Robust weights that optimally balance confounders for estimating the effect of binary and continuous treatments with time-to-event data

Michele Santacatterina*
Department of Biostatistics and Bioinformatics,
The George Washington University

October 16, 2020

Abstract

Covariate balance is crucial in obtaining unbiased estimates of treatment effects in observational studies. Methods based on Inverse Probability Weights (IPW) have been widely used to estimate treatment effects with observational data. Machine learning techniques have been proposed to estimate propensity scores. These techniques however target accuracy instead of covariate balance. Methods that target covariate balance have been successfully proposed and largely applied to estimate treatment effects on continuous outcomes. However, in many medical and epidemiological applications, the interest lies in estimating treatment effects on time-to-an-event outcomes. In this paper, we start by presenting robust orthogonality weights (ROW), a set of weights obtained by solving a quadratic constrained optimization problem that maximizes precision while constraining covariate balance defined as the sample correlation between confounders and treatment. By doing so, ROW optimally deal with both binary and continuous treatments. We then evaluate the performance of the proposed weights in estimating hazard ratios of binary and continuous treatments with time-to-event outcomes in a simulation study. We finally apply ROW on

*The author gratefully acknowledge Xiaojie Mao for valuable comments and help on this paper

the evaluation of the effect of hormone therapy on time to coronary heart disease and on the effect of red meat consumption on time to colon cancer among 24,069 postmenopausal women enrolled in the Women's Health Initiative observational study.

Keywords: survival analysis; covariate balance; Cox regression; optimization; continuous treatments; hazard ratio

1 Introduction

Covariate balance is crucial in obtaining unbiased estimates of treatment effects in observational studies. Weighted methods based on Inverse Probability Weights (IPW) have been used to estimate the effect of a treatment on an outcome using observational data. IPW are constructed as the inverse of the probability of a unit being assigned to a treatment conditional to pre-treatment covariates, *i.e.*, the propensity score (Rosenbaum and Rubin, 1983). Despite their wide use, IPW-based methods tediously rely on the correct specification of the propensity score model, which violations lead to biased estimates, and on the positivity assumption (Imbens and Rubin, 2015), which *practical* violations (Petersen et al., 2012) lead to extreme weights and erroneous inferences (Robins et al., 1995; Scharfstein et al., 1999; Robins et al., 2007; Kang and Schafer, 2007). Machine learning techniques, like the Super Learner (Van der Laan et al., 2007), have been proposed to improve propensity score estimation in the case of model misspecification (Lee et al., 2010; Pirracchio et al., 2015). These techniques however target accuracy instead of covariate balance.

Methods that mitigate model misspecification while targeting covariate balance have been proposed. Among others, Imai and Ratkovic (2014) proposed and extended (Fong et al., 2018) Covariate Balancing Propensity Score (CBPS), which uses generalized method of moments to estimate the logistic regression model that optimally balances covariates. Zubizarreta (2015) proposed Stable Balancing Weights (SBW), a set of weights with minimal sample variance that satisfy a list of approximate moment matching conditions. Hainmueller (2012) presented entropy balancing weights obtained by minimizing the entropy of the weights while satisfying balance conditions. The literature of covariate balance is extensive and many other methods have been developed in the recent years (Kallus and Santacatterina, 2019b,a; Kallus and Santacatterina, 2018; Hirshberg et al., 2019; Hirshberg and Wager, 2017; Zhao and Percival, 2017; Zhao et al., 2019; Wong and Chan, 2017; Visconti and Zubizarreta, 2018; Zubizarreta et al., 2014; Li et al., 2018; King et al., 2017; Tübbicke, 2020; Vegetabile et al., 2020; Wu et al., 2018; Yiu and Su, 2018, among others)

Although methods that target covariate balance have been shown to be robust to either model misspecification, practical positivity violation or both, these methodologies have been mainly developed and applied to estimate treatment effects on continuous outcomes. However, in many medical and epidemiological applications, the interest lies in estimating the effect of a treatment effects on time-to-an-event outcomes. Examples include, the evaluation of the impact of hormone therapy on time to coronary heart disease (CHD) (Hulley et al., 1998; Manson et al., 2003), and the impact of red meat consumption on time to colon cancer (Larsson et al., 2005). When estimating treatment effects with time-to-event data one of the most common causal estimand of interest is the hazard ratio of the Cox proportional hazard model (although other estimands may be of interest such as the survival difference or the mean survival (Mao et al., 2018)).

In this paper, we start by presenting robust orthogonality weights (ROW), which optimally and robustly balance covariates for estimating effects of binary and continuous treatments. We then evaluate its performance in estimating hazard ratios of binary and continuous treatments with time-to-event outcomes. The proposed weights are obtained by solving a convex constrained quadratic optimization problem which minimize the sample variance of the weights, thus controlling for extreme weights and maximizing precision, while constraining the sample correlation between treatment and covariates, thus optimally balance covariates with respect to either binary or continuous treatments. Similar to IPW, matching, CBPS and SBW, the set of ROW is obtained without the use of the outcome, thus emulating randomization. By minimizing the sample variance of the weights while controlling for a measure of balance, ROW is constructed in the spirit of SBW. However, differently from SBW, ROW can also be used to estimate effects of continuous treatments. Also, by using the sample correlation as a measure of balance, ROW can be seen as an extension of CBPS (Fong et al., 2018) in which precision is maximized while satisfying a constraint on the covariate balance. Finally, ROW can be seen as a specific case of the framework of Yiu and Su (2018) (See for instance Section 4.2 of Yiu and Su (2018) for continuous treatments).

Our contribution to this field of literature is twofold: to provide a set of weights of minimal variance that optimally balance confounders for estimating the effect of binary and continuous treatments; and to evaluate and apply the proposed weights to the estimation of hazard ratios for time-to-event data. In addition, we contribute to this literature, by providing a thorough comparison of the performance of several covariate balancing methods with time-to-event data in multiple simulation scenarios and in two case studies. Finally, we provide a `R` package for the computation of the weights available at https://github.com/michelesantacatterina/ROW. Code for simulations and case studies analyses is available at https://github.com/michelesantacatterina/ROW-time-to-event.

In the next Section we present ROW and provide some practical guidelines on their use. In Section 3, we evaluate the performance of ROW with respect to absolute bias, root mean squared error, covariate balance and computational time across levels of practical positivity violation, misspecification and censoring for both binary and continuous treatments. In Section 4, we apply ROW to the evaluation of the effect of hormone therapy on time to first occurrence of coronary heart disease and the impact of red meat consumption on time to colon cancer onset using data from 24,069 postmenopausal women enrolled in the Women's Health Initiative observational study (Women's Health Initiative, 1998). We provide some concluding remarks in Section 5.

2 Robust orthogonality weights

In this section, we first introduce some notations and assumptions, we then present the optimization problem used to obtain robust orthogonality weights and finally discuss some practical guidelines.

2.1 Notations and assumptions

Suppose we have a simple random sample of size n from a target population. For each unit $i = 1, \dots, n$, suppose we observe a binary or continuous treatment A_i, a set of pre-treatment covariates (also referred to as confounders) $\mathbf{X}_i$, and the observed time to an event, T_i. In addition, let C_i denote the i^{th} individual's censoring time, $\Delta_i = \mathbb{I}\left[T_i < C_i\right]$, the complete-case indicator and Y_i the observed event only if $\Delta_i = 1$. We define the potential (counterfactual) follow-up time and response as $T_i(a)$ and $Y_i(a)$ respectively. Throughout this paper, in addition to consistency and non-interference (Imbens and Rubin, 2015), we assume

Assumption 2.1. *Strong ignorability* $\{T_i(a), Y_i(a)\} \perp\!\!\!\perp A_i | \mathbf{X}_i$*,*

Assumption 2.2. *Noninformative censoring* $C_i \perp\!\!\!\perp \{T_i(a), Y_i(a)\}$*,*

Assumption 2.3. *Positivity* $\phi(\mathbf{X}_i) = P(T_i = t | \mathbf{X}_i) > 0$*,*

where $\phi(\mathbf{X}_i)$ is the classic propensity score if the treatment is binary and the generalized propensity score as presented in Hirano and Imbens (2004) if the treatment is continuous. The main causal estimand of interest is the marginal hazard ratio, *i.e.,* the θ parameter of the Cox proportional hazard model

$$\lambda_{T(a)}(t) = \lambda_0 \exp\left(\theta A\right).$$

To estimate θ, standard practice suggests to use an outcome model such as the conditional Cox regression model $\lambda(t|\mathbf{X}_i) = \lambda_0 \exp(\theta A_i + \beta \mathbf{X}_i)$ (a conditional estimator), or an IPW-marginal Cox model, *i.e.,* an IPW-weighted Cox regression, regressing only the treatment on the time to the event, with propensity scores estimated using regression or machine learning techniques (see Buchanan et al. (2014) for an applied example). These procedures provide biased and erroneous inferences when the outcome model or the propensity score model is misspecified or when the positivity assumption is practically violated. In addition, covariate balance is not targeted. In the next Section, we introduce a convex quadratic constrained optimization problem to obtain robust orthogonality weights that target covariate balance and are robust to practical positivity violations and misspecification.

2.2 A convex quadratic constrained optimization problem

A general measure of covariate balance is the correlation between treatment and covariates. When the correlation is equal to zero, the covariates are uncorrelated from the treatment. In the spirit of Zubizarreta (2015), we propose to find weights with minimum variance while satisfying contraints on the sample correlation between covariates and the treatment under study. We therefore propose to obtain ROW by solving the following quadratic linearly constrained optimization problem,

$$
\begin{aligned}
\underset{\mathbf{w}}{\text{minimize}} \quad & \|\mathbf{w} - e_n\|_2^2 && (2.1)\\
\text{subject to} \quad & |\rho_k(\mathbf{w})| \leq \delta,\ k = 1, \ldots, m, && (2.2)\\
& e^\top \mathbf{w} = 1, && (2.3)\\
& \mathbf{w} \geq 0, && (2.4)
\end{aligned}
$$

where $\rho_k(\mathbf{w}) = \sum_{i=1}^{n} w_i X^*_{i,k} T^*_{i,k}$ is the (weighted) mean of the products of the standardized covariates and treatment, *i.e.,* , the sample correlation, X^* and T^* are the scaled covariates and treatment variables, m is the total number of covariates, e is the unit vector, and e_n is the unit vector divided by the sample size n which, by construction, represent the mean of the weights. When a solution to optimization problem (2.1)-(2.4) exists, constraint (2.2) guarantees that the correlation between treatment and covariates is at most equal to δ (for each covariate), and constraints (2.3) and (2.4) guarantee that the weights sum up to 1 and are positive, respectively.

What does the set of ROW achieve? In Figure 1, we provide a simple scenario in which a binary treatment (left panel of Figure 1) and a continuous treatment (right panel) are generated as a function of a continuous covariate (x-axis), using a probit model and a simple regression model with normal errors, respectively. Blue lines represent the true relationship between the covariate and the treatment. For instance, in the continuous treatment scenario, the covariate has a positive impact on the treatment, *i.e.,* the regression coefficient is positive. The black lines represent the values of the regression coefficients after weighting by ROW (we first computed ROW by solving optimization problem (2.1)-(2.4) and then plug the obtained weights into a weighted probit model and ordinary square regressor, regressing the covariate on the treatment). Under the binary treatment scenario, the weighted coefficient equals 0.5, while it equals 0 in the continuous treatment scenario for each value of the covariate. In summary, these figures show that ROW orthogonalize covariate and treatment variables thus eliminating associations between them. By doing so, as shown in our simulations in Section 3 and in our case studies in Section 4 ROW maximize covariate balance. Figure 16 in Section 8 of the Supplementary Material shows that ROW orthogonalize covariates and treatment variables also when the true relationship between them is nonlinear quadratic, nonlinear cubic, nonlinear without correlation, sinusoidal, and when the treatment is right- and left-skewed. In addition Figure 16 shows that under independence between treatment and a covariate, ROW result in almost uniform weights.

2.3 Practical guidelines

In this section, we provide some practical guidelines on the choice of the parameter δ, and on standard error estimation for the marginal hazard ratio.

Optimization problem (2.1)-(2.4) depends on the parameter δ. This parameter set the upper bound for the absolute value of the sample correlation between treatment and covariates. Smaller values of δ induce smaller correlation, thus inducing higher balance and consequently lower bias. Figure 2 provides a graphical representation of the impact of δ on bias, mean squared error and balance. Precisely, it shows absolute bias (left panels), root mean square error (middle panels) of the hazard ratio estimated using a Cox regression model weighted by ROW, and mean covariate balance across four continuous covariates (right panels) for binary (upper panels) and continuous treatments (lower panels), across levels of the parameter δ, set equal to $0.0001, 0.025, 0.05, 0.075$, and 0.1 (Simulations details are provided in section 3). For the binary treatment, we considered the absolute standardized mean difference as a measure of balance while for the continuous treatment we considered the absolute correlation between treatment and covariates. Lower values of δ guarantee the lowest absolute bias, and balance. In addition, since optimization problem (2.1)-(2.4) constraints imbalance while controlling precision by minimizing the variance of the weights, lower values of δ also guarantee minimal root mean squared error for both types of treatments. We consequently suggest to set $\delta = 0.0001$ or $\delta = 0.001$.

As suggested by other authors (Hernán et al., 2001; Robins et al., 2000), we suggest to use the robust "sandwich" estimator (Freedman, 2006) for the estimation of the standard error. When computational resources are not limited, we also suggest to use bootstrap

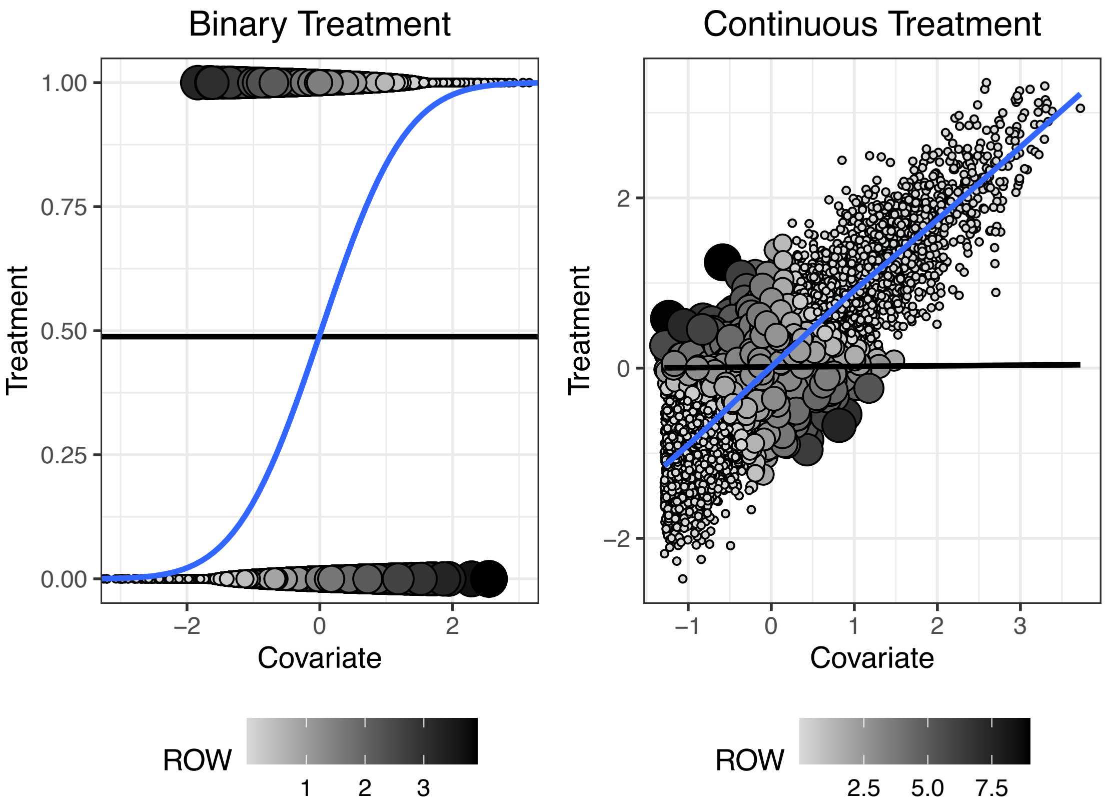

Figure 1: Graphical representation of ROW balancing a covariate (x-axis) across a binary (left panel) and continuous (right panel) treatments (y-axis). Blue lines represent the true relationships between the binary (a probit model) and the continuous (simple regression with normal errors and positive coefficient) treatment. Black lines represent the relationship between treatments and covariate after weighting for ROW. Size and color of the circles represent the individual weight assigned (the larger/darker the higher).

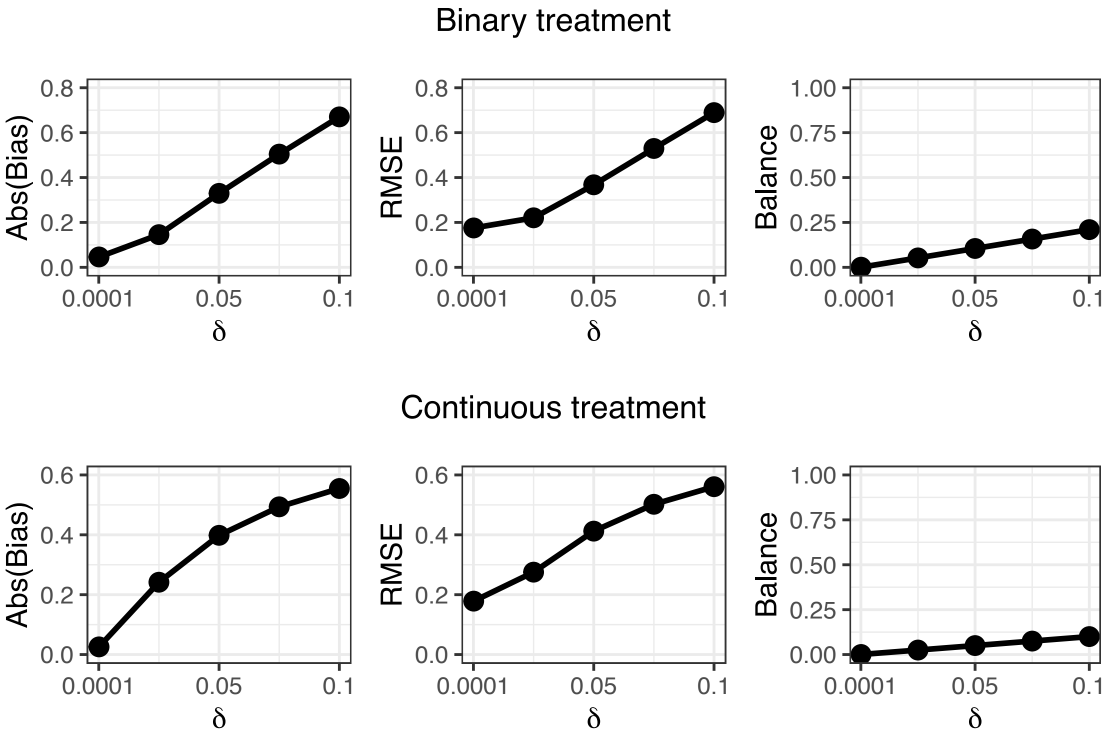

Figure 2: Absolute bias (left panels), root mean squared error (RMSE) (middle panels), absolute standardized mean difference (Balance)(top right panel), absolute correlation (bottom right panel), for the binary (top panels) and continuous (bottom panels) treatment across levels of the balance constraint δ (eq. (2.1)).

(Davison and Hinkley, 1997). We provide a comparison between the naive (the inverse of the observed Fisher information of $\hat{\theta}$, the estimated hazard ratio of the Cox hazard model introduced in Section 2.1), robust and bootstrap standard errors in Section 6.1 of the Supplementary Material. To compute the bootstrap standard error, we used non-parametric bootstrap with normal approximation confidence intervals (Davison and Hinkley, 1997).

Similarly to Zubizarreta (2015) and Santacatterina and Bottai (2018); Santacatterina et al. (2019), Lagrange multipliers can be used to evaluate the impact that a small decrease in the parameter δ would cause in the objective function (2.2). We refer to Section 3.3 of Zubizarreta (2015) and Section 3 and 3.1 of Santacatterina and Bottai (2018) and (Santacatterina et al., 2019), respectively for detail.

Many solvers are available to solve constrained convex quadratic optimization problems. We suggest using `Gurobi` (Gurobi Optimization, 2020).

3 Simulations

In this section we evaluate the performance of ROW with respect to, absolute bias, root mean square error, covariate balance and computational time, across levels of practical positivity violations, misspecification and censoring when estimating the marginal hazard ratio with a binary and continuous treatment. In summary, ROW performs well across all of the considered scenarios.

3.1 Setup

We considered a sample size of $N = 1,000$. We computed the expected survival time t by following the inverse probability method based on the Weibull distribution (Bender et al., 2005),

$$T_i = \left(- \frac{\log(u)}{\psi \exp\left(\theta A_i + \mathbf{X}_i^\top \beta\right)} \right)^{\frac{1}{\rho}},$$

where $u \sim \mathrm{Unif}(0,1)$, $\psi = 0.01$ (the scale parameter of the Weibull distribution), $\theta = 0.2$, $X_1 \sim \mathrm{N}(0.1,1), X_2 \sim \mathrm{N}(0.1,1), X_3 \sim \mathrm{logN}(0,0.5), X_4 \sim 5\mathrm{Beta}(3,1)$ X_5 is a random sample with replacement of size 4 with probabilities $0.35, 0.25, 0.05, 0.35$ respectively, and $X_6 \sim \mathrm{Binom}(0.25)$, $\beta = (0,1,0,1.4,1.4,1)$, and $\rho = 1$ (the shape parameter of the Weibull distribution). With the choice of these six covariates we wanted to reflect real-world populations in which the time to an event depends on binary, categorical and continuous (normal and non-normal) covariates. We generated the censoring times C_i using an exponential distribution, *i.e.*, $C_i \sim \mathrm{Exp}(\epsilon)$, with values for the rate parameter ϵ described in section 3.1.4. Finally, we obtained the observed (censored) survival times by taking the minimum between T_i and C_i. The causal estimand of interest is the hazard ratio (HR), $HR = \exp\theta = 1.22$. We provide detailed information on how we generated A_i in the following Section.

3.1.1 Estimating HR for binary and continuous treatments

To evaluate the performance of ROW, we considered estimating the hazard ratio under two scenarios: binary and continuous treatments. We refer to the first scenario as the *binary* scenario and the second as the *continuous* scenario. In the binary scenario, we considered $A_i \sim \mathrm{Binom}(\pi(\mathbf{X}_i))$, where $\pi(\mathbf{X}_i) = \left(1 + \exp(\gamma\left(\frac{\kappa}{\gamma} - \mathbf{X}_i^\top \mathbf{e}\right))\right)^{-1}$ and $\kappa = n^{-1}\sum_{i=1}^{n}\left(\mathbf{X}_i^\top \gamma\right)$. In the continuous scenario, we considered $A_i \sim \mathrm{logNorm}(\mu(\mathbf{X}_i), \eta)$, where $\mu(\mathbf{X}_i) = -\kappa + \mathbf{X}_i^\top \mathbf{e}$. The rate parameter ϵ was set equal to 0.01 (low percentage of censored observation - see section 3.1.4 for details) and the parameter η was set to be equal to 0 (null misspecification - see Section 3.1.3 for details). We provide detailed information on the parameters γ and η in the following Section.

3.1.2 Estimating HR across levels of practical positivity violations

To evaluate the performance of ROW across levels of practical positivity violations for the binary scenario, we considered five values for γ, from 0.1 to 2. We refer to $\gamma = 0.1$ as weak violation, $\gamma = 1$ as moderate violation and to $\gamma = 2$ as strong violation. The propensity score ranged from 0.13 to 0.61 under weak violation, 0.05 to 0.97 under moderate violation and 0.001 to 0.995 under strong violation (average of min/max propensities). In the continuous scenario, we considered five different values (0,0.1,0.5,0.6,0.7,0.9) for the parameter η of the log-Normal distribution, *i.e.,* the standard deviation of the random variable which higher values generate a more right-skewed distribution. We refer to $\eta = 0$ as weak violation, $\eta = 0.6$ as moderate violation and $\eta = 0.9$ as strong violation. The rate parameter ϵ for the censoring level was set equal to 0.01 (low percentage of censored observation - see Section 3.1.4 for details) and we considered correct specification (see Section 3.1.3 for details).

3.1.3 Estimating HR across levels of misspecification

We evaluated the performance of ROW across levels of misspecification. Specifically, for the binary scenario, we generated $Z_1 = (X_1+0.5)^2$, $Z_2 = ((X_1X_2)/5+1)^2$, $Z_3 = \exp(X_3/2)$, and $Z_4 = X_4(1+\exp(X_3))+1$. For the continuous scenario we generated $Z_1 = \exp(X_1/2)$, $Z_2 = X_2(1+\exp(X_1))+1$, $Z_3 = (X_1X_3/25+0.2)^3$, and $Z_4 = 2*\log(|X_4|)$. We then considered a convex combination between the correct variables (X_1, X_2, X_3, X_4) and the misspecified variables (Z_1, Z_2, Z_3, Z_4), *e.g.,* $X_1 = X_1(1-\tau) + Z_1\tau$, and let τ vary from 0 to 1 (0,0.25,0.5,0.75,1). We refer to $\tau = 0$ as null misspecification, $\tau = 0.5$ as moderate misspecification and $\tau = 1$ as strong misspecification. The rate parameter ϵ for the censoring level was set equal to 0.01 (low percentage of censored observation - see Section 3.1.4

for details) and we considered moderate pratical positivity violation (see Section 3.1.2 for details).

3.1.4 Estimating HR across levels of censoring

We also evaluated the performance of ROW across levels of censoring. We considered five values (0,1,10,100,1000) for the rate parameter ϵ of the exponential distribution used to generate the censoring times, resulting in 1, 7, 25, 52 and 78 percent of censored observation in the binary scenario and in 2, 10, 27, 53 and 76 percent of censored observation in the continuous scenario. We considered correct specification (see Section 3.1.3 for details) and moderate pratical positivity violation (see Section 3.1.2 for details).

3.1.5 Methods comparison

In addition to the standard of practice methods such as the conditional Cox proportional hazard model and IPW-Cox regression, we consider only methods that 1) use only the information of the treatment and covariates and not the outcome, 2) target covariate balance in some way and 3) their `R` implementation is readily available.

For the binary scenario, we compared ROW with IPW-Cox regression, where propensity scores were estimated by using SuperLearner (IPW) with the following library of algorithms: logistic regression model with only main effects, logistic regression with main effects and interactions, lasso-penalized logistic regression, random forest, bayesian logistic regression and extreme gradient boosting classifier; Balance SuperLearner (BalSL) as described in Pirracchio and Carone (2018) with the same library of algorithms as for IPW; and by using boosted logistic regression (GBM) (with absolute standardized mean difference as stopping criteria, interaction depth equal to 3, number of trees equal to 10,000, shrinkage equal to 0.01 and bag fraction equal to 1) (McCaffrey et al., 2004). In addition, we compared ROW with Propensity Score Matching (PSM), with propensity scores obtained as for IPW (Sekhon, 2008); Covariate Balancing Propensity Score (CBPS) containing only the covariate balancing conditions (exact identification); entropy balancing weights (EBAL) (Hainmueller, 2012); stable balance weights (SBW) (Zubizarreta, 2015) where we selected the degree of approximate covariates balance by following the tuning algorithm presented in Wang and Zubizarreta (2020), (we chose the grid of values for the tuning algorithm equal to 0.0001, 0.001, 0.002, 0.005, 0.01, 0.02, 0.05, and 0.1); outcome model (OM), a Cox proportional hazard model regressing confounders and treatment on the time to event; and a Cox regression model conditioning only on the binary treatment (naive).

For the continuous scenario, we compared ROW with IPW-Cox regression, where propensity scores were estimated by using SuperLearner (IPW) with the following library of algorithms: linear regression with only main terms, linear regression with main terms and interactions, lasso-penalized linear regression, random forest, Bayesian linear regression, local polynomial regression, and extreme gradient boosting regressor; Balance SuperLearner (BalSL) with the same library of algorithms as of IPW; and by using gradient boosted regression (GBM) (with Pearson correlation between covariates and treatment as stopping criteria, interaction depth equal to 4, number of trees equal to 20,000, shrinkage equal to 0.0005 and bag fraction equal to 1). The final IPW weights were obtained assuming Normal conditional density of the treatment as suggested by Robins et al. (2000). In addition, we compared ROW with Covariate Balancing Propensity Score (CBPS) containing only the covariate balancing conditions; non-parametric CBPS (npCBPS); outcome model (OM), a Cox proportional hazard model conditioned on confounders and treatment; and a Cox regression model conditioning only on the continuous treatment (naive).

3.2 Additional simulations

In addition to the simulations presented in section 3.1, we provide additional simulations to evaluate the impact of 1) practical positivity violation, misspecification and censoring on coverage of the 95% confidence interval; 2) sample sizes and 3) number of covariates included in the analysis, on absolute bias, root mean squared error, balance and computational time in seconds. We provide a summary of the results in section 3.3.3 and detailed results in Section 6 of the Supplementary Material.

3.3 Results

3.3.1 Binary treatment

In summary, ROW performed well across all simulation scenarios. Figure 3 shows absolute bias (left panels) and root mean squared error (RMSE) (right panels) across levels of practical positivity violations (top panels), misspecification (middle panels) and censoring (bottom panels) when estimating the hazard ratio of a binary treatment. ROW (black-solid line), EBAL (dark-blue dashed-dotted) and SBW (red dashed-dotted) performed well overall. IPW (blue dashed), BalSL (orange dashed), and GBM (light-blue dashed) performed moderately well across all levels of misspecification and censoring but showed higher bias and RMSE for moderate and strong violation of the positivity assumption. These results suggest that flexible models for the estimation of the propensity scores, may mitigate possible misspecification, but lead to erroneous inferences in the presence of lack of covariate overlap. Similar results were obtain for CBPS (yellow dotted-dashed). OM (purple dotted) outperformed all other methods across levels of positivity violation but performed worse across levels of misspecification. Contrary to previous literature (Austin, 2013), we found that all methods outperformed PSM (green dotted; values are outside figures).

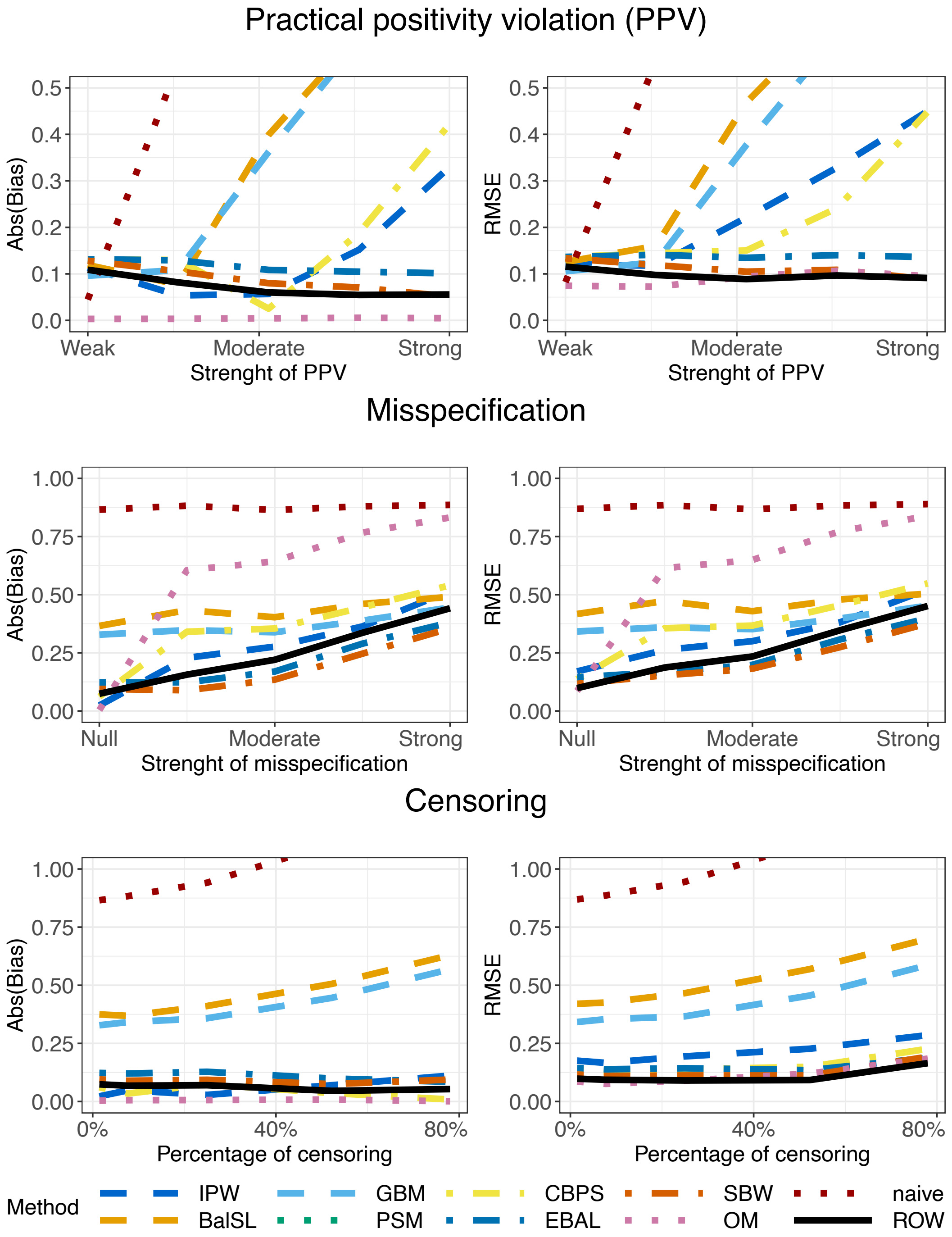

Figure 3: *Binary treatment*: Absolute bias (left panels) and root mean squared error (RMSE) (right panels) across levels of practical positivity violations (top panels), misspecification (middle panels) and censoring(bottom panels) when estimating the hazard ratio of a binary treatment. ROW: Robust Optimal Weights; IPW: Inverse Probability Weighting; GBM: propensity scores were estimated with Gradient Boosting Machine; CBPS: Covariate Balancing Propensity Score; SBW: Stable Balancing Weights; Naive: Cox proportional hazard model including only the treatment; BalSL: Balance SuperLearner; PSM: Propensity Score Matching (values outside figures); EBAL: Entropy Balancing; OM: (outcome model) Cox proportional hazard model including confounders and treatment.

3.3.2 Continuous treatment

ROW performed well across all simulation scenarios, especially across levels of misspecification. Figure 4 shows absolute bias (left panels) and root mean squared error (RMSE) (right panels) across levels of practical positivity violations (top panels), misspecification (middle panels) and censoring (bottom panels) when estimating the hazard ratio of a continuous treatment. As for the binary treatment scenario, IPW (blue dashed), BalSL (orange dashed), and GBM (light-blue dashed) performed well across levels of misspecification and censoring but were outperformed by all other methods across levels of practical positivity violation. CBPS (yellow dotted-dashed) and npCBPS (dark-blue dotted-dashed) performed well across levels of positivity violations and censoring and performed similarly to IPW, BalSL and GBM across levels of misspecification. OM performed worse across levels of misspecification.

3.3.3 Summary of additional simulations' results

By using either the robust or bootstrap standard error, ROW achieved desirable coverage under weak and moderate practical positivity violation and misspecification for both, binary and continuous treatment (Figures 9 and 10 in the Supplementary Material). Under strong practical positivity violation and misspecification, ROW showed undercoverage due to increased bias, regardless of the use of the robust, bootstrap or naive standard error. ROW achieved desirable levels under all censoring levels. Absolute bias and RMSE decreased while increasing the sample size (top panels of Figures 11 and 12 in the Supplementary Material). ROW achieved low balance (below 0.05 standardized absolute mean difference across confounders for the binary treatment scenario and below 0.025 absolute correlation across confounders for the continuous treatment scenario) across all sample sizes

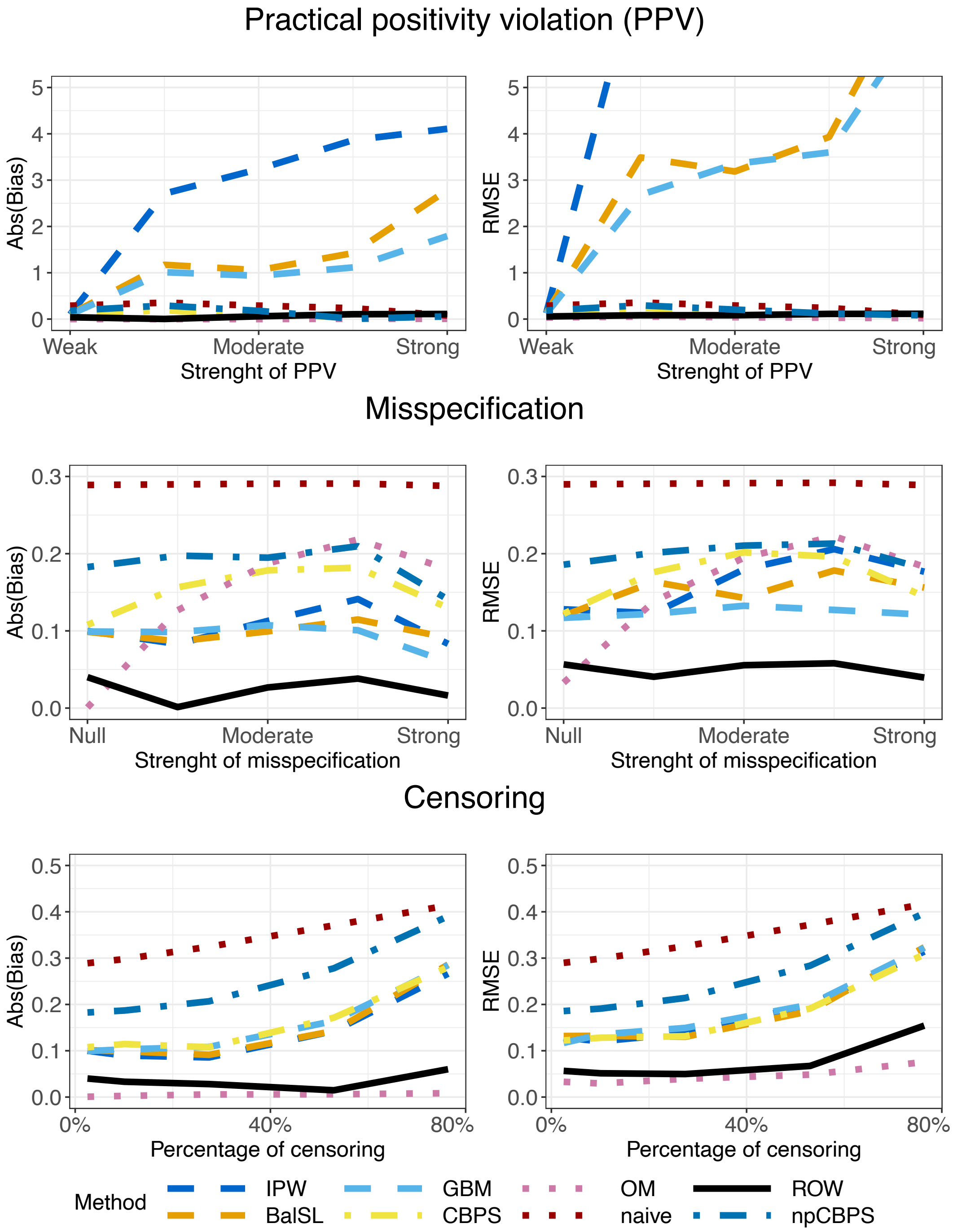

Figure 4: *Continuous treatment*: Absolute bias (left panels) and root mean squared error (RMSE) (right panels) across levels of practical positivity violations (top panels), misspecification (middle panels) and censoring(bottom panels) when estimating the hazard ratio of a binary treatment. ROW: Robust Optimal Weights; IPW: Inverse Probability Weighting (propensity scores were estimated with SuperLearner; GBM: propensity scores were estimated with Gradient Boosting Machine; OM: (outcome model) Cox proportional hazard model including confounders and treatment; BalSL: Balance SuperLearner; CBPS: Covariate Balancing Propensity Score; Naive: Cox proportional hazard model including only the treatment; npCBPS: non-parametric CBPS.

considered. Although balance was kept at low levels, absolute bias and RMSE increased while increasing the number of covariates (top panels and left bottom panels of Figures 13 and 14 in the Supplementary Material). This can be explained by the fact that we generated the data by setting the coefficient for each confounders equal to 1, regardless of the number of confounders considered, thus leading to a strongly confounded model when the number of confounders increased (detailed description of the simulation setting is provided in Section 6.3 of the Supplementary Material). Since we plan to apply ROW using observational data from medical registries, we were especially interested in the computational burden needed to find a solution for larger sample sizes and for an increased number of confounders. We found that for relatively large sample sizes, *e.g.,* $n = 10,000$ the solver could find a solution in a few seconds (bottom right panels of Figures 11 and 12 in the Supplementary Material). In our case-study presented in Section 7 in which we balanced 36 confounders on a population of $n = 24,069$ individuals, the solver found a solution in about 12 seconds.

4 Case studies

In this section, we apply ROW to the evaluation of the effect of hormone therapy on coronary heart disease and the impact of red meat consumption on colon cancer using data from the Women's Health Initiative observational study (Women's Health Initiative, 1998).

4.1 The effect of hormone therapy on coronary heart disease

The Women's Health Initiative (WHI) is a long-term study of postmenopausal women that focuses on best strategies for the prevention and treatment of heart diseases, breast and colon cancers and other chronic diseases. WHI is composed of a randomized clinical trial and an observational study. The WHI trial aimed at evaluating the health benefits and risks of hormone therapy when taken for chronic disease prevention among predominantly healthy postmenopausal women (Writing Group for the WHI Investigators and others, 2002). Specifically, one of the trial's component evaluated the impact of estrogen plus progestin therapy on the risk of coronary heart disease (CHD). Prior to this trials, large observational studies suggested that postmenopausal hormone users had a reduced risk of CHD (Stampfer and Colditz, 1991; Grady et al., 1992; Sidney et al., 1997; Psaty et al., 1994). In contrary, results from the WHI trial suggested an increased risk of CHD (Writing Group for the WHI Investigators and others, 2002). Precisely, in the original trial, the Authors found a statistically significant estimated hazard ratio equal to 1.29 (1.02-1.63).

In this section, we aim at evaluating the effect of estrogen plus progestin therapy on time to CHD among postmenopausal women aged 50-79 years using data from the WHI observational study (September 1993-September 2010). Following Hernán et al. (2008), we first mimic the design of the WHI trial as closely as possible in the WHI observational study. We then apply ROW to control for the non-randomization of the treatment. We also compare the estimated hazard ratio obtained by using ROW, with those obtained by using the methods presented in section 3.1.5.

4.1.1 Study population

We considered the target study population of postmenopausal women who in the WHI observational study had reported no use of estrogen therapy, progesterone therapy or their combination during 2-year prior the enrollment in the study. Baseline was defined as first follow-up visit and women were followed from baseline to diagnosis of CHD, loss to follow-up, death, or September 30, 2010, whichever occurred first. Out of the 93,676 women comprising the original WHI observational study 37,080 used any hormone therapy in the

2-year before the enrollment of the study while 30,960 lacked information about the number of days since enrollment, and 1,567 lacked information on time since menopause. The final study population was comprised of 24,069 women.

We considered the following 34 confounders: multivitamine without minerals use (yes, no), multivitamine with minerals use (yes, no), ethnicity (White, Black, Hispanic, Native American, Asian/Pacific Islander, Unknown), number of pregnancies (7 categories), bilateral oophorectomy (yes, no), age at menopause (numeric), breast cancer ever (yes, no), colon cancer ever (yes, no), endometrial cancer ever (yes, no), skin cancer ever (yes, no), melanoma cancer ever (yes, no), other cancer past 10 years (yes, no), deep vein thrombosis ever (yes, no), stroke ever (yes, no), myocardial infarction ever (yes, no), diabetes ever (yes, no), high cholesterol requiring pills ever (yes, no), osteoporosis ever (yes, no), cardiovascular disease ever (yes, no), coronary artery bypass graft (yes, no), atrial fibrillation ever (yes, no), aortic aneurysm ever (yes, no), angina (yes, no), hip fracture age 55 or older (yes, no), smoked at least 100 cigarettes ever (yes, no), alcohol intake (non drinker, past drinker, less than 1 drink per month, less than 1 drink per week, 1 to 7 drinks per week, 7+ drinks per week), fruits med serv/day (numeric), vegetables med serv/day (numeric), dietary energy (kcal), systolic blood pressure (numeric), diastolic blood pressure (numeric), body mass index (numeric), education (11 categories), income (10 categories).

Time since menopause has been recognized as an important factor for the risks and benefits of hormone therapy on CHD (Carrasquilla et al., 2015, 2017). We therefore evaluated the impact of estrogen plus progestin therapy on time to CHD by conducting a stratified analysis on three categories of time since menopause: 0-10 years, 11-20 years and 20+ years.

To impute missing values of the aforementioned confounders we used multiple imputation by chained equations (Buuren and Groothuis-Oudshoorn, 2010). Hazard ratio estimates were computed using the imputed dataset.

4.1.2 Models setup

We obtained ROW by solving optimization problem (2.1)-(2.4) setting $\delta = 0.0001$ in constraint (2.2) (for each confounder). To obtain the set of ROW we used the `R` interface of `Gurobi`. We compared ROW with IPW in which we estimated the propensity score by using SuperLearner with the following library of algorithms: logistic regression model with only main effects, lasso-penalized logistic regression, and random forest; BalSL with the same library of algorithms as that of IPW; GBM (with mean absolute standardized mean difference as stopping method, interaction depth equal to 3, number of trees equal to 10,000, shrinkage equal to 0.01 and bag fraction equal to 1); PSM with propensity scores estimated as for IPW and BalSL; CBPS containing only the covariate balancing conditions (exact identification); EBAL and SBW with balance tolerance in standard deviation set equal to 0.0001 (with the sample size of our dataset, the tuning algorithm presented in Wang and Zubizarreta (2020) used to choose the degree of approximate covariates balance for SBW considerably increased the computational burden of SBW and therefore was not performed), OM: outcome modelling, a Cox proportional hazard model including confounders and treatment; and Naive: Cox proportional hazard model including only the treatment. To obtain the set of SBW we also used `Gurobi`. Once the sets of weights were obtained, we plugged the weights into a weighted Cox regression regressing the treatment (estrogen plus progestin therapy) on the time to CHD. We computed robust (sandwich) standard errors.

4.1.3 Results

Covariate balance. Figures 5-7 show absolute mean differences (standardized for continuous variables, raw for binary variables) between each of the aforementioned 34 confounders and the binary treatment estrogen plus progestin therapy versus no therapy, before (grey dots) and after (black squares) weighting for ROW. ROW successfully balance all the confounders across the three strata of time since menopause. Table 1 shows the minimum, median and maximum absolute mean difference (standardized for continuous variables and raw for binary variables) between confounders and treatment across 34 confounders and across categories of time since menopause for each of the considered methods. ROW, SBW, EBAL, and CBPS achieved the lowest maximum absolute mean difference across confounders, with ROW obtaining the lowest for strata 0-10 years and 20+ years. Methods that target precision instead of balance, such as IPW with SuperLearner, balanced confounders worse than those methods that target covariate balance, such as ROW, SBW, CBPS or EBAL. SBW could not find a solution when setting the balance tolerance equal to 0.0001 (as described in the previous Section), 0.001, 0.01, and 0.1 in the 0-10 years and 20+ years category. We consequently do not report any results for SBW in these categories. Table 4 in the Supplementary Material, shows the computational time in seconds needed to obtain a solution across strata of time since menopause for each of the considered method. The computational time required by ROW was significantly lower than that needed by IPW, BalSL, GBP, SBW and PSM. Computational time slightly increased with sample size.

Outcome analysis. Table 2 shows hazard ratio estimates and 95% confidence intervals (computed using the robust standard error) for the effect of estrogen plus progestin therapy on time to CHD, and the robust standard error for each method and across categories of time since menopause. Most of the methods lead to similar results. Proportional hazards tests (Grambsch and Therneau, 1994) resulted in p-values greater than 0.05 for all models. Figure 15 in the Supplementary Material shows the Kaplan-Meier curves weighted by ROW for estrogen plus progestin (HT) across categories of time since menopause.

Based on assumptions 2.1-2.3 of Section 2.1, the simulation results presented in Section 3, and on the covariate balance performance of ROW presented in Figures 5-7 and Table 1, we conclude that estrogen plus progestin therapy has no statistically significant effect on time to CHD among $n = 24,069$ postmenopausal women aged 50-79 years enrolled in the WHI observational study (September 1993-September 2010), across three categories of time since menopause, *i.e.,* , $\hat{HR}$ and 95% confidence intervals equal to 1.26 (0.70;2.28), 1.33 (0.87;2.02), and 0.79 (0.45;1.39), for 0-10, 11-20 and 20+ years since menopause, respectively.

4.2 The effect of red meat consumption on colon cancer

Colon cancer is the third most common cause of cancer-related death in the United States. Although epidemiological studies have shown that excess consumption of red meat may be related to colon cancer (Larsson et al., 2005; Larsson and Wolk, 2006), its consumption in the United States has not been decreasing in the past few decades (Zeng et al., 2019).

In this section, using data from the WHI observational study, we aim at evaluating the effect of red meat consumption on time to colon cancer among postmenopausal women aged 50-79 years. Following Song et al. (2004), we defined red meat as the sum of beef, hamburger, lamb or pork as a main dish or a sandwich or mixed dish, and all processed red meat. We defined consumption as medium servings per day of red meat. In addition to the 34 confounders described in Section 4.1, we also consider the time since menopause (numeric) and the use or not of estrogen plus progestin therapy (yes, no) as additional confounders.

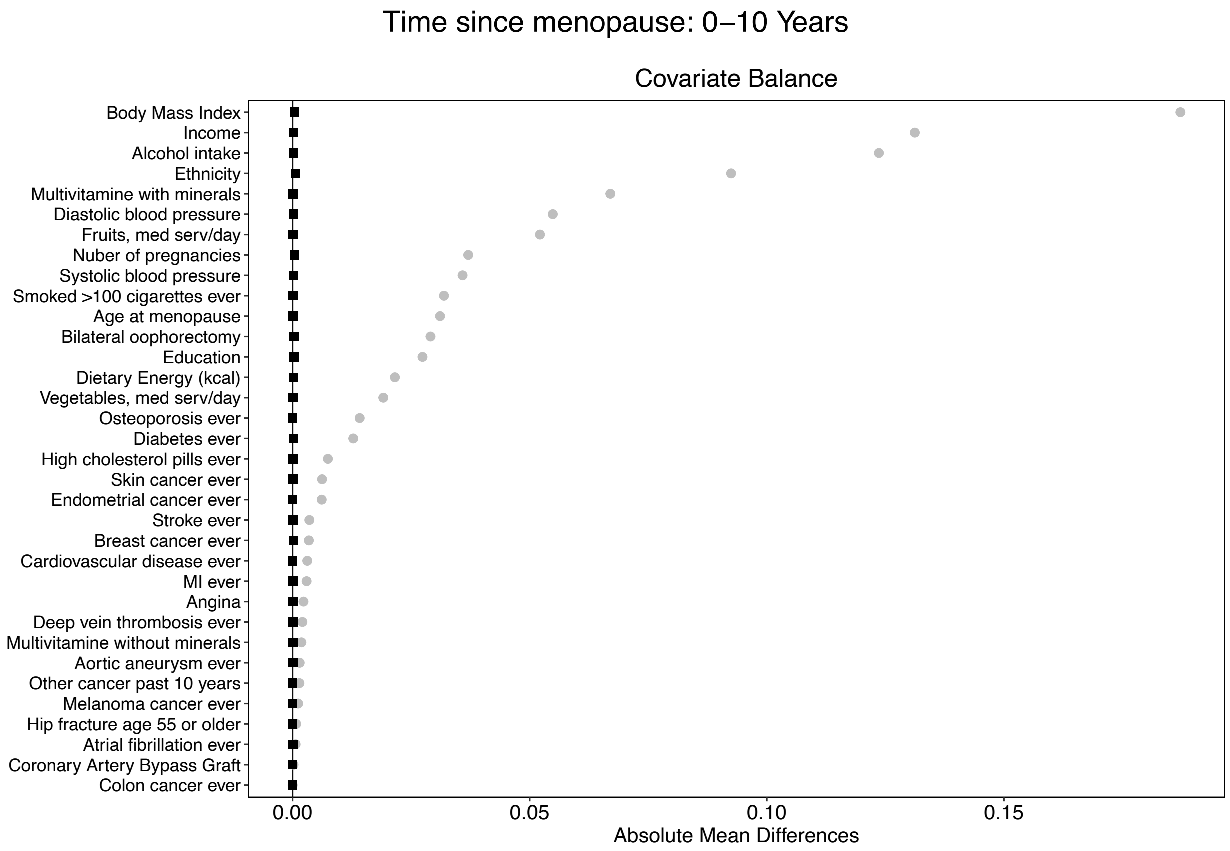

Figure 5: Adjusted (weighted by ROW) (black squares) and unadjusted (grey dots) absolute standardized mean differences between confounders and treatment (estrogen plus progestin).

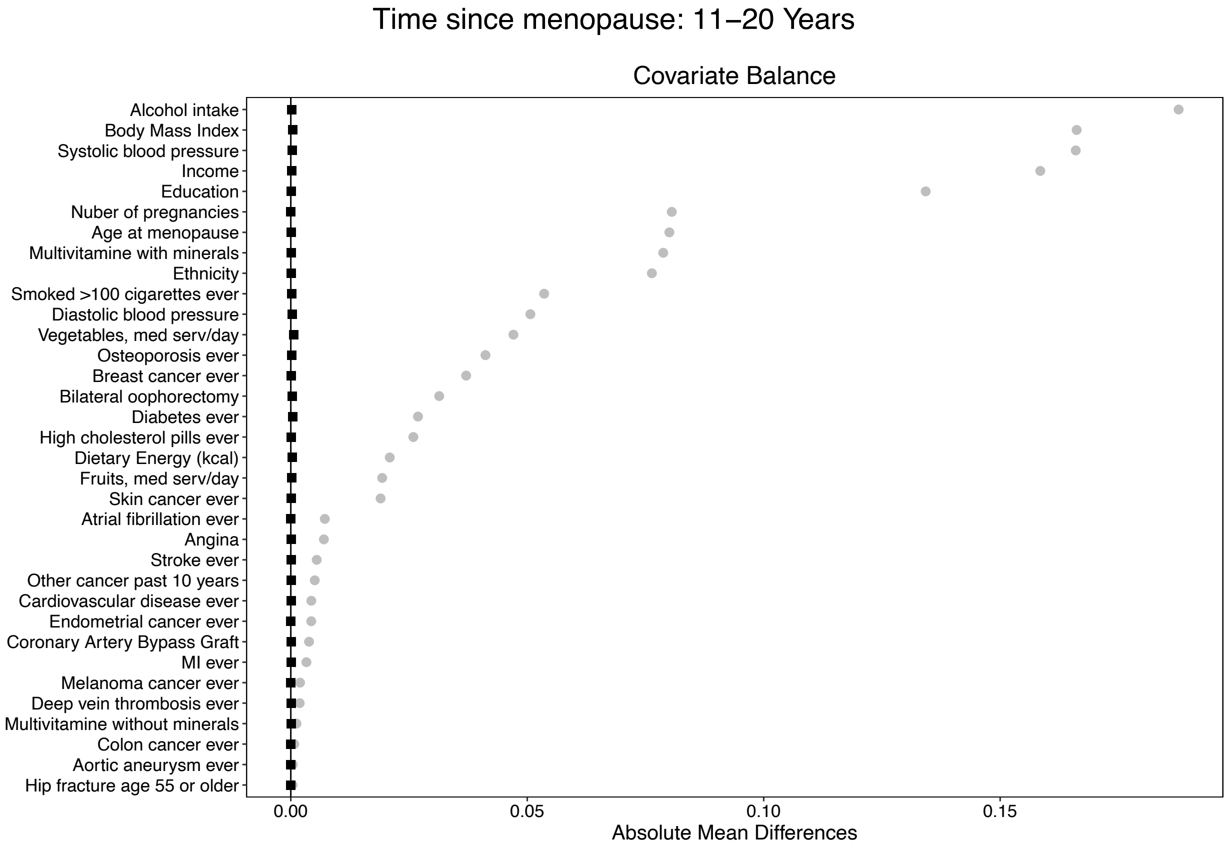

Figure 6: Adjusted (weighted by ROW) (black squares) and unadjusted (grey dots) absolute standardized mean differences between confounders and treatment (estrogen plus progestin).

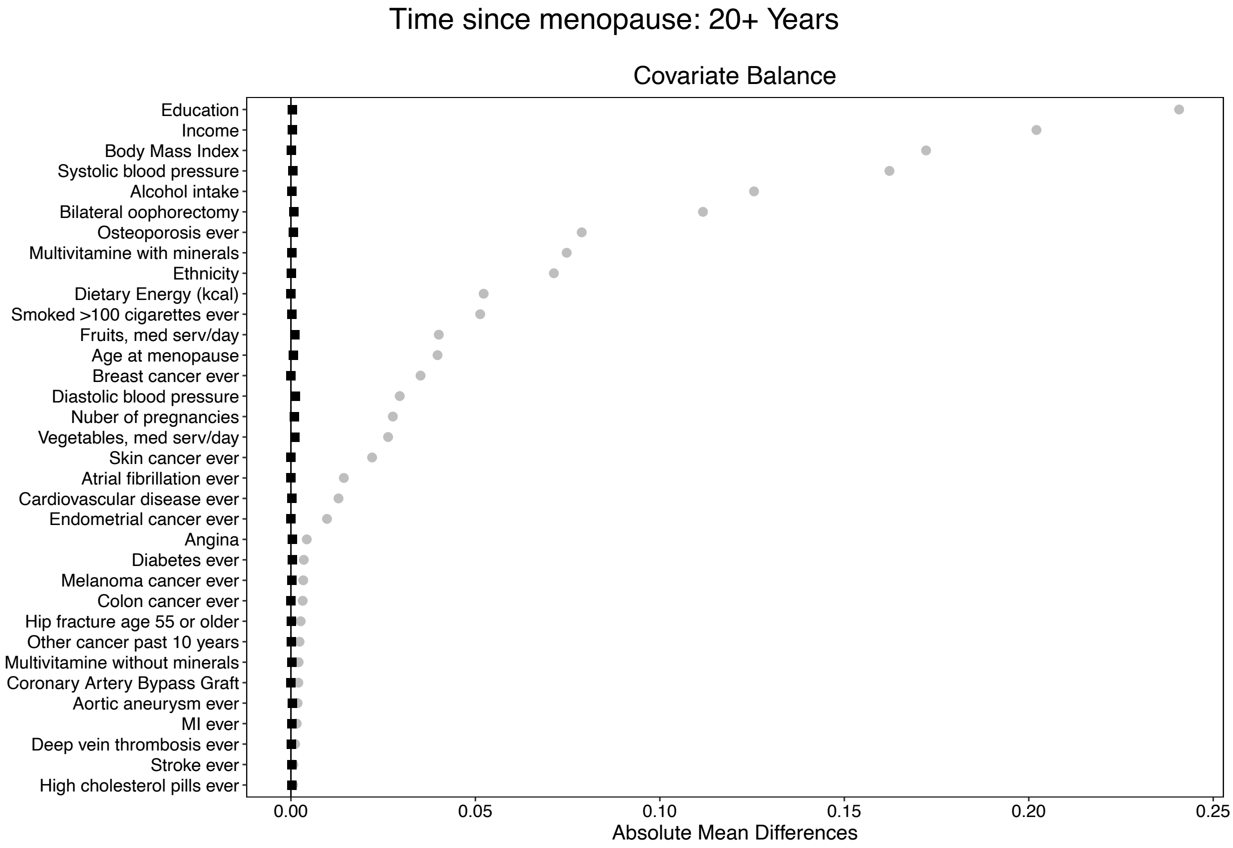

Figure 7: Adjusted (weighted by ROW) (black squares) and unadjusted (grey dots) absolute standardized mean differences between confounders and treatment (estrogen plus progestin).

Table 1: Minimum, median and maximum absolute mean difference (standardized for continuous variables, raw for binary variables) between confounders and treatment across 34 confounders and across categories of time since menopause (0-10; 11-20; 20+ years), WHI observational study, $n = 24,069$, 1993-2010.

	Time since menopause								
	0-10 Years			**11-20 Years**			**20+ Years**		
	$n = 6,661$			$n = 9,592$			$n = 7,816$		
	Abs Mean Diff			**Abs Mean Diff**			**Abs Mean Diff**		
	Min	**Median**	**Max**	**Min**	**Median**	**Max**	**Min**	**Median**	**Max**
Method									
ROW	<0.0001	<0.0001	0.0006	<0.0001	0.0001	0.0007	<0.0001	0.0003	0.0012
IPW	0.0001	0.0071	0.1239	0.0003	0.0124	0.0736	0.0002	0.0111	0.0609
BalSL	0.0002	0.0032	0.057	0.0007	0.0065	0.1195	0.0007	0.0065	0.1195
GBM	<0.0001	0.0035	0.0537	<0.0001	0.0054	0.0704	0.0007	0.0075	0.0689
CBPS	<0.0001	0.0002	0.0022	<0.0001	0.0002	0.0007	<0.0001	0.0004	0.0022
SBW	-	-	-	<0.0001	<0.0001	0.0002	-	-	-
EBAL	<0.0001	<0.0001	0.0012	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001	0.0017
PSM	0.0006	0.0185	0.3357	0.0002	0.0162	0.2411	0.0007	0.0227	0.1601
Naive	<0.0001	0.0101	0.1872	0.0004	0.0234	0.1877	0.0006	0.0242	0.2408

First column: Method implemented, ROW: Robust Optimal Weights; IPW: Inverse Probability Weighting (propensity scores were estimated with SuperLearner with linear regression model with only main effects, and random forest in the library of algorithms); BalSL: Balance SuperLearner; GBM: propensity scores were estimated with Gradient Boosting Machine; CBPS: Covariate Balancing Propensity Score; SBW: Stable Balancing Weights; EBAL: Entropy Balancing; PSM: Propensity Score Matching; OM: (outcome model) Cox proportional hazard model including confounders and treatment; Naive: Cox proportional hazard model including only the treatment. *Second, Fourth and Sixth columns*: Hazard ratio and 95% confidence interval (computed using robust standard error). *Third, Fifth and Seventh columns*: robust standard error.

Table 2: Hazard ratio estimate and 95% confidence intervals of the effect of estrogen plus progestin therapy on time to CHD among postmenopausal women between 50 and 79 across categories of time since menopause (0-10; 11-20; 20+ years), WHI observational study, $n = 24,069$, 1993-2010.

	Time since menopause					
	0-10 Years		**11-20 Years**		**20+ Years**	
	$n = 6,661$		$n = 9,592$		$n = 7,816$	
	$H\hat{R}$ (95% CI)	**SE**	$H\hat{R}$ (95% CI)	**SE**	$H\hat{R}$ (95% CI)	**SE**
Method						
ROW	1.26 (0.70;2.28)	0.301	1.33 (0.87;2.02)	0.216	0.79 (0.45;1.39)	0.285
IPW	1.23 (0.69;2.17)	0.292	1.24 (0.83;1.84)	0.202	0.79 (0.46;1.36)	0.275
BalSL	1.25 (0.70;2.23)	0.295	1.26 (0.85;1.89)	0.204	0.82 (0.48;1.38)	0.269
GBM	1.22 (0.68;2.2)	0.301	1.29 (0.85;1.94)	0.210	0.87 (0.5;1.52)	0.285
CBPS	1.26 (0.70;2.26)	0.300	1.33 (0.88;2.02)	0.213	0.77 (0.44;1.35)	0.283
SBW	-	-	1.31 (0.87;1.99)	0.212	-	-
EBAL	1.25 (0.70;2.26)	0.300	1.33 (0.88;2.02)	0.213	0.77 (0.44;1.34)	0.283
PSM	0.96 (0.19;4.73)	0.817	0.57 (0.14;2.39)	0.731	0.57 (0.14;2.39)	0.731
OM	1.24 (0.66;2.3)	0.318	1.35 (0.91;2.02)	0.203	0.93 (0.55;1.58)	0.271
Naive	1.23 (0.70;2.18)	0.291	1.17 (0.79;1.72)	0.198	0.82 (0.48;1.38)	0.267

First column: Method implemented, ROW: Robust Optimal Weights; IPW: Inverse Probability Weighting (propensity scores were estimated with SuperLearner with linear regression model with only main effects, and random forest in the library of algorithms); BalSL: Balance SuperLearner; GBM: propensity scores were estimated with Gradient Boosting Machine; CBPS: Covariate Balancing Propensity Score; SBW: Stable Balancing Weights; EBAL: Entropy Balancing; PSM: Propensity Score Matching; OM: (outcome model) Cox proportional hazard model including confounders and treatment; Naive: Cox proportional hazard model including only the treatment. *Second, Fifth and Eight columns*: Minimum absolute standardized mean difference between confounders and treatment across 33 confounders. *Third, Sixth and Ninth columns*: Median absolute standardized mean difference between confounders and treatment across 34 confounders. *Fourth, Seventh and Tenth columns*: Maximum absolute standardized mean difference between confounders and treatment across 33 confounders.

4.2.1 Models setup

We obtained ROW by solving optimization problem (2.1)-(2.4) setting $\delta = 0.001$ in constraint (2.2). To obtain the set of ROW we used the `R` interface of `Gurobi`. We compared ROW with IPW in which we estimated the generalized propensity scores by using SuperLearner with the following library of algorithms: linear regression model with only main effects, and random forest; BalSL with the same library of algorithms as for IPW; GBM (with mean Pearson correlation between covariates and treatment as stopping method, interaction depth equal to 4, number of trees equal to 20,000, shrinkage equal to 0.0005 and bag fraction equal to 1); CBPS and npCBPS containing only the covariate balancing conditions, and OM, a Cox proportional hazard model including confounders and red meat consumption. As suggested by Robins et al. (2000), to compute the generalized propensity scores for IPW, BalSL and GBM we assumed the conditional density of the treatment to be Normal. Once the sets of weights were obtained, we plugged the weights into a weighted Cox regression regressing red meat consumption on the time to colon cancer. We computed robust standard errors.

4.2.2 Results

Covariate balance. The first, second and third columns of Table 3 shows the minimum, median and maximum absolute correlation between confounders and red meat consumption across 36 confounders. ROW resulted in the lowest maximum absolute correlation, followed by npCBPS with a maximum absolute correlation more than 20 times higher than that of ROW. In addition, ROW required only 12 seconds to find a solution, compared with much higher computational times needed by the other methods (fourth column of Table 3). Figure 8 shows absolute correlations between the 36 confounders and red meat consumption. As shown in Table 3, ROW resulted in low absolute correlations across all confounders

Outcome analysis. The last two columns of Table 3 show the estimated hazard ratio, its 95% confidence interval and robust standard error of the impact of red meat consumption on time to colon cancer. Proportional hazards tests (Grambsch and Therneau, 1994) resulted in p-values greater than 0.05 for all models. Based on assumptions 2.1-2.3 of Section 2.1, our simulations results presented in Section 3, and the absolute correlations showed in Figure 8, we conclude that red meat consumption is statistically associated with higher risk of colon cancer among $n = 24,069$ postmenopausal women aged 50-79 years enrollend in the WHI observational study (September 1993-September 2010), *i.e.,* , $H\hat{R}$ and 95% confidence intervals equal to 1.68 (1.19;2.37).

5 Conclusions

Unbiased estimation of the effect of binary and continuous treatments using observational data is crucial for medical decision making. In this paper, we proposed a method based on a convex constrained quadratic optimization problem that finds weights with minimal variance, thus controlling for extreme weights, while satisfying constraints on the sample correlation between confounders and treatment, thus targeting covariate balance. ROW performed well across levels of practical positivity violation, misspecification and censoring for both binary and continuous treatments. In addition, in this paper we have shown that methods that target covariate balance, like ROW, CBPS, SBW and EBAL perfom well in terms of absolute bias and root mean squared error compared with propensity score methods, like IPW, regardless of what method is used to estimate the propensity scores.

In addition to hazard ratio, ROW can be used to estimate other survival causal parameter of interest like the survival difference or the mean survival (Mao et al., 2018). Also,

Table 3: Minimum, median and maximum absolute correlation between confounders and red meat consumption across 36 confounders, computational time, hazard ratio estimate, 95% confidence intervals and robust standard error of the effect of red meat consumption on time to colon cancer among postmenopausal women between 50 and 79 - WHI observational study, 1993-2010, $n = 24,069$.

	Absolute Correlation			**Time**		
Method	**Min**	**Median**	**Max**	**sec**	$\hat{HR}$ (95%CI)	SE
ROW	<0.001	0.0011	0.0070	12.0	1.68 (1.19;2.37)	0.17
IPW	<0.001	0.0097	0.2816	366.5	1.26 (1.05;1.5)	0.09
BalSL	<0.001	0.0101	0.2309	367.2	1.28 (1.06;1.54)	0.09
GBM	<0.001	0.0163	0.5924	778.3	1.17 (1.00;1.37)	0.08
CBPS	<0.001	0.0148	0.3489	64.5	1.23 (0.64;2.36)	0.33
npCBPS	<0.001	0.0094	0.1523	9724.1	1.22 (0.98;1.52)	0.11
OM	-	-	-	0.2	1.28 (1.00;1.63)	0.13
Naive	<0.001	0.0163	0.5915	0.1	1.17 (0.99;1.39)	0.08

First column: Method implemented, ROW: Robust Optimal Weights; IPW: Inverse Probability Weighting (propensity scores were estimated with SuperLearner with linear regression model with only main effects, and random forest in the library of algorithms); BalSL: Balance SuperLearner; CBPS: Covariate Balancing Propensity Score; npCBPS: non-parametric CBPS; OM: Cox proportional hazard model - outcome model. *Second to fourth columns*: minimum, median and maximum absolute correlation across confounders. *Fifth column*: Time in seconds. *Sixth column*: Hazard ratio estimate and 95% confidence intervals (computed with robust standard errors). *Seventh column*: robust standard error.

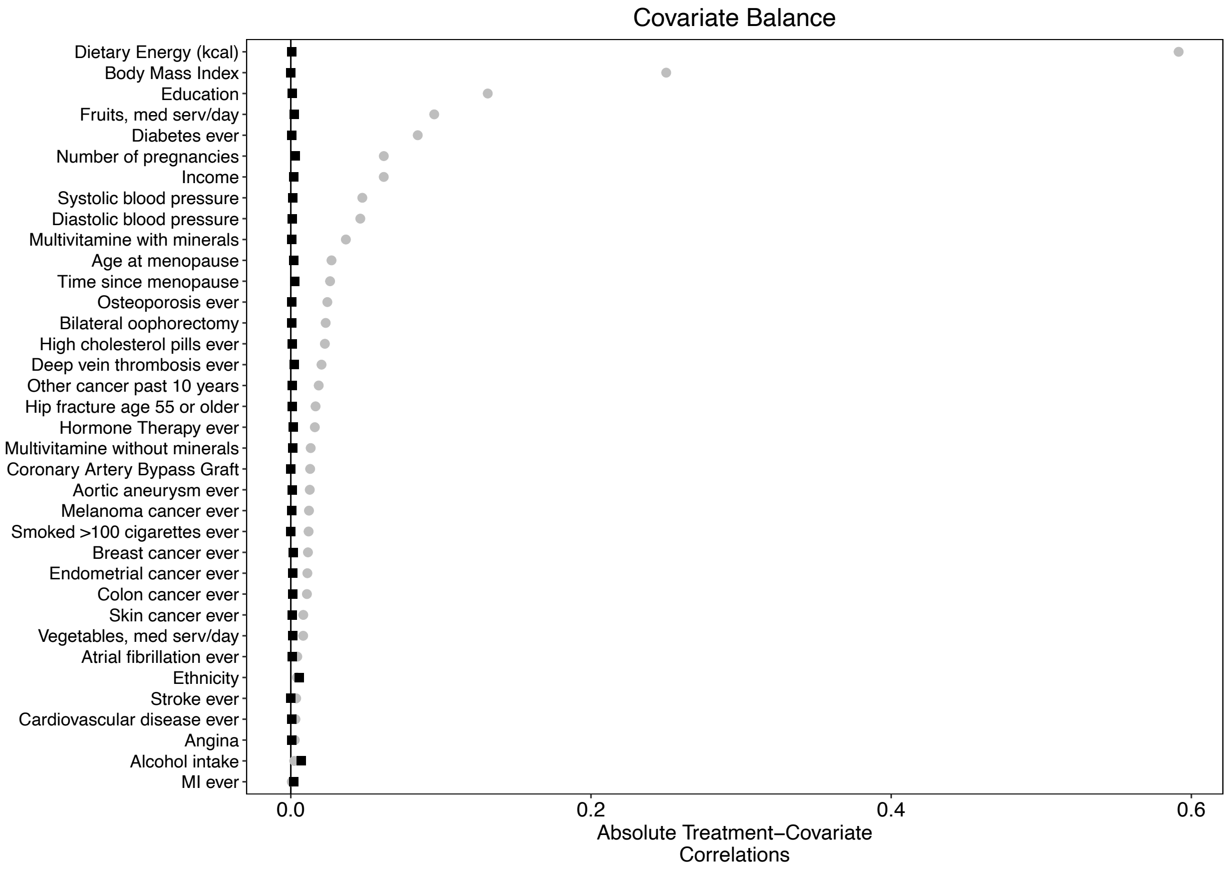

Figure 8: Adjusted (weighted by ROW) (black squares) and unadjusted (grey dots) absolute correlation between confounders and treatment (red meat consumption).

ROW can be used to estimate effects of binary and continuous treatments on binary outcomes, (marginal odds ratio), and continuous outcomes (average and quantile treatment effects). ROW can also be use to estimate effects of multi-value treatments. To do so, if the interest is to make inference on each of the treatment's contrasts, we suggest to run pair-wise comparisons by applying ROW as if it was a binary treatment. Alternatively, one can consider the multi-value treatment as a continuous treatment (as suggested by Fong et al. (2018)). Future research is needed to evaluate ROW with multi-value treatments.

As presented in our simulation and in our case studies, we suggest to considered categorical variables (such as education for instance) as numeric, and balance them accordingly. In addition, in our simulations and case studies we only considered balancing linear covariates. Quadratic or higher order and interaction terms can be balanced by adding them into constraint (2.2). If no solution to optimization problem (2.2)-(2.4) exists, we suggest to increase the parameter δ of (2.2) and re-run the solver. We also suggest to evaluate covariate balance, defined as the absolute standardized mean difference for a binary outcome and as the absolute correlation for a continuous outcome for each new set of ROW.

References

Austin, P. C. (2013). The performance of different propensity score methods for estimating marginal hazard ratios. *Statistics in medicine 32*(16), 2837–2849.

Bender, R., T. Augustin, and M. Blettner (2005). Generating survival times to simulate cox proportional hazards models. *Statistics in medicine 24*(11), 1713–1723.

Buchanan, A. L., M. G. Hudgens, S. R. Cole, B. Lau, A. A. Adimora, and W. I. H. Study (2014). Worth the weight: using inverse probability weighted cox models in aids research. *AIDS research and human retroviruses 30*(12), 1170–1177.

Buuren, S. v. and K. Groothuis-Oudshoorn (2010). mice: Multivariate imputation by chained equations in r. *Journal of statistical software*, 1–68.

Carrasquilla, G. D., C. Chiavenna, M. Bottai, P. K. Magnusson, M. Santacatterina, A. Wolk, G. Hallmans, J.-H. Jansson, G. Engstrom, C. Borgfeldt, et al. (2015). The association between menopausal hormone therapy and coronary heart disease depends on timing of initiation in relation to menopause onset. results based on pooled individual participant data from the combined cohorts of menopausal women-studies of register based health outcomes in relation to hormonal drugs (comprehend) study. In *Menopause: The Journal of the North American Menopause*, Volume 22, pp. 1373–1373.

Carrasquilla, G. D., P. Frumento, A. Berglund, C. Borgfeldt, M. Bottai, C. Chiavenna, M. Eliasson, G. Engström, G. Hallmans, J.-H. Jansson, et al. (2017). Postmenopausal hormone therapy and risk of stroke: A pooled analysis of data from population-based cohort studies. *PLoS medicine 14*(11), e1002445.

Davison, A. C. and D. V. Hinkley (1997). *Bootstrap methods and their application.* Number 1. Cambridge university press.

Fong, C., C. Hazlett, K. Imai, et al. (2018). Covariate balancing propensity score for a continuous treatment: Application to the efficacy of political advertisements. *The Annals of Applied Statistics 12*(1), 156–177.

Freedman, D. A. (2006). On the so-called "huber sandwich estimator" and "robust standard errors". *The American Statistician 60*(4), 299–302.

Grady, D., S. M. Rubin, D. B. Petitti, C. S. Fox, D. Black, B. Ettinger, V. L. Ernster, and S. R. Cummings (1992). Hormone therapy to prevent disease and prolong life in postmenopausal women. *Annals of internal medicine 117*(12), 1016–1037.

Grambsch, P. M. and T. M. Therneau (1994). Proportional hazards tests and diagnostics based on weighted residuals. *Biometrika 81*(3), 515–526.

Gurobi Optimization, L. (2020). Gurobi optimizer reference manual.

Hainmueller, J. (2012). Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies. *Political analysis*, 25–46.

Hernán, M. A., A. Alonso, R. Logan, F. Grodstein, K. B. Michels, M. J. Stampfer, W. C. Willett, J. E. Manson, and J. M. Robins (2008). Observational studies analyzed like randomized experiments: an application to postmenopausal hormone therapy and coronary heart disease. *Epidemiology (Cambridge, Mass.) 19*(6), 766.

Hernán, M. A., B. Brumback, and J. M. Robins (2001). Marginal structural models to estimate the joint causal effect of nonrandomized treatments. *Journal of the American Statistical Association 96*(454), 440–448.

Hirano, K. and G. W. Imbens (2004). The propensity score with continuous treatments. *Applied Bayesian modeling and causal inference from incomplete-data perspectives 226164*, 73–84.

Hirshberg, D. A., A. Maleki, and J. Zubizarreta (2019). Minimax linear estimation of the retargeted mean. *arXiv preprint arXiv:1901.10296*.

Hirshberg, D. A. and S. Wager (2017). Augmented minimax linear estimation. *arXiv preprint arXiv:1712.00038*.

Hulley, S., D. Grady, T. Bush, C. Furberg, D. Herrington, B. Riggs, E. Vittinghoff, Heart, E. R. S. H. R. Group, et al. (1998). Randomized trial of estrogen plus progestin for secondary prevention of coronary heart disease in postmenopausal women. *Jama 280*(7), 605–613.

Imai, K. and M. Ratkovic (2014). Covariate balancing propensity score. *Journal of the Royal Statistical Society: Series B: Statistical Methodology*, 243–263.

Imbens, G. W. and D. B. Rubin (2015). *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press.

Kallus, N. and M. Santacatterina (2018, Jun). Optimal Balancing of Time-Dependent Confounders for Marginal Structural Models. *arXiv e-prints*, arXiv:1806.01083.

Kallus, N. and M. Santacatterina (2019a). Kernel optimal orthogonality weighting: A balancing approach to estimating effects of continuous treatments. *arXiv preprint arXiv:1910.11972*.

Kallus, N. and M. Santacatterina (2019b). Optimal estimation of generalized average treatment effects using kernel optimal matching. *arXiv preprint arXiv:1908.04748*.

Kang, J. D. and J. L. Schafer (2007). Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical Science 22*(4), 523–539.

King, G., C. Lucas, and R. A. Nielsen (2017). The balance-sample size frontier in matching methods for causal inference. *American Journal of Political Science 61*(2), 473–489.

Larsson, S. C., J. Rafter, L. Holmberg, L. Bergkvist, and A. Wolk (2005). Red meat consumption and risk of cancers of the proximal colon, distal colon and rectum: the swedish mammography cohort. *International journal of cancer 113*(5), 829–834.

Larsson, S. C. and A. Wolk (2006). Meat consumption and risk of colorectal cancer: a meta-analysis of prospective studies. *International journal of cancer 119*(11), 2657–2664.

Lee, B. K., J. Lessler, and E. A. Stuart (2010). Improving propensity score weighting using machine learning. *Statistics in medicine 29*(3), 337–346.

Li, F., K. L. Morgan, and A. M. Zaslavsky (2018). Balancing covariates via propensity score weighting. *Journal of the American Statistical Association 113*(521), 390–400.

Manson, J. E., J. Hsia, K. C. Johnson, J. E. Rossouw, A. R. Assaf, N. L. Lasser, M. Trevisan, H. R. Black, S. R. Heckbert, R. Detrano, et al. (2003). Estrogen plus progestin and the risk of coronary heart disease. *New England Journal of Medicine 349*(6), 523–534.

Mao, H., L. Li, W. Yang, and Y. Shen (2018). On the propensity score weighting analysis with survival outcome: Estimands, estimation, and inference. *Statistics in medicine 37*(26), 3745–3763.

McCaffrey, D. F., G. Ridgeway, and A. R. Morral (2004). Propensity score estimation with boosted regression for evaluating causal effects in observational studies. *Psychological methods 9*(4), 403.

Petersen, M. L., K. E. Porter, S. Gruber, Y. Wang, and M. J. van der Laan (2012). Diagnosing and responding to violations in the positivity assumption. *Statistical methods in medical research 21*(1), 31–54.

Pirracchio, R. and M. Carone (2018). The balance super learner: A robust adaptation of the super learner to improve estimation of the average treatment effect in the treated based on propensity score matching. *Statistical methods in medical research 27*(8), 2504–2518.

Pirracchio, R., M. L. Petersen, and M. Van Der Laan (2015). Improving propensity score estimators' robustness to model misspecification using super learner. *American journal of epidemiology 181*(2), 108–119.

Psaty, B. M., S. R. Heckbert, D. Atkins, R. Lemaitre, T. D. Koepsell, P. W. Wahl, D. S. Siscovick, and E. H. Wagner (1994). The risk of myocardial infarction associated with the combined use of estrogens and progestins in postmenopausal women. *Archives of Internal Medicine 154*(12), 1333–1339.

Robins, J., M. Sued, Q. Lei-Gomez, and A. Rotnitzky (2007). Comment: Performance of double-robust estimators when" inverse probability" weights are highly variable. *Statistical Science 22*(4), 544–559.

Robins, J. M., M. A. Hernan, and B. Brumback (2000). Marginal structural models and causal inference in epidemiology.

Robins, J. M., A. Rotnitzky, and L. P. Zhao (1995). Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the american statistical association 90*(429), 106–121.

Rosenbaum, P. R. and D. B. Rubin (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika 70*(1), 41–55.

Santacatterina, M. and M. Bottai (2018). Optimal probability weights for inference with constrained precision. *Journal of the American Statistical Association 113*(523), 983–991.

Santacatterina, M., C. García-Pareja, R. Bellocco, A. Sönnerborg, A. M. Ekström, and M. Bottai (2019). Optimal probability weights for estimating causal effects of time-varying treatments with marginal structural cox models. *Statistics in medicine 38*(10), 1891–1902.

Scharfstein, D. O., A. Rotnitzky, and J. M. Robins (1999). Adjusting for nonignorable drop-out using semiparametric nonresponse models. *Journal of the American Statistical Association 94*(448), 1096–1120.

Sekhon, J. S. (2008). Multivariate and propensity score matching software with automated balance optimization: the matching package for r. *Journal of Statistical Software, Forthcoming*.

Sidney, S., D. B. Petitti, and C. P. Quesenberry Jr (1997). Myocardial infarction and the use of estrogen and estrogen-progestogen in postmenopausal women. *Annals of internal medicine 127*(7), 501–508.

Song, Y., J. E. Manson, J. E. Buring, and S. Liu (2004). A prospective study of red meat consumption and type 2 diabetes in middle-aged and elderly women: the women's health study. *Diabetes care 27*(9), 2108–2115.

Stampfer, M. J. and G. A. Colditz (1991). Estrogen replacement therapy and coronary heart disease: a quantitative assessment of the epidemiologic evidence. *Preventive medicine 20*(1), 47–63.

Tübbicke, S. (2020). Entropy balancing for continuous treatments. *arXiv preprint arXiv:2001.06281*.

Van der Laan, M. J., E. C. Polley, and A. E. Hubbard (2007). Super learner. *Statistical applications in genetics and molecular biology 6*(1).

Vegetabile, B. G., B. A. Griffin, D. L. Coffman, M. Cefalu, and D. F. McCaffrey (2020). Nonparametric estimation of population average dose-response curves using entropy balancing weights for continuous exposures. *arXiv preprint arXiv:2003.02938*.

Visconti, G. and J. R. Zubizarreta (2018). Handling limited overlap in observational studies with cardinality matching. *Observational Studies 4*, 217–249.

Wang, Y. and J. R. Zubizarreta (2020). Minimal dispersion approximately balancing weights: asymptotic properties and practical considerations. *Biometrika 107*(1), 93–105.

Women's Health Initiative (1998). Design of the women's health initiative clinical trial and observational study. *Controlled clinical trials 19*(1), 61–109.

Wong, R. K. and K. C. G. Chan (2017). Kernel-based covariate functional balancing for observational studies. *Biometrika 105*(1), 199–213.

Writing Group for the WHI Investigators and others (2002). Risks and benefits of estrogen plus progestin in healthy postmenopausal women: principal results from the women's health initiative randomized controlled trial. *Jama 288*(3), 321–333.

Wu, X., F. Mealli, M.-A. Kioumourtzoglou, F. Dominici, and D. Braun (2018). Matching on generalized propensity scores with continuous exposures. *arXiv preprint arXiv:1812.06575*.

Yiu, S. and L. Su (2018). Covariate association eliminating weights: a unified weighting framework for causal effect estimation. *Biometrika 105*(3), 709–722.

Zeng, L., M. Ruan, J. Liu, P. Wilde, E. N. Naumova, D. Mozaffarian, and F. F. Zhang (2019). Trends in processed meat, unprocessed red meat, poultry, and fish consumption in the united states, 1999-2016. *Journal of the Academy of Nutrition and Dietetics 119*(7), 1085–1098.

Zhao, Q. et al. (2019). Covariate balancing propensity score by tailored loss functions. *The Annals of Statistics 47*(2), 965–993.

Zhao, Q. and D. Percival (2017). Entropy balancing is doubly robust. *Journal of Causal Inference 5*(1).

Zubizarreta, J. R. (2015). Stable weights that balance covariates for estimation with incomplete outcome data. *Journal of the American Statistical Association 110*(511), 910–922.

Zubizarreta, J. R., R. D. Paredes, P. R. Rosenbaum, et al. (2014). Matching for balance, pairing for heterogeneity in an observational study of the effectiveness of for-profit and not-for-profit high schools in chile. *The Annals of Applied Statistics 8*(1), 204–231.

SUPPLEMENTARY MATERIAL

6 Simulations

In this section, we provide additional simulations' results evaluating 1) the naive (the inverse of the observed Fisher information of the coefficient), robust and bootstrap standard error estimator 6.1; 2) the impact of practical positivity violation, misspecification and censoring on coverage of 95% confidence intervals 6.1; 3) the impact of sample sizes 6.2 and 4) the impact of the number of covariates included in the analysis 6.3, on absolute bias, root mean squared error, balance and computational time in seconds.

6.1 Standard error and coverage

We considered the same simulation scenario as that described in Section 3 of the original manuscript. We used non-parametric bootstrap with normal approximation confidence intervals (Davison and Hinkley, 1997). Left panels of Figures 9 and 10 show the ratios between the standard deviation of the estimated hazard ratio across simulations and the bootstrap, robust and naive standard errors for the binary treatment (Figure 9) and for the continuous treatment (Figure 10) scenarios. The Naive estimator resulted in higher standard error compared with the empirical standard deviation (lower values of the ratio) and consequential overcoverage across all levels of practical positivity violation, misspecification and censoring for both binary and continuous treatments. Robust and bootstrap standard errors behaved similarly across most levels of practical positivity violation, misspecification and censoring and type of treatments. Under strong misspecification in the binary treatment scenario (middle panel of Figure 9), ROW using the robust standard error estimator substantially undercovered the 95% confidence interval, while it overcovered it while using bootstrap and the naive standard error. Overall, robust and bootstrap standard errors resulted in slightly higher standard errors compared with the standard deviation of the estimated hazard ratio, and consequently resulted in higher coverage. As expected, when standard errors are close to the standard deviation of the estimated hazard ratio and ROW has bias, ROW exhibits undercoverage (top right panel of Figure 10).

6.2 Sample size

We considered sample sizes equal to $50, 75, 100, 250, 500, 5,000$, and $10,000$. We computed the expected survival time t as described in Section 3.1 of the original manuscript, where $X_{1:4} \sim \text{MultiN}(\mu, \Sigma)$, $\mu = (0, 0, 0, 0)$ and $\Sigma = I_4$, where I_4 is the identity matrix. We considered $\beta = 1.4$ (the coefficients for the confounders in the outcome model - Section 3.1.5) and $\gamma = 1.5$ (the coefficients of the confounders in the treatment model - Section 3.1.1) for all four confounders. To reflect only the impact of sample size, we considered a moderate practical positivity violation scenario ($\gamma = 1$ and $\nu = 0.6$ as in Section 3.1.2), a null misspecification ($\tau = 0$ as in Section 3.1.5) and a low percentage of censored observation ($\epsilon = 0$ as in Section 3.1.4). Figure 11 shows the absolute bias (top lef panel), the root mean squared error (RMSE) (top right panel), absolute standardized mean difference across the four confounders (Balance) (left bottom panel) and the average computational time in seconds to obtain a solution (right bottom panel) across levels of sample sizes for the binary treatment scenario. Figure 11 shows those for the continuous treatment. For both treatments, absolute bias and RMSE decreased with increasing sample sizes. Balance was kept low across all sample sizes. The computational time in seconds increased up to 1.5 seconds for $n = 10,000$.

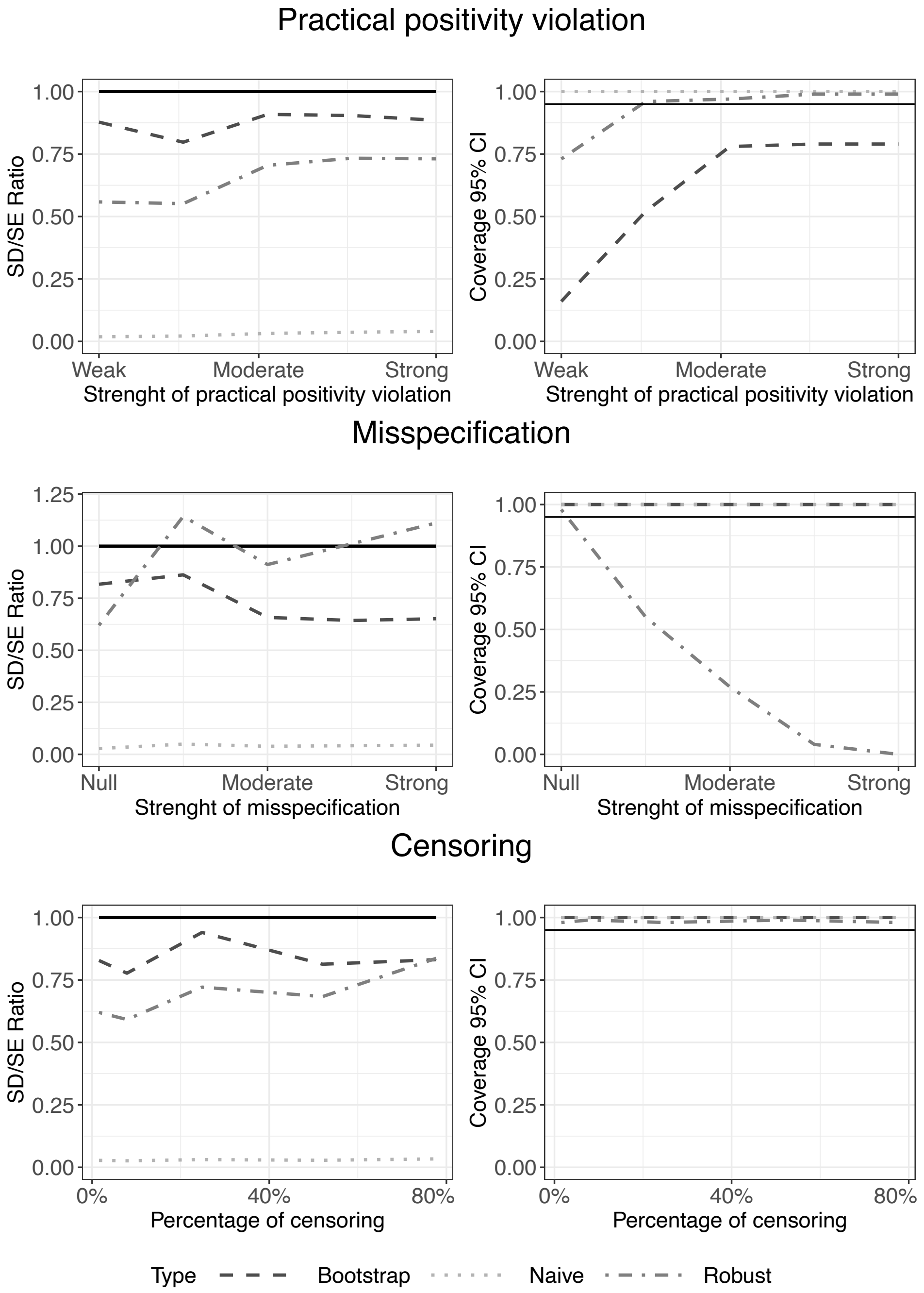

Figure 9: *Left panels*: ratios between the standard deviation of the estimated hazard ratio across simulations and the bootstrap (dashed), naive (the inverse of the observed Fisher information of the coefficient)(dotted) and robust (dotted-dashed) standard errors across levels of practical positivity violation (top panels), misspecification (middle panels) and censoring (bottom panels) for the binary treatment. *Right panels*: coverage of the 95% confidence interval using the bootstrap (with normal approximation confidence intervals)(dashed), naive (dotted) and robust (dotted-dashed) standard errors across levels of practical positivity violation (top panels), misspecification (middle panels) and censoring (bottom panels) for the binary treatment.

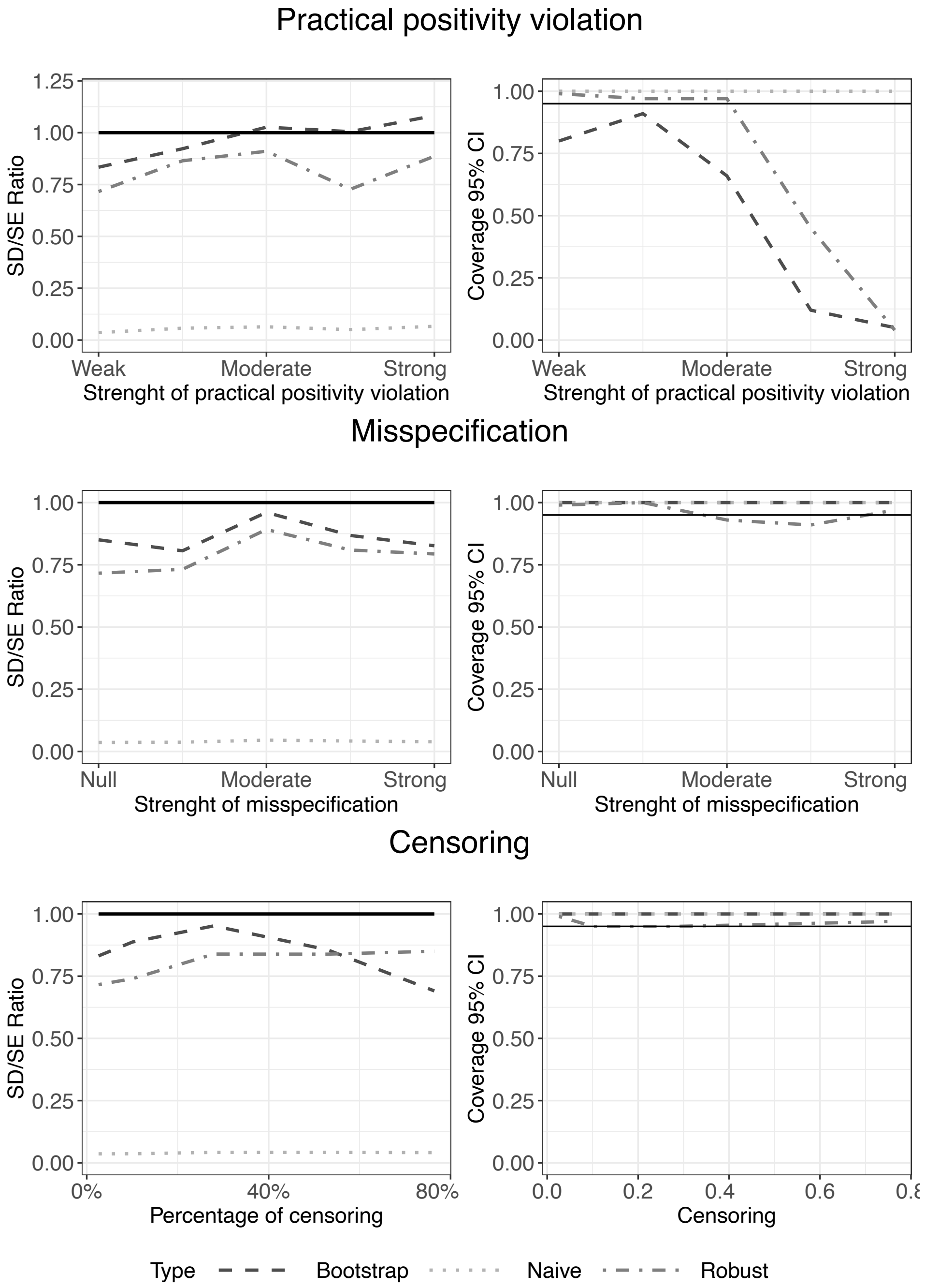

Figure 10: *Left panels*: ratios between the standard deviation of the estimated hazard ratio across simulations and the bootstrap (dashed), naive (the inverse of the observed Fisher information of the coefficient)(dotted) and robust (dotted-dashed) standard errors across levels of practical positivity violation (top panels), misspecification (middle panels) and censoring (bottom panels) for the continuous treatment. *Right panels*: coverage of the 95% confidence interval using the bootstrap (with normal approximation confidence intervals)(dashed), naive (dotted) and robust (dotted-dashed) standard errors across levels of practical positivity violation (top panels), misspecification (middle panels) and censoring (bottom panels) for the continuous treatment.

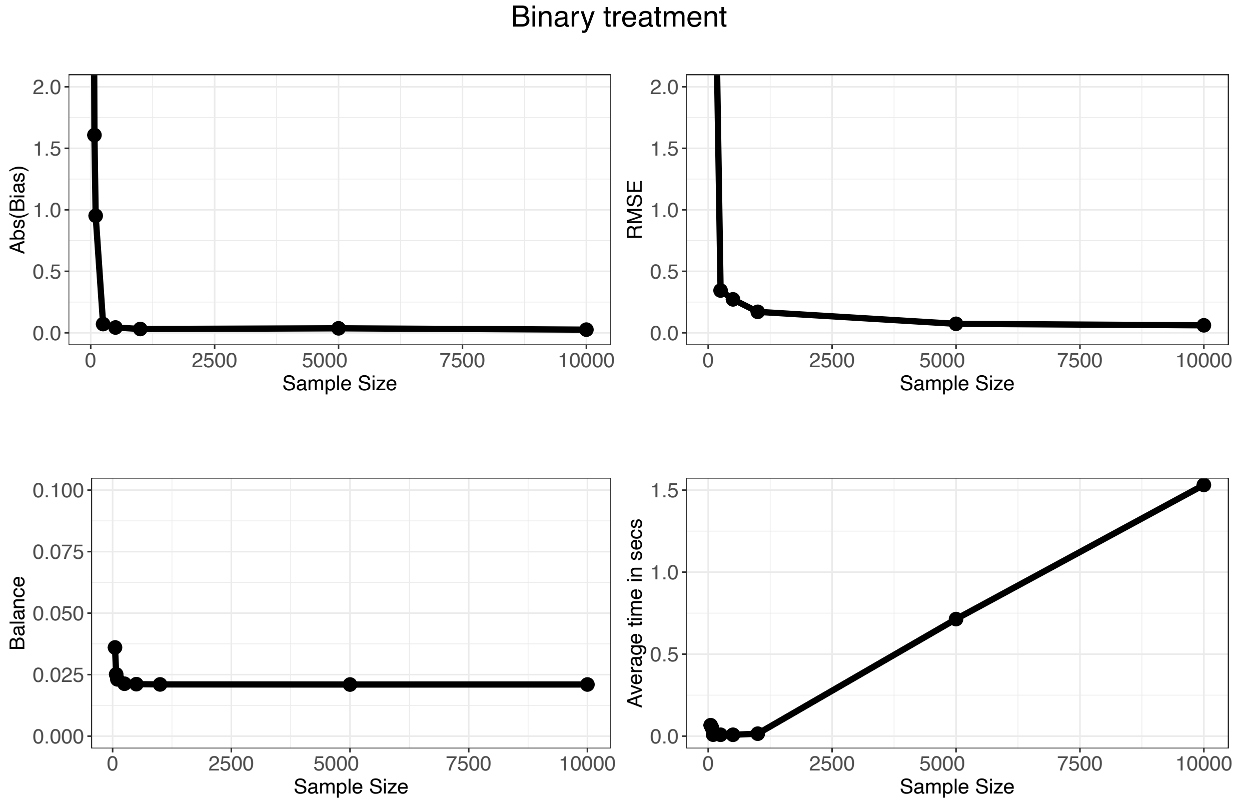

Figure 11: Absolute bias (top left panel), root mean squared error (top right panel), absolute standardized mean difference across four confounders (left bottom panel) and the average computational time in seconds to obtain a solution (right bottom panel) across levels of sample sizes for the binary treatment scenario.

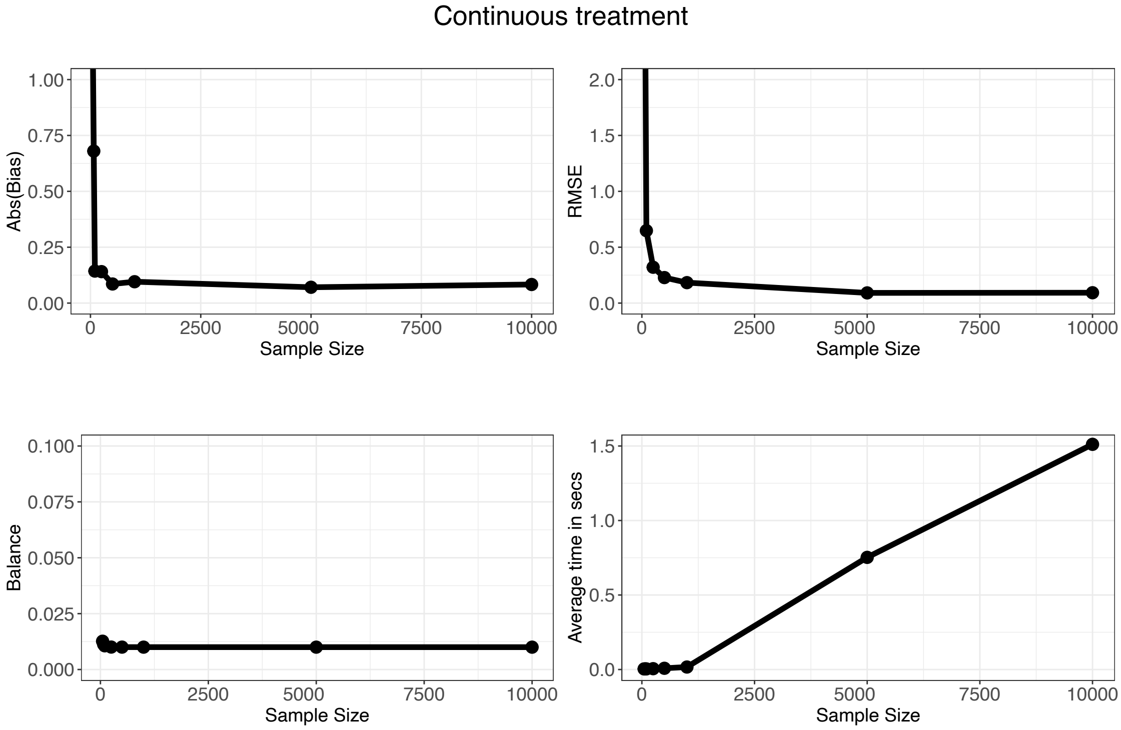

Figure 12: Absolute bias (top left panel), root mean squared error (top right panel), absolute standardized mean difference across four confounders (left bottom panel) and the average computational time in seconds to obtain a solution (right bottom panel) across levels of sample sizes for the continuous treatment scenario.

6.3 Number of confounders

We considered the following number of confounders $1, 5, 10, 20, 50$, and 100 and $n = 1,000$. We computed the expected survival time t as described in Section 3.1 of the original manuscript, where $X_{1:K} \sim \text{MultiN}(\mu, \Sigma)$, $\mu = \mathbf{0}$ and $\Sigma = I_K$, where, $\mathbf{0}$ and I_K are the vector of zero of dimension K, and the identity matrix of dimension $K \times K$, where K is total number of confounders. We considered $\beta = 0.1$ (the coefficients for the confounders in the outcome model - Section 3.1.5) and $\gamma = 0.1$ (the coefficients of the confounders in the treatment model - Section 3.1.1) for all confounders. As for the evaluation of the impact of sample sizes, we considered a moderate practical positivity violation scenario ($\gamma = 1$ and $\nu = 0.6$ as in Section 3.1.2), a null misspecification ($\tau = 0$ as in Section 3.1.5) and a low percentage of censored observation ($\epsilon = 0$ as in Section 3.1.4). Figure 13 shows the absolute bias (top lef panel), the root mean squared error (RMSE) (top right panel), absolute standardized mean difference across the four confounders (Balance) (left bottom panel) and the average computational time in seconds to obtain a solution (right bottom panel) across numbers of confounders for the binary treatment scenario. Figure 14 shows those for the continuous treatment. For both treatments, absolute bias and RMSE increased with increasing number of confounders. Balance was kept low across all levels. The computational time in seconds slightly increased from below 0.05 seconds to 0.20 seconds (100 confounders).

7 Case studies

In this Section, we provide additional results of the case studies presented in Section 4 in the original manuscript.

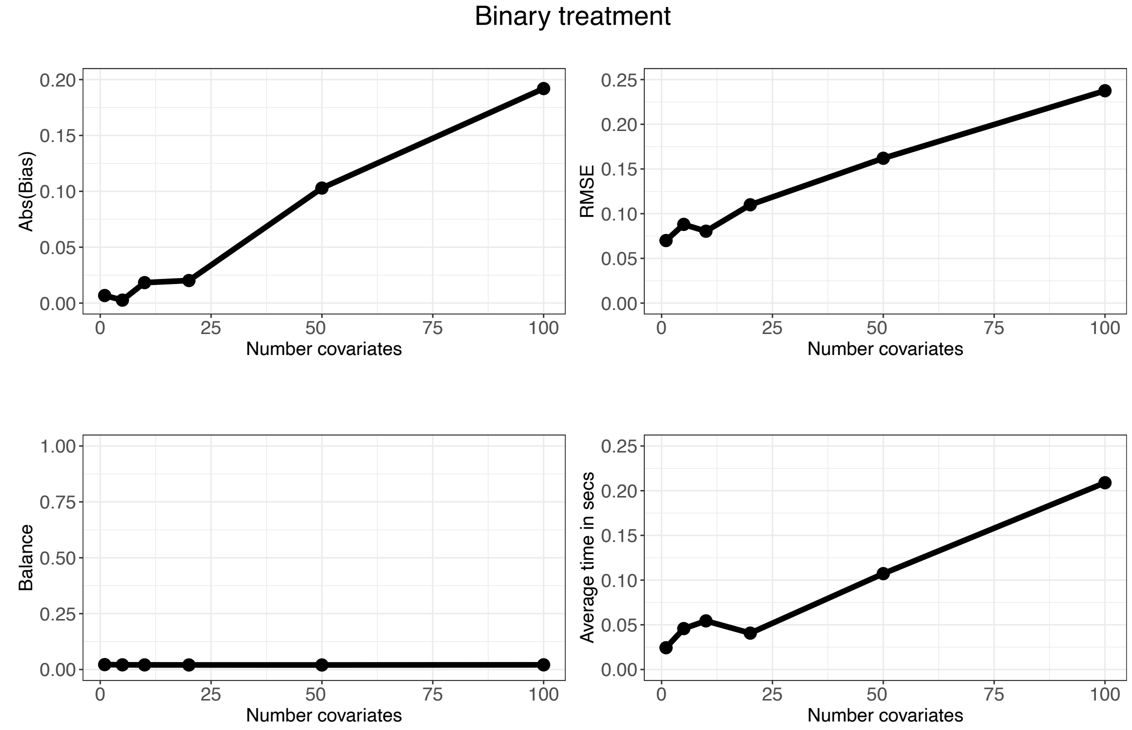

Figure 13: Absolute bias (top left panel), root mean squared error (top right panel), absolute standardized mean difference across four confounders (left bottom panel) and the average computational time in seconds to obtain a solution (right bottom panel) across number of confounders for the binary treatment scenario.

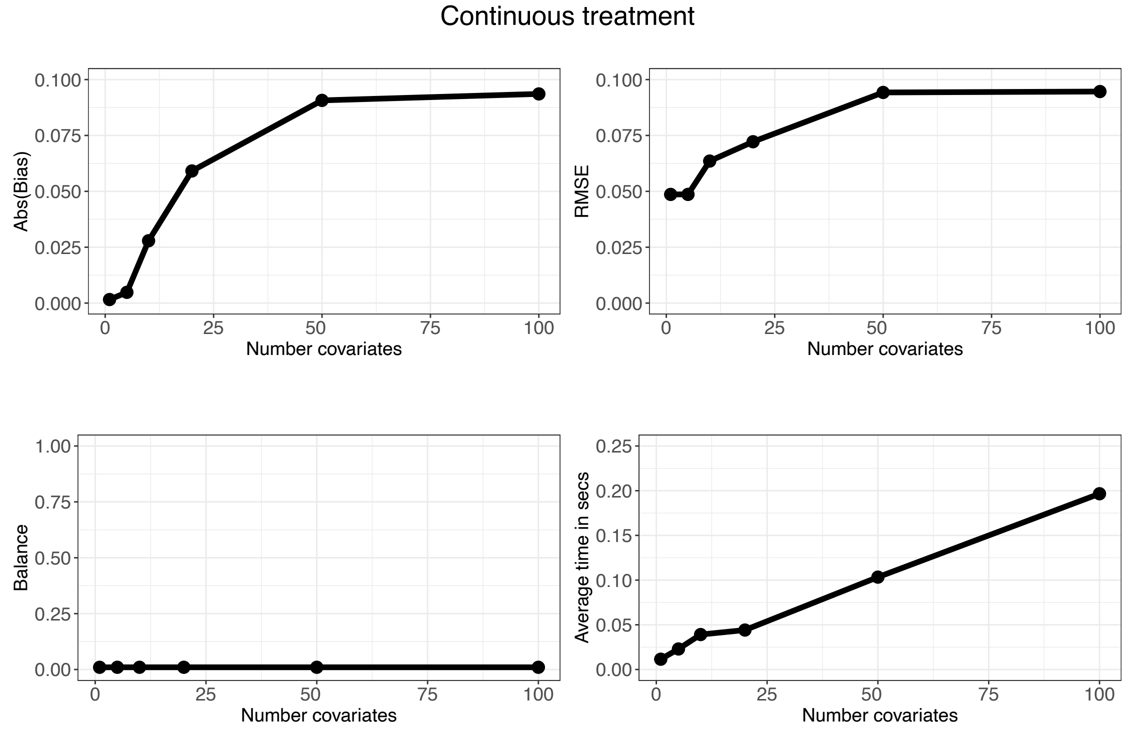

Figure 14: Absolute bias (top left panel), root mean squared error (top right panel), absolute standardized mean difference across four confounders (left bottom panel) and the average computational time in seconds to obtain a solution (right bottom panel) across number of confounders for the continuous treatment scenario.

Table 4: Computational time in seconds needed to obtain a solution across categories of time since menopause (0-10; 11-20; 20+ years), WHI observational study, $n = 24,069$, 1993-2010.

	Time since menopause		
	0-10 Years	**11-20 Years**	**20+ Years**
	$n = 6,661$	$n = 9,592$	$n = 7,816$
	Time (sec)	**Time (sec)**	**Time (sec)**
Method			
ROW	1.0	2.8	1.6
IPW	218.4	332.5	238.1
BalSL	213.6	312.2	225.6
GBM	83.7	115.2	93.3
CBPS	4.6	7.9	6.2
SBW	653.5*	2310.8	1313.8*
EBAL	1.1	1.6	1.3
PSM	236.0	338.1	229.1
OM	0.1	0.1	0.1
Naive	<0.1	<0.1	<0.1

First column: Method implemented, ROW: Robust Optimal Weights; IPW: Inverse Probability Weighting (propensity scores were estimated with SuperLearner with linear regression model with only main effects, and random forest in the library of algorithms); BalSL: Balance SuperLearner; GBM: propensity scores were estimated with Gradient Boosting Machine; CBPS: Covariate Balancing Propensity Score; SBW: Stable Balancing Weights; EBAL: Entropy Balancing; PSM: Propensity Score Matching; OM: (outcome model) Cox proportional hazard model including confounders and treatment; Naive: Cox proportional hazard model including only the treatment. *Second, Third and Fourth columns*: Computational time in seconds and sample size within each category of time since menopause. * Time at which SBW stopped without finding a solution with balance tolerance set equal to 0.0001.

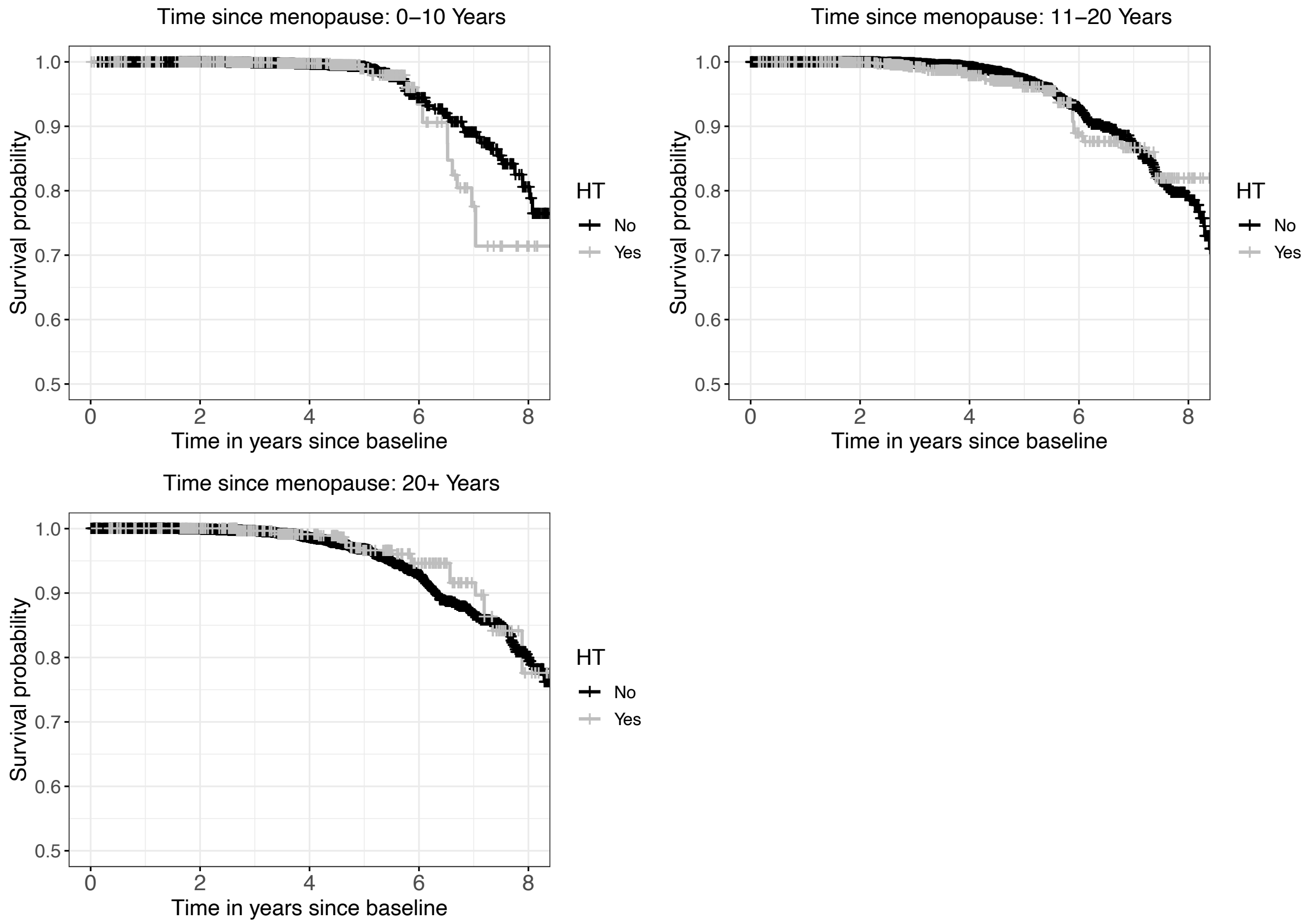

Figure 15: Kaplan-Meier curves weighted by ROW for estrogen plus progestin (HT; Yes versus No) across categories of time since menopause.

8 Additional results

In this Section, we investigate the distribution of ROW across different treatment, covariate relationships. We considered eight scenarios:

1. *Linear dependence*: the treatment, T has a linear dependence on the confounder, X as
$$X \sim N(0,1) \quad \text{and} \quad T \sim X + N(0,1)$$

2. *Nonlinear dependence quadratic*: the treatment, T has a quadratic dependence on the confounder, X as
$$X \sim N(0,1) \quad \text{and} \quad T \sim X + X^2 + N(0,1)$$

3. *Nonlinear dependence cubic*: the treatment, T has a cubic dependence on the confounder, X as
$$X \sim N(0,1) \quad \text{and} \quad T \sim 0.5(X+0.1)^3 + N(0,1)$$

4. *Nonlinear dependence without correlation*: the treatment, T has a lattice-dependence on the confounder, X as
$$X \sim Unif(-.5,.5) \quad \text{and} \quad T \sim \begin{cases} N(0,\frac{1}{3}) & \text{if } X \leq \|\frac{1}{6}\| \\ \frac{1}{2}\left(N(1,\frac{1}{3}) + N(-1,\frac{1}{3})\right) & \text{otherwise} \end{cases}$$

5. *Sinusoidal dependence*: the treatment, T has a sinusoidal dependence on the confounder, X as
$$X \sim N(0,4) \quad \text{and} \quad T \sim \sin X + N(0,0.1)$$

6. *Independence*: the treatment, T is independent on the confounder, X as

$$X \sim N(0, 1) \quad \text{and} \quad T \sim N(0, 1)$$

7. *Right-skewed*: the treatment, T is right-skewed and depends on the confounder, X as

$$X \sim Beta(1, 5) \quad \text{and} \quad T \sim 4X + LogN(0, 0.7)$$

8. *Left-skewed*: the treatment, T is left-skewed and depends on the confounder, X as

$$X \sim Beta(5, 1) \quad \text{and} \quad T \sim 4X + Beta(5, 1)$$

Figure 16 shows the relationships before (blue lines) and after (black lines) weighting for ROW across different covariate-treatment relationships. ROW was almost uniformly distributed under nonlinear dependence without correlation, sinusoidal and under independence. ROW presented larger weights under nonlinear cubic dependence. We balanced linear and terms for obtaining ROW under the quadratic scenario. We balanced linear, quadratic and cubic terms for obtaining ROW under the cubic scenario. Regardeless of the rigth or left skeweness of the treatment, ROW successfully eliminated the relationship between the covariate and the treatment. ROW could be consequentely used when continuous treatments are skweded, such as for example, when interested in evaluating treatment doses, number of cigarettes smoked in one day, or daily food consumption.

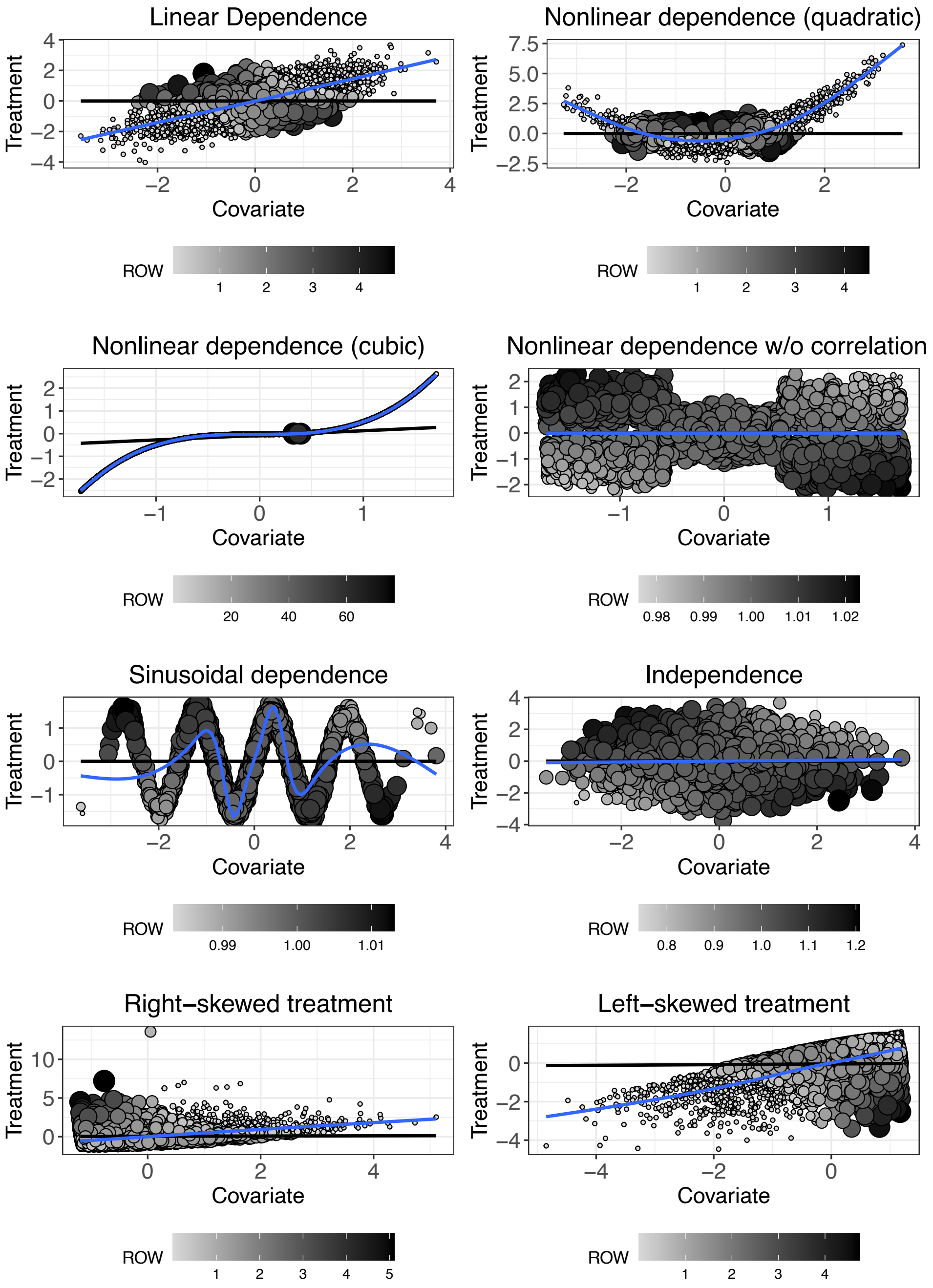

Figure 16: Graphical representation of ROW balancing a covariate (x-axis) and continuous treatment (y-axis) across different covariate-treatment relationships. Blue lines represent the true relationships between thebinary (a probit model) and the continuous (simple regression with normal errors and positive coefficient)treatment. Black lines represent the relationship between treatments and covariate after weighting for ROW. Size and color of the circles represent the individual weight assigned (the larger/darker the higher).