83 indicates poor overlap in the distributions of PSs in treated and comparator groups at both sites (Fig. A3.5). 7 Only 7,893 of 0,021 treated subjects were retained in the post-match dataset. Any bias in ATT estimates may therefore be partly attributable to this modification of the target population. Under variable ratio matching the number of comparator subjects increased to 23,187. Covariate balance in the pooled data from both sites as measured by the standardized mean difference improved markedly after 1:1 or 1:M matching (Table A2.2).

Figure A 2. 2 Distribution of PS Model 6 Propensity Scores in Treated (Bottom) and Comparator (Top) Populations at Site 1 (A) and Site 2 (B).



(A) (B)

Consistent with theory, stabilized IP weights had a mean of 1.00 for ATT (min = 0.00, max = 68.66, 99th percentile = 7.18%) and for ATE (min = 0.25, max = 75.69, 99th percentile = 4.07). Unstabilized weights had a mean equal to 1.00 for ATT (min = 0.00, max = 22.90), and 2.00 for ATE (min = 1, max = 302.65, 99th percentile = 13.27). Unstabilized ATE weights sum to approximately $2n$, while unstabilized ATT weights sum to $2n_{\text{treated}}$. After truncation stabilized ATT and ATE weights still had mean 1.00, while the mean of the truncated unstabilized ATE weights was 1.95.

Table A 2. 1 PS Model 5. Covariate Means, Standard Deviations (SD) and Standardized Mean Differences Before and After 1:1 and 1:M Matching.

	Before Matching					1:1 Matching					1:M Variable Ratio Matching				
	Comparator (n = 29,759)		Treatment (n = 10,021)			Comparator (n = 7,893)		Treatment (n = 7,893)			Comparator (n = 23,187)		Treatment (n = 7,893)		
Covariate	Mean	SD	Mean	SD	SMD	Mean	SD	Mean	SD	SMD	Mean	SD	Mean	SD	SMD
Age	76.58	8.62	69.92	10.28	0.702	72.22	8.62	72.28	8.64	0.008	72.24	7.46	72.28	8.64	0.006
Ambul_visits	14.89	12.79	9.87	7.67	0.476	10.30	8.04	10.30	7.87	0.000	10.36	6.16	10.30	7.87	0.009
Outpatient_visits	3.51	4.80	2.09	3.08	0.352	2.24	3.35	2.19	3.14	0.015	2.22	2.48	2.19	3.14	0.009
Inpatient_visits	0.82	1.02	0.60	0.87	0.234	0.59	0.89	0.59	0.87	0.004	0.59	0.68	0.59	0.87	0.003
ComorbidityScore	3.78	2.91	1.90	1.99	0.753	2.10	2.09	2.08	2.06	0.012	2.11	1.70	2.08	2.06	0.014
AtrialFib (y/n)	0.12	0.32	0.25	0.43	0.348	0.19	0.39	0.19	0.39	0.002	0.19	0.34	0.19	0.39	0.002
Diabetes (y/n)	0.42	0.49	0.26	0.44	0.341	0.28	0.45	0.29	0.45	0.008	0.29	0.35	0.29	0.45	0.003
GI_Bleed (y/n)	0.06	0.24	0.02	0.15	0.184	0.03	0.16	0.03	0.16	0.006	0.03	0.12	0.03	0.16	0.008
MI (y/n)	0.17	0.38	0.09	0.28	0.251	0.10	0.29	0.10	0.29	0.003	0.10	0.22	0.10	0.29	0.003
RenalDisease (y/n)	0.29	0.45	0.08	0.27	0.549	0.10	0.30	0.10	0.30	0.000	0.10	0.22	0.10	0.30	0.014
Diuretic (y/n)	0.44	0.50	0.24	0.43	0.415	0.27	0.44	0.27	0.44	0.002	0.27	0.34	0.27	0.44	0.006
NumDrugClasses	8.54	5.04	7.99	4.91	0.110	7.92	4.85	7.95	4.79	0.006	7.96	3.90	7.95	4.79	0.003
Prior_Ischemic (y/n)	0.16	0.37	0.08	0.27	0.241	0.09	0.28	0.09	0.29	0.019	0.09	0.21	0.09	0.29	0.006
Other_bleed (y/n)	0.10	0.29	0.04	0.19	0.244	0.04	0.19	0.04	0.20	0.010	0.04	0.14	0.04	0.20	0.006
Sex	0.55	0.50	0.55	0.50	0.000	0.55	0.50	0.54	0.50	0.006	0.54	0.40	0.54	0.50	0.003

References

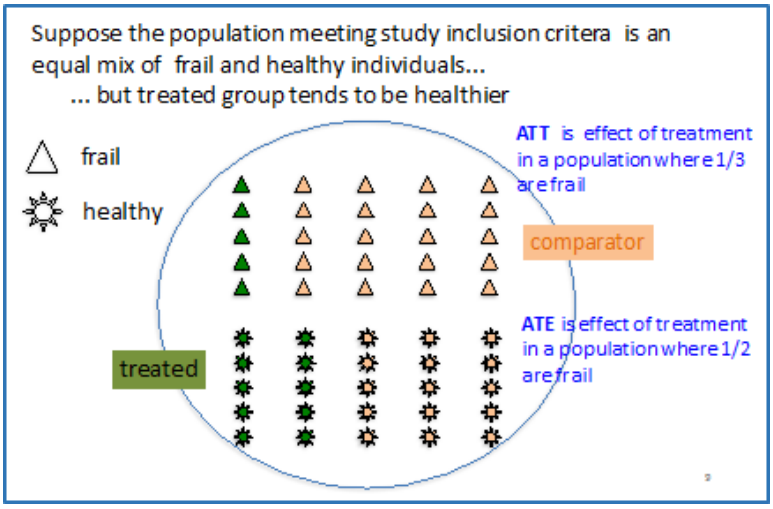
1. Westreich D, Cole SR, Funk M, Brookhart MA, Stürmer T. The role of the C-statistic in variable selection for propensity score models. *Pharmacoepidemiol Drug Saf.* 2011;20:317–320.
2. Evaluation of Propensity Score Based Methods in Sentinel Study Settings Using Simulation Experiments. *Sentinel Coordinating Center*. June 15, 2017. https://www.sentinelinitiative.org/sentinel/methods/evaluation-propensity-score-based-methods-sentinel-study-settings-using-simulation. Accessed March 28, 2018.
3. Austin PC, Stuart EA. Moving towards best practice when using inverse probability of treatment weighting (IPTW) using the propensity score to estimate causal treatment effects in observational studies. *Stat Med. 2015;*34:3661–3679.
4. Jackson JW. Diagnostics for confounding of time-varying and other joint exposures. *Epidemiology*. 2016;27(6):859-869.
5. Walker AM. Matching on provider is risky. *J Clin Epidemiol.* 2013;66(8):S65-68.
6. Pearl JM. Invited Commentary: Understanding Bias Amplification. *Am J Epidemiol. 2011;*174(11): 1223-1227.

Appendix 3. Technical Definitions

Average Treatment Effect (ATE) vs. Average Treatment Effect Among the Treated (ATT)

In observational settings people with certain covariate profiles may be more likely to be treated than people with other covariate profiles. When there is a homogeneous treatment effect the conditional hazard ratio is constant within all patient sub-populations. However, if background risk depends upon the covariate profile then the average, or marginal effect of treatment in the study population (ATE) will differ from the average effect of treatment in the treated population (ATT). For example, in the hypothetical population of healthy and frail people depicted in Figure A3.1 the ATT is the effect of treatment in a population where 1/3 are frail, and ATE is the effect of treatment in a population where 1/2 are frail. If frailty affects outcome risk, the ATT will not equal the ATE. Exceptions are when there is no effect of treatment, when the target of estimation is on the additive scale and the outcome is continuous, and when the distribution of covariates in the treated population is the same as in the entire study population.

Figure A 3. 1 Example of ATT and ATE target populations. Treated population (green), comparator population (orange) have different proportions of frail and healthy individuals.



Statistical Methods for HR Estimation for Time to Event Outcomes

Suppose there are K sites and each site $k = 1, \dots, K$, there are n_k observations $\{(\mathbf{X}_{i,k}, Z_{i,k}, T_{i,k}, \Delta_{i,k}): i = 1, \dots, n_k\}$ where $\mathbf{X}_{i,k}$ denote the vector of confounders, $Z_{i,k}$ denote the treatment variable (1 for treatment and 0 for comparator), $T_{i,k}$ denote the follow-up time (the minimum of survival and censoring times), and $\Delta_{i,k,}$ denote the censoring variable (1 for event and 0 for censoring). Let $e(\mathbf{X}_{i,k}) \equiv \Pr(Z_{i,k} = 1 | \mathbf{X}_{i,k})$ denote the PS. Let $(T_{1,i,k}, \Delta_{1,i,k})$ and $(T_{0,i,k}, \Delta_{0,i,k})$ denote the two pairs of potential outcomes on censoring variable and follow-up time for treatment and comparator respectively. A causal interpretation of the HR estimate rests on the assumption that $(T_{i,k}, \Delta_{i,k}) = (T_{1,i,k}, \Delta_{1,i,k})$ if $Z = 1$, $(T_{i,k}, \Delta_{i,k}) = (T_{1,i,k}, \Delta_{1,i,k})$ if $Z = 0$ (the *consistency* assumption). In addition, we assume that the PS is bounded away from both 0 and 1 ($0 < e(\mathbf{X}) < 1$, the *positivity* assumption). To derive valid causal inference from observational databases, the *assumption of no unmeasured confounding (NUC)* is required, i.e., $(\Delta_{1,i,k,} T_{1,i,k}, \Delta_{0,i,k,} T_{0,i,k}) \coprod Z \mid \mathbf{X}$ or $(\Delta_{1,i,k,} T_{1,i,k}, \Delta_{0,i,k,} T_{0,i,k}) \coprod Z \mid e(\mathbf{X})$.

There are two risk metrics of interest (RMI): the conditional hazard ratio (HR) and the marginal HR. We first introduce the two RMIs in the absence of censoring ($\Delta = 1$ for all subjects) and then discuss the impact of censoring under various censoring mechanisms. Let $\lambda(t|\mathbf{X},Z)$ denote the hazard rate for the survival time T conditional on $(\mathbf{X},Z)$. Suppose $\lambda(t|\mathbf{X},Z) = \lambda_0(t)\exp(h(\mathbf{X}) + \theta_c Z)$ following a Cox Proportional Hazards model, then $\exp(\theta_c)$ denotes the constant conditional HR given $\mathbf{X}$.

We define the marginal hazard ratio as the population mean hazard ratio averaged over the duration of the study. The definition of the marginal effect measure for survival outcomes is more complex than for binary outcomes. For a binary variable, the entire distribution is determined by the probability of success. For survival outcome, suppose $\lambda(t|\mathbf{X},Z) = \lambda_0(t)\exp(h(\mathbf{X}) + \theta_c Z)$, then the marginal hazard rates for the two potential survival times T_1 and T_0 are unlikely to follow a proportional hazards model. For each scenario in our simulation studies we estimated the marginal ATE HR ($\exp(\theta_m)$) empirically by repeating the following procedure 100 times and averaging the estimates.

1. sample covariates with replacement from each site in the claims data then pool the data.
2. artificially set administrative censoring time to C = 370 days (2 years)
3. for each simulated subject sample event times (T_1, T_0) under both counterfactual levels of treatment, $Z = 1$ and $Z = 0$, from the pre-specified Weibull distribution conditional on covariates.
4. output two rows per simulated subject with
    - $Z = 1$, T = min(C, T_1) and event status indicator $\Delta = 1$ if T < C, else 0
    - $Z = 0$, T = min(C, T_0) and event status indicator $\Delta = 1$ if T < C, else 0
5. Fit a Cox PH model regressing T on Z $(t|\mathbf{X},Z) = \lambda_0(t)\exp(\theta_m Z)$

The estimated coefficient for Z is the empirical estimate for the log of the "marginal HR", which we denote as θ_m. The marginal HR is the average, $\widehat{HR_m} = \frac{1}{100}\sum_{i=1}^{100}\exp(\theta_{m,i})$.

A slight modification allowed us to estimate the marginal ATT HR. When simulating the data we also generate an observed treatment indicator, Z_{obs}. Then fit the Cox model using observations contributed by people having $Z_{obs} = 1$, the treated population. Each treated person still contributes two counterfactual observations to the dataset, but now the distribution of event times corresponds to the distribution in the treated population rather than in the entire study population. The final estimate of the marginal ATT HR was the average over 100 iterations.

Due to differential depletion of high-risk individuals from the treatment groups, the marginal HR is not a constant over time in our baseline scenario where the conditional HR is constant by design. Furthermore, some PS-estimators which permit balanced comparison of treatment groups at baseline may increasingly diverge from $\exp(\theta_c)$ if outcome incidene is not low and the treatment groups are depleted differentially. SAS code for each of the estimators described next is available from the authors upon request.

Covariate-Adjusted Regression

This is the classical confounding adjustment method, included as a benchmark for comparing PS-based methods. Analysis 1.1 requires pooling patient-level data. Meta-analysis 1.2 requires site-specific parameter and variance estimators only.

Analysis 1.1

Fit a Cox PH regression model among the entire study population regressing (T,Δ) on Z and $\mathbf{X}$ with the correctly specified outcome regression model. Let $\hat{\theta}_{cr,1}$ denote the estimated coefficient for Z. The RMI is θ_c.

Analysis 1.2

Fit site-specific Cox PH regressions (with the same correct regression model) regressing (T, Δ) on Z and $\mathbf{X}$, combine site-specific results to construct an overall log(HR) estimator, $\hat{\theta}_{cr,2}$, and 95% CI, using the fixed-effects meta-analysis approach. The RMI is θ_c.

PS Regression

We consider 3 working models for the outcome regression adjusting for PS using polynomial terms $(\hat{e}(\mathbf{X}), \hat{e}(\mathbf{X})^2, \hat{e}(\mathbf{X})^3)$, a categorical variable defined by site-specific deciles, $d_s[\hat{e}(\mathbf{X})]$, and cubic B-splines with estimated quintiles as internal knots, $b_s[\hat{e}(\mathbf{X})]$, respectively. Analyses 2.1-2.3 correspond to the 3 working models respectively.

Analysis 2.1

Site-specific Cox PH regression model regressing (T, Δ) on $Z, \hat{e}(\mathbf{X}), \hat{e}(\mathbf{X})^2, \hat{e}(\mathbf{X})^3$, pool results across sites via a fixed-effects meta-analysis approach. let $\hat{\theta}_{r,2,1}$ denote the log(HR) estimator. The RMI is $\approx \theta_c$.

Analysis 2.2

Site-specific Cox PH regression model regressing (T, Δ) on $Z, d_s[\hat{e}(\mathbf{X})]$, pool results across sites via a fixed-effects meta-analysis approach. let $\hat{\theta}_{r,2,2}$ denote the log(HR) estimator. The RMI is $\approx \theta_c\theta_c$.

Analysis 2.3

Site-specific Cox PH regression model regressing (T, Δ) on $Z, b_s[\hat{e}(\mathbf{X})]$, pool results across sites via a fixed-effects meta-analysis approach. let $\hat{\theta}_{r,2,3}$ denote the log(HR) estimator. The RMI is $\approx \theta_c\theta_c$.

PS 1:1 Matching

We used FDA Sentinel's propensity score matching (PSM) tool which implements a nearest neighbor matching algorithm, matching on the probability scale (PS) with a caliper of 0.05. This creates a comparator population whose propensity score distribution matches the distribution in the treated population, naturally estimating the ATT. We did not investigate full matching to estimate an ATE.

Analysis 3.1

Fit a stratified Cox PH regression model among the matched population regressing (T, Δ) on Z, stratifying on both site and matched set. Let $\hat{\theta}_{fm,1,1}$ denote the estimated coefficient for Z. The RMI is $\approx \theta_c$.

Analysis 3.2

Fit a stratified Cox PH regression model among the matched population regressing (T, Δ) on Z, stratifying on site only. Let $\hat{\theta}_{fm,2,1}$ denote the estimated coefficient for Z. The RMI for this analysis is $\approx \theta_m$. Here the approximation is due to the adjustment of site.

Analysis 3.3

Fit a Cox PH regression model among the matched population regressing (T, Δ) on Z. Let $\hat{\theta}_{fm,3,1}$ denote the estimated coefficient for Z. The RMI is θ_m.

PS 1:*M* Variable Ratio Matching

We used Sentinel's PSM tool to implement 1:M variable ratio matching (M = 10) using a nearest neighbor matching algorithm, matching on PS with a caliper of 0.05. This creates a comparator population whose propensity score distribution matches the distribution in the treated population, naturally estimating the ATT. We did not investigate full matching to estimate an ATE.

Analysis 4.1

Fit a stratified Cox PH regression model among the matched population regressing (T, Δ) on Z, stratifying on both site and matched set. Let $\hat{\theta}_{vm,1,1}$ denote the estimated coefficient for Z. The RMI is $\approx \theta_c$.

Analysis 4.2

Fit a stratified Cox PH regression model among the matched population regressing (T, Δ) on Z, stratifying on both site and matching ratio m. Let $\hat{\theta}_{vm,2,1}$ denote the estimated coefficient for Z. The RMI for this analysis is $\approx \theta_m$.

Analyses 4.3-4.5

Incorporate ATT weights. The weight equals 1 for treated subjects and equals $1/m$ for the comparators where m denotes the number of comparators in the matched set. We did not investigate full matching to estimate the ATE.

Analysis 4.3

Fit a weighted Cox PH regression model among the matched population regressing (T, Δ) on Z, adjusting for site as a categorical covariate.. Let $\hat{\theta}_{vm,3,1}$ denote the estimated coefficient for Z_i. The RMI is $\approx \theta_m$. The model-based variance estimator incorrectly estimates the variance of $\hat{\theta}_{vm,3,1}$ and thus the robust, sandwich variance estimator is needed to construct confidence interval and conduct hypothesis testing.

Analysis 4.4

Fit a weighted Cox PH regression model among the matched population regressing (T, Δ) on Z. Let $\hat{\theta}_{vm,4,1}$ denote the estimated coefficient for Z_i. The RMI is θ_m. Use the robust, sandwich variance estimator.

Analysis 4.5

Fit site-specific weighted Cox PH regression models among the matched population regressing (T, Δ) on Z. Use the robust, sandwich variance estimator. Combine the results across sites using the fixed-effects meta-analysis approach. Let $\hat{\theta}_{vm,5,1}$ denote the estimated log(HR). The RMI for this analysis is $\approx \theta_m$.

PS Stratification

PS strata defined by site. We consider 3 stratification definitions, i.e., fixing the number of strata at each site at 10 and 20, as well as varying the number of strata based on sample size such that each stratum has 5 treated subjects (fine stratification, ATT) or 5 subjects, total (fine stratification, ATE). For a selected number of strata, PS strata are defined using estimated percentiles within the treated subjects to estimate the ATT, and within all subjects to estimate the ATE.

Analysis 5.1

Fit a stratified Cox PH regression model among the entire study population consisting of treated and comparator subjects regressing (T, Δ) on Z, stratifying on site and PS strata. Let $\hat{\theta}_{s,1,1}$ denote the estimated coefficient for Z. The RMI is $\approx \theta_c$.

For weighted analyses 5.2-5.4, the ATT weights equal 1 for treated subjects and $\frac{n_{1,s,k}}{n_{0,s,k}}$ for comparator subjects, where $(n_{1,s,k}, n_{0,s,k})$ denote the numbers of treated and comparator subjects in stratum s at site k. The ATE weights equal $1/n_{1,s,k}$ for treated subjects and $1/n_{0,s,k}$ for comparator subjects.

Analysis 5.2

Fit a weighted Cox PH regression model among the entire study population regressing (T, Δ) on Z, adjusting for site as a categorical covariate. Use the robust, sandwich variance estimator. Let $\hat{\theta}_{s,2,1}$ denote the estimated coefficient for Z. The RMI for this analysis is $\approx \theta_m$.

Analysis 5.3

Fit a weighted Cox PH regression model among the entire study population regressing (T, Δ) on Z. The weights are the same as in analysis 5.2. Use the robust, sandwich variance estimator. Let $\hat{\theta}_{s,3,1}$ denote the estimated coefficient for Z. The RMI for this analysis is θ_m.

Analysis 5.4

Fit site-specific weighted Cox PH regression models among the entire study population regressing (T, Δ) on Z. Use the robust, sandwich variance estimator. Combine the results across sites using the fixed-effects meta-analysis approach. Let $\hat{\theta}_{s,4,1}$ denote the estimated log(HR). The RMI for this analysis is $\approx \theta_m$.

IPTW

Unstabilized inverse probability weights for estimating the ATT for Analyses 6.1-6.3 equal 1 for treated subjects and $\frac{\hat{e}(\mathbf{X})}{1-\hat{e}(\mathbf{X})}$ for comparator subjects, where propensity score $\hat{e}(\mathbf{X})$ is estimated separately at each site. Unstabilized weights for estimating the ATE equal $\frac{1}{\hat{e}(\mathbf{X})}$ for treated subjects and $\frac{1}{1-\hat{e}(\mathbf{X})}$ for comparator subjects. For both ATT and ATE estimation weights can be stabilized in an effort to reduce variability. Weights for treated subjects are multiplied by $\bar{Z}$, the proportion of treated subjects in the population. Weights for untreated subjects are multiplied by $(1 - \bar{Z})$. For all IPTW analyses we used the robust, sandwich variance estimator.

Analysis 6.1

Fit a weighted Cox PH regression model regressing (T, Δ) on Z, adjusting for site as a categorical variable. Let $\hat{\theta}_{w,1}$ denote the estimated coefficient for Z_i, the RMI is $\approx \theta_m$.

Analysis 6.2

Fit a weighted Cox PH regression model regressing (T, Δ) on Z. Let $\hat{\theta}_{w,2}$ denote the estimated coefficient for Z_i, the RMI is θ_m.

Analysis 6.3

Fit site-specific weighted Cox PH regression models regressing (T, Δ) on Z. Combine the site-specific estimated coefficients for Z and the robust, sandwich variance estimators using a fixed-effects model to construct an overall estimator and variance. Let $\hat{\theta}_{w,3}$ denote the estimated coefficient for Z_i, the RMI is θ_m.

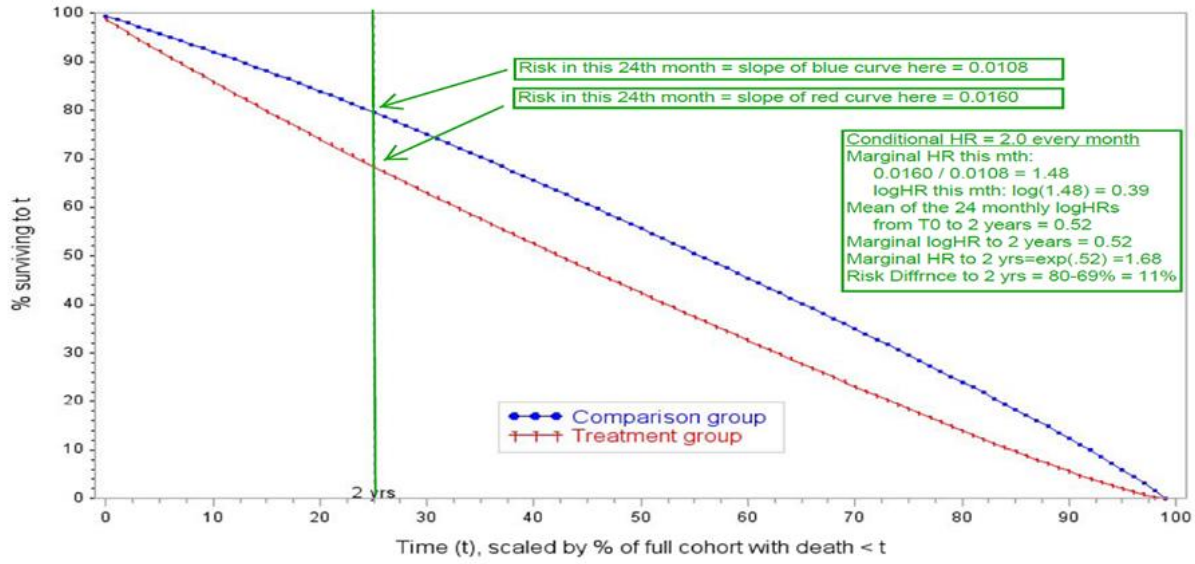
Appendix 4. Conditional and Marginal Hazard Ratios

In this appendix we show

- how the marginal hazard ratio, HR_m, diverges over time from the conditional hazard ratio, HR_c, toward the null (when the $HR_c \neq 1.0$)
- how PS-based estimators of the HR_c tend to be biased over time -- toward the HR_m -- by depletion from the cohort of people who have had outcome events.
- how PS-based estimators of the HR_m tend to be biased over time -- toward the HR_c -- by random censoring, if the cohort is depleted of higher risk individuals by outcome events.

We start by showing in Figure A4.1 what is meant by the marginal hazard ratio. The two survival curves show how the treatment and comparison groups are depleted over time by loss of the people who have an outcome event. Time in Figure A4.1 is scaled by the percentage of the entire cohort who have had an outcome event. At the outset, the cohort is balanced. It is a full counterfactual cohort, which means that it is comprised of pairs of "clones" such that each pair has a treated person and an untreated person with a matching covariate profile. The true HR_c in this simulation is 2.0: at every point in time the treatment doubles the risk of an outcome event in every individual who is still at risk. Over time the treatment group gets depleted at a faster rate than the untreated group – and we see the red curve in figure A4.1 diverge from the blue curve. (If the true HR_c was 1.0 instead of 2.0, the two groups would be depleted at the same rate and both survival curves would be precisely on the diagonal).

Figure A 4. 1 Survival by Treatment Arm in a Simulated Counterfactual Cohort. Treatment Assignment and Survival time are Driven by Three Covariates (Uniformly Distributed, Uncorrelated). Mechanisms Generating Treatment and Outcome: Logit(Tx) = $cov_1 + 3 \times cov_2 + 3 \times cov_3 - 3.5$. Exponential Survival Time with Linear Predictor = $ln(2) \times Tx + 1.5 \times cov_1 + 4.8 \times cov_2 + 1.3 \times cov_3$, Baseline Rate = 1/852.



After 24 months, 25% of the entire cohort has had an outcome event and is no longer in follow-up. The green text points out what we mean by the "instantaneous" HR_m at the 24th month and by the HR_m for the two years of follow-up until the end of the 24th month. During the 24th month 1.60% of the treated group had an outcome event compared with 1.08% of the untreated group. Thus, the slope of the red survival curve at the 24th month is about 0.0160, and the slope of the blue survival curve at that point is about 0.0108. At each timepoint, the instantaneous marginal HR is the ratio of the slope of the red curve

to the blue curve. The HR_m from T_0 until a 2-year endpoint amounts to the average (on the log scale) of instantaneous HRs over all times until the 2-year endpoint. The HR_m can also be found by Cox regression: it is the maximum likelihood estimate of the hazard ratio (on the log scale) that would be obtained from a Cox model, fitted to data on the full counterfactual cohort followed from T_0 until the 2-year endpoint, regressing time-to-event on treatment status without adjustment for covariates. In a counterfactual cohort – which is like an ideal RCT – the HR_m is the estimand of Cox regression without any adjustment for the baseline covariates, while the HR_c is the estimand of Cox regression that is adjusted for the baseline covariates.

Even if the HR_c is always 2.0, the HR_m diverges over time toward 1.0 as the cohort is depleted by outcome events. Table A4.1 shows this counterfactual cohort as entirely balanced at the outset of followup. The first row shows this initial balance: the mean of each covariate is 0.5 in each group, the mean PS in each group is 0.5, and the mean disease risk score (DRS) in each group is 3.8. This full cohort then loses its balance over time as the treated and untreated groups are differentially depleted of higher risk individuals. In the second row, profiling the cohort after it has been depleted by 5% who have had an outcome event, the "survivors" in each arm have a slightly lower risk profile with respect to each of the covariates. The treatment group lost more people due to outcome events than the untreated group and became slightly less risky – its average DRS decreased more. Over time the two arms become more unbalanced as they become more depleted by outcome events.

Table A 4. 1 Loss of Balance in a Full Counterfactual Cohort With no Censoring as Treatment (Trt) and Comparison (Comp) Groups are Differentially Depleted by Outcomes. Means of the Covariates, Propensity Score (PS), and Disease Risk Score (DRS) by Time[a]

Time	Covariate 1		Covariate 2		Covariate 3		PS		DRS	
Scale[a]	Trt	Comp	Trt	Comp	Trt	Comp	Trt	Comp	Trt	Comp
1	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500	3.800	3.800
5	0.494	0.497	0.484	0.492	0.495	0.497	0.487	0.493	3.706	3.751
10	0.487	0.492	0.463	0.480	0.488	0.494	0.470	0.484	3.588	3.685
20	0.475	0.484	0.422	0.455	0.478	0.486	0.437	0.463	3.362	3.542
30	0.465	0.475	0.380	0.423	0.470	0.478	0.404	0.438	3.131	3.365
40	0.456	0.466	0.336	0.388	0.462	0.471	0.371	0.410	2.898	3.173
50	0.446	0.458	0.293	0.349	0.453	0.464	0.337	0.380	2.665	2.967
60	0.435	0.449	0.254	0.307	0.444	0.456	0.306	0.348	2.449	2.741
70	0.417	0.438	0.211	0.262	0.429	0.446	0.270	0.313	2.198	2.496
80	0.392	0.421	0.171	0.218	0.407	0.432	0.232	0.276	1.938	2.237
90	0.346	0.389	0.128	0.168	0.367	0.404	0.185	0.229	1.612	1.915

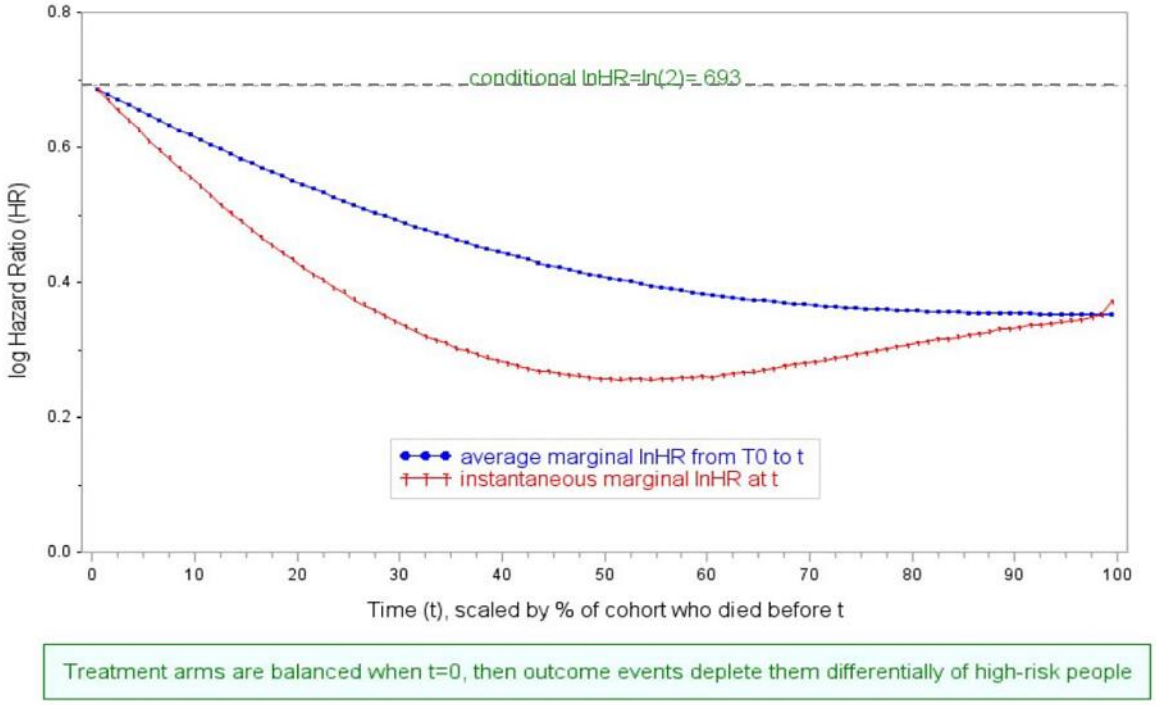
[a] Same cohort and time scale as Figure A4.1. When time = 10, 10% of the cohort has had an outcome event.

As the covariate profile of the treated survivors becomes less risky than that of the untreated survivors, the instantaneous HR (unadjusted for the covariates) diverges from 2.0, and the average of the instantaneous HRs – averaged over the followup period from T_0 to any endpoint of interest -- also diverges from 2.0, though somewhat more gradually. This pattern is shown in Figure A4.2.

In surveillance of a real cohort rather than this counterfactual cohort, we can emulate the Figure A4.2 analysis of the HR_m by using inverse probability of treatment weights (IPTW), or else by using our 1:1 PS-matched estimator with unstratified Cox regression. These estimators can consistently estimate the average HR_m in Figure A4.2 if there is no censoring. However, in drug safety surveillance there is often heavy early censoring, as many new-users quit their drug while others persist for years. Then long term estimates of the HR_m are derived disproportionately from the less censored outcomes earlier on the

timeline before there was much differential depletion of higher risk individuals, and consequently our IPTW and 1:1 matched estimators of the HR_m yield estimates that are biased away from their target toward the HR_c. This bias can arise from the amount and timing of the censoring even if who gets censored is entirely unrelated to treatment or risk. A way to avoid this bias in our PS-based estimators of the HR_m is to upweight the late under-represented risk sets.

Figure A 4. 2 Marginal Hazard Ratios Diverge From the Conditional Hazard Ratio Over Time in the Counterfactual Cohort (Same Cohort and Follow-Up Time Scale as Shown in Figure A4.1).



If we condition the analysis on the individual covariates (for example, by including them in a Cox regression model) it is possible to consistently estimate the HR_c despite the challenges from "depletion of susceptibles". However, if we condition on the baseline PS instead of the baseline covariates the HR estimate tends to land between the HR_c and the HR_m. As the cohort is depleted of susceptibles, there is change in the associations of the baseline covariates with treatment status (among the "survivors" who remain in followup), and so a PS derived from the cohort at baseline becomes less successful at adjusting analyses.

In Figure A4.3 the red curve shows how the HR estimates that were adjusted by the baseline PS fell between the true HR_c and the HR_m (on the log scale). This red curve shows average adjusted HR estimates (on the log scale) yielded by Cox models that were fit to the observed half of the same counterfactual cohort that was simulated for Figures A4.1 and A4.2. The correlation between the PS and the DRS is 0.87 in the scenario simulated for Figures A4.1-A4.3. If we modify the outcome-generating mechanism so that the DRS correlates nearly 1.0 with the PS, the PS-adjusted estimator will land very close to the target HR_c. This is shown in Figure A4.4. Conversely, if we modify the outcome-generating mechanism so that the PS is less correlated with the DRS, then our PS-regression estimator will land closer to the HR_m. Figure A4.5 shows the PS-regression estimates converging with the HR_m when the correlation of the PS with the DRS is nearly zero. Whether the PS-adjusted estimate lands closer to the HR_c or the HR_m depends on the size of the absolute value of this PS-DRS correlation, not on whether it is

positive or negative. In the baseline scenario examined in the main paper, the correlation between the PS and the DRS was -0.75.

For a PS-based method to consistently estimate the HR_c, the PS adjustment needs to keep the estimator on target as the treatment groups are depleted differentially. A Sentinel-sponsed workgroup lead by one of us (RW) is developing a method for calculating a time-varying PS that can achieve this goal.

Figure A 4. 3 Hazard Ratio (HR) Estimated by PS Regression Compared With True Conditional and Marginal HRs In Cohort Where the Correlation Between the True PS and the True Disease Risk Score (DRS) is 0.87. Red Curve at T = 10 Shows the PS-Adjusted Ln(HR) From T_0 Until the Time When 10% of the Full Cohort has had an Event. True Conditional HR = 2 (Shown at 0.693 on Vertical Axis, Log Scale).

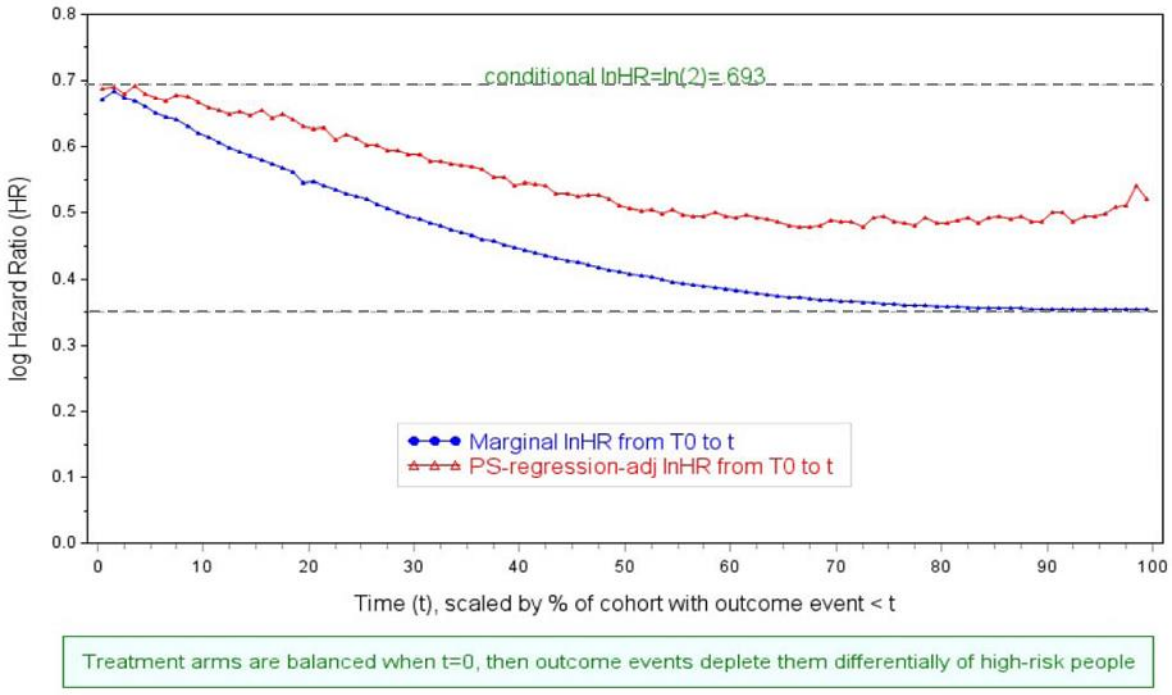


Figure A 4. 4 Hazard ratio (HR) Over Time Estimated by PS Regression Compared With True Conditional and Marginal HRs in Scenario Where the Correlation of the True PS with the Disease Risk score (DRS) is Near 1.0.

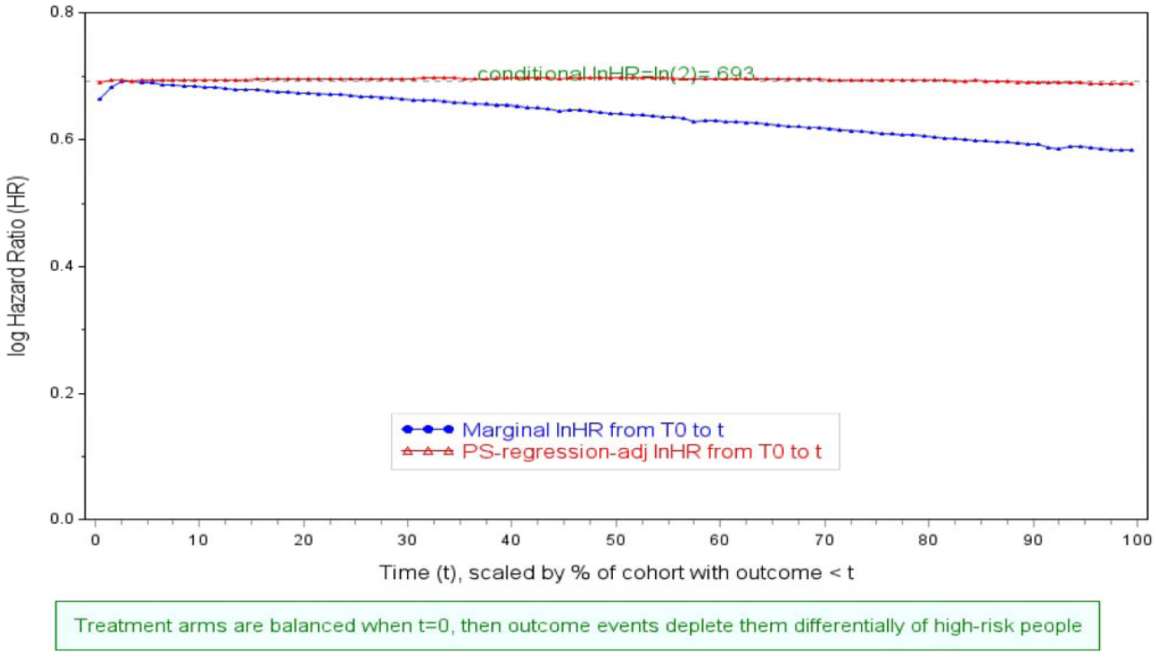


Figure A 4. 5 Hazard ratio (HR) Over Time Estimated by PS Regression Compared With True Conditional and Marginal HRs in Scenario Where the Correlation of the True PS with the Disease Risk score (DRS) is Near 0.