Received ; | Revised ; | Accepted

DOI: xxx/xxxx

RESEARCH ARTICLE

Estimating the Marginal Hazard Ratio by Simultaneously Using A Set of Propensity Score Models: A Multiply Robust Approach

Di Shu*[1] | Peisong Han[2] | Rui Wang[1,3] | Sengwee Toh[1]

[1]Department of Population Medicine, Harvard Medical School and Harvard Pilgrim Health Care Institute, Boston, Massachusetts, USA
[2]Department of Biostatistics, University of Michigan, Ann Arbor, Michigan, USA
[3]Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, Massachusetts, USA

Correspondence
*Di Shu, Department of Population Medicine, Harvard Medical School and Harvard Pilgrim Health Care Institute, 401 Park Drive, Suite 401 East, Boston, MA 02215, USA
Email: Di_Shu@harvardpilgrim.org

The inverse probability weighted Cox model is frequently used to estimate the marginal hazard ratio. Its validity requires a crucial condition that the propensity score model be correctly specified. To provide protection against misspecification of the propensity score model, we propose a weighted estimation method rooted in empirical likelihood theory. The proposed estimator is multiply robust in that it is guaranteed to be consistent when a set of postulated propensity score models contains a correctly specified model. Our simulation studies demonstrate satisfactory finite sample performance of the proposed method in terms of consistency and efficiency. We apply the proposed method to compare the risk of postoperative hospitalization between sleeve gastrectomy and Roux-en-Y gastric bypass using data from a large medical claims and billing database. We further extend the development to multi-site studies to enable each site to postulate multiple site-specific propensity score models.

KEYWORDS:
Cox model, inverse probability weighting, marginal hazard ratio, multiple robustness, propensity score.

1 | INTRODUCTION

In biomedical studies, marginal hazard ratios are commonly used to assess the effects of treatments on time-to-event outcomes by comparing the hazard functions of failure times between the treated and untreated individuals. In randomized controlled trials, fitting a Cox model relating the time-to-event outcome to only the treatment yields the estimated marginal hazard ratio. In observational studies, inverse probability weighted (IPW) Cox estimation provides one approach to estimating the marginal hazard ratio when measured confounders are adjusted for through weighting.[1,2,3,4,5,6,7,8,9] The weights are a function of the estimated propensity score, i.e., the probability of receiving treatment conditional on the measured baseline confounders.[10] Specifically, an IPW Cox model is a Cox model that relates the time-to-event outcome to only the treatment and weighs the individuals by the reciprocal of their estimated probabilities of receiving the observed treatment, i.e., the estimated propensity score (if treated) or 1 minus the estimated propensity score (if untreated). Like other standard propensity score methods, the validity of IPW Cox estimation requires correct specification of the propensity score model.

There is a growing interest in developing more robust methods to protect against potential misspecification of the propensity score model. For example, the doubly robust estimation methods provide consistent estimators if either the propensity score model or the outcome model is correctly specified.[11,12,13,14,15,16] Additional estimation methods have been developed to achieve multiply robustness, mainly in the context of missing data analysis.[17,18,19,20,21,22] An estimator is said to be multiply robust if it is consistent when a set of postulated propensity score or outcome models contains a correctly specified model.

To our knowledge, there are currently no doubly or multiply robust estimators for marginal hazard ratios. In this paper, we propose to estimate the marginal hazard ratio by fitting a weighted Cox model where the weights are obtained by adapting the method in Han and Wang,[17] which considered estimating a population mean for a non-survival outcome with missing values. Our method yields consistent marginal hazard ratio estimators as long as the set of postulated propensity score models contains a correctly specified model, thereby providing more protection against model misspecification than the commonly used IPW Cox estimation method. We further expand the proposed method to multi-site studies, where weighted Cox models stratified on data-contributing site will produce consistent estimators of the marginal hazard ratio when each site includes a correctly specified model in its set of propensity score models.

The rest of this paper is organized as follows. In Section 2, we describe the standard IPW Cox model framework for estimating the marginal hazard ratio using one (possibly misspecified) propensity score model. In Section 3, we discuss why propensity score model misspecification is a practical concern in observational studies and illustrate the need for doubly or multiply robust methods. In Section 4, we develop a multiply robust method to estimate marginal hazard ratios. In Sections 5 and 6, we conduct simulation studies to evaluate the finite sample performance of the proposed method, with and without including a correctly specified propensity score model in the set of postulated models, respectively. In Section 7, we apply the proposed method to analyze a real-world electronic healthcare dataset to compare the risk of postoperative hospitalization between sleeve gastrectomy and Roux-en-Y gastric bypass. In Section 8, we extend the development to multi-site studies. We conclude the paper with a discussion in Section 9.

2 | WEIGHTED ESTIMATION OF THE MARGINAL HAZARD RATIO USING ONE PROPENSITY SCORE MODEL

Let $\boldsymbol{X}$ be a vector of measured baseline covariates, A a binary treatment indicator ($A = 1$ if treated and $A = 0$ if untreated), and $T = \min(T^*, C)$ where T^* is the event time, C is the censoring time. Define $\delta = I(T^* \leq C)$ to be the event indicator, where $I(\cdot)$ is the indicator function.

Suppose we have an independent and identically distributed (i.i.d.) sample of size n. For the ith individual where $i = 1, \ldots, n$, the observed data are $(\boldsymbol{X}_i, A_i, T_i, \delta_i)$. We aim to use the observed data to estimate the log marginal hazard ratio θ of the model:

$$\lambda_a(t) = \lambda_0(t)\exp(\theta a), \tag{1}$$

where $\lambda_a(t)$ is the hazard function for T_a^* the event time for an individual that would have been observed had we set the treatment level $A = a$ for $a = 0$ or 1.

Provided that standard exchangeability, consistency, and positivity assumptions[9] hold, propensity score weighting effectively adjusts for confounding bias. Given that the weighted data emulate data that would have been collected from a randomized controlled trial, the IPW Cox models provide one approach to estimate marginal hazard ratios.

Based on propensity score $e(\boldsymbol{X}) = P(A = 1|\boldsymbol{X})$, the conventional inverse probability weights

$$w = w_{ipw} = \frac{A}{e(\boldsymbol{X})} + \frac{1-A}{1-e(\boldsymbol{X})} \tag{2}$$

and the stabilized weights

$$w = w_{stabilized} = P(A=1)\frac{A}{e(\boldsymbol{X})} + P(A=0)\frac{1-A}{1-e(\boldsymbol{X})} \tag{3}$$

are commonly used.[2,4,5] Since the treatment decision process is often unknown, one modeling strategy is to specify a parametric propensity score model for $e(\boldsymbol{X})$. We then estimate $e(\boldsymbol{X})$ by fitting the specified propensity score model relating the treatment indicator A to the baseline covariates $\boldsymbol{X}$. The treatment prevalence $P(A = 1)$ is estimated nonparametrically as the number of treated individuals divided by the total number of individuals in the study.

For $i = 1, \dots, n$, let $\widehat{w}_i$ denote the estimated weight for individual i under either the conventional inverse probability weights (2) or stabilized weights (3). The weighted partial likelihood score equation[23,24] for θ is

$$\sum_{i=1}^{n} \widehat{w}_i \delta_i \left\{ A_i - \frac{\sum_{l:l\in\mathfrak{R}_i} \widehat{w}_l \exp(A_l\theta)A_l}{\sum_{l:l\in\mathfrak{R}_i} \widehat{w}_l \exp(A_l\theta)} \right\} = 0, \tag{4}$$

where $\mathfrak{R}_i = \{l : T_l \geq T_i, \delta_i = 1\}$ is the risk set for uncensored individual i. Solving (4) for θ gives the IPW estimator of log hazard ratio, denoted by $\widehat{\theta}$.

The consistency of estimator $\widehat{\theta}$ requires the propensity score model be correctly specified. Misspecifying a propensity score model may result in severely biased results.

3 | CONCERNS ABOUT PROPENSITY SCORE MODEL MISSPECIFICATION IN OBSERVATIONAL STUDIES

Propensity score methods are commonly used in observational studies that investigate the effects of medical treatments, especially when there is more information to model the treatment decision process than the outcome process (e.g., drug safety studies with common exposures and rare outcomes). The validity of a typical propensity score-based analysis requires the propensity score model be correctly specified. However, treatment decision process is complex in clinical practice. Many studies collect high-dimensional data that may or may not fully capture factors that influence treatment decision, making it challenging to correctly specify the propensity score model.

Confounder selection has been extensively discussed. For example, Mickey and Greenland[25] and Maldonado and Greenland[26] examined various confounder selection strategies and found satisfactory performance of the "change-in-estimate" criterion. Pearl[27] and Greenland et al.[28] illustrated how to use known causal diagrams to identify covariates that should be measured and controlled for to eliminate confounding bias. Brookhart et al.[29] recommended including covariates related to the outcome and excluding covariates related only to the treatment but not the outcome. Schneeweiss et al.[30] developed a data-driven algorithm based on prioritizing covariates by their potential for controlling confounding unconditional on other covariates. VanderWeele[31] proposed to include covariates that are known causes of the exposure or the outcome, exclude instrumental variables, and include proxies for unmeasured variables that are common causes of the exposure and the outcome.

In practice, it may still be difficult to develop one final propensity score model based on these useful principles for various reasons. First, different covariate selection techniques may result in different sets of selected covariates. Second, even if the set of confounders were known, it is difficult to correctly specify their functional forms in the propensity score model (e.g., using the viral load measurement itself or its $log(\cdot)$ transformation). Third, different researchers (e.g., biostatisticians, clinicians, epidemiologists) in a multidisciplinary team may have different "best models" in mind and cannot reach a consensus about which model to use. Given the complexity of treatment decision process in observational studies, researchers may end up having multiple candidate models that all seem reasonable but difficult to choose from.

4 | MULTIPLY ROBUST ESTIMATION OF THE MARGINAL HAZARD RATIO USING A SET OF PROPENSITY SCORE MODELS

Using the empirical likelihood technique,[32,33] Han and Wang[17] proposed a multiply robust method for estimating population means for non-survival outcomes that are subject to non-response. We adapt their method to our survival context to estimate the marginal hazard ratio. Our method allows researchers to simultaneously postulate a set of propensity score models. The resulting estimator is consistent when this set contains a correctly specified model.

4.1 | Motivation for Empirical Likelihood Approach

In the context of non-survival outcomes where the interest is in the expectation of potential outcome under treatment (or control), Qin[34, p.357, 366-368] illustrated that generally the likelihood based on data from treated (or untreated) individuals is a biased version of the likelihood that would have been obtained had all individuals been treated (or untreated). To handle such biased sampling problem, he discussed a general empirical likelihood approach[35] that maximizes the biased sampling likelihood subject to a required constraint on propensity score and an optional constraint on a function of covariates. He further showed that the best choice for the optional constraint in terms of efficiency was the expectation of the potential outcome conditional on covariates. Han and Wang[17] and Han[19,20,21] extended this empirical likelihood approach to accommodate multiple choices for both required and optional constraints.

In our survival context where the interest is in the marginal hazard ratio, we propose to fit a weighted Cox model relating the time-to-event outcome to only the treatment, where the weights are obtained by adapting the work of Han and Wang[17] and Han.[19,20,21] While it is tempting to specify some optional constraints on conditional outcome models like these prior studies, in our survival context, inference results from a conditional Cox model may not be used to estimate a marginal Cox model because the proportional hazards assumption usually does not simultaneously hold for both the marginal and conditional Cox models. Given this challenge, here we drop the optional constraints.

4.2 | Multiply Robust Estimation

To increase the chance of correctly modeling $e(\boldsymbol{X})$, we allow a set of parametric models instead of using just one model. Suppose $\mathcal{E} = \{e^j(\boldsymbol{\gamma}^j; \boldsymbol{X}) : j = 1, \dots, J\}$ is a set of J postulated propensity score models for $e(\boldsymbol{X})$, where $\boldsymbol{\gamma}^j$ is the vector of parameters for the jth model. Let $\widehat{\boldsymbol{\gamma}}^j$ be the estimator of $\boldsymbol{\gamma}^j$ obtained by fitting the jth propensity score model. Write $\widehat{\boldsymbol{\gamma}} = (\widehat{\boldsymbol{\gamma}}^{1\mathrm{T}}, \dots, \widehat{\boldsymbol{\gamma}}^{J\mathrm{T}})^{\mathrm{T}}$.

Without loss of generality, let $i = 1, \dots, m$ be the indexes for treated individuals and $i = m+1, \dots, n$ the indexes for untreated individuals, where m is the size of the treated group. Define $\widehat{\mu}^j = n^{-1} \sum_{i=1}^n e^j(\widehat{\boldsymbol{\gamma}}^j; \boldsymbol{X}_i)$ for $j = 1, \dots, J$. For $i = 1, \dots, n$, define

$$\widehat{g}_i(\widehat{\boldsymbol{\gamma}}) = (e^1(\widehat{\boldsymbol{\gamma}}^1; \boldsymbol{X}_i) - \widehat{\mu}^1, \dots, e^J(\widehat{\boldsymbol{\gamma}}^J; \boldsymbol{X}_i) - \widehat{\mu}^J)^{\mathrm{T}}.$$

The proposed empirical likelihood weights for the treated individuals $i = 1, \dots, m$ are given by

$$\widehat{w}_i = \underset{w_i}{\operatorname{argmax}} \prod_{i=1}^m w_i$$

subject to constraints

$$\widehat{w}_i \geq 0, \sum_{i=1}^m \widehat{w}_i = 1, \quad \text{and} \quad \sum_{i=1}^m \widehat{w}_i \widehat{g}_i(\widehat{\boldsymbol{\gamma}}) = \boldsymbol{0},$$

which yields

$$\widehat{w}_i = \left\{ \frac{1}{1 + \widehat{\boldsymbol{\rho}}^{\mathrm{T}} \widehat{g}_i(\widehat{\boldsymbol{\gamma}})} \right\} \Big/ m \text{ for } i = 1, \dots, m \tag{5}$$

where $\widehat{\boldsymbol{\rho}} = (\widehat{\rho}_1, \dots, \widehat{\rho}_J)^{\mathrm{T}}$ is a $J \times 1$ vector obtained by solving the equation

$$\sum_{i=1}^m \frac{\widehat{g}_i(\widehat{\boldsymbol{\gamma}})}{1 + \boldsymbol{\rho}^{\mathrm{T}} \widehat{g}_i(\widehat{\boldsymbol{\gamma}})} = \boldsymbol{0}$$

for $\boldsymbol{\rho}$ with $\widehat{\boldsymbol{\gamma}}$ given. To get around possible multiple-root issues, we apply the computation method of Han[19] to obtain $\boldsymbol{\rho}$ by convex minimization.

Here we give an intuition for the proposed constraints. By definition, it is immediate that $\sum_{i=1}^n \widehat{g}_i(\widehat{\boldsymbol{\gamma}}) = \boldsymbol{0}$, which can be understood as an unweighted sample average of quantities $\widehat{g}_i(\widehat{\boldsymbol{\gamma}})$ that converges to zero. It follows that a weighted average using only data from the treated group, with suitable weights that make the biased sample representative of the target population, should also converge to zero, i.e., $\sum_{i=1}^m \widehat{w}_i \widehat{g}_i(\widehat{\boldsymbol{\gamma}}) = \boldsymbol{0}$.

By symmetry, the empirical likelihood weights for untreated individuals $i = m+1, \dots, n$ are given by

$$\widehat{w}_i = \left\{ \frac{1}{1 - \widehat{\boldsymbol{\eta}}^{\mathrm{T}} \widehat{g}_i(\widehat{\boldsymbol{\gamma}})} \right\} \Big/ (n - m) \text{ for } i = m+1, \dots, n, \tag{6}$$

where $\widehat{\boldsymbol{\eta}} = (\widehat{\eta}_1, \ldots, \widehat{\eta}_J)^{\mathrm{T}}$ is a $J \times 1$ vector solving the equation

$$\sum_{i=m+1}^{n} \frac{\widehat{g}_i(\widehat{\boldsymbol{\gamma}})}{1 - \boldsymbol{\eta}^{\mathrm{T}} \widehat{g}_i(\widehat{\boldsymbol{\gamma}})} = \mathbf{0}$$

for $\boldsymbol{\eta}$ with $\widehat{\boldsymbol{\gamma}}$ given.

Similar to the use of the stabilized weights (3) as an alternative to the conventional weights (2) in the IPW context, we also consider an alternative to empirical likelihood weights (5) and (6):

$$\widehat{w}_i = \widehat{P}(A = 1) \left\{ \frac{1}{1 + \widehat{\boldsymbol{\rho}}^{\mathrm{T}} \widehat{g}_i(\widehat{\boldsymbol{\gamma}})} \right\} \Big/ m \text{ for } i = 1, \ldots, m, \tag{7}$$

$$\widehat{w}_i = \widehat{P}(A = 0) \left\{ \frac{1}{1 - \widehat{\boldsymbol{\eta}}^{\mathrm{T}} \widehat{g}_i(\widehat{\boldsymbol{\gamma}})} \right\} \Big/ (n - m) \text{ for } i = m + 1, \ldots, n, \tag{8}$$

where $\widehat{\boldsymbol{\rho}}$ and $\widehat{\boldsymbol{\eta}}$ are the same as in (5) and (6), $\widehat{P}(A = 1) = m/n$, and $\widehat{P}(A = 0) = (n - m)/n$.

The proposed estimator of the log marginal hazard ratio is obtained by fitting a Cox model relating the time-to-event outcome to only the treatment with individuals weighted by empirical likelihood weights (5) and (6), or, their alternatives (7) and (8). Specifically, with the proposed empirical likelihood weights, solving the estimating equation (4) for θ gives the proposed estimator of the log marginal hazard ratio θ, denoted as $\widehat{\theta}$.

In Appendix, we establish the multiple robustness of $\widehat{\theta}$. That is, $\widehat{\theta}$ is a consistent estimator of the log marginal hazard ratio θ, if $\mathcal{E} = \{e^j(\boldsymbol{\gamma}^j; \boldsymbol{X}) : j = 1, \ldots, J\}$ contains a correctly specified model. Specifically, we show that the proposed weights are asymptotically equivalent to the IPW weights using the correctly specified propensity score model. The inverse probability weights from the correct model achieves covariate balancing after weighting, so do the proposed weights. Because weighting effectively eliminates confounding bias, the proposed weighted Cox model relating the time-to-event outcome to only the treatment provides consistent estimators of the marginal hazard ratio.

4.3 | Variance Estimation and Confidence Interval

In the setting of fitting an IPW Cox model with one propensity score model, Austin[7] suggested using the bootstrap method[36] for variance estimation. For each bootstrap sample, the weights are estimated using the same bootstrap sample rather than the original data. By doing so, the uncertainty in weight estimation is taken into account. His simulations demonstrated satisfactory performance of the bootstrap variance estimator with 200 bootstrap samples.

For our setting with multiple propensity score models, the bootstrap method can also be used for variance estimation. Specifically, we resample the data with replacement for B times to construct B bootstrap samples, each with the same size as the original data, where B is a user-specified number. For $b = 1, \ldots, B$, let $\widehat{\theta}_b$ denote the estimated log hazard ratio obtained from the bth bootstrap sample. Then the bootstrap variance estimator for $\widehat{\theta}$ is given by

$$\widehat{var}(\widehat{\theta}) = \frac{1}{B - 1} \sum_{b=1}^{B} \left(\widehat{\theta}_b - \frac{1}{B} \sum_{b=1}^{B} \widehat{\theta}_b \right)^2 . \tag{9}$$

A normality-based 95% confidence interval for θ is $\widehat{\theta} \pm 1.96 \cdot \sqrt{\widehat{var}(\widehat{\theta})}$.

4.4 | Balance Diagnostics

As noted in Section 4.2, the proposed method creates a weighted population in which the distribution of baseline covariates is expected to be the same between the treated and untreated individuals. Therefore, just like in an IPW analysis, it is essential to assess the balance of covariates between treatment groups in the weighted sample when using the proposed method. For an IPW analysis, Austin and Stuart[37] examined a suite of balance diagnostic tools that assess whether weighting balances measured covariates between the treated and untreated individuals in the weighted sample. These tools are also applicable to our weighting approach. For example, researchers can use the standardized difference to check for covariate balance in the weighted sample when using the proposed method. This measure compares the means of covariates between the treatment groups in units of the pooled standard deviation, and hence allows for fair comparisons of balance among covariates measured in different units.[37]

5 | SIMULATION STUDIES: WHEN ONE OF THE POSTULATED PROPENSITY SCORE MODELS IS CORRECTLY SPECIFIED

We conducted simulation studies to assess the finite sample performance of the proposed multiply robust method compared to the standard IPW Cox estimation.

5.1 | Data Generating Process

To simulate data that exactly followed model (1) with the true log marginal hazard ratio θ, we adapted the simulation method of Young et al.,[38] which was designed for time-varying treatment settings, to our one-time treatment studies. Specifically, for individuals $i = 1, \dots, n$, we simulated the following data:

Step 1: counterfactual control (i.e., untreated) group event time T_0^* that followed the unit exponential distribution.
Step 2: vector of covariates $\boldsymbol{X} = (X^{(1)}, X^{(2)}, X^{(3)}, X^{(4)}, X^{(5)}, X^{(6)})^{\mathrm{T}}$. The first three covariates were continuous, simulated as $X^{(1)} = -0.3{+}0.5T_0^*/(T_0^*{+}1){+}0.4Z_1$, $X^{(2)} = -0.3{+}\log(T_0^*{+}2){+}Z_2$, and $X^{(3)} = 1/(T_0^*{+}2){+}Z_1$, where Z_1 and Z_2 independently followed the uniform distribution ranging from -0.5 to 0.5. The other three covariates were binary, with $P(X^{(4)} = 1|T_0^*) = 0.2{+}0.6/(T_0^*{+}3)$, $P(X^{(5)} = 1|T_0^*) = 0.3{+}0.4/(0.5T_0^*{+}2)$, and $P(X^{(6)} = 1|T_0^*) = 1/(T_0^*{+}1)$.
Step 3: treatment indicator A that generated from the propensity score model

$$\text{logit } P(A = 1|\boldsymbol{X}) = \gamma_0 - 0.1\exp(X^{(1)}) - 0.3\exp(X^{(2)}) + 0.1\exp(X^{(3)}) + 0.6X^{(4)} + 0.4X^{(5)} + 0.5X^{(6)}, \quad (10)$$

where the parameter γ_0 was chosen to produce treatment prevalence of approximately 10%, 20%, 30%, 40% or 50%.
Step 4: actual true event time using formula $T^* = T_0^* \exp(-\theta A)$. We specified $\theta = \log(1.5)$ so that the true marginal hazard ratio was 1.5.
Step 5: event indicator $\delta = I(T^* \leq C)$ and $T = \min(T^*, C)$, where C followed an exponential distribution whose rate parameter was chosen to yield a censoring rate of about 30% or 60%.

The distribution and overlap of propensity scores between treatment groups are often examined prior to the estimation of treatment effects. Limited overlap may indicate substantial differences between treated and untreated individuals or violation of the positivity assumption. In that case, researchers may consider abandoning the analysis or restricting the analysis to a suitably chosen subsample.[39,40] For example, Crump et al.[41] proposed to select subsamples that can most precisely estimate the average treatment effect. For each treatment prevalence scenario in Step 3, Figure 1 visualizes the propensity score distribution and overlap between two treatment groups under the true propensity score model using density plots. Our simulation design produced a reasonable degree of overlap in the density and range of propensity scores.

[insert Figure 1 here]

5.2 | Specification of Propensity Score Models

In analyzing the simulated data, we considered $\mathcal{E} = \{e^j(\boldsymbol{\gamma}^j; \boldsymbol{X}) : j = 1, \ldots, 5\}$, a set of five postulated propensity score models:

$$\text{logit } P(A = 1|\boldsymbol{X}) = (1, X^{(1)}, X^{(2)})\boldsymbol{\gamma}^1, \tag{11}$$

$$\text{logit } P(A = 1|\boldsymbol{X}) = (1, X^{(4)}, X^{(5)}, X^{(6)})\boldsymbol{\gamma}^2, \tag{12}$$

$$\text{logit } P(A = 1|\boldsymbol{X}) = (1, \exp(X^{(1)}), X^{(5)}, X^{(6)})\boldsymbol{\gamma}^3, \tag{13}$$

$$\text{c loglog } P(A = 1|\boldsymbol{X}) = (1, X^{(3)}, X^{(5)}, X^{(3)} \cdot X^{(5)})\boldsymbol{\gamma}^4, \tag{14}$$

and

$$\text{logit } P(A = 1|\boldsymbol{X}) = (1, \exp(X^{(1)}), \exp(X^{(2)}), \exp(X^{(3)}), X^{(4)}, X^{(5)}, X^{(6)})\boldsymbol{\gamma}^5, \tag{15}$$

where “logit” was the logit link function, “c loglog” was the complementary log-log link function, and the $\boldsymbol{\gamma}^j (j = 1, \ldots, 5)$ were vectors of the associated propensity score model parameters.

Given that the true propensity score model was (10), the postulated models (11)-(14) were wrong, due to excluding certain covariates or using incorrect functional forms of covariates. The fifth postulated model (15) was correctly specified.

5.3 | Evaluation Criteria

We compared six estimators. The first five were IPW Cox estimators obtained from the individual propensity score models (11)-(15). The sixth estimator was the proposed multiply robust estimator obtained by simultaneously using the five models (11)-(15). We considered two types of weights. The first type was referred to as conventional weights, including the inverse probability weights (2) and the proposed empirical likelihood weights (5) and (6). The second type was referred to as stabilized weights, including the inverse probability weights (3) and the proposed empirical likelihood weights (7) and (8).

We considered sample sizes of 500 and 5000 and ran 1000 simulations for each parameter configuration. As in Austin,[7] we used 200 bootstrap samples for estimating the variance of each estimator. We used three criteria to evaluate the finite sample performance of each estimator. First, we examined the average empirical relative bias (in percent) for estimator $\widehat{\theta}$, defined as $(\widehat{\theta} - \theta)/\theta \times 100\%$, across 1000 simulation runs. Second, we examined the empirical coverage (in percent), defined as the percentage of 95% confidence intervals in 1000 simulation runs that covered the true log marginal hazard ratio θ. Third, we examined the average widths of the 95% confidence intervals across 1000 simulation runs.

5.4 | Results

5.4.1 | Empirical Relative Bias

Figures 2 and 3 report the empirical relative bias in percent using the six estimators with stabilized weights, under various combinations of censoring rate and treatment prevalence. The four IPW Cox estimators under the incorrectly specified propensity score models (11)-(14) produced substantially biased results due to model misspecification. As expected, the IPW Cox estimator under the correctly specified propensity score model (15) and the proposed multiply robust estimator using all five models (11)-(15) generally yielded negligible empirical bias, although a noticeable empirical bias for both estimators was seen when $n = 500$ with low treatment prevalence 10%.

[insert Figures 2 and 3 here]

5.4.2 | Empirical Coverage

Figures 4 and 5 report the empirical coverage using the six estimators with stabilized weights, under various combinations of censoring rate and treatment prevalence. Given that we used 1000 simulation runs for each parameter configuration, empirical coverage for a consistent point estimator with a reliable variance estimator was expected to fluctuate around 95% and roughly lie within the range of 93.65% to 96.35%. Therefore, as in Austin,[7] we drew three horizontal lines (at 93.65%, 95%, and 96.35%) to indicate a plausible range of coverage.

Due to model misspecification, the four IPW Cox estimators under the incorrectly specified propensity score models (11)-(14) resulted in severe undercoverage, and the performance worsened as sample size increased. The IPW Cox estimator under the correctly specified propensity score model (15) and the proposed multiply robust estimator using all five models (11)-(15) produced empirical coverage close to 95% and roughly within range, except that the IPW Cox estimator produced slight undercoverage when $n = 500$ with 30% censoring. These results showed that with 200 bootstrap samples, the bootstrap variance estimator performed reasonably well.

[insert Figures 4 and 5 here]

5.4.3 | Widths of 95% Confidence Intervals

Both the IPW Cox estimator under the correctly specified propensity score model (15) and the proposed multiply robust estimator using all five models (11)-(15) produced negligible empirical bias for estimating the log marginal hazard ratio, because of their consistency (Figures 2 and 3). We further compared their efficiency through examining their average widths of the 95% confidence intervals. Figures 6 and 7 summarize the results with stabilized weights, under various combinations of censoring rate and treatment prevalence. As seen from the overlapping lines for these two estimators, the widths of their 95% confidence intervals were almost the same under 60% censoring. Under 30% censoring, the proposed multiply robust estimator produced narrower 95% confidence intervals than the IPW Cox estimator under the correctly specified propensity score model (15). Therefore, the proposed multiply robust method not only provided protection against model misspecification, but also had better efficiency in some scenarios.

[insert Figures 6 and 7 here]

We observed similar simulation results using conventional weights (Figures S1-S6, Online Supplementary Material).

6 | SIMULATION STUDIES: WHEN NONE OF THE POSTULATED PROPENSITY SCORE MODELS IS CORRECTLY SPECIFIED

The validity of the proposed method requires a critical condition that the set of postulated propensity score models contains a correctly specified model. When all the proposed models are wrong, the proposed estimator is generally biased. To examine the performance of the proposed method when all models are wrong, we conducted simulations with a sample size of 5000 and replaced model (15) with the following incorrectly specified model:

$$\text{logit } P(A = 1|\boldsymbol{X}) = (1, X^{(1)}, X^{(2)}, X^{(3)}, X^{(4)}, X^{(5)}, X^{(6)})\boldsymbol{\gamma}^5. \tag{16}$$

We compared eight estimators. The first five were IPW Cox estimators obtained from the individual incorrectly specified propensity score models (11)-(14) and (16). The other three estimators were the proposed estimators obtained using various combinations of incorrectly specified models. Specifically, the sixth estimator used incorrect models (11)-(14) and (16). The seventh estimator used incorrect models (11)-(14). The eighth estimator used incorrect models (12)-(14). The sixth estimator represented a situation where the postulated models contained a nearly unbiased model, i.e., model (16). The seventh and eighth estimators represented a setting that all postulated models were severely misspecified.

Figure 8 reports the boxplots of the estimates obtained from the eight methods with stabilized weights, under various combinations of censoring rate and treatment prevalence. In each panel, the horizontal line indicates the true log marginal hazard ratio of log(1.5), and the sign "+" indicates the average of the estimates across 1000 simulation runs for each method. Since models (11)-(14) and (16) were all incorrectly specified, all the eight estimators generally produced biased results. In terms of empirical bias, the sixth estimator using the proposed method performed similarly to (and slightly better than) the fifth estimator under model (16) and outperformed the first four estimators under models

(11)-(14). This may not be surprising since the postulated models contained a nearly unbiased one. Interestingly, the seventh estimator, based on four seriously misspecified models (11)-(14), produced nearly unbiased results. The eighth estimator based on incorrect models (12)-(14), showed slightly worse but still comparable performance to the least biased IPW estimator among those estimators using models (12)-(14). It is also noted that the last three estimators using the proposed method were less or similarly variable than IPW estimators. Therefore, in our simulations, the proposed method was still a reasonable option even when all postulated propensity score models were wrong.

[insert Figure 8 here]

We observed similar results using conventional weights (Figures S7, Online Supplementary Material).

7 | APPLICATION TO REAL-WORLD DATA

We applied the existing IPW Cox estimation method in Section 2 and the proposed method in Section 4 to analyze the bariatric surgery dataset arising from the IBM® MarketScan® Research Databases, which contains de-identified patient-level healthcare claims information from a variety of contributors such as employers who are fully compliant with the Health Insurance Portability and Accountability Act (HIPAA). The dataset included 6690 patients who were 18 to 79 years of age and underwent either sleeve gastrectomy ($n = 4719$; 70.5%) or Roux-en-Y gastric bypass ($n = 1971$; 29.5%) between 1/1/2015 and 9/30/2015. The treatment indicator was 1 for sleeve gastrectomy and 0 for Roux-en-Y gastric bypass. The outcome was time to the first all-cause hospitalization during the first 30 post-surgical days. As a common feature of administrative databases with rare safety outcomes, the censoring rate was high (97%).

We classified the 30 researcher-identified baseline covariates into four categories **(a)** sex, age, and the Charlson/Elixhauser combined comorbidity score; **(b)** diagnosis of anxiety, cardiovascular disease, cancer, cerebrovascular disease, depression, diabetes, dyslipidemia, eating disorder, gastroesophageal reflux disease, hypertension, infertility, kidney disease, non-alcoholic fatty liver disease, osteoarthritis, polycystic ovary syndrome, psychosis, sleep apnea, substance use disorder, and tobacco use disorder; **(c)** number of emergency department visits, inpatient stays, non-acute institutional stays, outpatient visits, and other ambulatory visits; and **(d)** number of unique drug classes dispensed, unique generic medications dispensed, and outpatient pharmacy dispensing.

We first conducted IPW Cox estimation, separately using six postulated propensity score models. The first propensity score model (PS-1) was specified as a logistic regression model relating the treatment indicator to all 30 researcher-identified covariates, a commonly used strategy. In practice, it may be necessary to conduct covariate selection for propensity score models, so we specified the second propensity score model (PS-2) as a logistic regression model relating the treatment indicator to sex, age, comorbidity score and 15 other covariates (i.e., diagnosis of cardiovascular disease, depression, diabetes, dyslipidemia, hypertension, kidney disease, non-alcoholic fatty liver disease, psychosis, and sleep apnea; number of emergency department visits, inpatient stays, and outpatient visits; and number of unique drug classes dispensed, unique generic medications dispensed, and outpatient pharmacy dispensing) that were univariately statistically significant at the 5% level in their associations with the treatment (modeled via univariate logistic regression models). The last four logistic propensity score models reflected situations where only one category of covariates was available. Specifically, the third model (PS-3) contained demographic covariates: sex, age, and comorbidity score in category **(a)**, and three interaction terms between sex and age, sex and comorbidity score, and age and comorbidity score. The fourth (PS-4), fifth (PS-5), and sixth (PS-6) models contained the diagnosis covariates in category **(b)**, the health services utilization covariates in category **(c)**, and the drug dispensing covariates in category **(d)**, respectively.

Given that the true propensity score model was unknown, we applied the proposed multiply robust method to simultaneously use all six models. We conducted bootstrapping to estimate the variance using 200 bootstrap samples. Through the standardized difference, Figure 9 checks for covariate balance before and after weighting the sample using the proposed method. The proposed weights were found to well balance all the 30 measured covariates between the treated and untreated individuals in the weighted sample.

Table 1 summarizes the results. PS-1 and PS-2 produced similar results, suggesting that the exclusion of non-statistically significant covariates did not affect the log hazard ratio estimates and standard errors. PS-3, PS-4, and PS-5 produced slightly smaller (further from the null) hazard ratio estimates than PS-1 and PS-2. PS-6 produced larger (towards the null) hazard ratio estimates than PS-1 and PS-2, but the difference was negligible. The proposed method produced results similar to the results from IPW Cox estimation with PS-1, PS-2, and PS-6. The standard errors for all six IPW estimators and the proposed estimator were similar (around 0.14). For each method, the conventional and stabilized weights produced similar results.

All methods produced 95% confidence intervals for the marginal hazard ratio that excluded 1, suggesting a statistically significant lower risk of hospitalization 30-day postoperatively at the 5% level comparing sleeve gastrectomy to Roux-en-Y gastric bypass. The result was consistent with the findings from prior studies.[42,43]

[insert Figure 9 and Table 1 here]

8 | EXTENSION TO MULTI-SITE STUDIES

There is a growing number of studies that combine information from multiple data sources to help generate more statistically powerful and generalizable evidence. For example, the Sentinel System is a national electronic system funded by the U.S. Food and Drug Administration to monitor the safety of approved medical products using data from more than a dozen health plans and delivery systems.[44] The IPW Cox model stratified on data-contributing site provides one approach to estimate marginal hazard ratios in multi-site studies, where each site fits a site-specific propensity score model.[45,46,47] In this section, we extend the proposed multiply robust method in Section 4 to enable each participating site to postulate multiple site-specific propensity score models.

Suppose we have a sample of n individuals coming from K participating sites. For $k = 1, \dots, K$, let $\Omega_k = \{i :$ i in site k for $i = 1, \dots, n\}$ be the set of indexes for individuals that belong to the kth site. We consider a weighted Cox model stratified on site. By stratification, we assume the K sites have a common hazard ratio, but their baseline hazards are allowed to differ and be completely unspecified. The stratified weighted partial likelihood score equation is given by

$$\sum_{k=1}^{K} \sum_{i:i \in \Omega_k} \left\{ \widehat{\omega}_i \delta_i A_i - \widehat{\omega}_i \delta_i \frac{\sum_{l:l \in \mathfrak{R}_i(k)} \widehat{\omega}_l \exp(A_l \theta) A_l}{\sum_{l:l \in \mathfrak{R}_i(k)} \widehat{\omega}_l \exp(A_l \theta)} \right\} = 0, \tag{17}$$

where $\widehat{\omega}_i$ is the empirical likelihood weight for individual i, and $\mathfrak{R}_i(k) = \{l : T_l \geq T_i, l \in \Omega_k \, \delta_i = 1\}$ is the risk set for a noncensored individual i in site k.

Solving (17) for θ gives $\widehat{\theta}$, the estimate of the log hazard ratio θ. Equation (17) is an extension of the unstratified weighted partial likelihood score equation (4). When $K = 1$, (17) reduces to (4).

Instead of postulating a single site-specific propensity score model, each site can postulate a set of models to obtain empirical likelihood weights in Section 4 for their members. The resulting log hazard ratio estimator of the weighted Cox model stratified on site is multiply robust, as long as each site includes a correctly specified site-specific propensity score model in its set of candidate models. Below is the justification.

For each site k where $k = 1, \ldots, K$, the corresponding site-specific weighted partial likelihood score function

$$\sum_{i:i \in \Omega_k} \left\{ \widehat{\omega}_i \delta_i A_i - \widehat{\omega}_i \delta_i \frac{\sum_{l:l \in \mathfrak{R}_i(k)} \widehat{\omega}_l \exp(A_l \theta) A_l}{\sum_{l:l \in \mathfrak{R}_i(k)} \widehat{\omega}_l \exp(A_l \theta)} \right\} \tag{18}$$

is an unbiased estimating function for θ, where $\widehat{\omega}_i$ is the empirical likelihood weight for individual $i \in \Omega_k$ obtained from a set of site k-specific propensity score models. As the summation of these K unbiased estimating functions (18), the estimating function for the weighted Cox model stratified on site is also unbiased, i.e., solving estimating equation (17) for θ gives a consistent estimator of θ.

9 | DISCUSSION

In this paper, we proposed a multiply robust method for estimating marginal hazard ratios that can simultaneously accommodate a set of propensity score models. If one of these models is correctly specified, our method produces empirical likelihood weights that are asymptotically equivalent to the IPW weights from the correctly specified propensity score model and therefore guarantees estimation consistency. Compared to the IPW estimation method that relies on one propensity score model, the proposed method offers more protection against model misspecification and more model options for researchers. Our method is particularly useful when researchers have a difficult time developing or choosing only one propensity score model for their studies.

Our simulation studies showed that IPW Cox method can lead to severe bias and undercoverage when misspecifying the propensity score model. The proposed method showed satisfactory finite sample performance under various combinations of sample size, treatment prevalence, and censoring rate. The average widths of the 95% confidence intervals of the proposed method tended to be no wider than that of the IPW estimation method that used a correctly specified propensity score model, suggesting that our method achieved multiple robustness without losing efficiency (and sometimes even gained efficiency). Similar promising efficiency was also observed in simulations when all models were wrong. The reason that the proposed method leads to numerically more stable estimates than the IPW method is that it reduces the occurrence of extreme weights through maximizing $\prod_i w_i$ subject to the constraints, because this maximization results in more evenly distributed weights under the constraints.[19,20,21] For non-survival outcomes, Han and Wang[17] evaluated the efficiency of their multiply robust estimator when both the propensity score and the data distribution are correctly modeled, but a theoretical efficiency comparison to the IPW estimator is unclear if only a propensity score model is correct. Han[48] proposed estimators for which incorrect models can always help improve efficiency as long as the propensity score is correctly modeled, which thus are always more efficient than the IPW estimator. Future work will formally examine efficiency of the proposed estimator, which focuses on time-to-event outcomes.

Although the proposed method generally loses consistency if all postulated propensity score models are wrong, it was comparable to the best-performing IPW estimation method in simulation settings we considered. It would be useful to further investigate the theoretical properties of the proposed method when none of the postulated propensity score models is correctly specified. In practice, efforts should be made to increase the chance of including at least one model that gives consistent or nearly consistent estimators. Increasing the number of candidate models would increase the chance of achieving this goal. Theoretically, the proposed method allows for any finite number of models, but having too many models (i.e., high dimensional $\boldsymbol{\rho}$ and $\boldsymbol{\eta}$) may jeopardize its numerical performance.[19,20,21] For example, collinearity problems may arise when some candidate models are too similar or some constraints are highly correlated. These are well-known problems in empirical likelihood implementation, but formal solutions are lacking to our knowledge. An excessively large number of models also imposes heavy computational burden.

We recommend using both the subject-matter knowledge and reliable data-driven tools to carefully build a comprehensive set of reasonable but not too similar candidate models. Here we describe a four-step strategy for constructing propensity score models. Step 1: develop a large set of candidate models. Researchers can apply different guidelines to build various models.[25,26,27,28,29,30,31] In particular, important covariates determined based on subject-matter knowledge should be included. Step 2: refine the set of candidate models from Step 1. Researchers assess each of the models through criteria such as covariate balancing after weighting. They may try to improve bad-performing models by adding higher order terms or changing the functional forms. If it does not work, they may consider removing that model. Step 3: trim the set of candidate models from Step 2. Specifically, if numerical issues such as collinearity occur, remove models that are very similar to existing ones and hence offer no extra information for estimation. Continue reducing the model set until no numerical issues occur. Step 4: check for covariate balance between treatment groups in the weighted sample using the proposed method based on the set of candidate models from Step 3. Researchers should make sure the final weights achieve a satisfactory degree of covariate balancing. Otherwise, researchers may consider rolling back to Step 1.

We also extended our method to multi-site settings so that each participating site may postulate multiple site-specific propensity score models. It can be done in a privacy-protecting way using data-sharing methods of Shu et al.[47] Specifically, each site first calculates the empirical likelihood weights for its members using multiple propensity score models, and then obtain risk-set tables using the resultant empirical likelihood weights. Finally, instead of sharing individual-level data across sites, it suffices for sites to share their summary-level risk-set tables to the analysis center for making inference on the marginal hazard ratio.[47]

Although the current development focuses on marginal hazard ratios, the proposed weights can be directly used to conduct weighted estimation of other effect measures for survival outcomes, such as the difference in restricted mean survival times under treatment and control.[49] The resulting weighted estimators would be multiply robust, because the proposed weights are asymptotically equivalent to the inverse probability weights from a correctly specified propensity score model.

ACKNOWLEDGMENTS

The authors thank the reviewers for their constructive comments which led to an improved version of this paper. The authors also thank Qoua Her at Harvard Pilgrim Health Care Institute for the help with the bariatric surgery dataset. This work was funded in part by a Harvard Pilgrim Health Care Institute Robert H. Ebert Career Development Award. Toh was also funded by the National Institute of Biomedical Imaging and Bioengineering (U01EB023683) and the Agency for Healthcare Research and Quality (R01HS026214). Wang is also partially supported by R01 AI136947 from the National Institute of Allergy and Infectious Diseases. This study, including the secondary data analysis using the IBM® MarketScan® Research Databases that contain de-identified individual-level healthcare claims information (currently not available for public sharing), was approved by the Institutional Review Board at Harvard Pilgrim Health Care.

SUPPLEMENTARY MATERIAL

Supplementary material of additional simulation results and example R code for this article is available online.

References

1. Horvitz DG, Thompson DJ. A generalization of sampling without replacement from a finite universe. *J Am Stat Assoc.* 1952;47(260):663–685.

2. Rosenbaum PR. Model-based direct adjustment. *J Am Stat Assoc.* 1987;82(398):387–394.

3. Lunceford JK, Davidian M. Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. *Stat Med.* 2004;23(19):2937–2960.

4. Cole SR, Hernán MA. Adjusted survival curves with inverse probability weights. *Comput Methods Programs Biomed.* 2004;75(1):45–49.

5. Cole SR, Hernán MA. Constructing inverse probability weights for marginal structural models. *Am J Epidemiol.* 2008;168(6):656–664.

6. Austin PC. The performance of different propensity score methods for estimating marginal hazard ratios. *Stat Med.* 2013;32(16):2837–2849.

7. Austin PC. Variance estimation when using inverse probability of treatment weighting (IPTW) with survival analysis. *Stat Med.* 2016;35(30):5642–5655.

8. Seaman SR, White IR. Review of inverse probability weighting for dealing with missing data. *Stat Methods Med Res.* 2013;22(3):278–295.

9. Hernán MA, Robins JM. *Causal Inference: What If.* Boca Raton: Chapman & Hall/CRC; 2020.

10. Rosenbaum R, Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika.* 1983;70(1):41–55.

11. Robins JM, Rotnitzky A, Zhao LP. Estimation of regression coefficients when some regressors are not always observed. *J Am Stat Assoc.* 1994;89(427):846–866.

12. Scharfstein DO, Rotnitzky A, Robins M. Adjusting for nonignorable drop-out using semiparametric nonresponse models. *J Am Stat Assoc.* 1999;94(448):1096–1120.

13. Bang H, Robins JM. Doubly robust estimation in missing data and causal inference models. *Biometrics.* 2005;61(4):962–972.

14. Qin J, Shao J, Zhang B. Efficient and doubly robust imputation for covariate-dependent missing responses. *J Am Stat Assoc.* 2008;103(482):797–810.

15. Cao W, Tsiatis AA, Davidian M. Improving efficiency and robustness of the doubly robust estimator for a population mean with incomplete data. *Biometrika.* 2009;96(3):723–734.

16. Tan Z. Bounded, efficient and doubly robust estimation with inverse weighting. *Biometrika.* 2010;97(3):661–682.

17. Han P, Wang L. Estimation with missing data: beyond double robustness. *Biometrika.* 2013;100(2):417–430.

18. Chan KCG, Yam SCP. Oracle, multiple robust and multipurpose calibration in a missing response problem. *Stat Sci.* 2014;29(3):380–396.

19. Han P. A further study of the multiply robust estimator in missing data analysis. *J Stat Plan Inference.* 2014;148:101–110.

20. Han P. Multiply robust estimation in regression analysis with missing data. *J Am Stat Assoc.* 2014;109(507):1159–1173.

21. Han P. Combining inverse probability weighting and multiple imputation to improve robustness of estimation. *Scand J Stat.* 2016;43(1):246–260.

22. Chen S, Haziza D. Multiply robust imputation procedures for the treatment of item nonresponse in surveys. *Biometrika.* 2017;104(2):439–453.

23. Cox DR. Partial likelihood. *Biometrika.* 1975;62(2):269–276.

24. Binder DA. Fitting Cox's proportional hazards models from survey data. *Biometrika.* 1992;79(1):139–147.

25. Mickey RM, Greenland S. The impact of confounder selection criteria on effect estimation. *Am J Epidemiol.* 1989;129(1):125–137.

26. Maldonado G, Greenland S. Simulation study of confounder-selection strategies. *Am J Epidemiol.* 1993;138(11):923–936.

27. Pearl J. Causal diagrams for empirical research. *Biometrika.* 1995;82(4):669–688.

28. Greenland S, Pearl J, Robins JM. Causal diagrams for epidemiologic research. *Epidemiology.* 1999;10(1):37–48.

29. Brookhart M, Schneeweiss S, Rothman KJ, Glynn RJ, Avorn J, Stürmer T. Variable selection for propensity score models. *Am J Epidemiol.* 2006;163(12):1149–1156.

30. Schneeweiss S, Rassen JA, Glynn RJ, Avorn J, Mogun H, Brookhart MA. High-dimensional propensity score adjustment in studies of treatment effects using health care claims data. *Epidemiology.* 2009;20(4):512–522.

31. VanderWeele TJ. Principles of confounder selection. *Eur J Epidemiol.* 2019;34(3):211–219.

32. Qin J, Lawless J. Empirical likelihood and general estimating equations. *Ann of Stat.* 1994;22(1):300–325.

33. Owen AB. *Empirical Likelihood.* New York: Chapman & Hall/CRC; 2001.

34. Qin J. *Biased Sampling, Over-identified Parameter Problems and Beyond.* Singapore: Springer; 2017.

35. Qin J, Zhang B. Empirical-likelihood-based inference in missing response problems and its application in observational studies. *J R Stat Soc Series B.* 2007;69(1):101–122.

36. Efron B, Tibshirani RJ. *An Introduction to the Bootstrap.* New York: Chapman & Hall/CRC; 1993.

37. 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(28):3661–3679.

38. Young JG, Hernán MA, Picciotto S, Robins JM. Simulation from structural survival models under complex time-varying data structures. *JSM Proceedings, Section on Statistics in Epidemiology, Denver, CO: American Statistical Association.* 2008;:1–6.

39. Harder VS, Stuart EA, Anthony JC. Propensity score techniques and the assessment of measured covariate balance to test causal associations in psychological research. *Psychol Methods.* 2010;15(3):234–249.

40. Kang J, Chan W, Kim MO, Steiner PM. Practice of causal inference with the propensity of being zero or one: assessing the effect of arbitrary cutoffs of propensity scores. *Commun Stat Appl Methods.* 2016;23(1):1-20.

41. Crump RK, Hotz VJ, Imbens GW, Mitnik OA. Dealing with limited overlap in estimation of average treatment effects. *Biometrika.* 2009;96(1):187–199.

42. Young MT, Gebhart A, Phelan MJ, Nguyen NT. Use and outcomes of laparoscopic sleeve gastrectomy vs laparoscopic gastric bypass: analysis of the American College of Surgeons NSQIP. *J Am Coll Surg.* 2015;220(5):880–885.

43. Sippey M, Kasten KR, Chapman WHH, Pories WJ, Spaniolas K. 30-day readmissions after sleeve gastrectomy versus Roux-en-Y gastric bypass. *Surg Obes Relat Dis.* 2016;12(5):991–996.

44. Behrman RE, Benner JS, Brown JS, McClellan M, Woodcock J, Platt R. Developing the Sentinel System — a national resource for evidence development. *N Engl J Med.* 2011;364(6):498–499.

45. Yoshida K, Gruber S, Fireman BH, Toh S. Comparison of privacy-protecting analytic and data-sharing methods: a simulation study. *Pharmacoepidemiol Drug Saf.* 2018;27(9):1034–1041.

46. Toh S, Wellman R, Coley RY, et al. Combining distributed regression and propensity scores: a doubly privacy-protecting analytic method for multicenter research. *Clin Epidemiol.* 2018;10:1773–1786.

47. Shu D, Yoshida K, Fireman BH, Toh S. Inverse probability weighted Cox model in multi-site studies without sharing individual-level data. *Stat Methods Med Res.* 2020;29(6):1668–1681.

48. Han P. A further study of propensity score calibration in missing data analysis. *Stat Sin.* 2018;28(3):1307–1332.

49. Mao H, Li L, Yang W, Shen Y. On the propensity score weighting analysis with survival outcome: estimands, estimation, and inference. *Stat Med.* 2018;37(26):3745–3763.

APPENDIX: PROOF OF MULTIPLE ROBUSTNESS OF THE PROPOSED METHOD

Suppose a set of postulated propensity score models $\mathcal{E} = \{e^j(\boldsymbol{\gamma}^j; \boldsymbol{X}) : j = 1, \dots, J\}$ contains a correctly specified model, say, without loss of generality, the first model $e^1(\boldsymbol{\gamma}^1; \boldsymbol{X})$. Let $\boldsymbol{\gamma}_0^1$ be the true value of $\boldsymbol{\gamma}^1$, then $e^1(\boldsymbol{\gamma}_0^1; \boldsymbol{X}) = e(\boldsymbol{X})$.

By adapting the arguments of Han and Wang,[17] the proposed weights $\widehat{w}_i$ in (5) can be re-written as

$$\widehat{w}_i = \frac{1}{m} \frac{\widehat{\mu}^1 / e^1(\widehat{\boldsymbol{\gamma}}^1; \boldsymbol{X}_i)}{1 + \widehat{\boldsymbol{\lambda}}^{\mathrm{T}} \widehat{g}_i(\widehat{\boldsymbol{\gamma}}) / e^1(\widehat{\boldsymbol{\gamma}}^1; \boldsymbol{X}_i)} \text{ for } i = 1, \dots, m,$$

where $\widehat{\boldsymbol{\lambda}} = O_p(n^{-1/2})$ is the Lagrange multiplier and $\widehat{\mu}^1 = n^{-1} \sum_{i=1}^n e^1(\widehat{\boldsymbol{\gamma}}^1; \boldsymbol{X}_i)$. Then as $n \to \infty$,

$$1 + \widehat{\boldsymbol{\lambda}}^{\mathrm{T}} \widehat{g}_i(\widehat{\boldsymbol{\gamma}}) / e^1(\widehat{\boldsymbol{\gamma}}^1; \boldsymbol{X}_i) \xrightarrow{p} 1$$

and

$$\widehat{\mu}^1 \xrightarrow{p} E\{e^1(\boldsymbol{\gamma}_0^1; \boldsymbol{X})\}, \text{ which equals } P(A = 1).$$

As a nonparametric estimator of $P(A = 1)$, m/n well approximates $P(A = 1)$, where m is the number of individuals who receive the treatment. Therefore, the proposed weights for treated individuals $i = 1, \dots, m$ in (5) well approximate $1/\{n \cdot e^1(\widehat{\boldsymbol{\gamma}}^1; \boldsymbol{X}_i)\}$, which is equivalent to the conventional IPW weights for treated individuals using model $e^1(\boldsymbol{\gamma}^1; \boldsymbol{X})$.

By symmetry, we can show the proposed weights for untreated individuals $i = m+1, \dots, n$ in (6) well approximates $1/[n \cdot \{1 - e^1(\widehat{\boldsymbol{\gamma}}^1; \boldsymbol{X}_i)\}]$, which is equivalent to the conventional IPW weights for untreated individuals using the correct propensity score model $e^1(\boldsymbol{\gamma}^1; \boldsymbol{X})$.

Since $e^1(\boldsymbol{\gamma}^1; \boldsymbol{X})$ is the correctly specified model that can be used to consistently estimate the log marginal hazard ratio, the proposed weights (5) and (6), which are shown to be asymptotically equivalent to the conventional IPW weights (2), can also be used to consistently estimate the log marginal hazard ratio.

Note the proposed alternative weights (7) for treated individuals are defined as weights (5) multiplied by $\widehat{P}(A = 1)$, and the proposed alternative weights (8) for untreated individuals are defined as weights (6) multiplied by $\widehat{P}(A = 0)$, it is immediate that they are asymptotically equivalent to the stabilized weights (3) using model $e^1(\boldsymbol{\gamma}^1; \boldsymbol{X})$. Given that $e^1(\boldsymbol{\gamma}^1; \boldsymbol{X})$ is the correctly specified propensity score model which yields a consistent log marginal hazard ratio estimator, the proposed weights (7) and (8) also provide a consistent estimator of the log marginal hazard ratio.

How to cite this article:

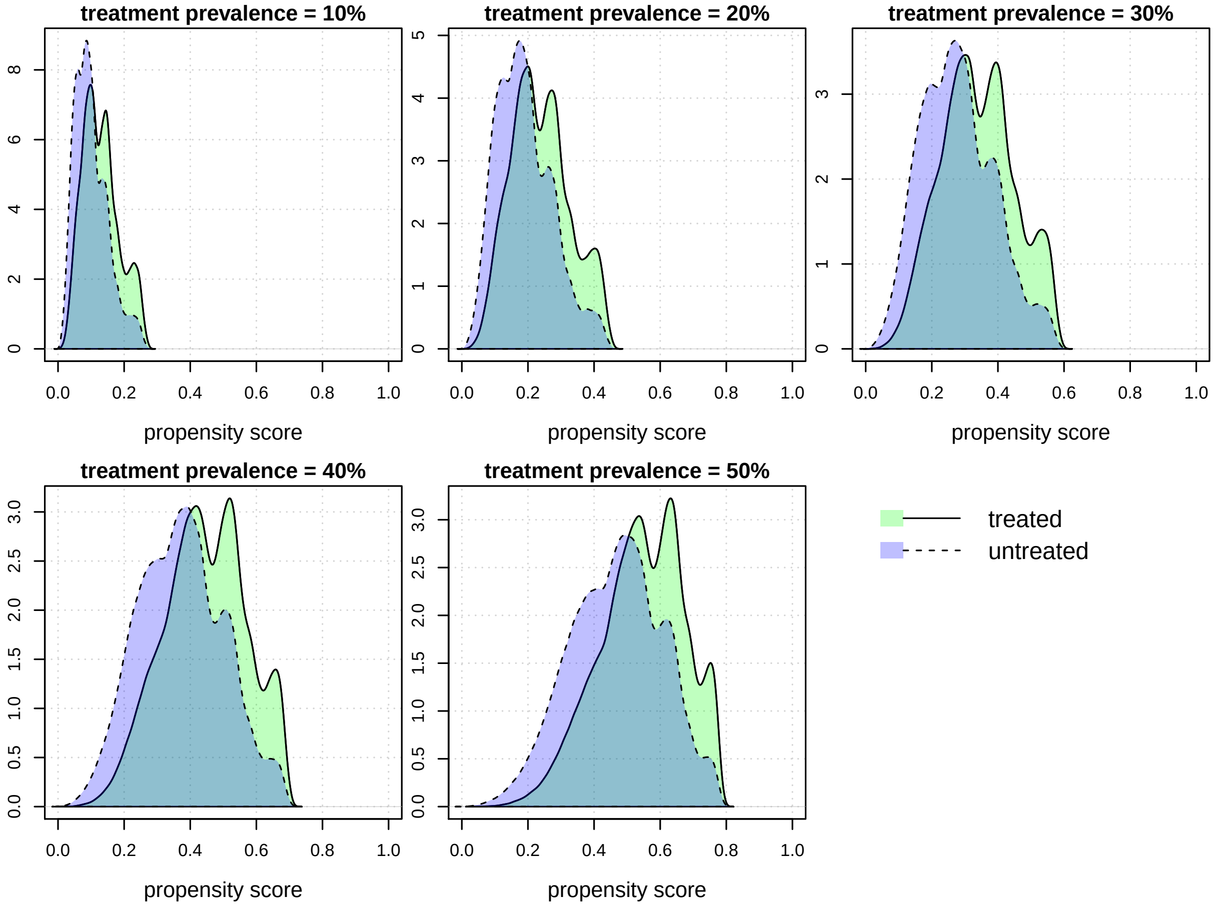


FIGURE 1 Propensity score distribution and overlap between two treatment groups under the true propensity score model: density plots.

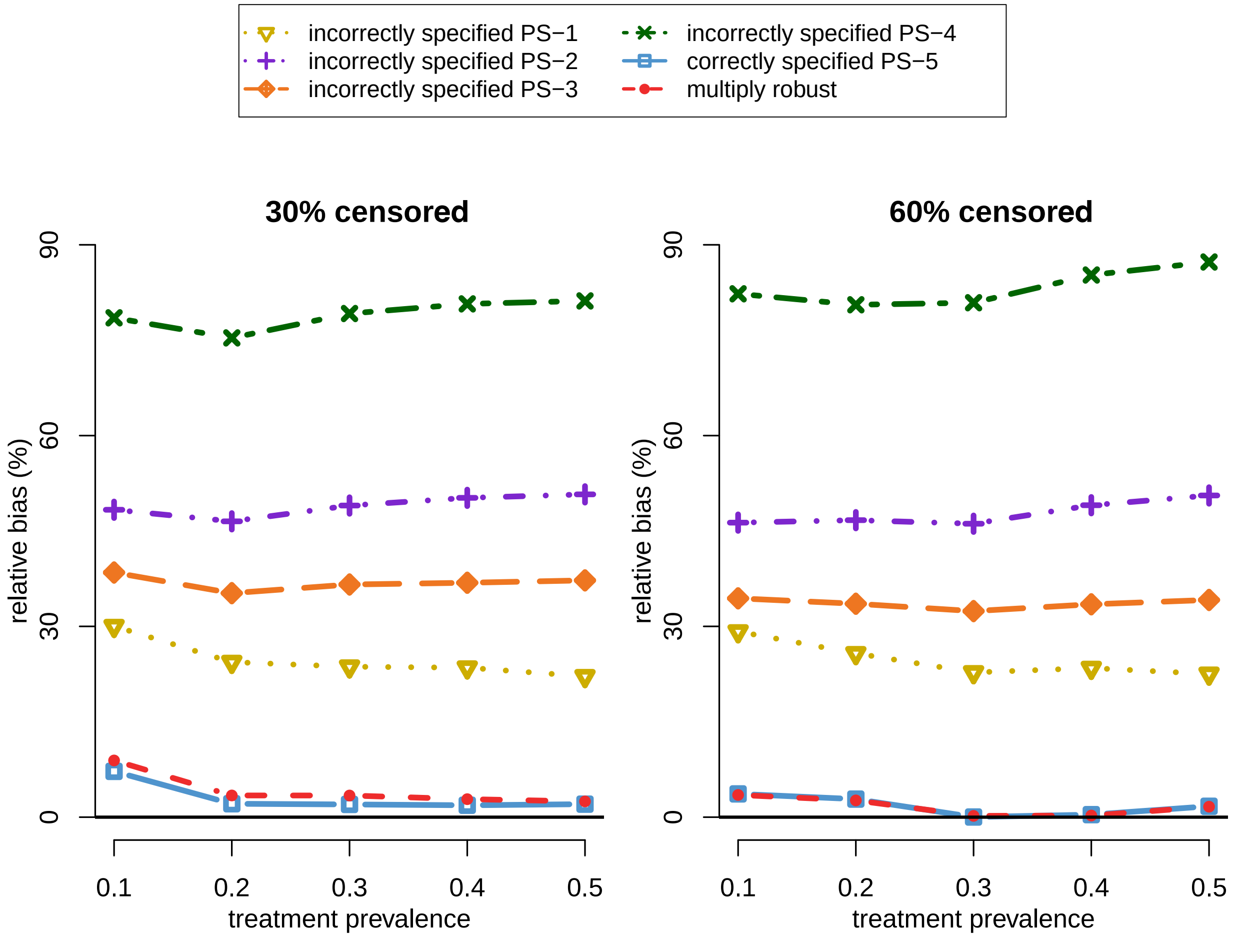


FIGURE 2 Empirical relative bias in percent using stabilized weights with $n = 500$.
incorrectly specified PS-1 to PS-4: IPW Cox estimators from four incorrectly specified propensity score models (11)-(14), respectively; correctly specified PS-5: IPW Cox estimator from a correctly specified propensity score model (15); multiply robust: the proposed multiply robust estimator using multiple models (11)-(15).

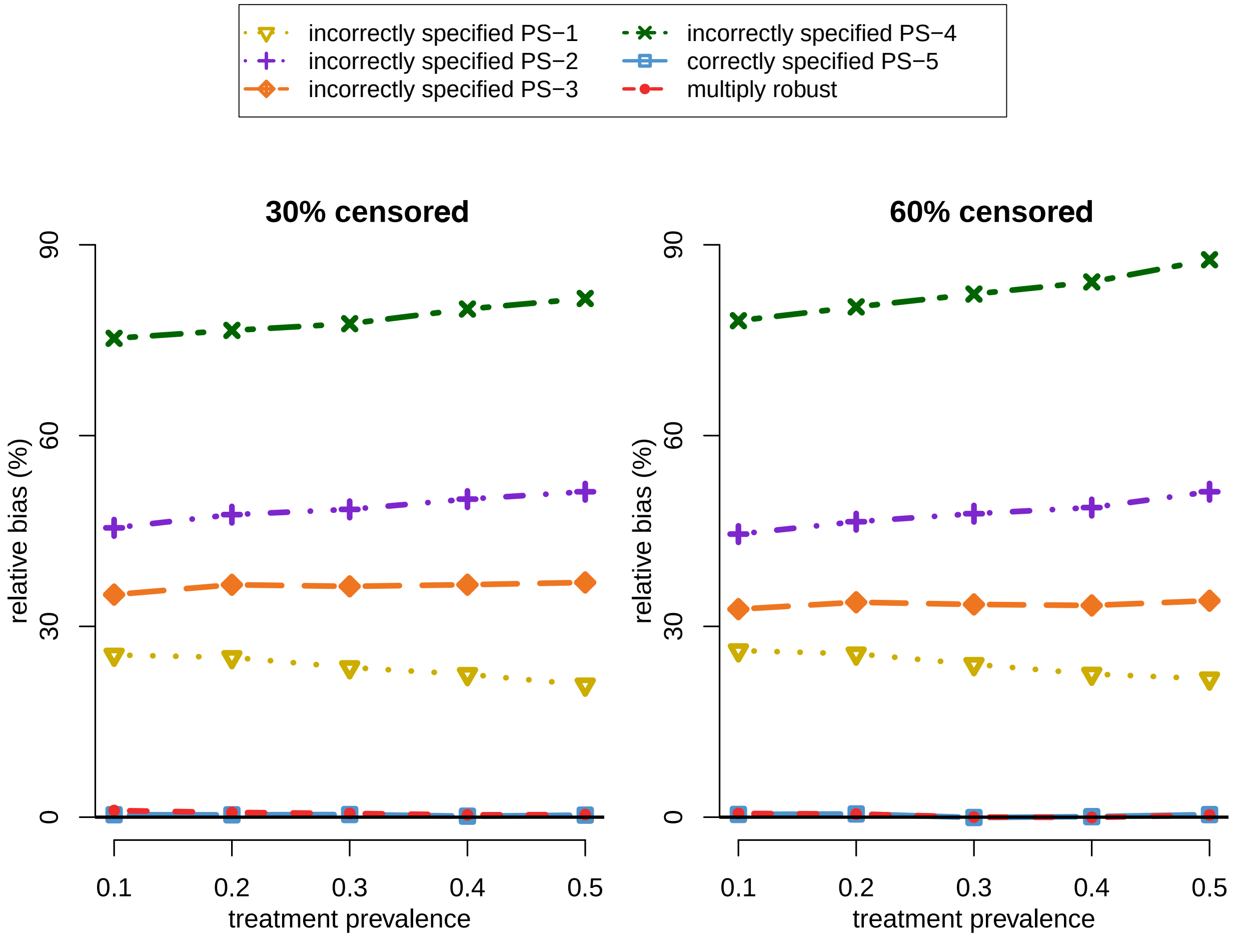


FIGURE 3 Empirical relative bias in percent using stabilized weights with $n = 5000$.
incorrectly specified PS-1 to PS-4: IPW Cox estimators from four incorrectly specified propensity score models (11)-(14), respectively; correctly specified PS-5: IPW Cox estimator from a correctly specified propensity score model (15); multiply robust: the proposed multiply robust estimator using multiple models (11)-(15).

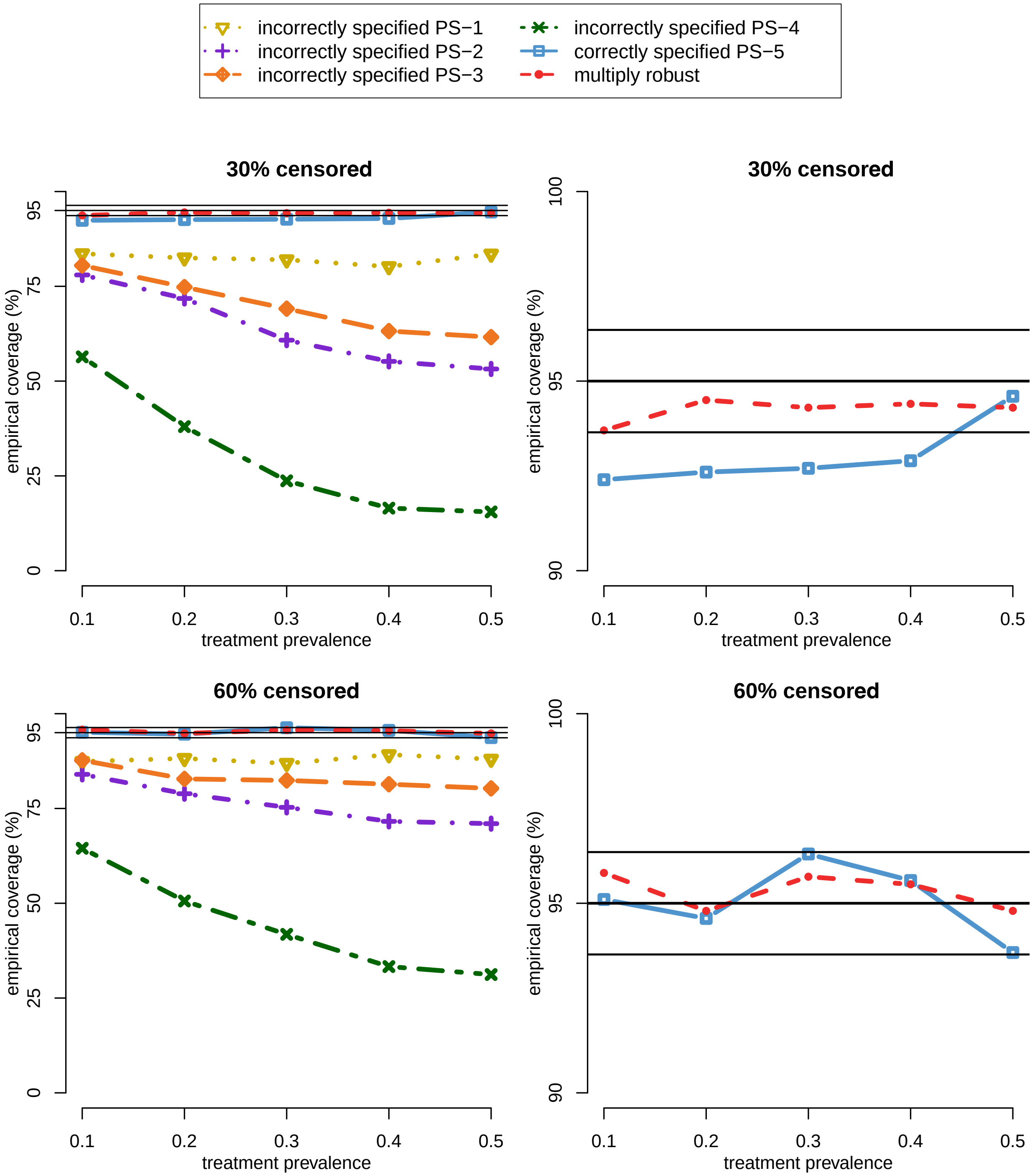


FIGURE 4 Empirical coverage in percent using stabilized weights with $n = 500$. The right panel shows a zoom-in version of the left panel.
incorrectly specified PS-1 to PS-4: IPW Cox estimators from four incorrectly specified propensity score models (11)-(14), respectively; correctly specified PS-5: IPW Cox estimator from a correctly specified propensity score model (15); multiply robust: the proposed multiply robust estimator using multiple models (11)-(15).

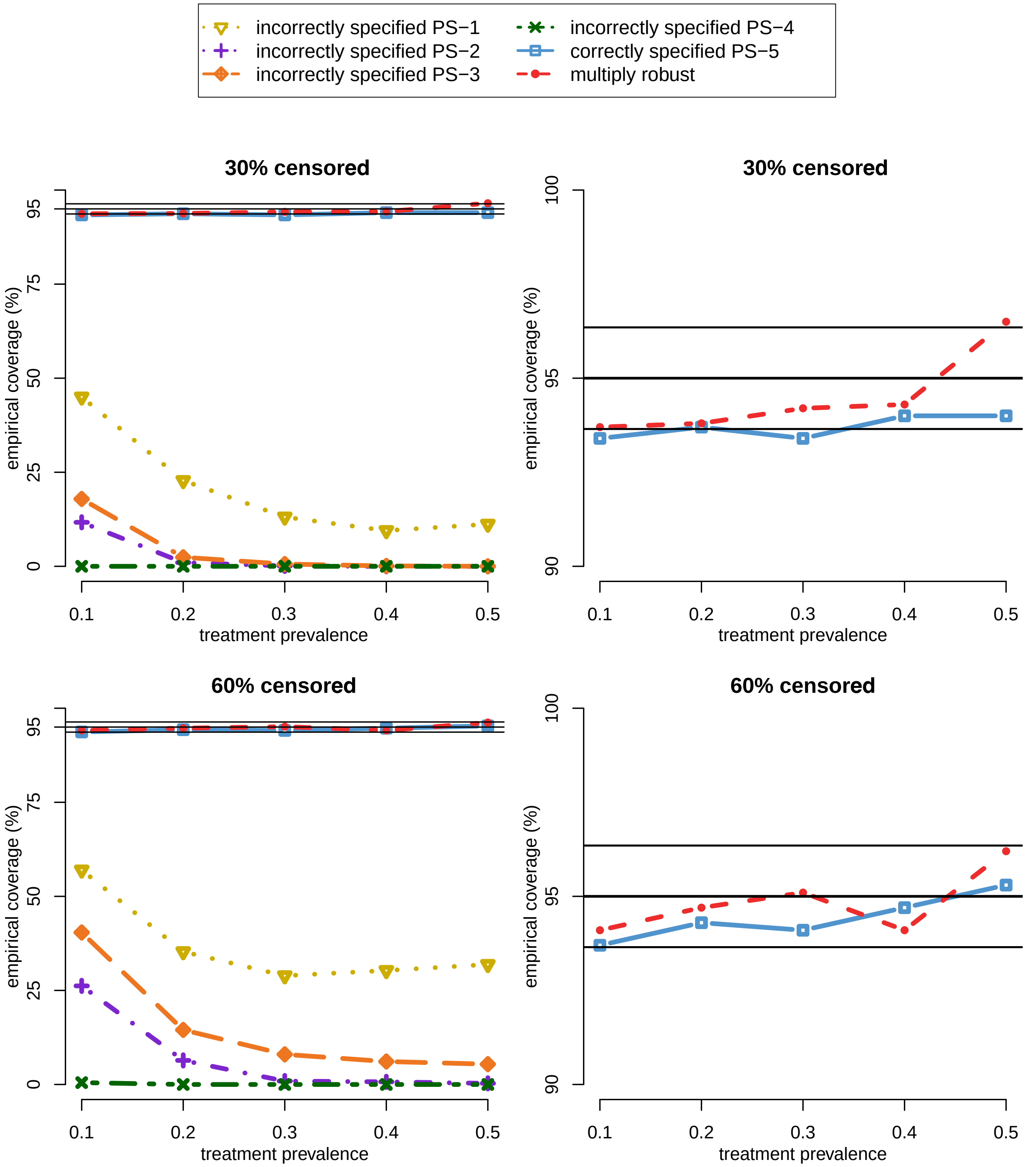


FIGURE 5 Empirical coverage in percent using stabilized weights with $n = 5000$. The right panel shows a zoom-in version of the left panel.
incorrectly specified PS-1 to PS-4: IPW Cox estimators from four incorrectly specified propensity score models (11)-(14), respectively; correctly specified PS-5: IPW Cox estimator from a correctly specified propensity score model (15); multiply robust: the proposed multiply robust estimator using multiple models (11)-(15).

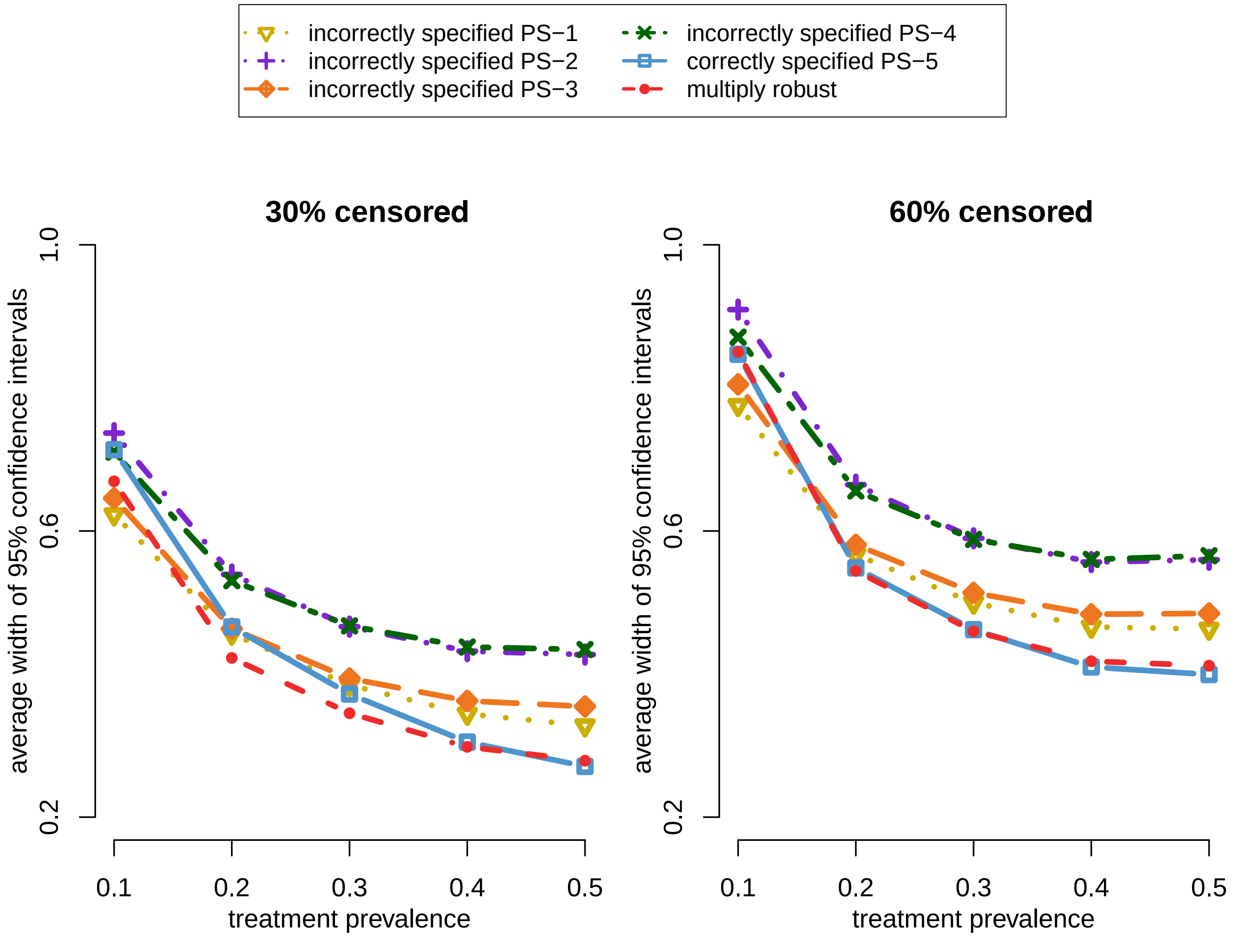


FIGURE 6 Average widths of 95% confidence intervals using stabilized weights with $n = 500$.
incorrectly specified PS-1 to PS-4: IPW Cox estimators from four incorrectly specified propensity score models (11)-(14), respectively; correctly specified PS-5: IPW Cox estimator from a correctly specified propensity score model (15); multiply robust: the proposed multiply robust estimator using multiple models (11)-(15).

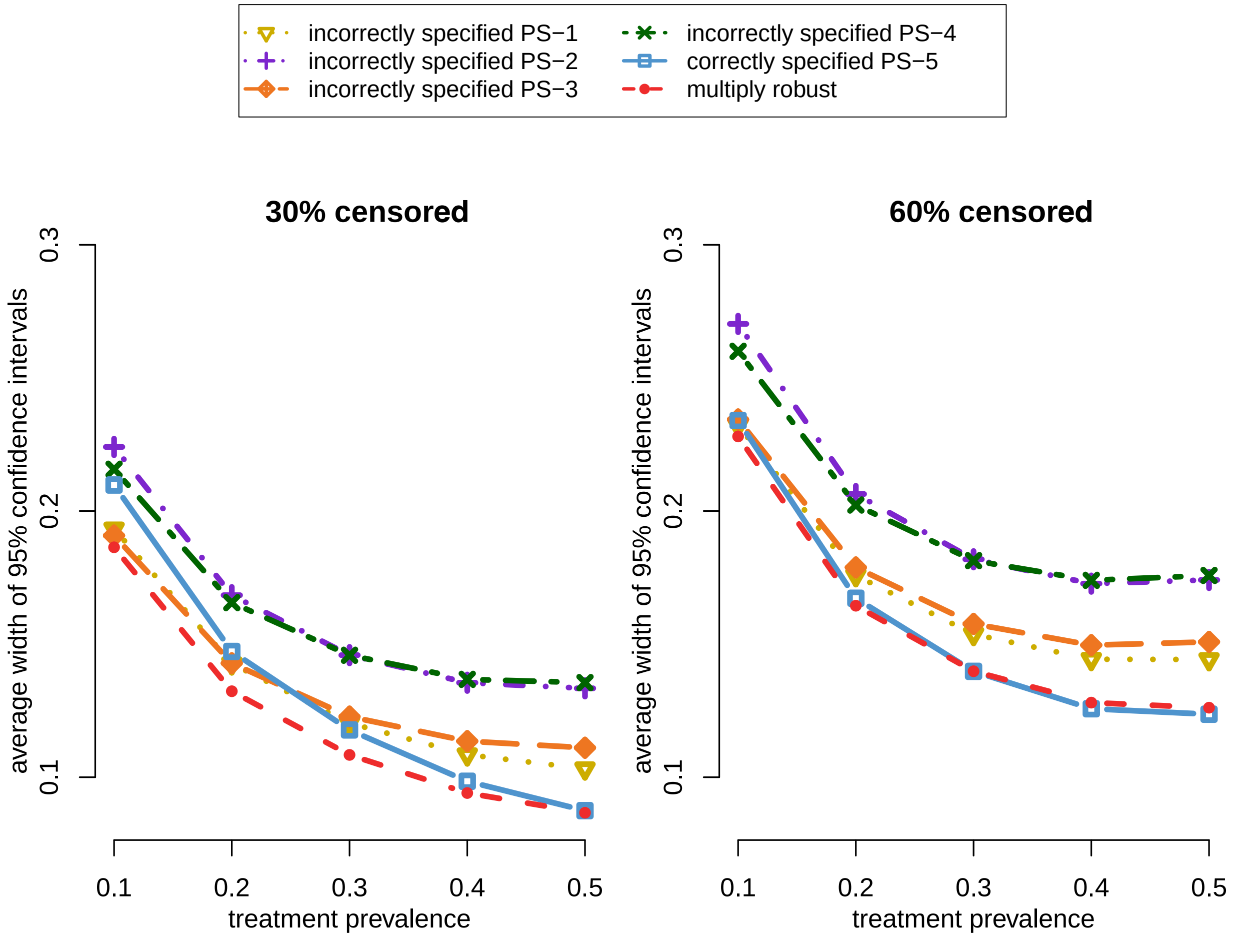


FIGURE 7 Average widths of 95% confidence intervals using stabilized weights with $n = 5000$. incorrectly specified PS-1 to PS-4: IPW Cox estimators from four incorrectly specified propensity score models (11)-(14), respectively; correctly specified PS-5: IPW Cox estimator from a correctly specified propensity score model (15); multiply robust: the proposed multiply robust estimator using multiple models (11)-(15).

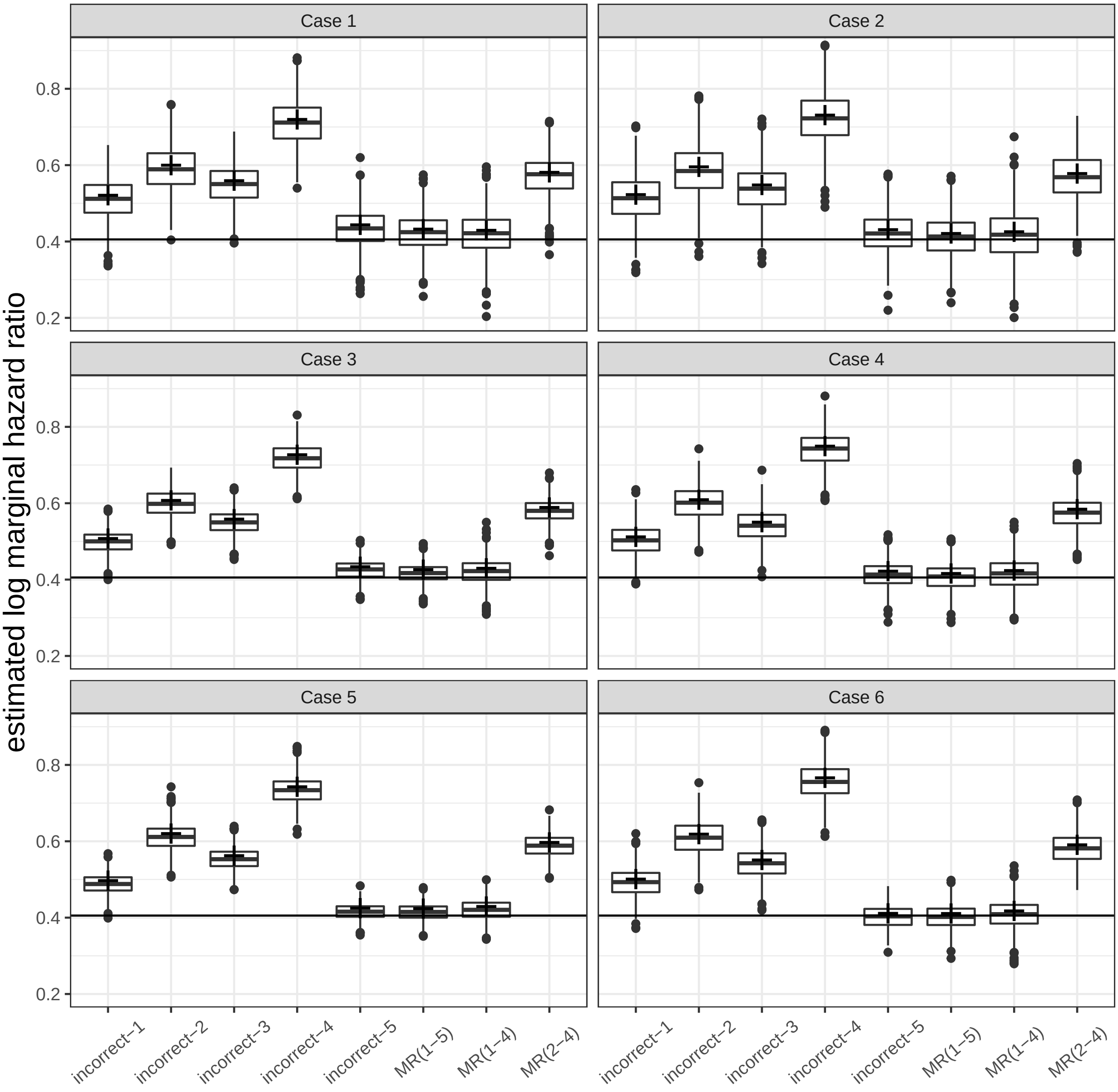


FIGURE 8 Simulation results when all postulated propensity score models are wrong using stabilized weights with $n = 5000$.

incorrect 1-5: IPW Cox estimators from five incorrectly specified propensity score models (11)-(14) and (16), respectively; MR(a-b): the proposed multiply robust estimator using incorrect models a to b;

Cases 1, 3, and 5: 30% censoring with treatment prevalence 10%, 30%, and 50%; Cases 2, 4, and 6: 60% censoring with treatment prevalence 10%, 30%, and 50%;

"+": average of estimates across 1000 simulation runs. The horizontal solid line indicates the true log marginal hazard ratio.

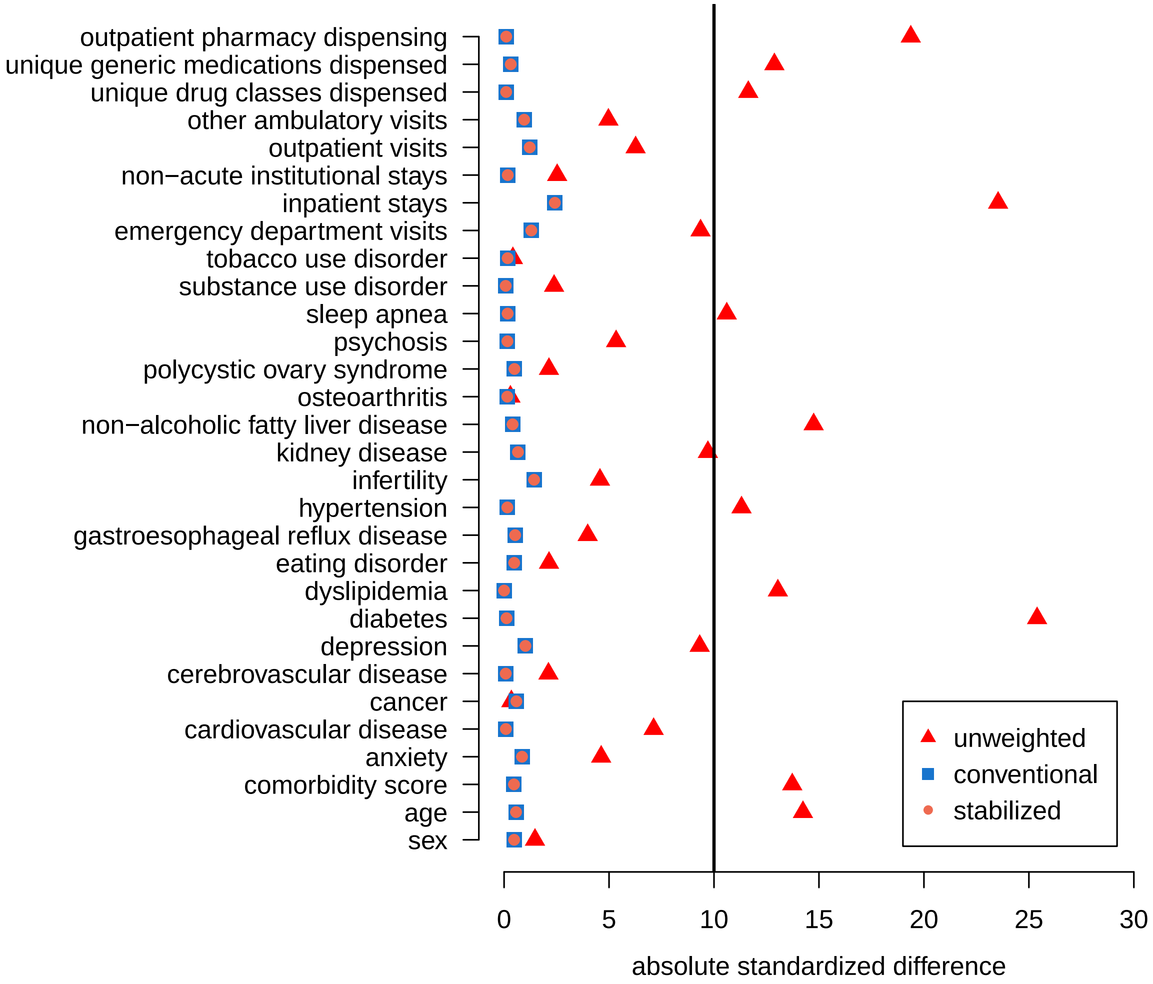
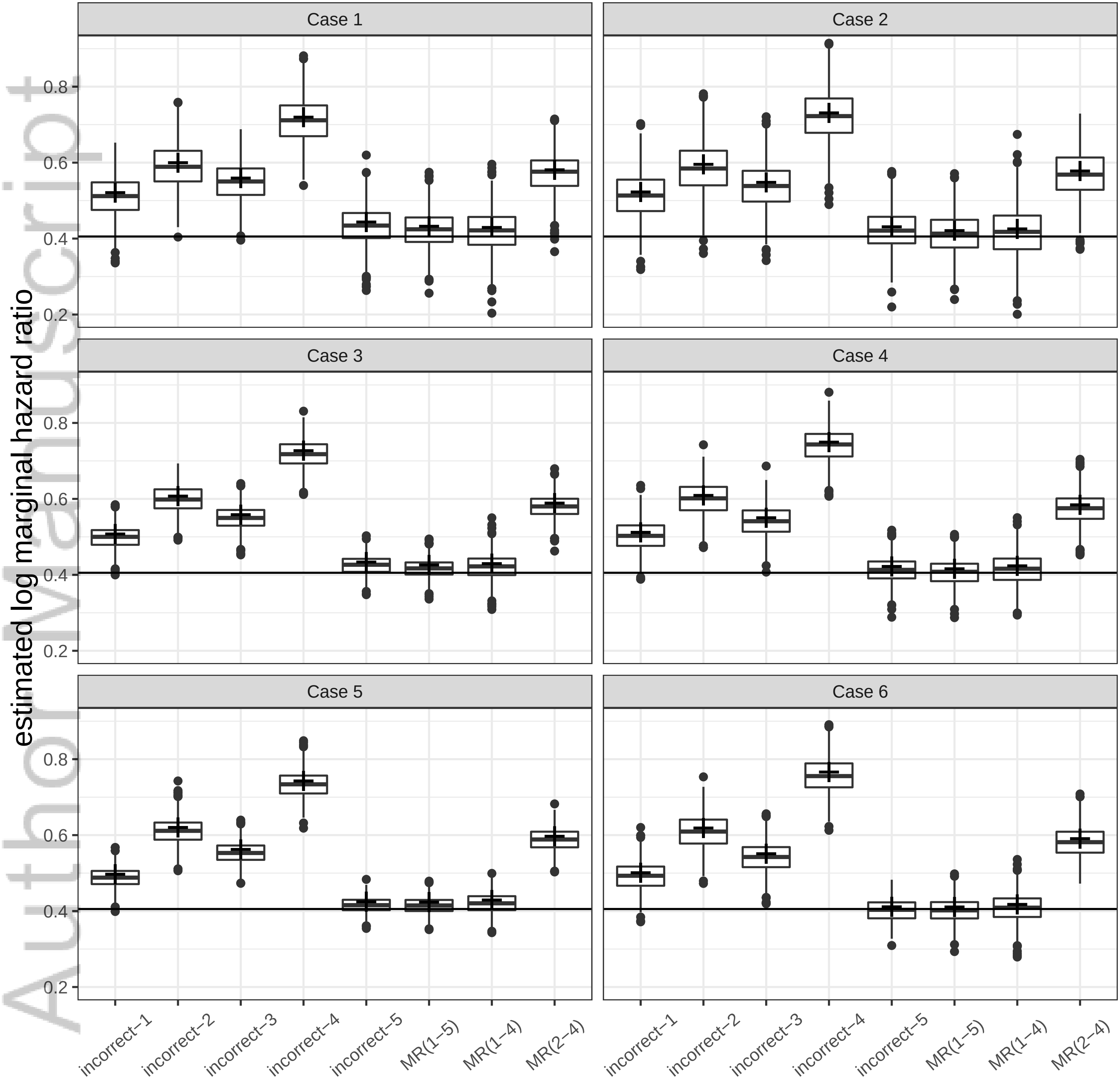


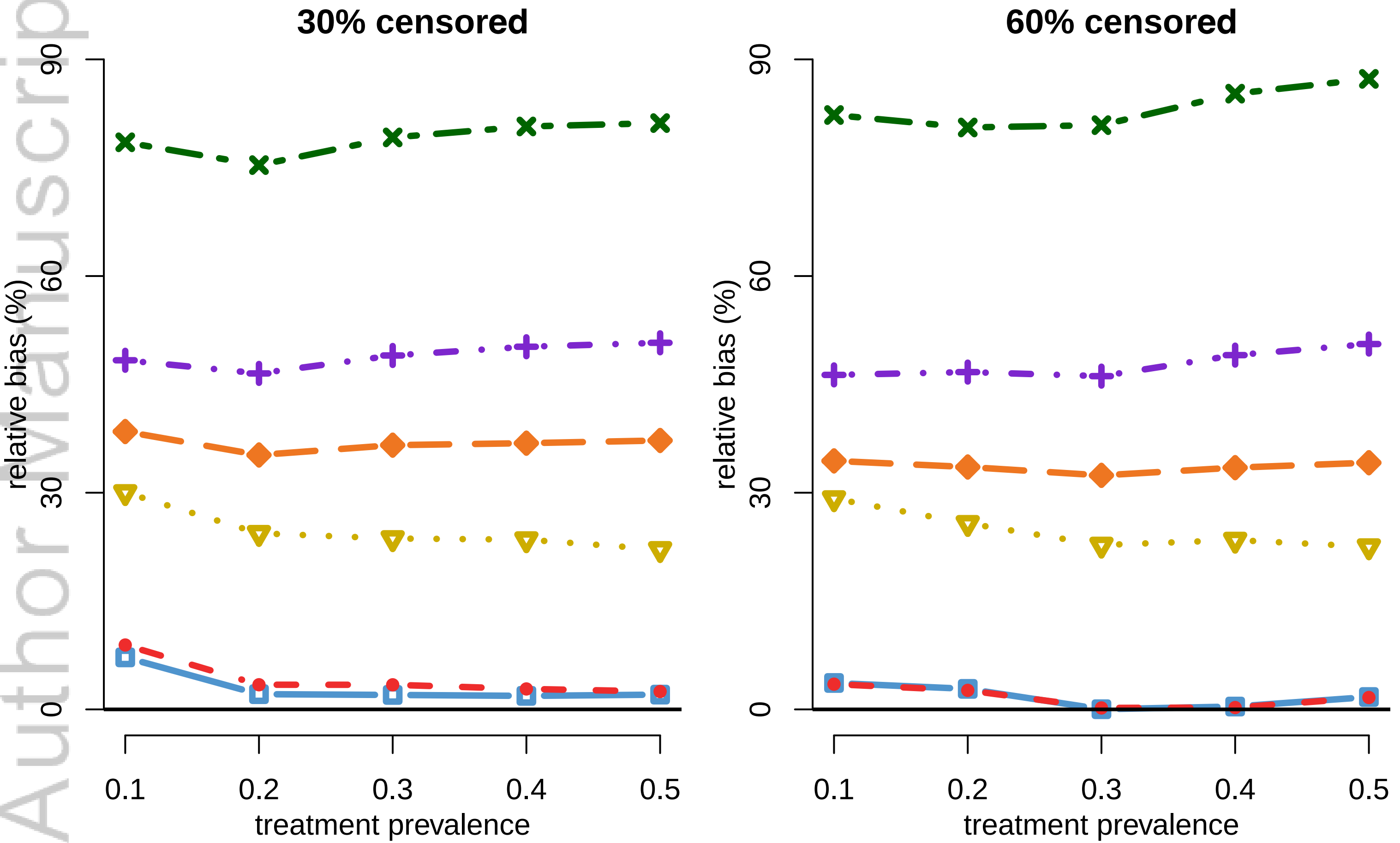
FIGURE 9 Absolute standardized differences (in percent) of baseline covariates in the original sample and the sample using the proposed weights (with both conventional and stabilized versions) for the bariatric surgery data. The vertical line denotes an absolute standardized difference of 10%, a threshold under which any covariate imbalance is generally considered negligible.[37]

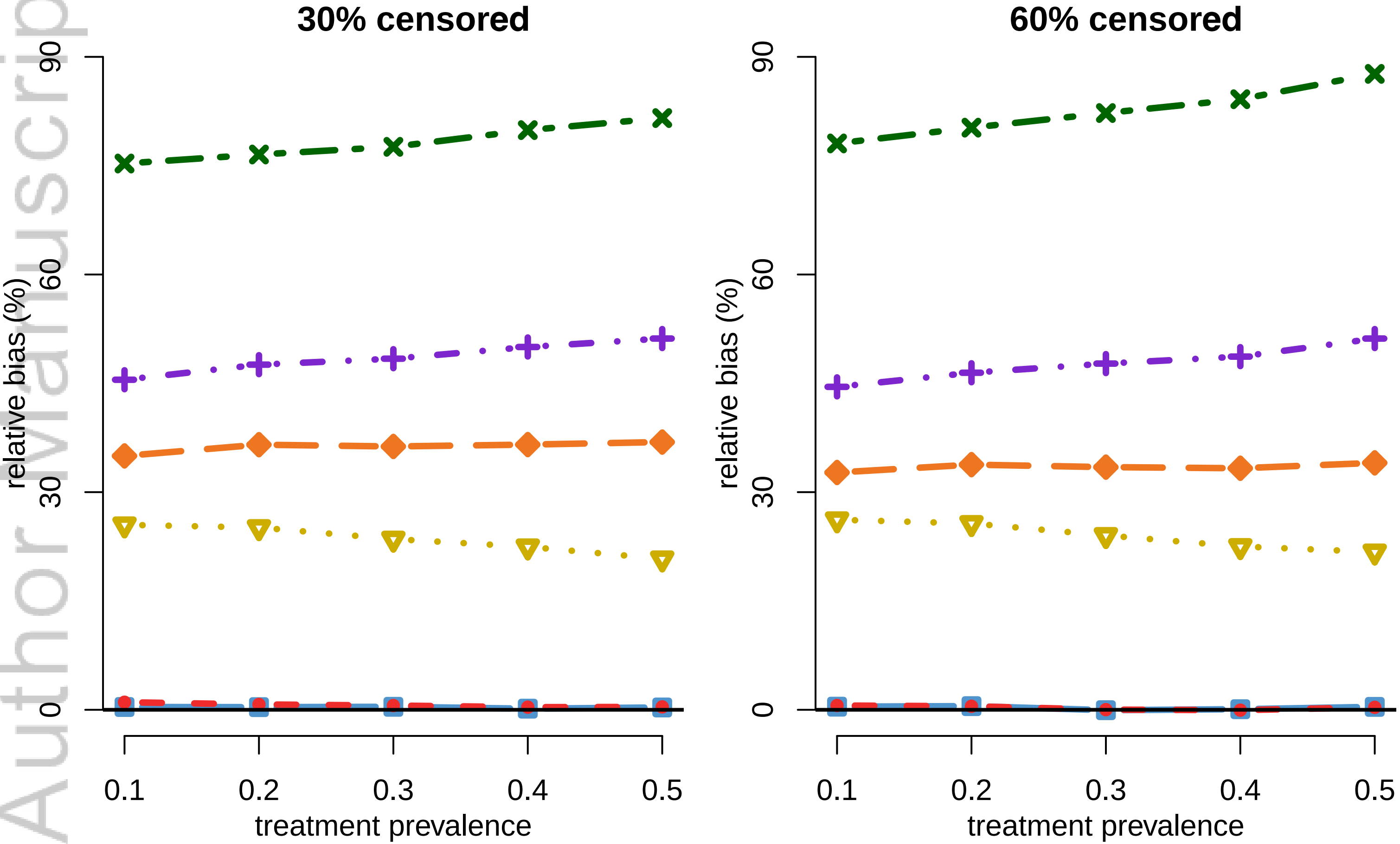
TABLE 1 Results from the real-world data analysis comparing the risk of hospitalization between sleeve gastrectomy and Roux-en-Y gastric bypass

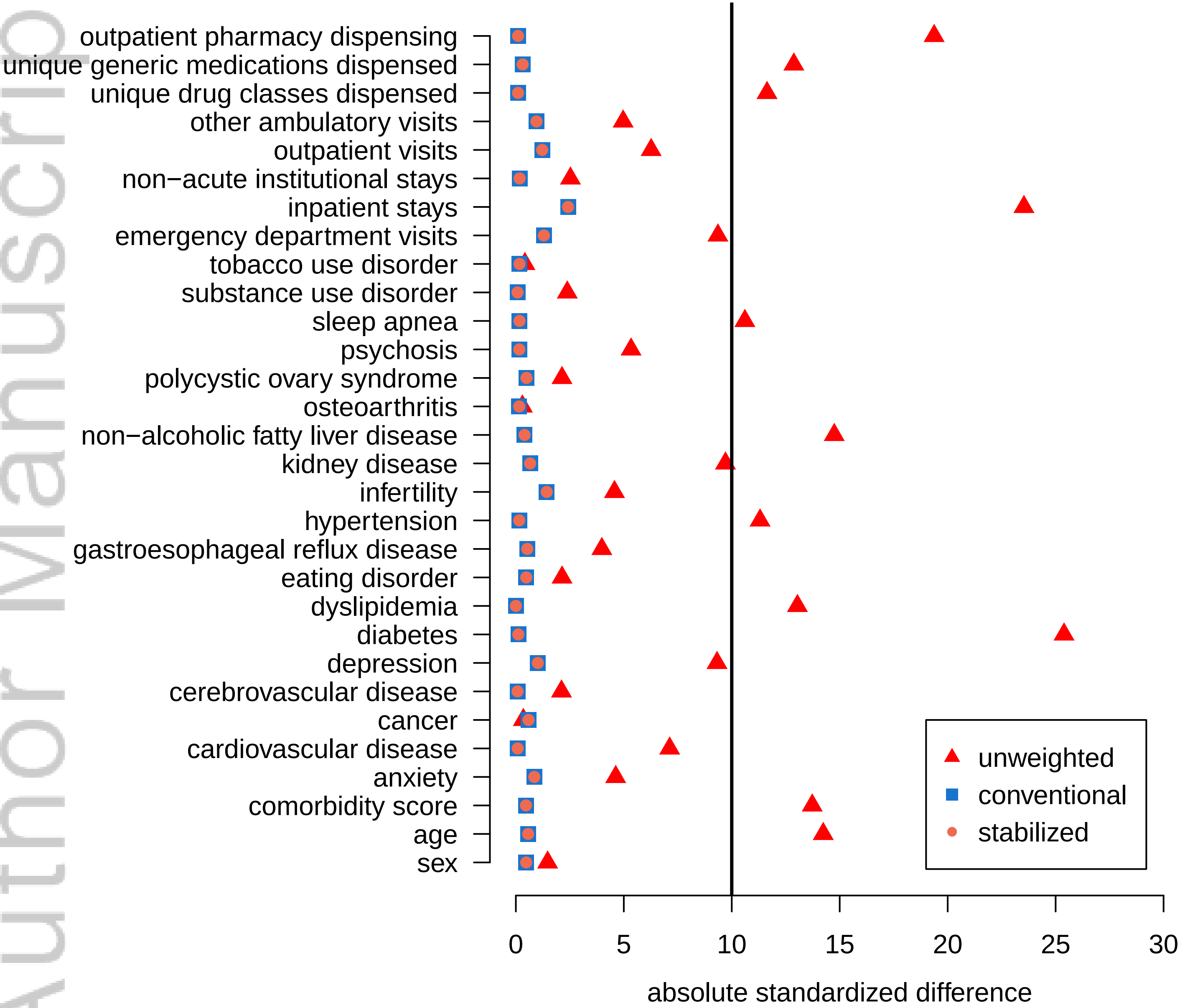
Weight	Method	Log hazard ratio	Standard Error	Hazard ratio	95% confidence interval
Conventional	PS-1	-0.372	0.143	0.689	(0.521, 0.912)
	PS-2	-0.376	0.142	0.686	(0.520, 0.906)
	PS-3	-0.435	0.137	0.647	(0.494, 0.847)
	PS-4	-0.416	0.139	0.660	(0.502, 0.866)
	PS-5	-0.404	0.140	0.668	(0.508, 0.878)
	PS-6	-0.368	0.139	0.692	(0.526, 0.909)
	MR	-0.364	0.141	0.695	(0.527, 0.916)
Stabilized	PS-1	-0.372	0.143	0.690	(0.521, 0.912)
	PS-2	-0.376	0.142	0.687	(0.520, 0.907)
	PS-3	-0.435	0.137	0.647	(0.495, 0.847)
	PS-4	-0.416	0.139	0.660	(0.503, 0.866)
	PS-5	-0.404	0.140	0.668	(0.508, 0.878)
	PS-6	-0.368	0.139	0.692	(0.526, 0.909)
	MR	-0.364	0.141	0.695	(0.527, 0.916)

PS-1: IPW Cox estimator using a logistic propensity score model including all 30 pre-specified covariates (see text for the list of covariates);
PS-2: IPW Cox estimator using a logistic propensity score model including sex, age, Charlson/Elixhauser combined comorbidity score and 15 other covariates that were univariately statistically significant at 5% level (see text for the list of selected covariates);
PS-3: IPW Cox estimator using a logistic propensity score model including sex, age, Charlson/Elixhauser combined comorbidity score, and three interaction terms between sex and age, sex and comorbidity score, and age and comorbidity score;
PS-4: IPW Cox estimator using a logistic propensity score model including diagnosis covariates in category **(b)** (see text for the list of diagnosis covariates);
PS-5: IPW Cox estimator using a logistic propensity score model including health services utilization covariates in category **(c)** (see text for the list of health services utilization covariates);
PS-6: IPW Cox estimator using a logistic propensity score model including drug dispensing covariates in category **(d)** (see text for the list of drug dispensing covariates);
MR: the proposed multiply robust estimator using all six propensity score models simultaneously.


Case 1
Case 2
Case 3
Case 4
Case 5
Case 6
estimated log marginal hazard ratio
0.2
0.4
0.6
0.8
incorrect−1
incorrect−2
incorrect−3
incorrect−4
incorrect−5
MR(1−5)
MR(1−4)
MR(2−4)


30% censored
60% censored
relative bias (%)
0
30
60
90
0.1
0.2
0.3
0.4
0.5
treatment prevalence


30% censored
60% censored
relative bias (%)
relative bias (%)
0
30
60
90
0
30
60
90
0.1
0.2
0.3
0.4
0.5
0.1
0.2
0.3
0.4
0.5
treatment prevalence
treatment prevalence


outpatient pharmacy dispensing
unique generic medications dispensed
unique drug classes dispensed
other ambulatory visits
outpatient visits
non−acute institutional stays
inpatient stays
emergency department visits
tobacco use disorder
substance use disorder
sleep apnea
psychosis
polycystic ovary syndrome
osteoarthritis
non−alcoholic fatty liver disease
kidney disease
infertility
hypertension
gastroesophageal reflux disease
eating disorder
dyslipidemia
diabetes
depression
cerebrovascular disease
cancer
cardiovascular disease
anxiety
comorbidity score
age
sex
▲ unweighted
■ conventional
● stabilized
0
5
10
15
20
25
30
absolute standardized difference

Dear Dr. Joel Greenhouse:

Thank you for your interest in our article SIM-19-0866, entitled "*Estimating Marginal Hazard Ratios by Simultaneously Using A Set of Propensity Score Models: A Multiply Robust Approach*." We appreciate the constructive comments from the reviewers and the opportunity to respond to these comments and revise the manuscript. We have revised the manuscript based on the comments. Enclosed please find the revised manuscript and a point-by-point response to all comments.

In preparing a revision, we make sure that the manuscript conforms to the Statistics in Medicine style guidelines, with references using the American Medical Association reference style.

All authors have reviewed and approved the revision. We look forward to hearing from you.

Sincerely,

Di Shu, PhD

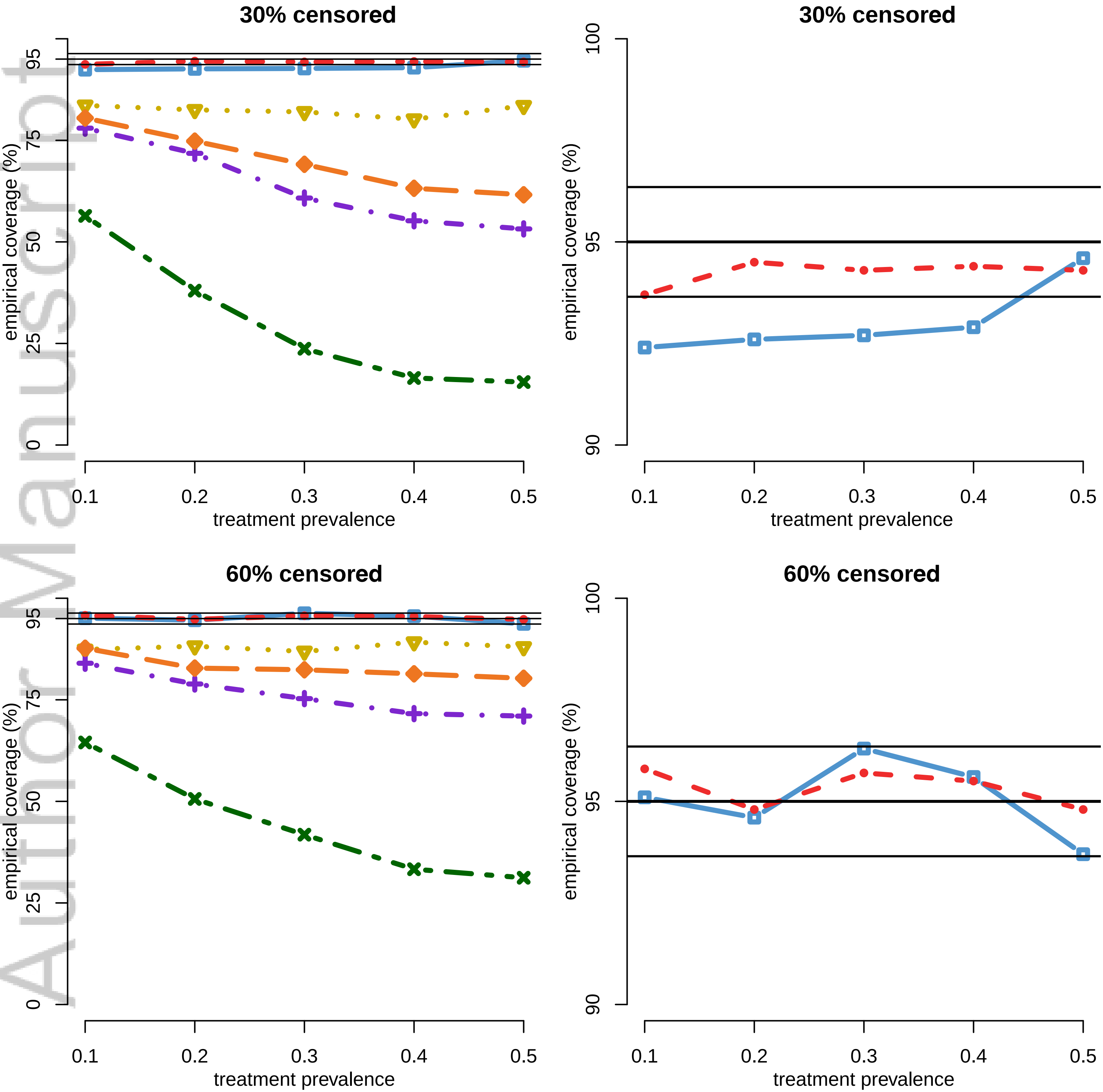
Department of Population Medicine

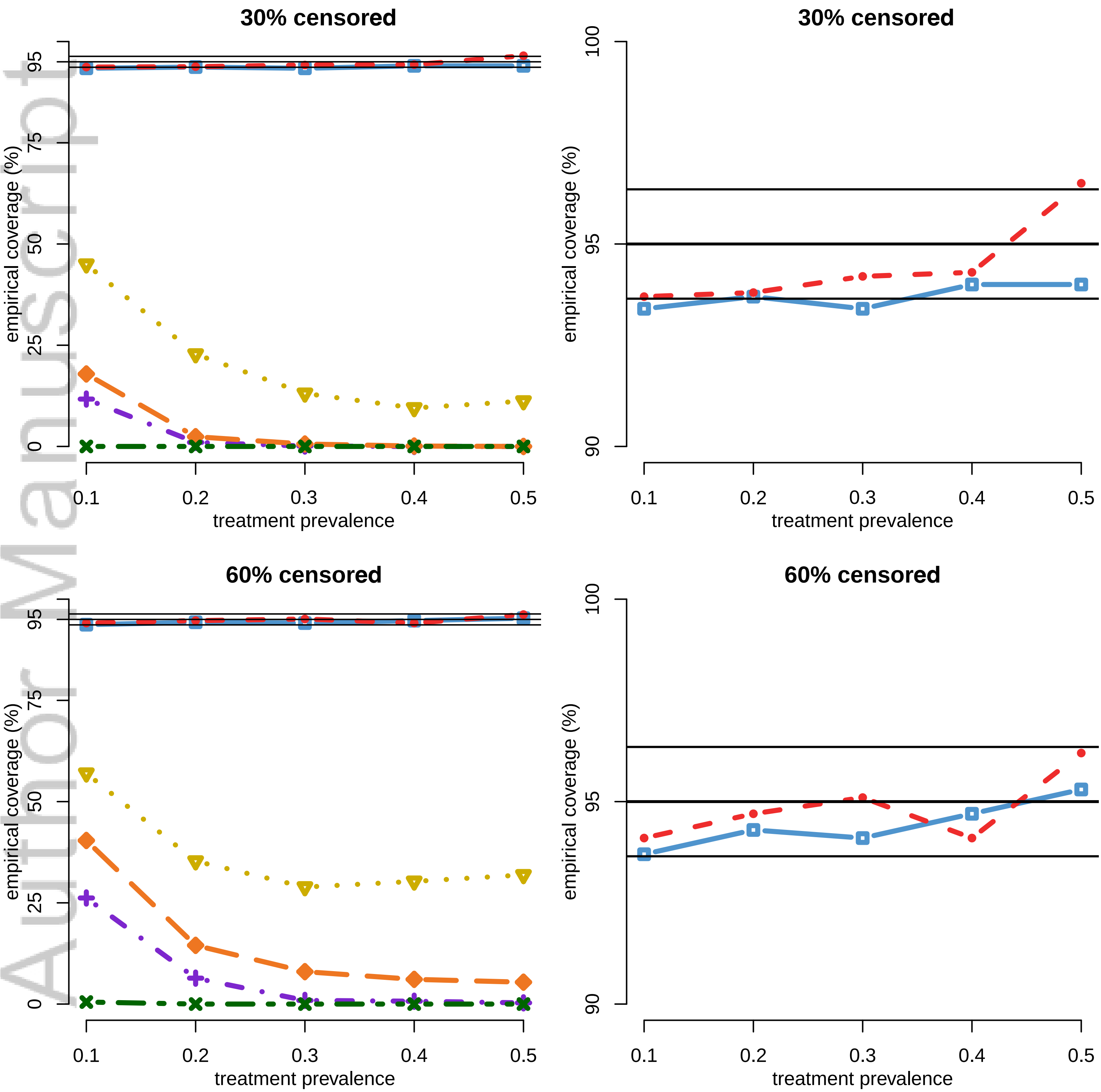
Harvard Medical School and Harvard Pilgrim Health Care Institute

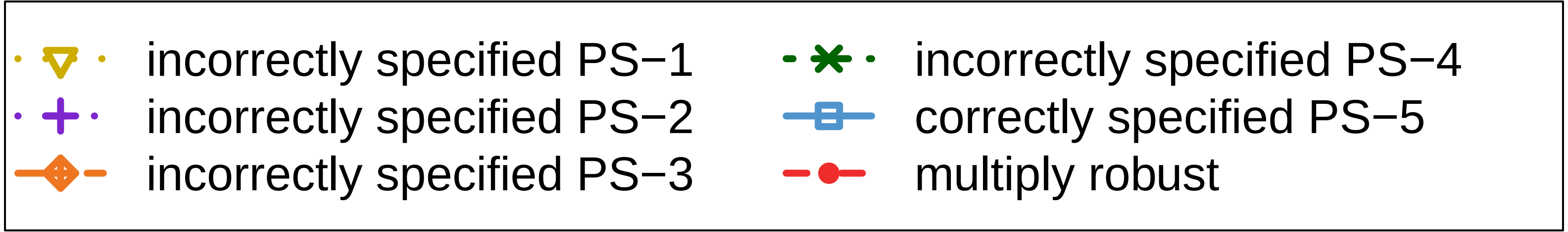
401 Park Drive, Suite 401, Boston, MA 02215

Email: Di_Shu@harvardpilgrim.org

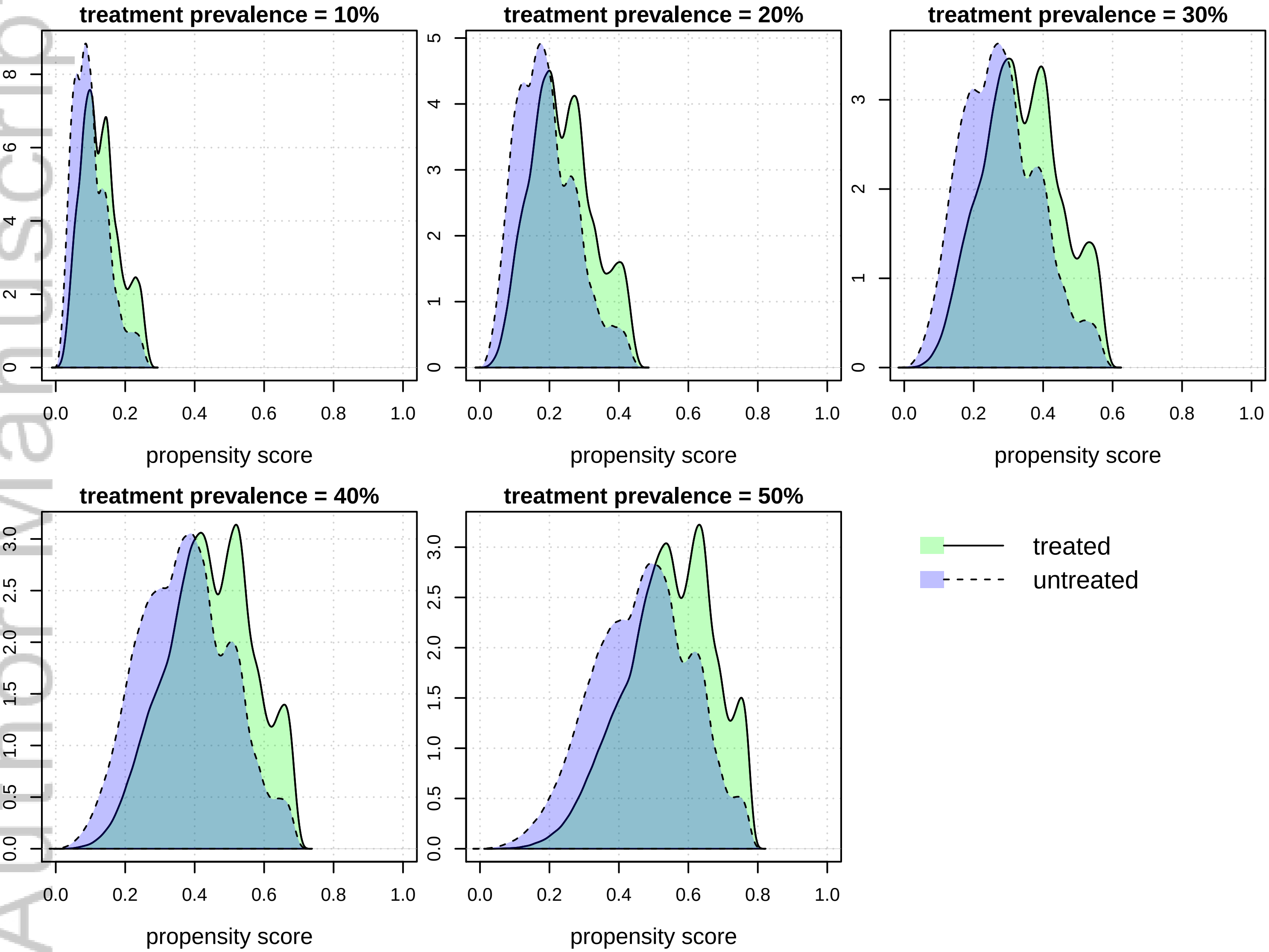
shudi1991@gmail.com

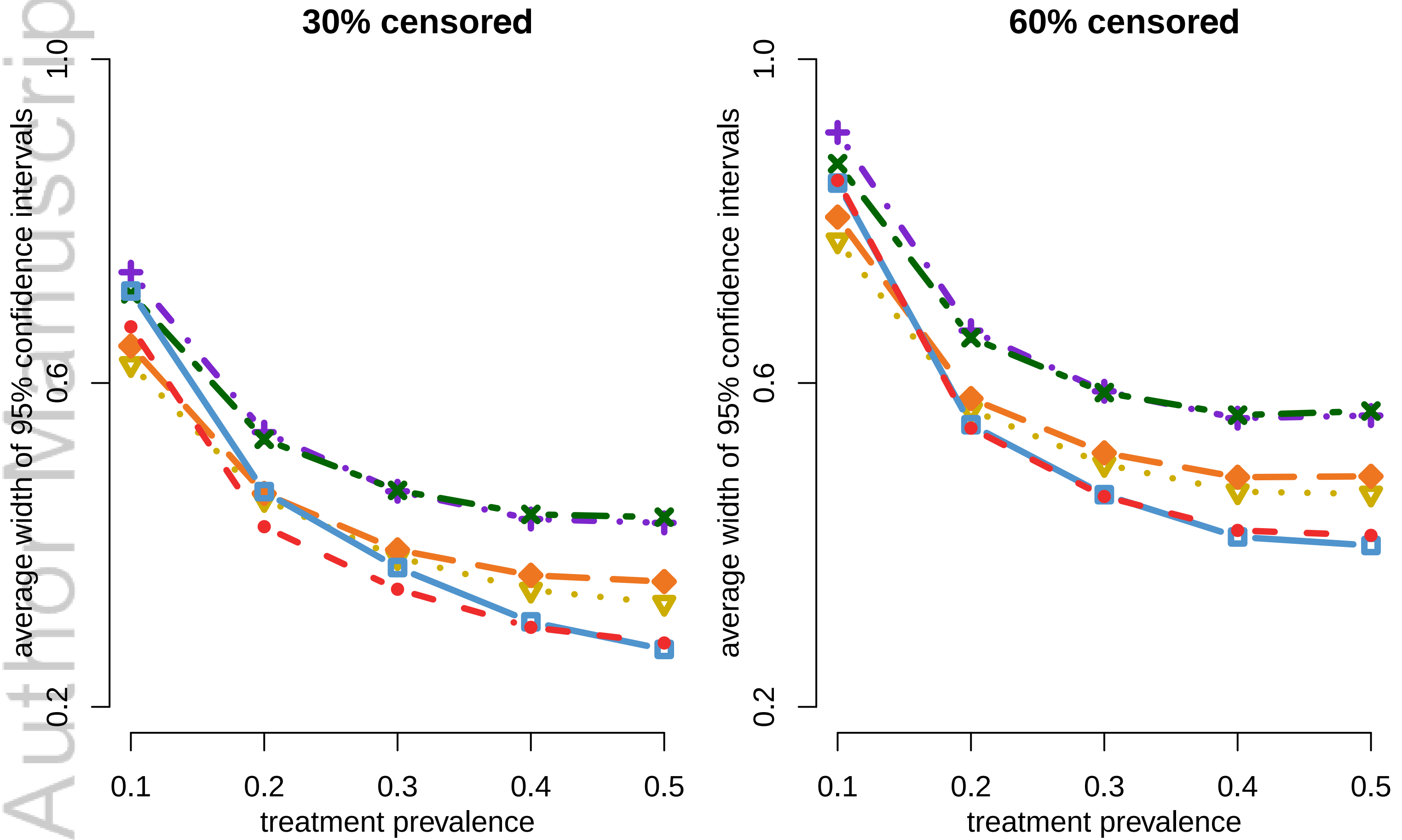

30% censored
empirical coverage (%)
0
25
50
75
95
0.1
0.2
0.3
0.4
0.5
treatment prevalence
30% censored
empirical coverage (%)
90
95
100
0.1
0.2
0.3
0.4
0.5
treatment prevalence
60% censored
empirical coverage (%)
0
25
50
75
95
0.1
0.2
0.3
0.4
0.5
treatment prevalence
60% censored
empirical coverage (%)
90
95
100
0.1
0.2
0.3
0.4
0.5
treatment prevalence

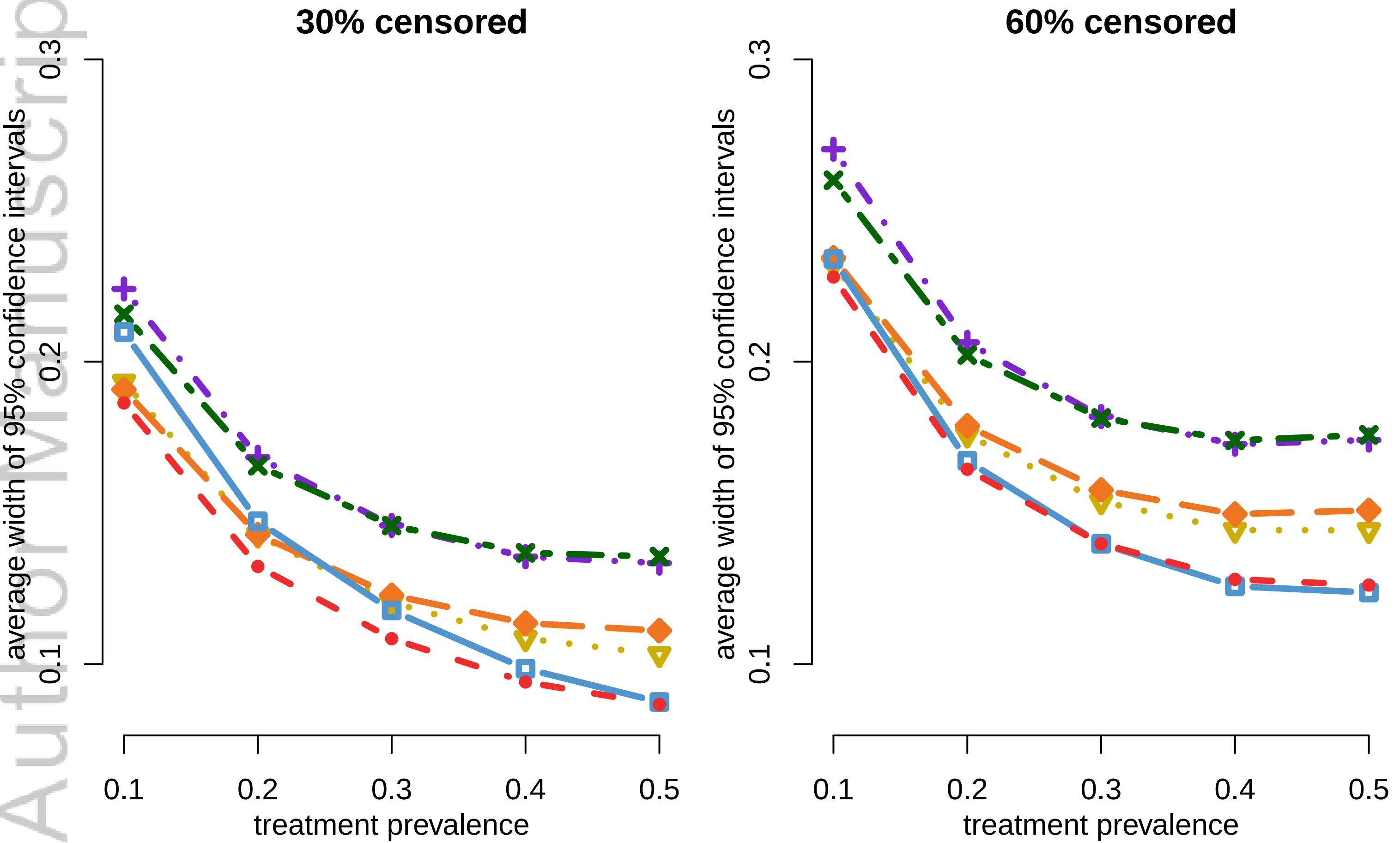

30% censored
empirical coverage (%)
0
25
50
75
95
0.1
0.2
0.3
0.4
0.5
treatment prevalence
30% censored
empirical coverage (%)
90
95
100
0.1
0.2
0.3
0.4
0.5
treatment prevalence
60% censored
empirical coverage (%)
0
25
50
75
95
0.1
0.2
0.3
0.4
0.5
treatment prevalence
60% censored
empirical coverage (%)
90
95
100
0.1
0.2
0.3
0.4
0.5
treatment prevalence



incorrectly specified PS−1
incorrectly specified PS−2
incorrectly specified PS−3
incorrectly specified PS−4
correctly specified PS−5
multiply robust


treatment prevalence = 10%
0
2
4
6
8
0.0
0.2
0.4
0.6
0.8
1.0
propensity score
treatment prevalence = 20%
0
1
2
3
4
5
0.0
0.2
0.4
0.6
0.8
1.0
propensity score
treatment prevalence = 30%
0
1
2
3
0.0
0.2
0.4
0.6
0.8
1.0
propensity score
treatment prevalence = 40%
0.0
0.5
1.0
1.5
2.0
2.5
3.0
0.0
0.2
0.4
0.6
0.8
1.0
propensity score
treatment prevalence = 50%
0.0
0.5
1.0
1.5
2.0
2.5
3.0
0.0
0.2
0.4
0.6
0.8
1.0
propensity score
treated
untreated


30% censored
60% censored
average width of 95% confidence intervals
0.2
0.6
1.0
0.1
0.2
0.3
0.4
0.5
treatment prevalence


30% censored
60% censored
average width of 95% confidence intervals
0.1
0.2
0.3
treatment prevalence
0.1
0.2
0.3
0.4
0.5