Value Function Guided Subgroup Identification via Gradient Tree Boosting: A Framework to Handle Multiple Outcomes for Optimal Treatment Recommendation

Pingye Zhang\textsuperscript{a}, Peng Liu\textsuperscript{b}, Junshui Ma\textsuperscript{a}, and Yue Shentu\textsuperscript{a}

\textsuperscript{a}MRL, Merck & Co., Inc., Rahway, NJ; \textsuperscript{b}Department of Biostatistics, University of Pittsburgh, Pittsburgh, PA

**ABSTRACT**

In randomized clinical trials, there has been an increasing interest in identifying subgroups with heterogeneous responses to study treatment based on baseline patient characteristics. Even though the benefit risk assessment of any patient population or subgroups is almost always a multi-facet consideration, the statistical literature of subgroup identification has largely been limited to a single clinical outcome. In the article, we propose a nonparametric method that searches for subgroup membership scores by maximizing a value function that directly reflects the subgroup-treatment interaction effect considering clinical priorities of multiple outcomes based on win difference. A gradient tree boosting algorithm is proposed to search for the individual subgroup membership scores. We conduct simulation studies to evaluate the performance of the proposed method and an application to a colon cancer adjuvant clinical trial is performed for illustration.

**ARTICLE HISTORY**

Received December 2020
Accepted August 2021

**KEYWORDS**

Gradient tree boosting; Nonparametric; Subgroup identification; Win difference

1. Introduction

It is widely recognized that randomized clinical trials (RCT) are the cornerstones for evidence-based medicine, but the benefit/risk assessment based on the entire trial population may not answer all the medically interesting questions. It is not appropriate to assume a priori that the patient populations studied in clinical trials are homogeneous in their response to treatments. An important aspect of regulatory and health technology assessment is to understand the potential heterogeneity in the treatment effect of a novel therapy (EMA 2019). On the other hand, a sponsor may also want to identify the patient population that benefits the most from the experimental treatment. In some cases, specific biomarker hypotheses provide the opportunity to optimize the development strategy in costly Phase III programs (Reck et al. 2016); in other cases, insights from subgroup analyses in completed trials in combination with advance in science have led to genuine breakthroughs. While the risk of inappropriate subgroup analysis and data dredging has long been recognized in the biostatistics literature (Assmann et al. 2000), some authors have cautioned against the dismissal of all exploratory subgroup analyses (Dmitrienko et al. 2016).

Traditional “fully stratified” subgroup analyses or treatment-by-factor interaction tests in linear or generalized linear models are known to suffer from both multiplicity and low power (Cui et al. 2002). Instead of focusing on a predefined list of subgroups, another perspective is to define a reproducible subgroup identification strategy so that its statistical properties can be appropriately assessed and reported (Foster et al. 2011). Interaction tree (Su et al. 2008), virtual twin (Foster et al. 2011), modified covariate method (Tian et al. 2014), SIDES (Lipkovich et al. 2011) and GUIDE (Loh et al. 2015) are only a few representative examples of the many subgroup identification methods that have been proposed in the last decade. See Lipkovich et al. (Lipkovich et al. 2017) for a comprehensive review. Another class of closely related methods fall under the umbrella of individualized treatment recommendation, or optimal treatment regime (Qian and Murphy 2011; Zhao et al. 2012; Zhang et al. 2012). Although the objective is to identify the optimal treatment for any given subjects, the result of optimal treatment regime modeling is often two complementary subgroups with qualitatively different treatment effects.

Despite the fact that the benefit risk assessment of any patient population or subgroups is almost always a multi-facet consideration, the statistical literature of subgroup identification and optimal treatment regime has largely been limited to a single outcome. Some recent research attempt to connect the efficacy and toxicity considerations by recasting it as a restricted optimization problem, or consider the tradeoff explicitly in a utility function (Schnell et al. 2017). In addition, with an ordinal composite outcome, one could also maximize a specified quantile of the ordinal composite outcome instead of assigning arbitrary numerical values and maximizing the mean (Liu et al. 2017). An alternative approach that incorporates multiple efficacy and safety endpoints may be formulated using win ratio (Pocock et al. 2012) or win difference (net benefit) (Luo et al. 2015; Péron et al. 2018). For any pair of patients from the treatment and control group in an RCT (possibly stratified by risk profiles), a “win,” “loss,” or “tie” is defined by the comparisons of a hierarchy of component endpoints in descending order. The component
endpoints may be efficacy or safety variables of any type, such as continuous, categorical, time-to-events, or even recurrent events. If the overall benefit risk assessment is based on the win ratio or win difference, then a natural question to ask is whether it is possible to identify subgroups with heterogeneous treatment effects based on these estimands.

Because win ratio/difference are based on pairwise comparisons between treatment groups, subgroup identification methods based on the modeling of individual outcomes cannot be directly implemented. Similarly, optimal treatment regime methods such as outcome weighted learning are not suitable. This is because a subgroup needs to be defined first before its win ratio/difference can be calculated, whereas the subgroups or the optimal treatment regime produced by the subgroup identification or optimization algorithms implicitly or explicitly maximize the difference in mean treatment effect or proportion of responders. In order to identify a subgroup with enhanced win ratio/difference, it may appear that only a direct search strategy is feasible. When there are only a few potential categorical baseline covariates, enumeration of all possible subgroups may be possible. However, with even a moderate number of candidate demographic variables and biomarkers (e.g., between 10 and 20 continuous or categorical variables), this strategy quickly breaks down due to the number of potential subgroups that need to be considered.

Recently, Zhang et al. proposed a nonparametric subgroup identification approach that optimizes a value function using gradient tree boosting (Zhang et al. 2020). Although their original work only focused on a restricted mean survival time value function for survival outcomes, the optimization procedure may be extended to other value functions. The gradient tree boosting algorithm essentially searches for the optimal classification into two complementary subgroups (treatment performing, treatment non-performing) along the gradients for each subject’s probability of belonging to the treatment performing subgroup. Since at every step of the boosting algorithm, a probabilistic classification of the two dueling subgroups exists, an estimand for each of these two “subgroups” may be estimated and they together define the value function.

In this article, we propose to extend the value function guided subgroup identification approach to win difference. This method provides the foundation for a more flexible benefit/risk optimization framework incorporating multiple outcomes with transparently defined hierarchy of clinical importance.

The remainder of the article is organized as follows. Value functions based on the win probabilities are introduced in Section 2, followed by a review of the gradient tree boosting algorithm for subgroup identification. In Section 3, the performance of the subgroup identification method is demonstrated in a simulation study that considers time to progression and overall survival (OS) as component efficacy outcomes in a win difference estimand. In Section 4, we implement the proposed method in a real-life dataset for a colon cancer adjuvant trial. The article is concluded with a discussion in Section 5.

2. Materials and Methods

Assume we have a two-armed randomized controlled trial in which treatment is equally assigned to N patients from a population of interest. We let A = 0 or 1 be the binary treatment indicator for control and experimental treatment, respectively. Suppose \( R = (Y_1, \ldots, Y_m) \) denote m clinical outcomes and we assume \( Z \) as a d-dimensional baseline patient characteristic vector. The observed data are then \( \{R_i = (Y_{1i}, \ldots, Y_{mi}), A_i, Z_i\} \) for \( i = 1, 2, \ldots, N \), which are assumed to be independently and identically distributed. Consider patients i and j from experimental treatment and control group, respectively, we use \( R_i(1) > R_j(0) \) to indicate that the overall clinical outcome for patient i in experimental treatment group is better than it for patient j in control group based on certain comparison rule prioritizing the m individual clinical outcomes; and use \( R_i(1) < R_j(0) \) to indicate that the overall clinical outcome for patient i is worse than it for patient j.

2.1. Proposed Value Function

For a single survival outcome, Zhang et al. (2020) proposed a nonparametric method that searches for subgroup membership scores by maximizing a value function that directly reflects the subgroup—treatment interaction effect based on restricted mean survival time. In this article, we follow the same strategy of directly searching for maximized differential treatment effect at the subgroup level while replacing the restricted mean survival time with the win difference (Pocock et al. 2012; Luo et al. 2015; Péron et al. 2018), which can be extended to quantify the treatment effect with multiple prioritized clinical outcomes. Specifically, we propose the below value function as a measure of the clinical interest associated with the subgroups to be identified.

\[
V(τ) = P(S) \left[ Pr\left[ R_i(1) > R_j(0) | S \right] - Pr\left[ R_i(1) < R_j(0) | S \right] \right] \\
- P(S^c) \left[ Pr\left[ R_i(1) > R_j(0) | S^c \right] - Pr\left[ R_i(1) < R_j(0) | S^c \right] \right] 
\]

(1)

where \( S \) is a subset of a patient population set \( U \) in which experimental treatment leads to better prioritized clinical outcomes (treatment performing subgroup); \( S^c \) is the complementary subgroup in which experimental treatment does not lead to better prioritized clinical outcomes (treatment nonperforming subgroup). The subgroup win probability \( Pr\left[ R_i(1) > R_j(0) | S \right] \) is the probability for a patient i in experimental treatment group that has better prioritized clinical outcomes than a patient j in the control group in the treatment performing subgroup; \( Pr\left[ R_i(1) < R_j(0) | S \right] \) is the probability to lose in the treatment performing subgroup and \( WD_D(S) = Pr\left[ R_i(1) > R_j(0) | S \right] - Pr\left[ R_i(1) < R_j(0) | S \right] \) is the win difference with a positive value indicating a treatment effect in favor of the experimental treatment. Similarly, \( WD_D(S^c) = Pr\left[ R_i(1) > R_j(0) | S^c \right] - Pr\left[ R_i(1) < R_j(0) | S^c \right] \) is the win difference in the treatment nonperforming subgroup. The proposed value function can be viewed as a measure of differential experimental treatment effects, based on prioritized clinical outcomes, across subgroups weighted by the subgroup prevalence.

Pocock et al. (2012) proposed an approach to estimate the win and loss probability. They compare every patient on experimental treatment with every patient on control and each time noting who win and who lose. Specifically, for two patients i and j, the win indicator is \( W_{ij} = 1 \left[ R_i > R_j \right] \)
and the loss indicator is $L_{ij} = I [R_i < R_j]$. Let $N_w$ and $N_i$ be the number of winners and losers for experimental treatment in which $N_w = \sum_{r=1}^N \sum_{j=1}^N A_r (1 - A_j) W_{ij}$ and $N_i = \sum_{r=1}^N \sum_{j=1}^N A_r (1 - A_j) L_{ij}$. The win ratio is defined as $N_w/N_i$, and win difference is defined as $N_w - N_i$. Following the same strategy in Zhang et al. (2020), we propose to estimate the win difference in the treatment performing subgroup and treatment nonperforming subgroup as,

$$
\hat{W}_D (S) = \sum_{i=1}^N \sum_{j=1}^N p (Z_i) p (Z_j) A_r (1 - A_j) \left( W_{ij} - L_{ij} \right),
$$

$$
\hat{W}_D (S') = \sum_{i=1}^N \sum_{j=1}^N \left[ 1 - p (Z_i) \right] \left[ 1 - p (Z_j) \right] A_r (1 - A_j) \left( W_{ij} - L_{ij} \right),
$$

where $p (Z_i)$ and $p (Z_j)$ as the subgroup membership score quantifying the likelihood of belonging to the treatment performing subgroup for $i$ and $j$, respectively. The value function in (1) can then be estimated as follows:

$$
\hat{V} (\tau) = \left( \sum_{i=1}^N p_i \right) \hat{W}_D (S) - \left( \sum_{i=1}^N (1 - p_i) \right) \hat{W}_D (S').
$$

The goal is to estimate the subgroup membership score $p (Z_i)$ for patients $i = 1, 2, \ldots, N$ by maximizing the value function estimator in (3). Once $p (Z_i)$ are estimated, we can stratify the overall population to treatment performing subgroup with patients with $p (Z_i) > c$, and to treatment nonperforming subgroup with patients with $p (Z_i) \leq c$, where $c$ is a cutoff. For example, $c$ can be 0.5. Note that we use $p (Z_i)$ instead of $I [p (Z_i) > c]$ in $\hat{W}_D (S)$ when estimating win difference in treatment performing subgroup; and use $1 - p (Z_i)$ instead of $I [p (Z_i) \leq c]$ in $\hat{W}_D (S')$ when estimating win difference in treatment nonperforming subgroup. As a result, the value function estimator (3) is a smooth and differentiable approximation to the value function (1) and the gradient of (3) can be used to solve the optimization problem.

### 2.2. Estimation Via Gradient Tree Boosting

The gradient tree boosting is an ensemble method that constructs a predictive model by additive expansions of decision trees (Friedman 2001). For the proposed method, we define $F_i$ as a logit function of $p (Z_i)$ for patient $i = 1, \ldots, n$.

$$
F_i = \log \frac{p (Z_i)}{1 - p (Z_i)}.
$$

The final prediction $\hat{F}_i$ of $F_i$ follows an additive expansion of K base tree functions.

$$
\hat{F}_i = \sum_{k=1}^K \hat{f}_k (R_i, A_i, Z_i),
$$

where $\hat{f}_k$ is the kth base tree function and $K$ is the number of trees. To learn the set of base tree functions, the following regularized objective function is minimized at each of the K iterations (Chen and Guestrin 2016):

$$
\text{obj}(k) = -\hat{V} \left( \left[ \hat{F}_1^{(k-1)} + f_k (R_1, A_1, Z_1), \ldots, \hat{F}_n^{(k-1)} + f_k (R_n, A_n, Z_n) \right]^T \right) + \Omega
$$

where $\hat{F}_i^{(k-1)} = \sum_{j=1}^{k-1} f_j (R_i, A_i, Z_i)$ is the prediction up to the $(k - 1)$th base tree function and $\Omega$ is a penalty function that penalizes the complexity of the tree functions. We use the same penalty function as defined in “xgboost” (Chen and Guestrin 2016). Note that the loss function, the negative value function, is not a direct summation of individual loss measuring counterfactual treatment effect at individual level. Instead, the loss function is a measure of differential treatment effect at the subpopulation level. The loss function is differentiable with respect to $F_i$ and the first-order approximation can be used to optimize the objective function. Specifically, at the kth iteration, the first-order gradient of the loss function evaluated at $\hat{F}_i^{(k-1)}$, $g_i^k = \left( \partial (-\hat{V}) / \partial \hat{F}_i \right)_{F_i=\hat{F}_i^{(k-1)}}$, can be used as the individual target label in the kth tree to construct the prediction of $f_k (R_i, A_i, Z_i)$. We provide the derivation of the first-order gradient in the supplementary material. The goal is to optimize the final prediction for $F_i$ to minimize the objective function and we can classify patients based on their estimated subgroup membership score $p (Z_i)$. The numerical optimization can be implemented with commonly used software such as “xgboost.”

### 3. Simulation Studies

We conduct simulation studies to evaluate the performance of the proposed method. Specifically, we consider a two-armed 1:1 RCT with a uniform enrollment of 12 months followed by an 18 months study follow-up period (Figure 1). There are two survival endpoints of interest: time to progression (TTP) and OS. TTP is defined as the time from randomization to disease progression. OS is defined as the time from randomization to death due to any cause, which is the clinically more important endpoint. In this simulation study, OS is a more accurate measurement of clinical benefit but subject to censoring because death may not occur during the trial for many patients. TTP is less informative, however, it can be a surrogate measurement of clinical benefit for treatment when OS is censored.

The clinical endpoints OS and TTP are simulated as follows: We use $T_{\text{enroll}}$ to denote the time from the start of trial to the patient’s entry (randomization) in which $T_{\text{enroll}} \sim \text{unif} (0, 12)$. $T_{PD}$ is the time from patient’s entry to the onset of disease progression. $T_{PDtoDeath}$ denotes the time from disease progression to Death. We assume $T_{PD}$ and $T_{PDtoDeath}$ follows log normal distribution: $T_{PD} = \exp (\beta_0 + \beta_1 + A * g (S) + \epsilon)$, where $\beta_0 = \sqrt{3}$ and $\epsilon \sim N (0, 1)$, and $T_{PDtoDeath} = \exp (\beta_0 + \beta_1 * A * g (S) + \epsilon)$, where we set $\beta_0 = \sqrt{3}, \sqrt{6}, \sqrt{10}$ or $\sqrt{14}$ to generate increasing level of censoring for OS, and let $\beta_1 < \beta_2$ such that $T_{PDtoDeath}$ carries more information toward treatment by subgroup interaction. Specifically, in the four scenarios described below, we let $\beta_1 = 0.7$ and $\beta_2 = 0.8$ in Scenario 1; $\beta_1 = 0.8$ and $\beta_2 = 1.3$ in Scenario 2; $\beta_1 = 1.2$ and $\beta_2 = 1.5$ in Scenario 3; $\beta_1 = 0.8$ and $\beta_2 = 1.1$ in Scenario 4. $A$ is the treatment indicator and is randomly generated as 1 or 0 with equal probability. $g (S)$ is a function of predictive variables $S = \{S_1, S_2\}$, which are randomly generated along with the unrelated covariates $Z = \{Z_1, Z_2, \ldots, Z_q\}$ with $q = 20$ from a mean zero multivariate
normal distribution with a compound symmetric variance covariance matrix $\Sigma = (1 - \rho) I_{(q+2)} + \rho 1 1^T$, where $\rho = 1/3$. The sign of $g(S)$ determines the underlying patients subgroup membership, whereas covariates $Z$ are not associated with the outcomes nor the treatment effect. $T_{OS} = T_{PD} + T_{PD\text{toDeath}}$ is the time from entry to death. $C_{PD}$ is the censoring time for TTP which is randomly generated from an exponential distribution assuming a yearly dropout rate of 20%. Censoring of TTP and OS is by the following rules: TTP is censored if $T_{PD} > C_{PD}$ or $T_{PD} + T_{enroll} > 30$ (the duration of the trial). OS is only censored if $T_{OS} + T_{enroll} > 30$. Here we assume that a patient censored on TTP will still be followed until death or end of the trial.

Based on the experimental treatment effect, the patients could be divided into two subgroups: treatment performing subgroup ($g(S) > 0$) and the treatment nonperforming subgroup ($g(S) < 0$). As shown in Figure 2, we simulated four scenarios with different boundaries of subgroups defined by variable $S_1$ only (Scenario 1 and 3) or $S_1$ and $S_2$ jointly (Scenarios 2 and 4). For each scenario, we generate $n = 600$ independent survival time samples based on the following formulas of $g(S)$, respectively:

- **Scenario 1**: $g(S_1) = S_1$
- **Scenario 2**: $g(S_1, S_2) = S_1 - S_2$
- **Scenario 3**: $g(S_1) = I[-0.67 \leq S_1 \leq 0.67] \left(e^{-S_1^2} - 0.4\right) + I[S_1 < -0.67 \cup S_1 \geq 0.67](e^{-S_1^2} - 0.8)$
- **Scenario 4**: $g(S_1, S_2) = 2I((-1.07 \leq S_1 \leq 1.07) \cap (-1.07 \leq S_2 \leq 1.07)) - 1$

In Scenario 1, the two subgroups can be perfectly separated by the linear boundary $S_1 = 0$. When $S_1 > 0$, greater value in $S_1$ leads to greater experimental treatment benefit, when $S_1 < 0$, treatment effect becomes harmful and the harmfulness increases as the value of $S_1$ decreases. In Scenario 2, the linear boundary $S_1 = S_2$ can perfectly separate the two subgroups. The magnitude of experimental treatment benefit or harmfulness changes along both the $S_1$ axis and $S_2$ axis. In Scenario 3, the two subgroups can be separated by a nonlinear “U” shaped
boundary. The magnitude of experimental treatment benefit increases when $S_1$ gets closer to 0 and the magnitude of experimental treatment harmfulness increases when $S_1$ gets farther away from 0.67 or $-0.67$. In scenario 4, treatment performing subgroup is “enclave” by the treatment nonperforming subgroup.

The treatment effect for a subgroup of patients is measured by win-difference in our proposed method. To calculate win-difference, we compare every patient in treatment arm with every patient in control arm. The rule for comparing two patients’ clinical endpoints is that: for a pair of patients $i$ from treatment arm and $j$ from control arm, the win ($W_{ij}$) or lose ($L_{ij}$) first depend on who has death first. For example, if patient $i$ dies first, then treatment lose, $W_{ij} = 0$ and $L_{ij} = 1$. If comparison on death cannot be made or tied, only then the values of $W_{ij}$ and $L_{ij}$ depend on who has a progressive disease (PD) first. If neither death nor PD can tell who wins, they are considered tied/unknown, in this case $W_{ij} = L_{ij} = 0$. In other words, we recognize OS as the clinically more important outcome and compare it first, and only if OS is tied, TTP will be compared.

We compare our proposed win-difference based method (Win-diff) with the RMST-based method (Zhang et al. 2020). The RMST based method searches for the maximal difference of difference in areas under experimental treatment and control survival curves between subgroups. Note that the RMST-based method is limited to a single outcome. Specifically, according to the outcome used, there are five approaches, namely: (1) RMST for a single outcome TTP (RMST_TTP), (2) RMST for a single outcome OS (RMST_OS), (3) Win-diff for a single outcome TTP (Win-diff_TTP), (4) Win-difference for a single outcome OS (Win-diff_OS) and (5) Win-difference for two outcomes TTP and OS (Win-diff_TTP+OS). For all four scenarios, we generate 500 replicated datasets. In each replicated dataset, we simulate 600 patients as the training set and an independent validation set with 5000 patients. In each simulated dataset, since we know the true subgroup membership for each patient, we calculate the accuracy rate as the percentage of correctly predicted subgroup memberships in the validation set. In addition, sensitivity is calculated as the percentage of correctly predicted subgroup memberships among patients truly belong to treatment performing subgroup; and specificity is calculated as the percentage of correctly predicted subgroup memberships among patients truly belong to treatment nonperforming subgroup. A cutoff of 0.5 is used in all simulation settings.

The accuracy results are summarized in Figure 3. Either for OS only or TTP only, Win-diff based approaches

Figure 3. Classification accuracy based on validation set from simulation study.
always outperform RMST based approaches. Among all five approaches, Win-diff_OS+TTP achieves the highest accuracy across all four simulation scenarios. In addition, as the level of censoring of OS increases, the accuracy of methods relying solely on OS decrease rapidly, but the accuracy for Win-diff_TTP+OS only slightly decreases, mitigating the information loss due to censoring on OS by using the information from both clinical outcomes. Sensitivity and specificity are also calculated based on the validation set and are summarized in Figures S1 and S2 in supplementary material. In each simulation setting, we also calculate the average of hazard ratio (HR) in subgroups identified by each method across replicated datasets. In general, a higher classification accuracy in terms of subgroup membership leads to a greater magnitude of difference in treatment effects across identified subgroups. As Win-diff_TTP+OS provides the highest accuracy across all simulation scenarios, it also consistently identifies subgroups with the greatest difference in treatment effects. The results are summarized on Table S1 in the supplementary material.

4. Real Data Analysis

We apply the proposed method to a colon cancer adjuvant trial for the illustration purpose. In this trial, 315, 310, and 304 patients with stage B2 or stage C colon cancer received observation (Obs), levamisole alone (Lev), and levamisole combined with fluorouracil (Lev + 5-FU), respectively. levamisole combined with fluorouracil was found highly effective for prolonging survival ($P = 0.006$) and reducing risk of disease recurrence ($P < 0.0001$). The therapy with levamisole alone produced no detectable effect (Moertel et al. 1990). In this article, we focus on the comparison between the observation and levamisole combined with fluorouracil. We include 9 baseline covariates in addition to the treatment indicator. The covariates include 1 continuous variable: age; 6 binary variables: sex (male, female), obstruction of colon by tumor (yes, no), perforation of colon (yes, no), adherence to nearby organs (yes, no), number of lymph nodes with detectable cancer ($\geq 4$ nodes, $<4$ nodes), time from surgery to registration (short, long); and 2 categorical variables: differentiation of tumor (well, moderate, poor), Extent of local spread (submucosa, muscle, serosa, contiguous structures). For simplicity, we exclude 13 patients with missing data on at least one of the 9 baseline covariates and there are 298 and 308 patients in the Lev + 5-FU and Obs treatment groups, respectively. There are two time to event endpoints: OS and recurrence-free survival (RFS), and we consider OS as the clinically more important endpoint. The goal is to stratify the overall population into two subgroups that patients may or may not benefit from Lev + 5-FU relative to Obs.

To reduce overfitting, we use cross-validation to estimate the patients’ optimal treatment. Specifically, we partition the data into 5 roughly equal-sized sets based on original order of the observations in the dataset. At each iteration, we pick a different set as the validation set and perform the proposed method
to the remaining 4 sets of the data to develop the subgroup membership prediction model. We then predict the subgroup memberships for patients in the validation set and stratify them into Lev + 5-FU performing (Lev + 5-FU > Obs) subgroup and Lev + 5-FU nonperforming (Lev + 5-FU < = Obs) subgroup. A cutoff c of 0.6 is used. This way patients are not involved in the training process for the prediction model used to estimate their subgroup memberships. The majority of the 606 patients, 533 patients (88% of overall population), are classified to the Lev + 5-FU > Obs subgroup. Interestingly, there are 73 patients are classified to the Lev + 5-FU < = Obs subgroup given Lev + 5-FU leads to a significant benefit over Obs in the overall population. The Kaplan–Meier curves for the two treatment arms in the overall population and two subgroups identified by the proposed method are plotted in Figure 4 for OS and in Figure 5 for RFS. Enhanced treatment effect is observed in the Lev + 5-FU > Obs subgroup, with a HR of 0.66 for OS and a HR of 0.58 for RFS, compared to a HR of 0.69 for OS and a HR of 0.60 for RFS in the overall population (Table 1). In the Lev + 5-FU < = Obs subgroup, the HR is 0.9 for OS and 0.74 for RFS (Table 1). Finally, we fit the model to the overall sample and plot the Kaplan–Meier curves for the two identified subgroups in Figures S3 and S4, for OS and RFS respectively. The variable importance from the tree boosting is summarized in Figure S5 in the supplementary material. Age is the top one variable identified to determine the subgroup membership. Lev + 5-FU was found highly effective in the overall population without any crossing of the survival curves (Moertel et al. 1990). In our search for subgroups, the small subgroup size and the less than one HR observed in the identified Lev + 5-FU nonperforming subgroup suggest that the Lev + 5-FU may indeed works for everyone and there is not a sensible subgroup not benefitting from it.

### Table 1. Summary of treatment effect across subgroups identified by the proposed method based on a 5-fold cross-validation for colon cancer adjuvant study.

| Group                    | N  | OS HR (95% CI) | RFS HR (95% CI) |
|--------------------------|----|----------------|-----------------|
| Overall                  | 606| 0.69 (0.55, 0.88) | 0.60 (0.48, 0.76) |
| Lev + 5-FU > Obs         | 533| 0.66 (0.52, 0.86) | 0.58 (0.45, 0.75) |
| Lev + 5-FU < = Obs       | 73 | 0.90 (0.46, 1.75) | 0.74 (0.40, 1.38) |

5. Discussion

Win ratio/difference can be used to combine multiple outcomes and prioritize clinically more important outcome in a RCT. In this article, we explore the possibility of identifying subgroups with heterogeneous treatment effects based on the win difference for survival outcomes. The simulation studies suggest that using multiple outcomes while prioritizing the clinically more important one by win difference could increase the accuracy of subgroup identification. Especially, when censoring rate is high in the clinically more important outcome, borrowing subgroup
information encoded in the clinically less important outcome could help to identify the subgroups. For example, in an oncology adjuvant study, death events could be rare and borrowing information from time to disease recurrence when time to death is unknown could increase the chance to identify the underlying subgroups with heterogeneous response to study treatment.

The proposed value function in (3) is based on win difference and one can also use win ratio to quantify the treatment effect. We did simulation studies and found that the classification accuracies are very close between the two estimands. The pre-specified subgroup membership score \( \hat{p}(Z) \) is a quantity on the continuous scale in the range of 0–1 with a greater value indicating greater likelihood this patient belongs to the treatment performing subgroup. We simply used a fixed cutoff \( c = 0.5 \) to stratify the overall population into subgroups in our simulation studies to demonstrate the decent prediction accuracy the proposed method could provide. But in real data application, users can treat \( c \) as a hyperparameter and search for the optimal value using cross validation. Alternatively, one could look into the distribution of the estimated subgroup membership scores to determine the cutoff value \( c \) to stratify the overall population into two or more than two subgroups. For example, users could only claim subgroup membership (treatment performing vs. non-performing) for a patient with an estimated subgroup membership score well deviated from 0.5 and group patients with an estimated subgroup membership score close to 0.5 into a third subgroup and claim no treatment is better than the other for them. A R package "SubgroupBoost" was developed with two functions, each for win difference or win ratio, respectively. The R functions currently can be used to incorporate two survival outcomes with prioritized clinical importance.

As pointed by Pocock et al. (2012) in their original article, it would be preferred to use a risk score or risk stratification to select matched pair of patients with similar prognosis, thereby making all paired comparisons fair. A stratified version of the value function in Equation (3) stratifying on prognosis homogeneous strata may be considered if known prognostic factors are collected in the trial. Although we focus on combining two survival outcomes in this article the proposed framework can be used to incorporate multiple more than two outcomes with different types, such as continuous or categorical efficacy and safety outcomes. As long as one can form a clinically meaningful comparison rule with a hierarchy of clinical importance for benefit/risk assessment, the proposed value function can be used to conduct pairwise comparison per the user defined rule and search for the subgroups. These extensions will be the subjects of our future research. Lastly the proposed method is not immune to common problems in statistical learning such as overfitting and it may not be easy to interpret the subgroups identified as the prediction model could be based on multiple trees. The proposed method is recommended to be used as an exploratory tool and it is crucial to have an independent validation set to verify the findings.

References
Assmann, S. F., Pocock, S. J., Enos, L. E., and Kasten, L. E. (2000), “Subgroup Analysis and Other (mis)uses of Baseline Data in Clinical Trials,” Lancet (London, England), 355, 1064–1069. [523]
Chen, T., and Guestrin, C. (2016), “XGBoost: A Scalable Tree Boosting System,” in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, New York: ACM, pp. 785–794. [525]
Cui, L., Hung, H. M., Wang, S. J., and Tsong, Y. (2002), “Issues Related to Subgroup Analysis in Clinical Trials,” Journal of Biopharmaceutical Statistics, 12, 347–358. [523]
Dmitrienko, A., Muyers, C., Fritsch, A., and Lipkovich, I. (2016), “General Guidance on Exploratory and Confirmatory Subgroup Analysis in Late-Stage Clinical Trials,” Journal of Biopharmaceutical Statistics, 26, 71–98. [523]
European Medicines Agency Committee for Medicinal Products for Human Use (EMA/CHMP). (2019), “Guideline on the Investigation of Subgroups in Confirmatory Clinical Trials,” available at https://www.ema.europa.eu/en/documents/scientific-guideline/guideline-investigation-subgroups-confirmatory-clinical-trials_en.pdf. [523]
Foster, J. C., Taylor, J. M., and Ruberg, S. J. (2011), “Subgroup Identification From Randomized Clinical Trial Data,” Statistics in Medicine, 30, 2867–2880. [523]
Friedman, J. (2001), “Greedy Function Approximation: A Gradient Boosting Machine,” The Annals of Statistics, 29, 1189–1232. [525]
Lipkovich, I., Dmitrienko, A., and D’Agostino Sr. B. R. (2017), “Tutorial in Biostatistics: Data-Driven Subgroup Identification and Analysis in Clinical Trials,” Statistics in Medicine, 36, 136–196. [523]
Liu, W., Zhang, Z., Nie, L., and Soon, G. (2011), “Subgroup Identification Based on Differential Effect Search–A Recursive Partitioning Method for Establishing Response to Treatment in Patient Subpopulations,” Statistics in Medicine, 30, 2601–2621. [523]
Luo, X., Tian, H., Mohanty, S., and Tsai, W. Y. (2015), “An Alternative Approach to Confidence Interval Estimation for the Win Ratio Statistic,” Biometrics, 71, 139–145. [523,524]
Luo, X., Tian, H., Mohanty, S., and Tsai, W. Y. (2015), “An Alternative Approach to Confidence Interval Estimation for the Win Ratio Statistic,” Biometrics, 71, 139–145. [523,524]
Moe rt e l , C. G., Fleming, T. R., Macdonald, J. S., Haller, D. G., Laurie, J. A., Goodman, P. J., Ungerleider, J. S., Emerson, W. A., Tormey, D. C., and Glick, J. H. (1990), “Levamisole and Fluorouracil for Adjuvant Therapy of Resected Colon Carcinoma,” The New England Journal of Medicine, 322, 352–358. [528,529]
Loh, W. Y., He, X., and Man, M. (2015), “A Regression Tree Approach to Identifying Subgroups With Differential Treatment Effects,” Statistics in Medicine, 34, 1818–1833. [523]
Luo, X., Tian, H., Mohanty, S., and Tsai, W. Y. (2015), “An Alternative Approach to Confidence Interval Estimation for the Win Ratio Statistic,” Biometrics, 71, 139–145. [523,524]
Moe rt e l , C. G., Fleming, T. R., Macdonald, J. S., Haller, D. G., Laurie, J. A., Goodman, P. J., Ungerleider, J. S., Emerson, W. A., Tormey, D. C., and Glick, J. H. (1990), “Levamisole and Fluorouracil for Adjuvant Therapy of Resected Colon Carcinoma,” The New England Journal of Medicine, 322, 352–358. [528,529]
Penon, J., Buyse, M., Ozemne, B., Roche, L., and Roy, P. (2018), “An Extension of Generalized Pairwise Comparisons for Prioritized Outcomes in the Presence of Censoring,” Statistical Methods in Medical Research, 27, 1230–1239. [523,524]
Pocock, S. J., Ariti, C. A., Collier, T. J., and Wang, D. (2012), “The Win Ratio: A New Approach to the Analysis of Composite Endpoints in Clinical Trials Based on Clinical Priorities,” European Heart Journal 33, 176–182. [523,524,530]
Qian, M., and Murphy, S. A. (2011), “Performance Guarantees for Individualized Treatment Rules,” Annals of Statistics, 39, 1180–1210. [523]
Reck, M., Rodriguez-Abrue, D., Robinson, A. G., Hui, R., Cásószí, T., Fülöp, A., Gottfried, M., Peled, N., Tafreshi, A., Cuffe, S., O’Brien, M., Rao, S., Hotta, K., Leiby, M. A., Lubiniecki, G. M., Shentu, Y., Rangwala, R., Brahmer, J. R., and KEYNOTE-024 Investigators. (2016), “Pembrolizumab versus Chemotherapy for PD-L1-Positive Non-Small-Cell Lung Cancer,” The New England Journal of Medicine, 375, 1823–1833. [523]
Schnell, P. M., Tang, Q., Muller, P., and Carlin, B. P. (2017), “Subgroup Inference for Multiple Treatments and Multiple Endpoints in an Alzheimer’s Disease Treatment Trial,” The Annals of Applied Statistics, 11, 949–966. [523]
Su, X., Zhou, T., Yan, X., Fan, J., and Yang, S. (2008), "Interaction Trees With Censored Survival Data," *The International Journal of Biostatistics*, 4, 2. [523]

Tian, L., Alizadeh, A. A., Gentles, A. J., and 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, 1517–1532. [523]

Zhang, P., Ma, J., Chen, X., and Shentu, Y. (2020), "A Nonparametric Method for Value Function Guided Subgroup Identification Via Gradient Tree Boosting for Censored Survival Data," *Statistics in Medicine*, 39, 4133–4146. [524,525,527]

Zhang, B., Tsiatis, A. A., Laber, E. B., and Davidian, M. (2012), "A Robust Method for Estimating Optimal Treatment Regimes," *Biometrics*, 68, 1010–1018. [523]

Zhao, Y., Zeng, D., Rush, A. J., and Kosorok, M. R. (2012), "Estimating Individualized Treatment Rules Using Outcome Weighted Learning," *Journal of the American Statistical Association*, 107, 1106–1118. [523]