Journal of Data Science 0 (0), 1–17
??? 2023
DOI: 10.6339/23-JDS1107
Statistical Data Science

Identification of Optimal Combined Moderators for Time to Relapse

Bang Wang[1], Yu Cheng[1,2,*], and Michele D. Levine[3]

[1]*Department of Statistics, University of Pittsburgh, Pittsburgh, PA, 15213, USA*
[2]*Department of Biostatistics, University of Pittsburgh, Pittsburgh, PA, 15213, USA*
[3]*Department of Psychiatry, University of Pittsburgh, Pittsburgh, PA, 15213, USA*

Abstract

Identifying treatment effect modifiers (i.e., moderators) plays an essential role in improving treatment efficacy when substantial treatment heterogeneity exists. However, studies are often underpowered for detecting treatment effect modifiers, and exploratory analyses that examine one moderator per statistical model often yield spurious interactions. Therefore, in this work, we focus on creating an intuitive and readily implementable framework to facilitate the discovery of treatment effect modifiers and to make treatment recommendations for time-to-event outcomes. To minimize the impact of a misspecified main effect and avoid complex modeling, we construct the framework by matching the treated with the controls and modeling the conditional average treatment effect via regressing the difference in the observed outcomes of a matched pair on the averaged moderators. Inverse-probability-of-censoring weighting is used to handle censored observations. As matching is the foundation of the proposed methods, we explore different matching metrics and recommend the use of Mahalanobis distance when both continuous and categorical moderators are present. After matching, the proposed framework can be flexibly combined with popular variable selection and prediction methods such as linear regression, least absolute shrinkage and selection operator (Lasso), and random forest to create different combinations of potential moderators. The optimal combination is determined by the out-of-bag prediction error and the area under the receiver operating characteristic curve in making correct treatment recommendations. We compare the performance of various combined moderators through extensive simulations and the analysis of real trial data. Our approach can be easily implemented using existing R packages, resulting in a straightforward optimal combined moderator to make treatment recommendations.

Keywords *counterfactual outcomes; matched pair; personalized medicine; smoking cessation*

1 Introduction

Substantial heterogeneity of treatment effectiveness exists in some clinical studies, and identifying treatment effect modifiers plays an essential role in improving the treatment efficacy. Here treatment effect modifiers (or moderators) are defined as variables measured at baseline that exhibit an interactive effect with treatment on outcomes (Kraemer et al., 2002). In practice, existing trial data and observational studies are often utilized to search for moderators by looking for all possible interactions with treatment through regression-based methods. However,

*Corresponding author. Email: yucheng@pitt.edu.

Received August 28, 2022; Accepted May 29, 2023

the power to detect the treatment effect modifiers could be reduced by limited sample sizes, modest interaction effects, or the strong main effect that explains much of the variability in the outcome (Kraemer, 2013). On the other hand, exploratory analyses that examine one moderator per statistical model are known for the tendency of finding spurious interactions, especially when a long list of variables is tested for moderation effects. Thus, extant literature has focused on recommending an appropriate treatment by estimating effect modification via a systematic approach (Kraemer, 2013; Tian et al., 2014; Chen et al., 2017; Song et al., 2017; Liang and Yu, 2022; Yadlowsky et al., 2021; Park et al., 2022).

The goal of this article is to propose an intuitive and readily implementable approach to discovering treatment effect modifiers and making treatment recommendations for time-to-event outcomes. The work is motivated by one of our randomized controlled trials (RCTs), Strategies to Avoid Returning to Smoking (STARTS), with the goal of preventing postpartum smoking relapse (Levine et al., 2013). In STARTS, a cognitive behavioral treatment (CBT) was compared to a supportive behavioral treatment (SBT), and no significant differences were found in time to relapse during one-year postpartum (Levine et al., 2016). We are curious, given null effects, particularly, in studies with two active treatments, about whether one condition might benefit a subgroup more than another and thus search for moderators.

We start with the contrast function

$$\Delta(M) = E[Y|D = 1, M] - E[Y|D = 0, M],$$

where $E[Y|D, M]$ is the expected outcome Y given an intervention D and a set of moderators M. Adopting the potential outcome framework in causal inference, let (Y^1, Y^0) denote the potential outcomes if a participant received a new treatment and a standard treatment, respectively. Then, under regular causal inference assumptions, $\Delta(M)$ is the conditional average treatment effect (CATE) and can be interpreted as a causal effect modifier (Rubin, 1974, 2005), i.e.,

$$\text{CATE} = E[Y^1 - Y^0|M] = E[Y^1|M] - E[Y^0|M] = E[Y|D = 1, M] - E[Y|D = 0, M] = \Delta(M).$$

Kraemer (2013) developed a parametric framework based on matched pairs of treated and untreated subjects and modeled the difference in the paired outcomes as a linear combination of moderators. Similarly, based on the causal interpretation of the moderator effect, Tian et al. (2014) developed a framework that posited working models for estimating the moderator effect in RCT studies by directly modeling the outcome on modified moderators. Mo and Liu (2022) recently proposed an efficient learning framework for continuous outcomes, which includes the model by Tian et al. (2014) as a special case under homogeneous variance. However, the implementation of efficient learning is not straightforward.

In this work, we focus on developing an intuitive method to be used in practice. Kraemer's framework has been frequently implemented in psychiatric studies to detect treatment effect modifiers for eating disorders, anxiety, and depressive disorder, among others (Wallace et al., 2013, 2018; Wallace and Smagula, 2018; Smagula et al., 2016; Kaneriya et al., 2016; Niles et al., 2017; Hildebrandt et al., 2020; Chin Fatt et al., 2020). We will extend Kraemer's framework from a continuous outcome to the time-to-event setting, and construct a composite moderator from a list of candidates as an optimal causal effect modifier for time to relapse in STARTS. CATE is often the focus in detecting treatment effect modifiers for survival outcomes and is modeled based on standard survival models like Cox proportional hazard (PH) model to handle censoring (Tian et al., 2014; Yadlowsky et al., 2021). In this work, we adopt Tian's interpretation by modeling CATE as the causal effect modifier and use inverse-probability-of-censoring weighting (IPCW)

to handle censoring, as it can be flexibly combined with different frameworks without changing the model interpretations (Goldberg and Kosorok, 2012; Zhao et al., 2015; Cui et al., 2017).

We will model the CATE based on matched pairs to minimize the impact of misspecifying main effects and avoid complex modeling as in Mo and Liu (2022). The 1:1 nearest neighbor matching (NNM) algorithm is usually used to estimate the conditional average treatment effect on treated (CATT) because it matches control individuals to the treated and discards unselected controls. Different from the CATE, the CATT is the conditional expectation over the subpopulation of treated people of the treatment effect.

$$\text{CATT} = E[Y^1 - Y^0 | D = 1, M].$$

However, under random assignment, the CATT is equivalent to the CATE. Besides, given similar sample sizes in each group, the 1:1 NNM algorithm discards few observations and thus has a limited reduction in power (Stuart, 2010). To reduce matching bias, it is recommended to convert categorical covariates to a series of binary indicators and standardize continuous covariates before matching (Kraemer, 2013).

After matching, we regress the weighted difference in the observed outcomes from a matched pair on the differences in potential effect modifiers, with or without their corresponding average scores, and select important factors to be included in the final linear combination using Z scores, Lasso, and random forest. As subsequent modeling depends on matching, it is important to explore the impact of matching on estimating causal effect modifiers. Two matching metrics are considered in this study: Mahalanobis distance (MD) and propensity score (PS). King et al. (2011) showed that using PS could degrade causal inferences as compared to unmatched methods if the two groups are already well balanced, while using the MD would achieve a lower imbalance. Thus, we will compare these two metrics via simulation studies under different scenarios to assess the impact of matching on our estimators.

The rest of the article is organized as follows: In Section 2, we propose various matched-weighting (MW) estimators for the causal effect modifier. In Section 3, simulation studies under different scenarios are conducted to compare the performance of MW estimators to comparative methods in estimating the treatment effect modification and making treatment recommendations. In Section 4, we illustrate the utility of our modeling framework by applying it to STARTS. Finally, the conclusion and discussion are provided in Section 5.

2 Matched-Weighting Estimators

In the following, for individual i, let $\tilde{T}_i$ be the minimum of event time T_i and independent censoring time C_i. Denote the event status as $\delta_i = \mathbf{1}\{T_i \leqslant C_i\}$. Let M_i be a p-dimensional vector of all potential moderators. The independent censoring assumption can be relaxed to be conditional independence given moderators. In addition, we center the treatment allocation D_i, which equals to 0.5 if individual i is in the treatment group and -0.5 if individual i belongs to the control group. Then, the observed outcome can be denoted by n independent and identically distributed (i.i.d.) replications of $(\tilde{T}, \delta, D, M)$, such that $\{(\tilde{T}_i, \delta_i, D_i, M_i), i = 1, \ldots, n\}$.

In general, we assume that given any variable M, the treatment assignment D is independent of the potential outcomes and is not deterministic, i.e.,

$$D \perp (T^1, T^0) \mid M$$

and

$$0 < \Pr\{D = d \mid M\} < 1,$$

for all d and M (Strong Ignorability, SI). Furthermore, the Stable Unit Treatment Value Assumption (SUTVA) is assumed, where the potential outcomes for any individual are not affected by the treatment assigned to other individuals, and there is only one form of treatment for each treatment level.

2.1 A Model

Now we adopt the framework of Kraemer (2013) based on the matched pairs and extend it to survival outcomes and handle censoring by introducing IPCW weights w_j to adjust for the bias caused by censoring. Consider a survival model for the event time T of an individual with treatment D and a vector of potential treatment modifiers M

$$h(T) = \theta_0 + \theta_d D + \boldsymbol{\theta}_{ma}^{\intercal} M + \boldsymbol{\theta}_{mo}^{\intercal} D \times M + \epsilon, \tag{1}$$

where h is some monotone function of T, ϵ is a mean-zero error term, θ_0 is the intercept, θ_d is the treatment effect, and $\boldsymbol{\theta}_{ma}^{\intercal}$ and $\boldsymbol{\theta}_{mo}^{\intercal}$ are the transposes ($\intercal$) of two p-dimensional vectors, referring to the main and moderator effects, respectively. When $h = \log$, it becomes the familiar accelerated failure time (AFT) model. Then under the SUTVA assumption, we have

$$\begin{aligned}
\text{CATE} &= E[\Delta h(T) \mid M] \\
&= E[h(T^1) - h(T^0) \mid M] \\
&= E[h(T^1) \mid D = 0.5, M] - E[h(T^0) \mid D = -0.5, M] && \text{SI} \\
&= E[h(T) \mid D = 0.5, M] - E[h(T) \mid D = -0.5, M] && \text{SUTVA} \\
&= \theta_d + \boldsymbol{\theta}_{mo}^{\intercal} M,
\end{aligned}$$

where θ_d is the treatment effect when $M = 0$, and the moderator effect $\boldsymbol{\theta}_{mo}$ becomes the coefficient for the causal effect modifier. Motivated by the above relationship, we now consider a matched pair of a treated subject and their control with event times and moderators (T_1, M_1, T_0, M_0). With a perfect match, the difference in the potential moderators $dM = 0$. Thus, if one works on the matched pairs and regresses the difference in the matched outcome on the average of two moderators aM, the slope will be an unbiased estimator of $\boldsymbol{\theta}_{mo}$.

For time-to-event data, not all outcomes are observed. To maintain a relatively large number of matched pairs, all censored observations from each treatment will be excluded before matching. For each matched pair j, denote the paired event time as (T_{1j}, T_{0j}) and the paired patient profile as (M_{1j}, M_{0j}), for $j = 1, \ldots, n_p$ and n_p is the number of pairs. After matching, we start modeling the paired contrast,

$$\Delta h(T_j) = h(T_{1j}) - h(T_{0j}).$$

Then, based on model (1) and the relationship revealed above, we have:

$$\sqrt{w_j}\Delta h(T_j) = \sqrt{w_j}[\theta_d + \boldsymbol{\theta}_{mo}^{\intercal} aM_j + d\epsilon_j], \tag{2}$$

where $\theta_d \in \mathbf{R}$, $\boldsymbol{\theta}_{mo} \in \mathbf{R}^p$, $d\epsilon_j$ are i.i.d. mean zero error terms. The modified moderators,

$$aM_j = (M_{1j} + M_{0j})/2 \in \mathbf{R}^p,$$

which is due to the centering of the treatment allocation. The IPCW is calculated in terms of the paired event times and the survival probability of the censoring time S_c,

$$w_j = \frac{\delta_{1j}\delta_{0j}}{S_c(T_{1j}-)S_c(T_{0j}-)}.$$

In practice, S_c is typically unknown and can be estimated by the Kaplan Meier estimator (Kaplan and Meier, 1958) or from a Cox proportional hazards model (Cox, 1972), denoted as $\hat{S}_c$.

Based on model (2), we propose an Ordinary Least Squares (OLS)-typed matched-weighting (MW) estimator for the causal effect modifier, $\hat{\boldsymbol{\theta}}_a$:

$$\hat{\boldsymbol{\theta}}_a := \underset{\boldsymbol{\theta}_{mo}}{\arg\min} \sum_{j=1}^{n_p} \hat{w}_j \left[\Delta h(T_j) - \theta_d - \boldsymbol{\theta}_{mo}^{\intercal} a M_j\right]^2, \tag{3}$$

where

$$\hat{w}_j = \frac{\delta_{1j}\delta_{0j}}{\hat{S}_c(T_{1j}-)\hat{S}_c(T_{0j}-)}.$$

Thus, the "A model" indicates that the modified outcome is fitted on the paired average only.

2.2 DA Model

However, in practice, the misspecification of the statistical model or the covariate set when calculating matching metrics may lead to imbalanced baseline characteristics after matching. Therefore, considering the impact of matching imbalance on estimating the causal effect modifier, we adjust for the paired difference term dM and name it as the "DA model":

$$\sqrt{w_j}\Delta h(T_j) = \sqrt{w_j}[\theta_d + \boldsymbol{\theta}_{ma}^{\intercal} dM_j + \boldsymbol{\theta}_{mo}^{\intercal} a M_j + d\epsilon_j], \tag{4}$$

where the main effect $\theta_{ma} \in \mathbf{R}^p$,

$$dM_j = M_{1j} - M_{0j} \in \mathbf{R}^p.$$

Subsequently, we propose another MW estimator $\hat{\theta}_{da}$:

$$\hat{\boldsymbol{\theta}}_{da} := \underset{\boldsymbol{\theta}_{mo}}{\arg\min} \sum_{j=1}^{n_p} \hat{w}_j \left[\Delta h(T_j) - \theta_d - \boldsymbol{\theta}_{mo}^{\intercal} a M_j - \boldsymbol{\theta}_{ma}^{\intercal} dM_j\right]^2. \tag{5}$$

2.3 Moderator Selection

In Kraemer (2013), an optimal treatment effect modifier was constructed as a linear combination of modified continuous moderators, where those moderators were selected based on their correlations with the paired difference of continuous outcomes. A selection threshold was set after the univariate analysis, but the correlations calculated in practice are generally small. Therefore, we propose to use the standardized estimated coefficient of aM in each univariate analysis as the selector and use the critical value of the corresponding distribution of the estimated coefficient as the threshold. In addition, to minimize the impact of imbalance caused by matching, dM could also be screened and selected into the composite moderator based on its standardized estimated coefficient. These screening procedures yield the other two MW estimators, $\hat{\boldsymbol{\theta}}_{Sa}(\alpha)$ and $\hat{\boldsymbol{\theta}}_{Sda}(\alpha)$,

$$\hat{\boldsymbol{\theta}}_{Sa}(\alpha) := \underset{\boldsymbol{\theta}_{mo}(\alpha)}{\arg\min} \sum_{j=1}^{n_p} w_j \left[\Delta h(T_j) - \theta_d(\alpha) - \boldsymbol{\theta}_{mo}^{\intercal}(\alpha) a M_j\right]^2, \tag{6}$$

where S stands for screening, and α denotes the screening threshold. $\hat{\boldsymbol{\theta}}_{Sda}(\alpha)$ is defined similarly as $\hat{\boldsymbol{\theta}}_{Sa}(\alpha)$, with the additional adjustment $\boldsymbol{\theta}_{ma}^{\intercal}(\alpha) dM_j(\alpha)$ in the equation.

Similarly, L_1 penalized (Lasso) estimators proposed by Tibshirani (1996) can be applied to this weighted matching framework to select important causal effect modifiers, with or without adjusting for the imbalance captured by dM. The penalized MW estimators with a shrinkage parameter λ are denoted as $\hat{\boldsymbol{\theta}}_{La}(\lambda_a)$ and $\hat{\boldsymbol{\theta}}_{Lda}(\lambda_{da})$, without and with dM, respectively, which can be calculated by minimizing:

$$\frac{1}{n_p}\sum_{j=1}^{n_p} w_j \left[\Delta h(T_j) - \theta_d - \boldsymbol{\theta}_{mo}^{\intercal} aM_j\right]^2 + \lambda_a \|\boldsymbol{\theta}_{mo}\|_1, \tag{7}$$

$$\frac{1}{n_p}\sum_{j=1}^{n_p} w_j \left[\Delta h(T_j) - \theta_d - \boldsymbol{\theta}_{ma}^{\intercal} dM_j - \boldsymbol{\theta}_{mo}^{\intercal} aM_j\right]^2 + \lambda_{da} \|\{\boldsymbol{\theta}_{ma}, \boldsymbol{\theta}_{mo}\}\|_1, \tag{8}$$

where $\|\cdot\|_1$ is the L_1 norm. If the focus is on making treatment recommendations rather than estimating treatment effect modifiers, one could adopt machine learning techniques like random forest or the decision tree under the MW framework (Breiman, 2001; Liaw et al., 2002).

2.4 Evaluation

To evaluate the performance of our proposed MW estimators, we will tabulate the sample mean, the average of estimated standard errors (ASE), the empirical standard deviation (ESD), and the empirical coverage rate (CVRT) of the coefficients for causal effect modifiers. On the other hand, the out-of-bag prediction error (OOBPE) and the out-of-bag area under the receiver operating characteristic curve (OOBAUC) will be used to evaluate the performance in making personalized recommendations. The MW estimators with relatively larger OOBAUC and smaller OOBPE will be considered optimal ones to make treatment recommendations. In addition, we also calculated PE and area under the curve (AUC) under a two-sample (TS) setting to check the robustness of the OOB metrics, where the original simulated sample is treated as the training data, and another independent sample is simulated as the testing dataset.

3 Numerical Studies

3.1 Simulation Setting

In this section, we performed numerical studies to investigate the finite sample performance of the proposed MW estimators in various settings. The causal effect modifiers were estimated under the MW framework in combination with the OLS (MW.O), Lasso (MW.L), and random forest (MW.RF) methods.

Here our method was evaluated and compared with existing approaches in estimating moderator effects, including the AFT model with prior knowledge of error distribution and the Cox PH model. Both approaches were fitted on treatment allocation, each (U) or all possible moderators, and their interactions with treatment. As a competitive method, we also adopted a random survival forest (RSF) model with the log-rank score splitting rule and a Kaplan-Meier (KM) based OOB ensemble estimator (Ishwaran et al., 2008; Ishwaran and Kogalur, 2007). Although Cox could not provide a comparative estimation of the moderator effect when the PH assumption is violated, and the RSF predicts the survival probability only, one could still calculate the pair difference in the predicted hazard risks or survival probability and evaluate their performance in making treatment recommendations.

The outcome was generated from parametric survival models with log-linear representation:

$$\log(T_i) = \alpha + \beta D_i + \boldsymbol{\theta}_{ma}^{\intercal} M_i + \boldsymbol{\theta}_{mo}^{\intercal} D_i M_i + \sigma e_i,$$

where $e_i \sim F$ and σ is the scale parameter, $i = 1, \ldots, 300$. Two treatment arms with equal group sizes were specified. We assumed that the first 5 of 15 moderator candidates had an interactive effect, such that $(\alpha, \beta) = (-6, 0.10)$, $\boldsymbol{\theta}_{ma}^{\intercal} = (0.15, -0.20, -0.50, 0.25, -0.20, 0.50, 0.25, 0, 0, -0.20, -0.20, 0, 0, 0, 0)$ and $\boldsymbol{\theta}_{mo}^{\intercal} = (-0.75, 0.50, 0.25, -1.25, 1, 0, \ldots, 0)$. The baseline covariates were generated independently from either the standard normal distribution or a Bernoulli distribution with mean of 0.5. To study the properties of the proposed MW estimators, we considered different simulation scenarios with the following key aspects of interest: (1.) The 1:1 NNM algorithm with two different metrics, MD and PS, was implemented using the R package "MathIt" (Ho et al., 2011); (2.) We considered two error distributions: extreme value distribution (EV) and standard logistic distribution (Logistic); (3.) Two scales of error variance, $\sigma = \frac{1}{2}, \frac{1}{6}$, were used to determine the noise of the data; (4.) We assumed the independent censoring time to follow a 50-50 mixture distribution of $\exp(\lambda_1)$ and $\exp(\lambda_2)$ and considered two censoring rates 15% and 25%. One thousand simulations were performed under each simulation setting. Due to the space limitation, in this section, we only present the results from representative scenarios and refer the reader to Supplementary Material for the remaining results.

3.2 Simulation Results

We first compared the matching performance using MD and PS to study how the matching bias impacts our proposed methods. Ho et al. (2007) pointed out that, one should try as many matching solutions as possible and choose the one that yields the best balance. Consequently, the inclusion of covariates depends not only on factors like the covariate distribution, covariate effect, sample size, etc., but also on the objective of matching and how the optimal balance is defined. In Stuart (2010), the method that achieves optimal balance can be defined as follows: (1.) The one yields the smallest standardized mean difference across the largest covariates. (2.) The one minimizes the standardized difference of means of a few, particularly prognostic covariates. (3.) The one results in the fewest number of "large" standardized differences of means.

In this study, we aim to create matched pairs with similar characteristics. Thus, we included all covariates and evaluated matching performance by the standard pair difference (SPD) for each moderator, i.e., the average absolute within-pair difference of each covariate after matching (Ho et al., 2011). In addition, the proportion of "perfect match" is reported for categorical variables for both MD and PS. Shown in Table 1, the SPDs of moderators matched by the PS are generally larger than the MD, especially for categorical moderators, resulting in smaller

Table 1: Standard pair difference ("prefect matching" proportion) under the setting: 15% censoring rate, EV, $\sigma = 1/6$; using Mahalanobis distance (MD) and propensity score (PS) metrics.

	Type	MD	PS
M1	cont.	.818	1.08
M2	cont.	.816	1.09
M3	cont.	.817	1.10
M4	cate.	.42(78.8%)	.92(53.5%)
M5	cate.	.42(78.8%)	.92(53.1%)

Table 2: Estimated moderator effects under the setting: MD $\sigma = 1/6$ 15% censoring rate; for each estimator of the five non-zero moderators, the cell above shows the estimated moderator effect (and coverage rate), the cell below shows the average standard error (and empirical standard deviation), and $k = 1, 2, 3$ in MW.SDAk and MW.SAk refer to the threshold value used to select meaningful moderators.

Moderator	***M1***		***M2***		***M3***		***M4***		***M5***	
TRUE	***−0.75***		***0.50***		***0.25***		***−1.25***		***1***	
Error Dist.	***EV***	***Logistic***	***EV***	***Logistic***	***EV***	***Logistic***	***EV***	***Logistic***	***EV***	***Logistic***
MW.ODA.U	−.76(.95)	−.77(.96)	.50(.95)	.52(.95)	.25(.95)	.26(.96)	−1.26(.96)	−1.28(.96)	1.02(.96)	1.01(.95)
	.14(.14)	.14(.14)	.15(.14)	.15(.15)	.15(.14)	.15(.14)	.28(.26)	.29(.28)	.29(.28)	.30(.30)
MW.OA.U	−.76(.95)	−.77(.95)	.50(.96)	.52(.95)	.25(.94)	.26(.95)	−1.26(.96)	−1.28(.96)	1.02(.95)	1.01(.95)
	.14(.14)	.14(.14)	.15(.15)	.16(.15)	.16(.16)	.16(.16)	.28(.26)	.29(.28)	.29(.28)	.30(.30)
MW.ODA	−.75(.97)	−.75(.97)	.50(.98)	.50(.96)	.25(.97)	.25(.97)	−1.25(.97)	−1.26(.94)	1.00(.95)	1.00(.96)
	.04(.04)	.06(.05)	.04(.04)	.06(.05)	.04(.04)	.06(.05)	.08(.07)	.11(.11)	.08(.07)	.11(.10)
MW.OA	−.75(.94)	−.75(.94)	.49(.93)	.50(.94)	.24(.91)	.24(.93)	−1.24(.94)	−1.24(.94)	.99(.92)	.98(.93)
	.10(.12)	.11(.12)	.10(.11)	.11(.12)	.11(.12)	.11(.12)	.20(.21)	.21(.22)	.20(.22)	.21(.23)
MW.ODA.S1	−.74(.95)	−.74(.95)	.48(.92)	.49(.93)	.19(.72)	.19(.74)	−1.22(.94)	−1.23(.95)	.98(.93)	.97(.94)
	.04(.04)	.06(.06)	.04(.07)	.05(.08)	.03(.11)	.04(.12)	.08(.08)	.10(.12)	.08(.12)	.10(.15)
MW.OA.S1	−.74(.92)	−.74(.93)	.48(.92)	.48(.91)	.19(.67)	.20(.67)	−1.22(.91)	−1.23(.91)	.97(.92)	.96(.90)
	.10(.11)	.11(.12)	.10(.12)	.10(.13)	.07(.15)	.08(.15)	.19(.21)	.20(.23)	.19(.23)	.20(.25)
MW.ODA.S2	−.72(.91)	−.72(.92)	.44(.86)	.44(.86)	.10(.37)	.10(.34)	−1.19(.90)	−1.20(.93)	.92(.90)	.89(.86)
	.05(.06)	.06(.07)	.05(.14)	.06(.16)	.02(.13)	.02(.13)	.10(.17)	.12(.19)	.09(.24)	.11(.29)
MW.OA.S2	−.74(.91)	−.75(.93)	.46(.86)	.46(.85)	.11(.31)	.11(.29)	−1.22(.92)	−1.23(.92)	.95(.89)	.92(.86)
	.10(.12)	.11(.12)	.09(.18)	.10(.19)	.03(.16)	.03(.17)	.19(.25)	.20(.27)	.19(.30)	.19(.34)
MW.ODA.S3	−.71(.91)	−.70(.92)	.33(.65)	.33(.64)	.03(.10)	.03(.08)	−1.11(.87)	−1.12(.87)	.71(.69)	.66(.64)
	.08(.12)	.08(.12)	.05(.24)	.05(.24)	.01(.09)	.01(.09)	.13(.33)	.15(.36)	.10(.47)	.10(.49)
MW.OA.S3	−.74(.91)	−.74(.92)	.35(.62)	.35(.60)	.04(.07)	.04(.06)	−1.17(.88)	−1.18(.88)	.76(.68)	.71(.62)
	.11(.15)	.11(.15)	.07(.27)	.07(.28)	.01(.12)	.01(.12)	.19(.37)	.20(.40)	.14(.51)	.14(.54)
MW.LDA	−.73(.95)	−.73(.95)	.48(.94)	.47(.93)	.23(.92)	.22(.94)	−1.22(.93)	−1.21(.92)	.97(.92)	.95(.94)
	.04(.04)	.06(.05)	.04(.04)	.06(.05)	.04(.04)	.06(.05)	.07(.07)	.10(.11)	.07(.07)	.10(.10)
MW.LA	−.68(.90)	−.68(.91)	.42(.90)	.42(.90)	.17(.83)	.17(.86)	−1.11(.90)	−1.10(.90)	.86(.89)	.84(.89)
	.11(.12)	.11(.12)	.11(.12)	.11(.12)	.10(.12)	.11(.12)	.20(.21)	.21(.23)	.20(.23)	.21(.24)
Cox.U	−.12(.00)	−.12(.00)	.08(.00)	.08(.00)	.04(.00)	.03(.00)	−.19(.00)	−.19(.00)	.15(.00)	.15(.00)
	.02(.02)	.02(.02)	.02(.02)	.02(.02)	.02(.03)	.02(.03)	.04(.04)	.04(.04)	.04(.04)	.04(.04)
AFT.U	−.75(.87)	−.78(.94)	.50(.90)	.52(.94)	.22(.87)	.25(.93)	−1.23(.88)	−1.30(.93)	.98(.88)	1.03(.93)
	.12(.15)	.12(.12)	.12(.15)	.13(.13)	.12(.15)	.12(.12)	.24(.30)	.25(.27)	.25(.31)	.26(.27)
Cox	−.82(.76)	−.48(.00)	.54(.81)	.32(.01)	.27(.88)	.16(.15)	−1.36(.78)	−.81(.00)	1.09(.79)	.64(.00)
	.05(.05)	.03(.05)	.04(.04)	.03(.04)	.03(.03)	.03(.04)	.09(.09)	.06(.09)	.08(.08)	.06(.08)
AFT	−.75(.94)	−.76(.95)	.50(.93)	.50(.92)	.25(.93)	.25(.95)	−1.25(.93)	−1.26(.93)	1.00(.92)	1.00(.90)
	.02(.02)	.04(.04)	.02(.02)	.04(.04)	.02(.02)	.04(.04)	.04(.05)	.07(.08)	.04(.05)	.07(.08)

proportions of "perfect match." When estimating the causal effect modifiers (comparing Table 2 to Table 3 under $\sigma = 1/6$), the estimated and empirical standard deviations increase as the matching bias increase. A slightly larger bias also appears in the estimators matched by the PS if the selection is involved. We observe similar results from Tables S1 and S2 in Supplementary Material under $\sigma = 1/2$.

When making treatment recommendations, comparing Table 4 to Table 5, a larger matching bias would degrade the performance of A models, as shown by smaller AUC and larger PE. For example, in Tables 4–5, with 25% censoring, an EV error distribution, and a smaller variance ($\sigma = 1/6$), the OOB (TS) AUC of the MW estimator with the OA model using the MD is 0.73 (0.75), which is larger than the AUC using PS, 0.66 (0.67). While for the MW.ODA estimators,

Table 3: Estimated moderator effects under the setting: PS $\sigma = 1/6$ 15% censoring rate; for each estimator of the five non-zero moderators, the cell above shows the estimated moderator effect (and coverage rate), the cell below shows the average standard error(and empirical standard deviation), and $k = 1, 2, 3$ in MW.SDAk and MW.SAk refer to the threshold value used to select meaningful moderators.

Moderator	***M1***		***M2***		***M3***		***M4***		***M5***	
TRUE	***−.75***		***.50***		***.25***		***−1.25***		***1***	
Error Dist.	***EV***	***Logistic***	***EV***	***Logistic***	***EV***	***Logistic***	***EV***	***Logistic***	***EV***	***Logistic***
MW.ODA.U	−.81(.93)	−.82(.93)	.54(.91)	.55(.92)	.27(.91)	.28(.93)	−1.35(.94)	−1.37(.93)	1.09(.91)	1.10(.92)
	.19(.20)	.19(.19)	.19(.21)	.20(.21)	.18(.21)	.19(.20)	.38(.39)	.39(.41)	.39(.41)	.40(.44)
MW.OA.U	−.81(.93)	−.82(.93)	.54(.91)	.55(.93)	.27(.92)	.28(.92)	−1.35(.94)	−1.37(.93)	1.10(.92)	1.10(.92)
	.19(.20)	.19(.19)	.20(.21)	.20(.22)	.21(.23)	.21(.23)	.38(.39)	.39(.41)	.39(.41)	.40(.44)
MW.ODA	−.75(.98)	−.75(.97)	.50(.97)	.50(.97)	.25(.97)	.25(.97)	−1.25(.96)	−1.26(.95)	1.00(.96)	1.00(.96)
	.05(.04)	.07(.06)	.05(.05)	.07(.06)	.05(.05)	.07(.07)	.10(.09)	.14(.13)	.10(.09)	.14(.13)
MW.OA	−.76(.94)	−.76(.95)	.49(.95)	.49(.95)	.25(.94)	.25(.94)	−1.26(.96)	−1.26(.96)	1.00(.95)	1.00(.95)
	.17(.17)	.17(.16)	.17(.17)	.17(.18)	.17(.17)	.18(.19)	.33(.32)	.34(.34)	.34(.35)	.34(.35)
MW.ODA.S1	−.74(.95)	−.74(.95)	.47(.89)	.47(.89)	.17(.67)	.17(.66)	−1.22(.93)	−1.22(.93)	.93(.89)	.94(.89)
	.05(.05)	.07(.07)	.05(.12)	.06(.13)	.03(.12)	.04(.13)	.10(.16)	.13(.16)	.09(.23)	.12(.26)
MW.OA.S1	−.74(.95)	−.75(.93)	.47(.90)	.47(.88)	.19(.61)	.20(.61)	−1.21(.94)	−1.22(.93)	.96(.89)	.96(.90)
	.16(.16)	.16(.16)	.15(.19)	.15(.20)	.10(.19)	.11(.20)	.31(.33)	.32(.35)	.30(.38)	.31(.38)
MW.ODA.S2	−.72(.93)	−.72(.91)	.38(.73)	.38(.72)	.09(.32)	.08(.29)	−1.12(.87)	−1.13(.86)	.76(.72)	.76(.72)
	.06(.11)	.08(.12)	.05(.21)	.06(.22)	.02(.13)	.02(.13)	.12(.34)	.14(.35)	.10(.42)	.11(.44)
MW.OA.S2	−.74(.94)	−.74(.93)	.41(.74)	.41(.72)	.11(.26)	.11(.25)	−1.17(.89)	−1.19(.88)	.84(.74)	.84(.74)
	.16(.19)	.16(.19)	.12(.26)	.12(.27)	.04(.18)	.04(.19)	.29(.43)	.30(.46)	.24(.52)	.25(.53)
MW.ODA.S3	−.67(.87)	−.67(.86)	.24(.44)	.23(.41)	.03(.08)	.03(.08)	−.90(.70)	−.89(.66)	.48(.44)	.45(.41)
	.09(.23)	.09(.24)	.04(.26)	.04(.27)	.01(.10)	.01(.10)	.13(.57)	.14(.60)	.08(.53)	.08(.54)
MW.OA.S3	−.71(.86)	−.71(.86)	.27(.41)	.27(.41)	.04(.07)	.04(.07)	−1.00(.70)	−.98(.66)	.58(.43)	.54(.41)
	.15(.28)	.15(.28)	.07(.32)	.07(.32)	.01(.13)	.01(.14)	.23(.66)	.23(.69)	.15(.65)	.14(.64)
MW.LDA	−.73(.95)	−.73(.95)	.48(.93)	.47(.94)	.22(.92)	.22(.93)	−1.21(.94)	−1.20(.94)	.96(.94)	.95(.93)
	.05(.04)	.07(.06)	.05(.05)	.07(.06)	.05(.05)	.07(.07)	.09(.09)	.13(.13)	.09(.09)	.13(.13)
MW.LA	−.64(.92)	−.64(.92)	.37(.89)	.37(.88)	.14(.86)	.15(.87)	−1.02(.93)	−1.02(.92)	.77(.89)	.76(.89)
	.17(.18)	.18(.17)	.17(.18)	.17(.18)	.15(.15)	.15(.16)	.34(.34)	.35(.35)	.33(.36)	.34(.37)

their AUCs are quite close to each other: 0.91 (0.92) and 0.91 (0.90). This suggests that adjusting D terms could provide a more robust result in the presence of a larger matching bias. When we change censoring rates, we observe by comparing Table 2 to Table 7, and comparing Table 4 to Table S3 in Supplementary Material that the MW estimators have a robust performance in estimating causal effect modifiers and making treatment recommendations as the censoring rate increases from 15% to 25%. The IPCW appears effective in overcoming the censoring issue when the censoring rate is modest.

As illustrated by Tables 2–3 and 6–7, if the error term follows a logistic distribution, our proposed estimators' SDs become larger, as the logistic distribution has a heavier tail than the EV error in our setting. When the selection procedure is involved, the skewer the error distribution, the larger the bias. In terms of prediction, a logistic error yields a smaller AUC and a larger PE than the EV error, according to Tables 4–5. Furthermore, when the scale parameter of the error term σ increases, i.e., the data become noisier, the performance of all methods degrades, especially for a heavier tail error, resulting in a larger bias, larger ASE, empirical SD, and PE, and a smaller AUC.

Table 4: Prediction result under the setting: MD, $\sigma = 1/2, 1/6$, 25% censoring rate.

	OOB(TS) AUC				***OOB(TS) PE***			
Error Dist	***EV***		***Logistic***		***EV***		***Logistic***	
Scale σ	***1/2***	***1/6***	***1/2***	***1/6***	***1/2***	***1/6***	***1/2***	***1/6***
MW.ODA	.76(.79)	.90(.92)	.68(.72)	.87(.89)	1.28(1.08)	.42(.36)	1.73(1.54)	.59(.51)
MW.OA	.67(.70)	.73(.75)	.65(.67)	.73(.74)	1.46(1.38)	1.06(1.03)	1.73(1.70)	1.10(1.08)
MW.ODA.S1	.74(.79)	.89(.91)	.68(.72)	.86(.88)	1.21(1.06)	.48(.41)	1.62(1.49)	.61(.53)
MW.OA.S1	.72(.70)	.82(.76)	.67(.67)	.80(.75)	1.31(1.36)	.74(1.01)	1.66(1.68)	.83(1.06)
MW.ODA.S2	.73(.76)	.85(.87)	.68(.70)	.83(.85)	1.23(1.14)	.62(.56)	1.59(1.56)	.72(.68)
MW.OA.S2	.69(.70)	.77(.75)	.65(.66)	.76(.74)	1.37(1.37)	.92(1.01)	1.66(1.69)	.98(1.07)
MW.ODA.S3	.70(.71)	.80(.80)	.65(.65)	.78(.78)	1.33(1.32)	.82(.83)	1.64(1.70)	.90(.91)
MW.OA.S3	.66(.67)	.73(.73)	.63(.63)	.72(.72)	1.44(1.44)	1.04(1.08)	1.69(1.75)	1.08(1.13)
MW.LDA	.75(.79)	.90(.92)	.69(.72)	.87(.89)	1.17(1.04)	.40(.36)	1.55(1.47)	.56(.50)
MW.LA	.67(.70)	.73(.75)	.64(.66)	.73(.74)	1.41(1.36)	1.03(1.02)	1.66(1.67)	1.07(1.07)
MW.RFDA	.61(.63)	.62(.64)	.59(.61)	.62(.64)	1.47(1.48)	1.16(1.20)	1.67(1.75)	1.20(1.24)
MW.RFA	.58(.59)	.59(.60)	.57(.58)	.59(.60)	1.54(1.57)	1.25(1.31)	1.73(1.82)	1.29(1.35)
AFT	.79(.81)	.91(.93)	.73(.75)	.89(.90)	1.03(.95)	.37(.32)	1.41(1.36)	.49(.46)
Cox	.79(.81)	.92(.93)	.72(.74)	.88(.90)	-	-	-	-
RSF	.70(.73)	.76(.79)	.67(.69)	.75(.78)	-	-	-	-

Table 5: Prediction result under the setting: PS, $\sigma = 1/2, 1/6$, 25% censoring rate.

	OOB(TS) AUC				***OOB(TS) PE***			
Error Dist	***EV***		***Logistic***		***EV***		***Logistic***	
Scale σ	***1/2***	***1/6***	***1/2***	***1/6***	***1/2***	***1/6***	***1/2***	***1/6***
MW.ODA	.77(.75)	.91(.90)	.71(.69)	.88(.87)	1.20(1.45)	.40(.50)	1.63(2.01)	.56(.68)
MW.OA	.62(.64)	.66(.67)	.61(.62)	.66(.67)	1.66(1.62)	1.36(1.33)	1.88(1.90)	1.39(1.37)
MW.ODA.S1	.77(.79)	.90(.90)	.71(.73)	.87(.88)	1.16(1.13)	.45(.47)	1.55(1.55)	.59(.59)
MW.OA.S1	.72(.64)	.80(.68)	.68(.62)	.79(.67)	1.35(1.60)	.85(1.31)	1.67(1.87)	.92(1.35)
MW.ODA.S2	.76(.78)	.87(.87)	.70(.72)	.85(.85)	1.19(1.15)	.59(.61)	1.54(1.55)	.69(.70)
MW.OA.S2	.67(.64)	.73(.67)	.64(.61)	.72(.67)	1.49(1.60)	1.11(1.31)	1.74(1.87)	1.16(1.35)
MW.ODA.S3	.72(.73)	.82(.81)	.67(.67)	.80(.79)	1.29(1.33)	.79(.85)	1.61(1.70)	.87(.92)
MW.OA.S3	.63(.62)	.68(.65)	.61(.59)	.68(.65)	1.57(1.65)	1.25(1.36)	1.79(1.91)	1.29(1.40)
MW.LDA	.77(.79)	.92(.91)	.71(.73)	.88(.88)	1.13(1.08)	.38(.41)	1.50(1.51)	.53(.56)
MW.LA	.62(.64)	.66(.67)	.61(.61)	.66(.67)	1.60(1.59)	1.31(1.31)	1.80(1.86)	1.34(1.35)
MW.RFDA	.65(.68)	.69(.71)	.63(.65)	.68(.70)	1.46(1.49)	1.17(1.20)	1.67(1.76)	1.21(1.24)
MW.RFA	.57(.58)	.58(.59)	.56(.56)	.58(.59)	1.64(1.69)	1.40(1.45)	1.82(1.94)	1.42(1.48)
AFT	.80(.82)	.92(.94)	.74(.76)	.90(.91)	1.00(.95)	.36(.32)	1.38(1.35)	.48(.45)
Cox	.80(.82)	.93(.94)	.74(.76)	.89(.90)	-	-	-	-
RSFALL	.72(.75)	.78(.80)	.69(.71)	.77(.79)	-	-	-	-

Table 6: Estimated moderator effects under the setting: MD $\sigma = 1/2$ 25% censoring rate; for each estimator of the five non-zero moderators, the cell above shows the estimated moderator effect (and coverage rate), the cell below shows the average standard error(and empirical standard deviation), and $k = 1, 2, 3$ in MW.SDAk and MW.SAk refer to the threshold value used to select meaningful moderators.

Moderator	***M1***		***M2***		***M3***		***M4***		***M5***	
TRUE	***−.75***		***.50***		***.25***		***−1.25***		***1***	
Error Dist.	***EV***	***Logistic***	***EV***	***Logistic***	***EV***	***Logistic***	***EV***	***Logistic***	***EV***	***Logistic***
MW.ODA.U	−.77(.95)	−.79(.95)	.50(.94)	.53(.95)	.25(.95)	.24(.97)	−1.25(.96)	−1.29(.93)	1.03(.95)	1.02(.94)
	.18(.18)	.21(.21)	.19(.19)	.22(.22)	.19(.19)	.23(.21)	.36(.34)	.43(.44)	.37(.37)	.44(.44)
MW.OA.U	−.77(.96)	−.79(.95)	.50(.94)	.53(.95)	.24(.96)	.24(.96)	−1.25(.96)	−1.29(.93)	1.03(.95)	1.02(.94)
	.18(.18)	.21(.21)	.19(.19)	.22(.22)	.20(.20)	.23(.22)	.36(.34)	.43(.44)	.37(.37)	.44(.44)
MW.ODA	−.76(.96)	−.76(.97)	.49(.97)	.50(.97)	.25(.97)	.24(.98)	−1.24(.97)	−1.25(.96)	1.01(.96)	.98(.97)
	.14(.12)	.19(.18)	.14(.12)	.19(.18)	.14(.12)	.19(.16)	.26(.23)	.36(.35)	.26(.24)	.36(.33)
MW.OA	−.76(.95)	−.76(.94)	.49(.95)	.50(.94)	.24(.94)	.23(.96)	−1.24(.94)	−1.25(.93)	1.00(.94)	.98(.95)
	.16(.16)	.19(.20)	.16(.16)	.19(.20)	.16(.16)	.19(.19)	.30(.31)	.36(.40)	.30(.32)	.37(.37)
MW.ODA.S1	−.72(.95)	−.71(.94)	.45(.89)	.45(.88)	.17(.61)	.17(.55)	−1.17(.94)	−1.17(.92)	.95(.92)	.88(.87)
	.12(.11)	.17(.17)	.11(.16)	.15(.20)	.08(.16)	.10(.18)	.23(.24)	.31(.38)	.22(.28)	.29(.40)
MW.OA.S1	−.74(.94)	−.73(.93)	.47(.88)	.47(.87)	.18(.57)	.18(.52)	−1.19(.92)	−1.21(.91)	.97(.92)	.91(.87)
	.15(.15)	.18(.20)	.14(.19)	.17(.22)	.09(.18)	.10(.20)	.28(.31)	.34(.42)	.27(.34)	.32(.43)
MW.ODA.S2	−.70(.92)	−.68(.90)	.36(.70)	.34(.62)	.08(.22)	.07(.17)	−1.09(.87)	−1.06(.81)	.80(.77)	.66(.61)
	.12(.14)	.16(.23)	.09(.24)	.11(.28)	.03(.15)	.03(.16)	.22(.37)	.27(.52)	.19(.44)	.20(.56)
MW.OA.S2	−.73(.94)	−.72(.90)	.39(.69)	.37(.60)	.08(.19)	.08(.16)	−1.15(.89)	−1.14(.81)	.86(.77)	.72(.60)
	.15(.17)	.17(.25)	.11(.27)	.11(.31)	.03(.17)	.03(.18)	.27(.41)	.30(.56)	.23(.49)	.22(.60)
MW.ODA.S3	−.65(.85)	−.60(.74)	.22(.36)	.19(.26)	.02(.03)	.01(.01)	−.88(.69)	−.80(.54)	.50(.42)	.37(.27)
	.12(.26)	.14(.36)	.05(.28)	.05(.31)	.00(.09)	.00(.08)	.19(.60)	.20(.74)	.12(.58)	.10(.60)
MW.OA.S3	−.69(.87)	−.63(.73)	.24(.35)	.19(.25)	.02(.03)	.01(.01)	−.94(.68)	−.85(.52)	.55(.41)	.40(.27)
	.14(.27)	.14(.38)	.06(.31)	.05(.32)	.00(.10)	.00(.10)	.21(.64)	.21(.78)	.13(.63)	.11(.64)
MW.LDA	−.68(.93)	−.64(.91)	.42(.91)	.38(.89)	.17(.86)	.13(.87)	−1.10(.92)	−1.01(.90)	.87(.91)	.74(.88)
	.13(.12)	.18(.18)	.13(.12)	.17(.17)	.12(.12)	.15(.13)	.24(.23)	.33(.35)	.24(.25)	.33(.35)
MW.LA	−.65(.93)	−.61(.91)	.38(.88)	.35(.88)	.14(.84)	.12(.88)	−1.02(.90)	−.96(.88)	.79(.89)	.70(.87)
	.16(.17)	.20(.21)	.16(.17)	.19(.20)	.14(.15)	.16(.15)	.30(.31)	.37(.42)	.30(.34)	.36(.40)
Cox.U	−.34(.00)	−.31(.00)	.22(.02)	.19(.02)	.10(.43)	.09(.38)	−.54(.00)	−.49(.00)	.43(.02)	.38(.01)
	.07(.07)	.07(.07)	.06(.07)	.06(.07)	.07(.07)	.07(.07)	.13(.13)	.13(.13)	.13(.14)	.13(.14)
AFT.U	−.77(.88)	−.79(.94)	.50(.90)	.52(.93)	.23(.88)	.25(.96)	−1.26(.88)	−1.31(.91)	1.01(.88)	1.03(.94)
	.14(.17)	.16(.16)	.15(.17)	.17(.17)	.14(.18)	.16(.16)	.29(.35)	.33(.36)	.29(.37)	.33(.35)
Cox	−.83(.86)	−.52(.26)	.54(.91)	.35(.49)	.27(.92)	.17(.80)	−1.37(.89)	−.86(.34)	1.10(.88)	.67(.46)
	.08(.09)	.08(.09)	.08(.08)	.08(.08)	.07(.08)	.07(.08)	.16(.16)	.16(.16)	.15(.17)	.15(.17)
AFT	−.76(.94)	−.77(.94)	.50(.93)	.51(.93)	.25(.93)	.26(.94)	−1.26(.93)	−1.27(.93)	1.01(.92)	1.00(.91)
	.07(.07)	.11(.11)	.07(.07)	.11(.12)	.07(.08)	.11(.12)	.14(.15)	.22(.25)	.14(.16)	.22(.24)

Among all MW estimators, those with variable selection (MW.ODA.S, MW.OA.S and MW.LDA, MW.LA) have slightly larger biases and lower coverage probabilities than the estimators using all possible moderators (MW.ODA, MW.OA). Additionally, selections based on each moderator's standardized coefficient tend to underestimate the variability (i.e., smaller than the empirical standard deviation) as the threshold increases, while for the Lasso-based selections, the averaged standard errors are close to the empirical ones. When making treatment recommendations, ODA and LDA estimators seem to have larger AUCs and smaller PEs than other estimators across all scenarios, where the random forest method has the worst AUC. The MW estimators accounting for all possible moderators have a better performance than the "univariate" analysis (ODA.U, OA.U, AFT.U, and Cox.U).

Table 7: Estimated moderator effects under the setting: MD $\sigma = 1/6$ 25% censoring rate; for each estimator of the five non-zero moderators, the cell above shows the estimated moderator effect (and coverage rate), the cell below shows the average standard error(and empirical standard deviation), and $k = 1, 2, 3$ in MW.SDAk and MW.SAk refer to the threshold value used to select meaningful moderators.

Moderator	***M1***		***M2***		***M3***		***M4***		***M5***	
TRUE	***−.75***		***.50***		***.25***		***−1.25***		***1***	
Error Dist.	***EV***	***Logistic***	***EV***	***Logistic***	***EV***	***Logistic***	***EV***	***Logistic***	***EV***	***Logistic***
MW.ODA.U	−.77(.96)	−.78(.94)	.50(.96)	.51(.96)	.25(.95)	.25(.97)	−1.26(.96)	−1.28(.96)	1.02(.95)	1.02(.95)
	.15(.14)	.16(.15)	.16(.16)	.17(.16)	.16(.16)	.17(.14)	.30(.28)	.31(.30)	.31(.31)	.32(.31)
MW.OA.U	−.77(.95)	−.78(.94)	.50(.96)	.51(.97)	.25(.94)	.25(.97)	−1.26(.96)	−1.28(.96)	1.02(.95)	1.02(.95)
	.15(.14)	.16(.15)	.16(.16)	.17(.16)	.17(.17)	.18(.16)	.30(.28)	.31(.30)	.31(.31)	.32(.31)
MW.ODA	−.75(.98)	−.75(.98)	.50(.97)	.50(.97)	.25(.98)	.25(.98)	−1.25(.96)	−1.26(.97)	1.00(.96)	1.00(.97)
	.05(.04)	.07(.06)	.05(.04)	.06(.06)	.05(.04)	.07(.05)	.09(.08)	.12(.11)	.09(.08)	.12(.11)
MW.OA	−.75(.94)	−.76(.92)	.48(.93)	.49(.94)	.25(.93)	.24(.92)	−1.24(.95)	−1.25(.94)	.99(.93)	.99(.94)
	.12(.12)	.12(.13)	.12(.12)	.12(.13)	.12(.13)	.12(.13)	.22(.23)	.23(.24)	.22(.25)	.23(.25)
MW.ODA.S1	−.73(.95)	−.74(.94)	.48(.93)	.48(.93)	.18(.70)	.18(.69)	−1.22(.94)	−1.23(.94)	.97(.94)	.97(.94)
	.05(.05)	.06(.06)	.04(.09)	.06(.09)	.03(.12)	.04(.13)	.09(.10)	.12(.13)	.09(.12)	.12(.16)
MW.OA.S1	−.75(.94)	−.74(.93)	.48(.91)	.48(.93)	.19(.64)	.19(.63)	−1.22(.94)	−1.23(.93)	.97(.92)	.96(.91)
	.11(.12)	.12(.12)	.11(.14)	.11(.14)	.07(.16)	.08(.16)	.21(.23)	.22(.24)	.21(.24)	.22(.26)
MW.ODA.S2	−.72(.93)	−.72(.92)	.42(.83)	.42(.82)	.09(.32)	.08(.30)	−1.18(.92)	−1.19(.93)	.86(.84)	.84(.83)
	.06(.07)	.07(.09)	.05(.17)	.06(.18)	.02(.13)	.02(.13)	.12(.17)	.14(.20)	.10(.31)	.12(.35)
MW.OA.S2	−.75(.93)	−.75(.93)	.44(.82)	.44(.82)	.10(.26)	.09(.24)	−1.22(.93)	−1.23(.93)	.91(.85)	.90(.83)
	.11(.12)	.12(.13)	.10(.20)	.10(.21)	.03(.16)	.03(.16)	.21(.26)	.22(.28)	.19(.37)	.20(.40)
MW.ODA.S3	−.70(.91)	−.70(.90)	.29(.56)	.28(.54)	.02(.06)	.02(.06)	−1.06(.85)	−1.07(.85)	.62(.60)	.59(.56)
	.09(.14)	.09(.17)	.05(.25)	.05(.26)	.00(.08)	.01(.09)	.15(.40)	.16(.42)	.10(.50)	.10(.52)
MW.OA.S3	−.74(.92)	−.73(.91)	.30(.52)	.30(.51)	.02(.04)	.03(.05)	−1.13(.86)	−1.14(.85)	.68(.58)	.65(.56)
	.12(.16)	.12(.19)	.06(.29)	.06(.29)	.01(.10)	.01(.10)	.20(.44)	.21(.46)	.14(.56)	.13(.57)
MW.LDA	−.73(.95)	−.72(.92)	.48(.93)	.42(.82)	.22(.93)	.08(.30)	−1.21(.93)	−1.19(.93)	.96(.94)	.84(.83)
	.04(.04)	.07(.09)	.04(.04)	.06(.18)	.05(.04)	.02(.13)	.08(.08)	.14(.20)	.08(.08)	.12(.35)
MW.LA	−.68(.91)	−.75(.93)	.41(.88)	.44(.82)	.16(.82)	.09(.24)	−1.09(.90)	−1.23(.93)	.85(.91)	.90(.83)
	.12(.13)	.12(.13)	.12(.12)	.10(.21)	.11(.13)	.03(.16)	.22(.23)	.22(.28)	.22(.25)	.20(.40)
Cox.U	−.13(.00)	−.12(.00)	.08(.00)	.08(.00)	.04(.00)	.04(.00)	−.20(.00)	−.19(.00)	.16(.00)	.15(.00)
	.02(.03)	.02(.03)	.02(.02)	.02(.02)	.02(.03)	.02(.03)	.05(.05)	.05(.05)	.05(.05)	.05(.05)
AFT.U	−.76(.87)	−.78(.94)	.50(.89)	.51(.94)	.23(.87)	.25(.94)	−1.26(.88)	−1.29(.93)	1.01(.88)	1.03(.94)
	.13(.16)	.13(.13)	.13(.16)	.13(.14)	.12(.16)	.12(.12)	.26(.33)	.26(.29)	.26(.33)	.26(.28)
Cox	−.83(.73)	−.50(.00)	.55(.80)	.33(.02)	.27(.89)	.17(.24)	−1.37(.75)	−.83(.00)	1.10(.77)	.66(.01)
	.05(.06)	.04(.05)	.04(.05)	.03(.05)	.03(.04)	.03(.04)	.09(.10)	.07(.10)	.08(.09)	.06(.09)
AFT	−.75(.94)	−.76(.94)	.50(.93)	.50(.93)	.25(.93)	.25(.94)	−1.25(.93)	−1.26(.92)	1.00(.91)	1.00(.91)
	.02(.02)	.04(.04)	.02(.03)	.04(.04)	.02(.03)	.04(.04)	.05(.05)	.07(.09)	.05(.05)	.07(.08)

As the true model in our simulation studies, the AFT model performs well in general, with a smaller bias, estimated/empirical SD, PE, and a larger AUC. Even though the Cox PH model tends to have a larger bias, slightly underestimate the variability, and consequently have a lower coverage probability than the AFT model and our proposed MW estimators, it still seems robust enough to classify patients to suitable treatments when the proportional hazards assumption is violated. The RSF approach yields a smaller AUC than both AFT and Cox, especially when the error variance is small. Compared to competitive methods, when making recommendations, the ODA and LDA estimators have similar AUCs and PEs as the AFT and Cox PH models, where the MW estimators are more sensitive to the increase in noise. When estimating the causal effect modifiers, the AFT and Cox have a smaller estimated standard error than the

empirical standard deviation, and thus have a smaller coverage probability than ODA and LDA estimators. In general, the ODA estimator and the AFT model have a smaller bias.

4 Real Data Application

The Strategies to Avoid Returning to Smoking (STARTS) study was conducted on 300 women from September 2007 to June 2014 in Pittsburgh, PA, USA (Levine et al., 2013). It was a randomized controlled trial aiming to assess the effect of a 24-week cognitive behavioral therapy (CBT) on postpartum smoking relapse prevention, as compared to a standard supportive behavioral therapy (SBT) with fewer interventions. The primary endpoint was the biochemically confirmed sustained tobacco abstinence within 52 weeks postpartum. Then, the time to relapse was determined by counting the number of days between delivery and the first day of 7 consecutive days of smoking.

To illustrate the use of our proposed methodology, we chose thirteen baseline variables as the moderator candidates, including age in years, motivation to stay quit, the number of previous quit attempts, Fagerstrom (FAGR) test score for nicotine dependence, Smoking Self-Efficacy Questionnaire (SEQ-12) score, smoking year to age ratio, the number of cigarettes smoked daily, Edinburgh Postnatal Depression Scale (EPDS) (higher vs. not), Perceived Stress Scale (PSS) (higher vs. not), race (black vs. Others), income level (household income below \$30k/yr vs. not), parity and education background (High school or equivalent vs. not), after considering clinical rationales, missing data, and substantial collinearity with others.

Among 268 women with complete data, the censoring rate is 22.8%. Then, 103 matched pairs were created via the 1:1 NNM algorithm with MD, as it yields a more negligible matching bias. The analysis results on STARTS data, including the OOB PE/AUC of Cox, RSF, MW estimators combined with different methods and their causal effect estimators, are tabulated in Table 8 and Table 9, respectively.

Based on Table 8, we observe that all MW estimators have similar PE and AUC. When compared with Cox and RSF, MW methods have slightly larger AUCs. The AUCs from all are generally around 0.60, indicating that the data are noisy. According to Tables 8 and 9, Cox fails to select any significant moderator. At the same time, MW estimators, in general, reveal that

Table 8: Prediction result of MW estimators on STARTS data.

OOB	**PE**	**AUC**
Methods	**Mean(SD)**	**Mean(SD)**
MW.ODA	1.59(.20)	.62(.08)
MW.OA	1.46(.16)	.61(.07)
MW.ODA.S1	1.52(.18)	.61(.08)
MW.ODA.S2	1.49(.17)	.57(.08)
MW.OA.S1	1.50(.17)	.61(.08)
MW.OA.S2	1.49(.17)	.57(.07)
MW.LDA	1.43(.18)	.61(.07)
MW.LA	1.42(.16)	.60(.07)
Cox	-	.54(.08)
RSFALL	-	.57(.08)

Table 9: Analysis of STARTS data using the Cox and MW estimators.

Moderator ($\hat{\theta}_{mo}$, **p**)	Cox	AllA	AllDA	SA1	SDA1	SA2	SDA2	LA	LDA
Age	−.12(.55)	−.11(.60)	−.03(.88)	−.05(.79)	−.02(.90)				
Motivation	.19(.46)	**.69(.01)**	**.71(.01)**	**.67(.01)**	**.76(.00)**			.49	.60
No. of Quit	.07(.63)	**−.33(.01)**	**−.46(.00)**	**−.30(.03)**	**−.37(.01)**	**−.26(.06)**		−.22	−.30
FAGR	−.25(.22)	−.16(.42)	−.19(.36)	−.13(.53)	−.17(.40)			−.12	−.14
SEQ12	−.06(.70)	−.05(.74)	−.04(.80)						
Smoking year(%)	.25(.19)	**−.40(.06)**	**−.40(.05)**	**−.38(.07)**	**−.35(.07)**	**−.35(.03)**	**−.41(.01)**	−.31	−.34
No.cigarettes	.25(.28)	−.16(.60)	−.14(.64)	−.20(.51)	−.16(.58)			−.03	−.04
EPDS(high)	−.66(.11)	**.71(.07)**	**.68(.10)**					.24	.36
PSS(high)	.40(.28)	**−.98(.01)**	**−1.34(.00)**	**−.65(.04)**	**−1.00(.00)**	**−.81(.01)**	**−.89(.00)**	−.62	−.95
Black	.38(.29)	−.46(.16)	**−.64(.06)**	−.51(.11)	**−.63(.04)**			−.29	−.37
Income (⩽30k/yr)	−.04(.93)	−.21(.61)	.18(.68)	−.09(.83)	.30(.44)				
Nulliparous	−.12(.71)	−.20(.52)	−.26(.41)						−.01
⩽Highschool	.25(.46)	−.12(.70)	−.35(.29)						−.08

women with stronger motivation, fewer quit attempts, shorter lengths of smoking concerning their age, higher EPDS screening scores, and milder perceived stress and those who were not identified as African American would benefit more from the CBT than the SBT. If we used the combined MW estimator from the LDA model, the one with relatively larger OOBAUC and smaller OOBPE, as our optimal estimator to make treatment recommendations, then, for the 103 matched pairs, 47 of them would be assigned to the CBT group and the rest to SBT. Furthermore, the mean (SD) of time to smoking relapse among the 103 CBT-treated patients is 18.2 (15.3) weeks before re-assignment. After assigning the rest to SBT, the mean (SD) of the modified time to smoking relapse becomes 24.5 (16.1). The roughly six weeks improvement suggests the usefulness of the recommendation by our proposed method.

5 Discussion

In this paper, we proposed an intuitive and readily implementable framework for estimating causal effect modifiers and making treatment recommendations for a study with survival outcomes. Our approach can be easily applied using well-established R packages, and the resulting optimal combined moderator has a clear and straightforward interpretation. Our framework is built upon matching, which might yield a non-negligible bias. We explored the impact of matching imbalance on the performance of our estimators of causal effect modifiers. With a larger matching imbalance, the bias and estimated and empirical standard deviations also increase. When making treatment recommendations, a larger matching bias could degrade the performance with a smaller AUC and a larger predicted error. However, adjusting the paired differences (DM) in the model provides a more robust result in the presence of a larger matching bias, and matching bias has a limited impact on the performance of our estimators. In the literature, there are other methods that do not require matching, e.g., the method in Tian et al. (2014). However, those methods are more complicated and less straightforward to interpret. Our goal was to find an intuitive method to be used by practitioners to make personalized recommendations for survival outcomes. Thus, we made the trade-off between some bias and the simplicity of the method.

In general, modeling the CATE on a composite of moderator candidates provides higher precision and a more significant effect than exploring the moderator effect univariately. The optimal MW estimator could achieve similar performance as the results of the AFT model with prior knowledge of the error distribution. We also observe that even though the PH assumption is violated, the Cox PH model is often robust enough to make treatment recommendations.

The proposed methods can also be adapted to scenarios with nonlinear effects. If only the main effect exhibits nonlinearity, it will impact the estimation of the intercept θ_d but not the moderator effect θ_{mo}, and our proposed methods remain valid. If both the main effect and the interactions are nonlinear, the matching framework simplifies the detection of the nonlinear pattern, as one can plot the residuals from a linear model versus a potential moderator. However, interpreting a nonlinear moderation effect can be notably challenging. In practice, a dichotomized moderator is often employed as a workaround.

One limitation of the MW estimator is that it fails to make precise treatment recommendations with considerable noise, a common problem faced by traditional methods as well. The other disadvantage of our MW framework is that it is subject to matching performance, where the imbalance can be enormous in a high-dimensional setting. Therefore, future studies could adopt high-dimensional matching methods with penalization methods like Lasso to extend the MW framework to a high-dimensional setting. Nevertheless, the proposed methods provide a straightforward and intuitive framework for practitioners to explore heterogeneous treatment effects. More importantly, although we used an RCT study as our data example, the matching framework is more useful for observational studies to draw any causal inference on CATE.

Supplementary Material

Some additional simulation results and a compressed folder with the code to simulate the settings with 5 moderators (5M), implement our proposed methods, and some existing methods are provided as the online Supplementary Material.

Funding

This work was partially supported by the National Science Foundation Division of Mathematical Sciences (1916001 to Y.C.) and by the University of Pittsburgh Center for Research Computing, RRID:SCR-022735, through the resources provided. Specifically, this work used the H2P cluster, which is supported by NSF award number OAC-2117681.

References

Breiman L (2001). Random forests. *Machine Learning*, 45: 5–32. https://doi.org/10.1023/A:1010933404324

Chen S, Tian L, Cai T, Yu M (2017). A general statistical framework for subgroup identification and comparative treatment scoring. *Biometrics*, 73(4): 1199–1209. https://doi.org/10.1111/biom.12676

Chin Fatt CR, Jha MK, Cooper CM, Fonzo G, South C, Grannemann B, et al. (2020). Effect of intrinsic patterns of functional brain connectivity in moderating antidepressant treatment response in major depression. *The American Journal of Psychiatry*, 177(2): 143–154. https://doi.org/10.1176/appi.ajp.2019.18070870

Cox DR (1972). Regression models and life-tables. *Journal of the Royal Statistical Society, Series B, Methodological*, 34(2): 187–202.

Cui Y, Zhu R, Kosorok M (2017). Tree based weighted learning for estimating individualized treatment rules with censored data. *Electronic Journal of Statistics*, 11(2): 3927–3953.

Goldberg Y, Kosorok MR (2012). Q-learning with censored data. *The Annals of Statistics*, 40(1): 529–560. https://doi.org/10.1214/12-AOS968

Hildebrandt T, Michaeledes A, Mayhew M, Greif R, Sysko R, Toro-Ramos T, et al. (2020). Randomized controlled trial comparing health coach-delivered smartphone-guided self-help with standard care for adults with binge eating. *The American Journal of Psychiatry*, 177(2): 134–142. https://doi.org/10.1176/appi.ajp.2019.19020184

Ho D, Imai K, King G, Stuart EA (2011). Matchit: Nonparametric preprocessing for parametric causal inference. *Journal of Statistical Software*, 42(8): 1–28. https://doi.org/10.18637/jss.v042.i08

Ho DE, Imai K, King G, Stuart EA (2007). Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political Analysis*, 15(3): 199–236. https://doi.org/10.1093/pan/mpl013

Ishwaran H, Kogalur UB (2007). Random survival forests for R. *R News*, 7(2): 25–31.

Ishwaran H, Kogalur UB, Blackstone EH, Lauer MS (2008). Random survival forests. *Annals of Applied Statistics*, 2(3): 841–860. https://doi.org/10.1214/08-AOAS169

Kaneriya SH, Robbins-Welty GA, Smagula SF, Karp JF, Butters MA, Lenze EJ, et al. (2016). Predictors and moderators of remission with aripiprazole augmentation in treatment-resistant late-life depression: An analysis of the IRL-GRey randomized clinical trial. *JAMA Psychiatry*, 73(4): 329–336. https://doi.org/10.1001/jamapsychiatry.2015.3447

Kaplan EL, Meier P (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53(282): 457–481. https://doi.org/10.1080/01621459.1958.10501452

King G, Nielsen R, Coberley C, Pope JE, Wells A (2011). Comparative effectiveness of matching methods for causal inference. Cambridge, MA. Unpublished manuscript.

Kraemer HC (2013). Discovering, comparing, and combining moderators of treatment on outcome after randomized clinical trials: A parametric approach. *Statistics in Medicine*, 32(11): 1964–1973. https://doi.org/10.1002/sim.5734

Kraemer HC, Wilson GT, Fairburn CG, Agras WS (2002). Mediators and moderators of treatment effects in randomized clinical trials. *Archives of General Psychiatry*, 59(10): 877–883. https://doi.org/10.1001/archpsyc.59.10.877

Levine MD, Cheng Y, Marcus MD, Kalarchian MA, Emery RL (2016). Preventing postpartum smoking relapse: A randomized clinical trial. *JAMA Internal Medicine*, 176(4): 443–452. https://doi.org/10.1001/jamainternmed.2016.0248

Levine MD, Marcus MD, Kalarchian MA, Cheng Y (2013). Strategies to avoid returning to smoking (STARTS): A randomized controlled trial of postpartum smoking relapse prevention interventions. *Contemporary Clinical Trials*, 36(2): 565–573. https://doi.org/10.1016/j.cct.2013.10.002

Liang M, Yu M (2022). A semiparametric approach to model effect modification. *Journal of the American Statistical Association*, 117(538): 752–764. https://doi.org/10.1080/01621459.2020.1811099

Liaw A, Wiener M, et al. (2002). Classification and regression by randomforest. *R News*, 2(3): 18–22.

Mo W, Liu Y (2022). Efficient learning of optimal individualized treatment rules for heteroscedastic or misspecified treatment-free effect models. *Journal of the Royal Statistical Society, Series B, Statistical Methodology*, 84(2): 440–472. https://doi.org/10.1111/rssb.12474

Niles AN, Wolitzky-Taylor KB, Arch JJ, Craske MG (2017). Applying a novel statistical method to advance the personalized treatment of anxiety disorders: A composite moderator of comparative drop-out from cbt and act. *Behaviour Research and Therapy*, 91: 13–23. https://doi.org/10.1016/j.brat.2017.01.001

Park H, Petkova E, Tarpey T, Ogden RT (2022). A sparse additive model for treatment effect-modifier selection. *Biostatistics*, 23(2): 412–429. https://doi.org/10.1093/biostatistics/kxaa032

Rubin DB (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5): 688–701. https://doi.org/10.1037/h0037350

Rubin DB (2005). Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association*, 100(469): 322–331. https://doi.org/10.1198/016214504000001880

Smagula SF, Wallace ML, Anderson SJ, Karp JF, Lenze EJ, Mulsant BH, et al. (2016). Combining moderators to identify clinical profiles of patients who will, and will not, benefit from aripiprazole augmentation for treatment resistant late-life major depressive disorder. *Journal of Psychiatric Research*, 81: 112–118. https://doi.org/10.1016/j.jpsychires.2016.07.005

Song R, Luo S, Zeng D, Zhang HH, Lu W, Li Z (2017). Semiparametric single-index model for estimating optimal individualized treatment strategy. *Electronic Journal of Statistics*, 11(1): 364–384.

Stuart EA (2010). Matching methods for causal inference: A review and a look forward. *Statistical science: A Review Journal of the Institute of Mathematical Statistics*, 25(1): 1–21. https://doi.org/10.1214/09-STS313

Tian L, Alizadeh AA, Gentles AJ, Tibshirani R (2014). A simple method for estimating interactions between a treatment and a large number of covariates. *Journal of the American Statistical Association*, 109(508): 1517–1532. https://doi.org/10.1080/01621459.2014.951443

Tibshirani R (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B, Methodological*, 58(1): 267–288.

Wallace ML, Banihashemi L, O'Donnell C, Nimgaonkar VL, Kodavali C, McNamee R, et al. (2018). Using optimal combined moderators to define heterogeneity in neural responses to randomized conditions: Application to the effect of sleep loss on fear learning. *NeuroImage*, 181: 718–727. https://doi.org/10.1016/j.neuroimage.2018.07.051

Wallace ML, Frank E, Kraemer HC (2013). A novel approach for developing and interpreting treatment moderator profiles in randomized clinical trials. *JAMA Psychiatry*, 70(11): 1241–1247. https://doi.org/10.1001/jamapsychiatry.2013.1960

Wallace ML, Smagula SF (2018). Promise and challenges of using combined moderator methods to personalize mental health treatment. *The American Journal of Geriatric Psychiatry*, 26(6): 678–679. https://doi.org/10.1016/j.jagp.2018.02.001

Yadlowsky S, Pellegrini F, Lionetto F, Braune S, Tian L (2021). Estimation and validation of ratio-based conditional average treatment effects using observational data. *Journal of the American Statistical Association*, 116(533): 335–352. https://doi.org/10.1080/01621459.2020.1772080

Zhao YQ, Zeng D, Laber EB, Song R, Yuan M, Kosorok MR (2015). Doubly robust learning for estimating individualized treatment with censored data. *Biometrika*, 102(1): 151–168. https://doi.org/10.1093/biomet/asu050