Original Research Article

SMMR
STATISTICAL METHODS IN MEDICAL RESEARCH

Statistical Methods in Medical Research
2024, Vol. 33(6) 1021–1042
© The Author(s) 2024
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/09622802241247728
journals.sagepub.com/home/smm
Sage

Causal rule ensemble method for estimating heterogeneous treatment effect with consideration of prognostic effects

Mayu Hiraishi[1,2], Ke Wan[3], Kensuke Tanioka[4], Hiroshi Yadohisa[5] and Toshio Shimokawa[3]

Abstract

We propose a novel framework based on the RuleFit method to estimate heterogeneous treatment effect in randomized clinical trials. The proposed method estimates a rule ensemble comprising a set of prognostic rules, a set of prescriptive rules, as well as the linear effects of the original predictor variables. The prescriptive rules provide an interpretable description of the heterogeneous treatment effect. By including a prognostic term in the proposed model, the selected rule is represented as an heterogeneous treatment effect that excludes other effects. We confirmed that the performance of the proposed method was equivalent to that of other ensemble learning methods through numerical simulations and demonstrated the interpretation of the proposed method using a real data application.

Keywords

Heterogeneous treatment effect, RuleFit, randomized clinical trial, machine learning, metalearner

1 Introduction

Randomized controlled clinical trials have been conducted to verify the effect of new treatments, and the average treatment effect is commonly used to evaluate the difference in outcomes between new and existing treatments.[1,2] However, the treatment effect is not always homogeneous in the overall population and varies according to some individual characteristics. Heterogeneous treatment effect (HTE) has received wide attention in recent years; it focuses on a subgroup that exhibits specific characteristics as a result of receiving treatment. Various machine learning methodologies have been developed for estimating the HTE. Tree-based methods are well-suited to handling large-scale data and they enable flexible modeling with various levels of covariates, compared to conventional statistical methods.[3] For example, a regression tree construction based on Classification and Regression Tree (CART)[4] has been proposed.[5,6] Moreover, in a forest-based algorithm, random forest[7] has been extended to causal effect estimation.[3,8] Furthermore, three methods were proposed by Powers et al.[9] in a framework of conditional outcome differences. Among Bayesian approaches, Bayesian additive regression trees

[1] Clinical Study Support Center, Wakayama Medical University Hospital, Wakayama, Japan
[2] Graduate School of Culture and Information Science, Doshisha University, Kyoto, Japan
[3] Department of Medical Data Science, Graduate School of Medicine, Wakayama Medical University, Wakayama, Japan
[4] Department of Biomedical Sciences and Informatics, Doshisha University, Kyoto, Japan
[5] Department of Culture and Information Science, Doshisha University, Kyoto, Japan

Corresponding author:
Mayu Hiraishi, Wakayama Medical University Hospital, Doshisha University, 811-1 Kimiidera, Wakayama, Kyoto 641-8510, Japan.
Email: m-hira@wakayama-med.ac.jp

(BARTs)[10] was developed for HTE estimation.[11,12] Although ensemble learning models have demonstrated significant predictive capabilities, the interpretability of variable contributions to the predicted values lacks clarity, known as the black box problem.

To address this challenge, model-interpretable methods have been proposed. The RuleFit method introduced by Friedman and Popescu[13] uses a nonparametric tree-based ensemble technique that can be expressed as a linear combination of base functions. This method generates functions based on "rules" from the paths from each root to the terminal nodes in each decision tree, and these functions can be expressed as a form of rule, that is, "weight $>$ 64.8 kg and height $\leq$ 173.6 cm." In this example, weight and height are covariate variables, and this rule function includes the subgroup whose weight is greater than 64.8 kg and height is under 173.6 cm. These rules help to determine the relationships among the characteristics of subgroups and the effects of the treatments. The RuleFit method has been applied to the framework of HTE in several studies. Bargagli-Stoffi et al.[14] proposed the causal rule ensemble (CRE), which uses RuleFit to extract interpretable HTE as the form of rule after estimating the HTE using another ensemble method. However, this method does not use RuleFit for the estimation of the HTE itself.

In this study, we propose a novel framework based on RuleFit to estimate interpretable HTE. The proposed framework assumes that the estimated HTE can be expressed as a linear combination of coefficients and rules to interpret the HTE between target treatment group and control group. Then, we interpret the characteristics represented by the obtained rules, specifically how they affect HTE. The proposed method includes a prognostic term in addition to prescriptive terms to express the HTE as a linear combination. If the prognostic term is not included in the model, the estimated value may contain both the prognostic effect and the HTE, making it difficult to accurately evaluate the specific effects of the treatment. Including a prognostic term in the model allows the HTE to be interpreted in the form of rules.

Moreover, to obtain an interpretable HTE, that is, to represent the estimated HTE as a linear combination, we incorporate the idea of a shared basis proposed by Powers et al.[9] into the proposed framework. In the shared-basis concept, the models for the two treatment groups assume that the base functions for the conditional mean regression are the same. The comparability of an HTE is not assured if the two treatment groups do not share the base functions. In our proposed method, by sharing the same rules related to the HTE between the target treatment group and control group, the calculation results of HTE can be described as a linear combination of coefficients and rules. To accomplish this, we use group lasso[15] instead of the lasso[16] used in the conventional RuleFit method. Wan et al.[17] proposed a RuleFit-based method to estimate the HTE; however, it does not consider the prognostic effect. By contrast, the proposed method considers the prognostic effect in estimating HTE, thereby allowing for a more refined interpretation of the treatment effect. We have developed R code for the proposed method,[18] in which some of the functions of the proposed method partially use the codes of the R package pre.[19]

In Section 2, we explain the HTE and conventional RuleFit in relation to the proposed method. Then, we introduce the framework and algorithm of the proposed method in Section 3. In Section 4, we demonstrate the efficiency of the proposed method through numerical simulations, and in Section 5, we describe the application of the proposed method to real genetic data related to breast cancer. Based on the obtained results, Section 6 concludes the article.

2 Related works

We apply our proposed method by extending the RuleFit method[13] to estimate the HTE in a randomized clinical trial. Before presenting our method, we explain HTE and RuleFit.

2.1 Heterogeneous treatment effect

In randomized controlled clinical trials, the target treatments are compared with the standard treatments to test the effectiveness of new treatments. The average treatment effect (ATE) is typically used for estimation. However, the ATE cannot detect the subgroups in which the new treatment is more effective than the standard treatment owing to the average of the population. To identify subgroups, the HTE focuses on the variability in treatment effects that may be attributed to patient factors.[2] Let Y_i ($i = 1, 2, \dots, n$) be the outcome variable, where n is the number of subjects, $X_i = (X_{i1}, X_{i2}, \dots, X_{ip})^\top$ be random covariate vectors, where p is the number of variables and $\cdot^\top$ denotes the transpose, and $Z_i \in \{0, 1\}$ be the allocation group in two levels, where $Z_i = 1$ and $Z_i = 0$ are the target treatment group and control treatment group, respectively.

Herein, we describe the settings used in this study. Each subject exhibited only one response to treatment.

The treatment effect considering heterogeneity is defined as

$$\tau(\boldsymbol{x}_i) = \mu_1(\boldsymbol{x}_i) - \mu_0(\boldsymbol{x}_i) \tag{1}$$

where

$$\mu_1(\boldsymbol{x}_i) = \mathrm{E}(Y_i|X_i = \boldsymbol{x}_i, Z_i = 1),\ \mu_0(\boldsymbol{x}_i) = \mathrm{E}(Y_i|X_i = \boldsymbol{x}_i, Z_i = 0) \tag{2}$$

Here, $\boldsymbol{x}_i \in \mathbb{R}^p$ is the observed covariate vector, and the HTE is the difference in the conditional mean functions between the two treatment groups. $\mu_1(\boldsymbol{x}_i)$ and $\mu_0(\boldsymbol{x}_i)$ denote the expected respected responses when subject i assigned to the target treatment group and standard treatment group, respectively. In this study, equations (1) and (2) are used to estimate the HTE.

2.2 RuleFit

RuleFit is a rule-based ensemble method[13] that is designed to handle cases in which the relationships between the outcomes and covariates are even nonlinear. This method also allows the interpretation of results using estimated rules. Given covariates $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \dots x_{ip})^\top$, the model of the RuleFit is defined as

$$F_{RF}(\boldsymbol{x}_i) = \beta_0 + \sum_{k=1}^{K} \beta_k r_k(\boldsymbol{x}_i) + \sum_{j=1}^{p} \alpha_j l_j(\boldsymbol{x}_i) \tag{3}$$

where $\beta_0 \in \mathbb{R}$ denotes the intercept, $\beta_k \in \mathbb{R}(k = 1, 2, \dots, K)$ denotes the coefficients of the rule terms, and $\alpha_j \in \mathbb{R}$ $(j = 1, 2, \dots, p)$ denotes the coefficient of the linear terms. As seen in equation (3), the RuleFit method consists of a rule term and a linear term. The rule term of the kth rule is defined as the following function:

$$r_k(\boldsymbol{x}_i) = \prod_{j=1}^{p} I(x_{ij} \in S_{jk})$$

where S_j is the set of all possible values of the covariates $x_{ij}(x_{ij} \in S_j)$ and $S_{jk} \subset S_j$, and $I(\cdot)$ is an indicator function that returns 1 if $x_{ij} \in S_{jk}$ is true, otherwise it returns 0. S_{jk} can be defined from the interval $(x_{jk}^-, x_{jk}^+]$ $(x_{jk}^-, x_{jk}^+ \in \mathbb{R})$ when $\boldsymbol{x}_j$ is an ordinal or scale variable. Rule- or tree-based ensembles have difficulty approximating linear structures, particularly when the number of training samples is insufficient. As a result, the rules for estimating the appropriate model may be insufficiently generated.[13] To improve accuracy and interpretability, the RuleFit model adds a linear term as an additional basis function based on j. To reduce the influence of the outliers of the covariates, the linear function $l_j(\cdot)$ is substituted for the "Winsorized" version to provide robustness. The "Winsorized" version of the linear function is defined as

$$l_j(\boldsymbol{x}_i) = \min\left(\delta_j^+, \max(\delta_j^-, x_{ij})\right) \tag{4}$$

where δ_j^- and δ_j^+ are the thresholds of the outliers, which are in the $q \in (0, 1)$ and $(1 - q)$ quantiles of variable j. Friedman and Popescu[13] recommended $q \simeq 0.025$, and it was adopted in this study. The coefficient vector of the linear term also depends on the scale. Therefore, equation (4) is normalized as

$$l_j(\boldsymbol{x}_i) \leftarrow 0.4 \cdot \frac{l_j(\boldsymbol{x}_i)}{std(l_j(\boldsymbol{x}_i))}$$

where $std(l_j(\boldsymbol{x}_i))$ is the standard deviation of $l_j(\boldsymbol{x}_i)$. Here, 0.4 is the average standard deviation of the rule under certain conditions.[13,19]

3 Proposed method

In this section, we present the framework of the proposed method and its calculation of HTE. Then, we explain the details of the algorithm.

3.1 Framework of the proposed method

We define a model of the proposed method, with explanation of the four steps for estimating the HTE based on the proposed method.

Let $y_i \in \mathbb{R}$ be the continuous outcome variable, $\boldsymbol{x}_i \in \mathbb{R}^p$ be the covariates, $z_i \in \{0,1\}$ be the treatment group, $r^{\dagger}_{k^{\dagger}}(\cdot)$ ($k^{\dagger} = 1,2,\ldots,K^{\dagger}$) be the rule function for the prognostic effect, $r^{*}_{k^{*}}(\cdot)$ ($k^{*} = 1,2,\ldots,K^{*}$) be the rule function for the prescriptive effect, and $l_j(\cdot)$ be the linear function. Given $y_i, \boldsymbol{x}_i$, and z_i, the model of the proposed method is defined as

$$\begin{aligned} F(\boldsymbol{x}_i, z_i) =& \beta_0 + \sum_{k^{\dagger}=1}^{K^{\dagger}} \beta_{k^{\dagger}} r^{\dagger}_{k^{\dagger}}(\boldsymbol{x}_i) + \sum_{j=1}^{p} \alpha_j l_j(\boldsymbol{x}_i) \\ &+ \sum_{k^{*}=1}^{K^{*}} \beta^{(a)}_{k^{*}} r^{*}_{k^{*}}(\boldsymbol{x}_i) \cdot I(z_i = 1) + \sum_{k^{*}=1}^{K^{*}} \beta^{(c)}_{k^{*}} r^{*}_{k^{*}}(\boldsymbol{x}_i) \cdot I(z_i = 0) \end{aligned} \tag{5}$$

where β_0 is the intercept, $\beta_{k^{\dagger}}$ ($k^{\dagger} = 1,2,\ldots,K^{\dagger}$) is the coefficients of the rule term of the prognostic effect, and α_j is the coefficients of the linear term of the prognostic effect.

$\beta^{(a)}_{k^{*}}$ ($k^{*} = 1,2,\ldots,K^{*}$) is the coefficient of the rule term of the prescriptive effect in the target treatment group, $z = 1$, whereas $\beta^{(c)}_{k^{*}}$ ($k^{*} = 1,2,\ldots,K^{*}$) is that for the control treatment group, $z = 0$. $I(\cdot)$ denotes the indication function. The first, second, and third terms in equation (5) are associated with the prognostic effect and do not depend on the treatment. By contrast, the fourth and fifth terms in equation (5) are related to treatment.

To calculate the HTE, the proposed method uses the following four steps:

STEP 1: Generation of a base function.

Given $(y_i, (\boldsymbol{x}_i, z_i))$ where y_i is a continuous response variable, the kth base function explaining y_i is generated as

$$r_k(\boldsymbol{x}_i, z_i) = I(z_i \in V_{zk}) \prod_{j=1}^{p} I(x_{ij} \in S_{jk}) \tag{6}$$

where $V_{zk} \subset \{1,0\}$ is a subset of all possible values of the allocation groups. The right-hand side of equation (6) is the same as the rule function of the original RuleFit. We denote the model of this rule function to emphasize that z_i is included in the covariate values. Thus, equation (6) can be generated using the same procedure as the original RuleFit model.

STEP 2: Rule Sorting.

The set of all rule functions generated in **STEP 1** is denoted as *Rule*. In this step, all generated K rules are divided into $K^{\dagger}$ rules associated with the prognostic effect and K^{*} rules associated with the HTE as

$$Rule^{\dagger} = \{r_k(\cdot,\cdot) | V_{zk} = \{1,0\}\} \tag{7}$$

$$Rule^{*} = \{r_k(\cdot,\cdot) | V_{zk} = \{1\} \vee V_{zk} = \{0\}\} \tag{8}$$

where $\#(Rule^{\dagger}) = K^{\dagger}$ and $\#(Rule^{*}) = K^{*}$, and # denotes the cardinality of the set. $Rule^{\dagger}$ is the set of base functions related to the prognostic effect and is denoted by $r^{\dagger}_{k^{\dagger}}(\boldsymbol{x}_i) = r_k(\boldsymbol{x}_i, z_i) \in Rule^{\dagger}$. Additionally, $Rule^{*}$ is a set of base functions related to the prescriptive effects. We set $r^{*}_{k^{*}}(\boldsymbol{x}_i) = \prod_{j=1}^{p} I(x_{ij} \in S_{jk^{*}})$, and $r^{*}_{k^{*}}(\boldsymbol{x}_i) \cdot I(z_i \in V_{zk^{*}}) = r_k(\boldsymbol{x}_i, z_i) \in Rule^{*}$ as a rule definition. Here, $Rule^{\dagger} \cup Rule^{*} = Rule$ and $Rule^{\dagger} \cap Rule^{*} = \phi$. In other words, $Rule^{\dagger}$ in equation (7) is a set of rules in which each element does not contain z, whereas $Rule^{*}$ in equation (8) is a set in which each element includes z. An example of this step is introduced after STEP 4.

STEP 3: Estimation of regression coefficients.

Based on the base functions in **STEP 2**, equation (5) is constructed. The optimization problem of the proposed method for **STEP 2** is defined as

$$\begin{aligned} &L\big(\beta_0, \{\beta_{k^{\dagger}}\}_{k^{\dagger}=1}^{K^{\dagger}}, \{\alpha_j\}_{j=1}^{p}, \{\boldsymbol{\beta}^{\ddagger}_{k^{*}}\}_{k^{*}=1}^{K^{*}}\big) \\ &= \frac{1}{2}\sum_{i=1}^{n}(y_i - F(\boldsymbol{x}_i, z_i))^2 + \lambda\left(\sum_{k^{\dagger}=1}^{K^{\dagger}} |\beta_{k^{\dagger}}| + \sum_{j=1}^{p} |\alpha_j| + \sqrt{2}\sum_{k^{*}=1}^{K^{*}} \|\boldsymbol{\beta}^{\ddagger}_{k^{*}}\|_2\right) \to \min \end{aligned} \tag{9}$$

where $\|\cdot\|_2$ is the L2 norm. $\beta_{k^{\dagger}}$ ($k^{\dagger} = 1,2,\ldots,K^{\dagger}$) and α_j are the coefficient vectors of the rule term and the linear term effects, respectively, for the prognostic effect. $\boldsymbol{\beta}^{\ddagger}_{k^{*}} = (\beta^{(a)}_{k^{*}}, \beta^{(c)}_{k^{*}})^{\top}$ ($k^{*} = 1,2,\ldots,K^{*}$) is the set of coefficient vectors related to the treatment. $\beta^{(a)}_{k^{*}}$ and $\beta^{(c)}_{k^{*}}$ are the parameters of the target treatment group and the control group, respectively. The terms related to the prognostic effect of $F(\boldsymbol{x}_i, z_i)$ are the first, second, and the third terms of equation (5), which are not dependent on the prescriptive effect. Here, in the fourth and fifth terms of equation (5), $r^*_{k^*}(\boldsymbol{x}_i) \cdot I(z_i = 1)$ and $r^*_{k^*}(\boldsymbol{x}_i) \cdot I(z_i = 0)$ are treated as one group. The target treatment group and the control group have the same rule function $r^*_{k^*}(\boldsymbol{x}_i)$ to reflect the concept of a shared basis.[9] Each pair of $\beta^{(a)}_{k^*}$ and $\beta^{(c)}_{k^*}$ represents the coefficients within the same group. To calculate the HTE under identical base functions between the two treatment arms, the proposed method uses group lasso[15] to prune the rule terms for HTE, whereas the conventional RuleFit method uses lasso[16] to prune the base learners. This allows the selection of the same rules between the base functions in the two treatment groups and is expected to ensure comparability of the two treatment groups.

STEP 4: Calculation of HTE.

Using the parameters estimated in **STEP 3**, the HTE is computed based on the model and allocation groups.

The rules generated in **STEP 1** are not separately estimated for the prognostic effect and the prescriptive effect. It needs to classify all K rules into rules for treatment effect and those for prognostic effect to estimate HTE. Next, we explain equations (7) and (8) in **STEP 2** with an example. First, in equation (7), $Rule^\dagger$ is the set of rules in that each element does not include z. For $r^\dagger_{k^\dagger}(\boldsymbol{x}_i)$, we have "age ≥ 53 & hemoglobin < 9.2," where age and hemoglobin are covariate variables. If a subject is > 53 years old and has a hemoglobin value under 9.2, $r^\dagger_{k^\dagger}(\boldsymbol{x}_i)$ returns 1, otherwise, it returns 0. Here, if $r^\dagger_{k^\dagger}(\boldsymbol{x}_i) = 1$, subject i belongs to this rule; thus, if $r^\dagger_{k^\dagger}(\boldsymbol{x}_i) = 0$, the subject i does not belong to this rule. On the other hand, in equation (8), $Rule^*$ is the set of rules in that each element includes z, such as "age $\geq$ 53 & $z > 0.5$ & hemoglobin < 9.2." Equation (8) is then interpreted the same way as equation (7). From this classification process, we obtain $Rule^\dagger$ for the prognostic effect, and $Rule^*$ for the prescriptive effect. In **STEP 3**, rule "age $\geq$ 53 & $z > 0.5$ & hemoglobin < 9.2" can relate to the prescriptive effect of both treatment group and control group as it includes treatment indicator z. These rule functions and each treatment indicator are regarded as one group to allow the HTE to be calculated using a linear combination of regression coefficients related to the prescriptive effect in **STEP 4**.

3.2 Algorithm

In this subsection, we explain rule generation, rule sorting, the regression coefficient estimation, and the HTE calculation using the algorithm specified in the previous section.

Rule generation related to both prognostic effect and prescriptive effect

The algorithm for generating rules is the same as that by Friedman and Popescu.[13] Here, the base function $\{r_k(\cdot)\}_{k=1}^K$ is generated from the covariate $\boldsymbol{x}$ and allocation group z of the training data $\{y_i, (\boldsymbol{x}_i, z_i)\}_{i=1}^n$, where n is the number of subjects in the training data. The details are presented in Algorithm 1. Model H is formed as $H(\boldsymbol{x}_i, z_i) = \{h_m(\boldsymbol{x}_i, z_i)\}_{m=1}^M$ $(i = 1, 2, \dots, n;\ m = 1, 2, \dots, M)$, where M is the number of tree-based learners h_m. M, $\bar{L}$ $(\bar{L} \geq 2)$, ν $(\nu \simeq 0.01)$, and $\lfloor \eta \rfloor$ $(\eta = \min(n/2, 100 + 6\sqrt{n})$) represent the number of tree-based learners, mean depth of the tree-based learners, the shrinkage rate, and the the number of subsamples for each tree-based learner in training, respectively, which are given as hyperparameters. To update model H, we use the gradient boosting tree (GBT) method.[20] H is successively updated by the regression tree model[4] h_m using a greedy stagewise approach.

In lines 1–3 of Algorithm 1, the model is initialized as $H_0(\boldsymbol{x}_i, z_i)$. Next, for each m $(m = 1, 2, \dots, M)$, the pseudo-residual ξ_{im} is calculated as shown in line 6 of Algorithm 1. Subsequently, in line 8, the number of terminal nodes for the mth tree-based learner t_m is calculated as

$$t_m = 2 + \text{floor}(u) \quad u \sim \exp(-u/(\bar{L} - 2))/(\bar{L} - 2)$$

where floor$(\cdot)$ is the floor function and $\bar{L}$ $(\bar{L} \geq 2)$ is the mean depth of tree-based learners.[13] This random setting of the number of terminal nodes for each tree enables the production of trees of different sizes. Then, a regression tree providing the disjoint terminal regions R_{qm} $(q = 1, 2, \dots, t_m; m = 1, 2, \dots, M)$ is fitted to the pseudo-residual ξ_{im}. In line 9 of Algorithm 1, different optimal constants γ_{qm} exist in each region. γ_{qm} represents the mean of ξ_{im} in the qth node of the mth tree. For calculations from line 9, we used the R package rpart.[21] With these values, H_m is updated as shown in line 10. After generating M regression trees, K rule functions are constructed from them, as shown in line 13. Here, K is the total number of rules generated from all trees, which can be calculated as

$$K = \sum_{m=1}^{M} 2(t_m - 1)$$

where t_m denotes the number of terminal nodes in the mth tree. In line 15, the K rules are combined.

Algorithm 1. Rule generation.

```
Require: training set $\{y_i, (\boldsymbol{x}_i, z_i)\}_{i=1}^n$, number of tree-based learners $M$, mean depth of tree-based learners $\bar{L}$, shrinkage rate $v$, and
    training sample for each tree-based learner $\lfloor \eta \rfloor$
for $i = 1$ to $n$ do
    Set the initial model $H_0(\boldsymbol{x}_i, z_i) \leftarrow \bar{y}$
end for
for $m = 1$ to $M$ do
    for $i = 1$ to $n$ do
        Compute the pseudo-residual
            $\xi_{im} \leftarrow \tilde{y}_i - H_{m-1}(\boldsymbol{x}_i, z_i)$
    end for
    Calculate the number of terminal nodes for the tree-based learner
        $t_m = 2 + \text{floor}(u)$, where $u \sim \exp(-u/(\bar{L}-2))/(\bar{L}-2)$
    Fit a regression tree to the pseudo residual $\xi_{im}$, giving terminal regions $R_{qm} \quad (q = 1, 2, \dots, t_m)$
    and region-specific means $\hat{\gamma}_{qm}$
    Update $H_m(\boldsymbol{x}_i, z_i) \leftarrow H_{m-1}(\boldsymbol{x}_i, z_i) + v h_m(\boldsymbol{x}_i, z_i)$
    where $h_m(\boldsymbol{x}_i, z_i) = \sum_{q=1}^{t_m} \hat{\gamma}_{qm} I(\boldsymbol{x}_i \in R_{qm})$
end for
for $m = 1$ to $M$ do
    Compose rules $\{r_{k_m}(\boldsymbol{x}_i, z_i)\}_{k_m=1}^{K_m}$ from $h_m(\boldsymbol{x}_i, z_i)$
end for
Collect all rule sets $\{r_{k_1}(\boldsymbol{x}_i, z_i)\}_{k_1=1}^{K_1}, \dots, \{r_{k_M}(\boldsymbol{x}_i, z_i)\}_{k_M=1}^{K_M}$ as $\{r_k(\boldsymbol{x}_i, z_i)\}_{k=1}^{K} \quad (K = \sum_{m=1}^{M} K_m)$
```

*Algorithm 1 follows the algorithm of rule generation provided in Friedman and Popescu.(2008)

Rule ensemble and parameter estimation using group lasso

This step provides the two advantages of the proposed method. First, the rule term function $r_k(\cdot)$ generated in 3.2.1 is divided into rules related to prescriptive effects $r^*_{k^*}(\cdot)$ and others $r^\dagger_{k^\dagger}(\cdot)$, thereby indicating that the model in equation (5) contains the base functions relevant to the prescriptive effect and that of prognostic effect, respectively. This enables the estimation of the prescriptive effect for nonlinear relationships while considering the prognostic effects. Second, to select rules that contribute to the outcome, the proposed method uses group lasso[15] to interpret the treatment effects based on the selected rules. The conventional RuleFit method uses lasso[16] to prune the generated rules. In the case of lasso, if a rule is selected for only one of the two treatment groups, it does not specify whether the rule affects the outcomes. The necessity of this concept is referred to by Powers et al.[9] as the shared basis for both the target treatment group and control group.

The details are described in Algorithm 2. As mentioned in the previous section, a linear term is introduced in lines 1 to 6 of Algorithm 2. In line 7, the r_k generated in 3.2.1 is divided into rules related to prescriptive effects $r^*_{k^*}$ and the others $r^\dagger_{k^\dagger}$, and the model in equation (5). In line 7, the K rules are divided into K^* rules for prescriptive effects and $K^\dagger$ rules for the others. To estimate the parameters using group lasso, the group information of the rule terms is introduced as

$$\mathscr{C} = \{D_1, D_2, \dots, D_{K^\dagger}, G_1, G_2, \dots, G_{K^*}\} \tag{10}$$

where the singleton set of rules is related to the prognostic effects $D_{k^\dagger} = \{k^\dagger\} (k^\dagger = 1, 2, \dots, K^\dagger)$ and the set of two pairs $G_{k^*} = \{(k^*, z = 1), (k^*, z = 0)\}$ $(k^* = 1, 2, \dots, K^*)$ includes the prescriptive effect.

The R package grpreg[22] was used to estimate parameters, and hyperparameter λ $(\lambda > 0)$ was selected by cross-validation using this package. The rule terms of the prognostic effects and the linear term in equation (5) are the common terms for both treatment groups, indicating that these parameters do not belong to group; their regularization is treated as a traditional lasso.

Then, regression parameters $\beta_0, \beta_{k^\dagger}, \alpha_j$, and $\boldsymbol{\beta}^\ddagger_{k^*} = (\beta^{(a)}_{k^*}, \beta^{(c)}_{k^*})^\top$ such that equation (9) is minimized. Here, $\hat{\beta}_0$ is the estimated intercept, $\hat{\beta}_{k^\dagger}$ is the estimated coefficients relevant to the prognostic effect, and $\hat{\alpha}_j$ is the estimated coefficients of the linear term. Additionally, $\hat{\boldsymbol{\beta}}^\ddagger_{k^*} = (\hat{\beta}^{(a)}_{k^*}, \hat{\beta}^{(c)}_{k^*})^\top$ is the estimated coefficients relevant to the HTE, where $\hat{\beta}^{(a)}_{k^*}$ and $\hat{\beta}^{(c)}_{k^*}$ correspond to the target treatment group and the control group, respectively.

Algorithm 2. Rule ensemble and parameter estimation

```
Require: ${y_i, (\boldsymbol{x}_i, z_i)}_{i=1}^n$, estimated rules ${r_k(\boldsymbol{x}_i, z_i)}_{k=1}^K$, values of Winsorized margin $(\delta_j^-, \delta_j^+)$ $(j = 1, \ldots, p)$, $\lambda$ $(\lambda > 0)$
for $i = 1$ to $n$ do
    for $j = 1$ to $p$ do
        $l_j(\boldsymbol{x}_i) = \min(\delta_j^+, \max(\delta_j^-, x_{ij}))$
        $l_j(\boldsymbol{x}_i) \leftarrow 0.4 \cdot l_j(\boldsymbol{x}_i)/std(l_j(\boldsymbol{x}_i))$
    end for
end for
Divide rule terms ${r_k(\boldsymbol{x}_i, z_i)}_{k=1}^K$ into those related to treatment effects ${r_{k^*}^*(\boldsymbol{x}_i) \cdot I(z_i \in V_{zk^*})}_{k^*=1}^{K^*}$ and the others ${r_{k^\dagger}^\dagger(\boldsymbol{x}_i)}_{k^\dagger=1}^{K^\dagger}$
Create group information to apply group lasso
    $\mathscr{C} = \{D_1, D_2, \ldots, D_{K^\dagger}, G_1, G_2, \ldots, G_{K^*}\}$
Estimate coefficient vectors using group lasso

    $$(\hat{\beta}_0, \{\hat{\beta}_{k^\dagger}\}_{k^\dagger=1}^{K^\dagger}, \{\hat{\alpha}_j\}_{j=1}^p, \{\hat{\boldsymbol{\beta}}_{k^*}^\ddagger\}_{k^*=1}^{K^*}) = \underset{\left(\beta_0, \{\beta_{k^\dagger}\}_{k^\dagger=1}^{K^\dagger}, \{\alpha_j\}_{j=1}^p, \{\hat{\boldsymbol{\beta}}_{k^*}^\ddagger\}_{k^*=1}^{K^*}\right)}{\operatorname{argmin}} \frac{1}{2}\sum_{i=1}^N (y_i - F(\boldsymbol{x}_i, z_i))^2 + \lambda\left(\sum_{k^\dagger=1}^{K^\dagger} |\beta_{k^\dagger}| + \sum_{j=1}^p |\alpha_j| + \sqrt{2}\sum_{k^*=1}^{K^*} \|\hat{\boldsymbol{\beta}}_{k^*}^\ddagger\|_F^2\right)$$

    where $\hat{\boldsymbol{\beta}}_{k^*}^\ddagger = (\beta_{k^*}^{(a)}, \beta_{k^*}^{(c)})$.
return $\hat{\beta}_0, \{\hat{\beta}_{k^\dagger}\}_{k^\dagger=1}^{K^\dagger}, \{\hat{\alpha}_j\}_{j=1}^p, \{\hat{\boldsymbol{\beta}}_{k^*}^\ddagger\}_{k^*=1}^{K^*}$
```

HTE calculation

In **STEP 3**, we estimate each parameter of both treatment groups using the common base function to calculate the HTE. From the model in equation (5), equation (2) in our proposed method for the target treatment group $z = 1$ and the control group $z = 0$ can be expressed as follows:

$$\hat{\mu}_1(\boldsymbol{x}_i) = \hat{\beta}_0 + \sum_{k^\dagger=1}^{K^\dagger} \hat{\beta}_{k^\dagger} r_{k^\dagger}^\dagger(\boldsymbol{x}_i) + \sum_{j=1}^{p} \hat{\alpha}_j l_j(\boldsymbol{x}_i) + \sum_{k^*=1}^{K^*} \hat{\beta}_{k^*}^{(a)} r_{k^*}^*(\boldsymbol{x}_i) \quad \text{and} \tag{11}$$

$$\hat{\mu}_0(\boldsymbol{x}_i) = \hat{\beta}_0 + \sum_{k^\dagger=1}^{K^\dagger} \hat{\beta}_{k^\dagger} r_{k^\dagger}^\dagger(\boldsymbol{x}_i) + \sum_{j=1}^{p} \hat{\alpha}_j l_j(\boldsymbol{x}_i) + \sum_{k^*=1}^{K^*} \hat{\beta}_{k^*}^{(c)} r_{k^*}^*(\boldsymbol{x}_i) \tag{12}$$

Regardless of the treatment, the prognostic effects and linear terms of both treatment groups in equations (11) and (12) are the same. From equations (11) and (12), the HTE in equation (1) is calculated as follows:

$$\begin{aligned}
\hat{\tau}(\boldsymbol{x}_i) &= \left(\hat{\beta}_0 + \sum_{k^\dagger=1}^{K^\dagger} \hat{\beta}_{k^\dagger} r_{k^\dagger}(\boldsymbol{x}_i) + \sum_{j=1}^{p} \hat{\alpha}_j l_j(\boldsymbol{x}_i) + \sum_{k^*=1}^{K^*} \hat{\beta}_{k^*}^{(a)} r_{k^*}^*(\boldsymbol{x}_i)\right) \\
&\quad - \left(\hat{\beta}_0 + \sum_{k^\dagger=1}^{K^\dagger} \hat{\beta}_{k^\dagger} r_{k^\dagger}(\boldsymbol{x}_i) + \sum_{j=1}^{p} \hat{\alpha}_j l_j(\boldsymbol{x}_i) + \sum_{k^*=1}^{K^*} \hat{\beta}_{k^*}^{(c)} r_{k^*}^*(\boldsymbol{x}_i)\right) \\
&= \sum_{k^*=1}^{K^*} \hat{\beta}_{k^*}^{(a)} r_{k^*}^*(\boldsymbol{x}_i) - \sum_{k^*=1}^{K^*} \hat{\beta}_{k^*}^{(c)} r_{k^*}^*(\boldsymbol{x}_i) \\
&= \sum_{k^*=1}^{K^*} \left(\hat{\beta}_{k^*}^{(a)} - \hat{\beta}_{k^*}^{(c)}\right) r_{k^*}^*(\boldsymbol{x}_i)
\end{aligned} \tag{13}$$

This indicates that the HTE can be calculated using equation (1) with terms for each treatment arm. Therefore, the HTE of the proposed method can be estimated using the difference in the predicted values of each treatment arm, considering the prognostic effects of the estimation.

4 Numerical simulation

Numerical simulations were conducted to evaluate the performance of the proposed method. We expected the results of the proposed method to be equivalent to those of the compared methods. In this section, we explain the simulation design and present the results.

4.1 Simulation design

Our simulation was designed based on the settings described by Powers et al.[9] First, we generated covariate matrix $\boldsymbol{X} = (x_{ij})$. x_{ij} was randomly distributed from $N(0, 1)$, where N is a normal distribution. Our setting was a two-armed randomized controlled trial; therefore, we set the treatment group variable as z_i for the treatment arm, where $z_i = 1$ and $z_i = 0$ signify the target treatment group and the control group, respectively. The treatment group indicator z_i was generated based on Bernoulli distribution, $z_i \sim B(0.5)$.

Using $\boldsymbol{x}_i$ and z_i, the outcome variable was randomly generated as

$$y_i = \psi(\boldsymbol{x}_i) + \left(z_i - \frac{1}{2}\right) \tau(\boldsymbol{x}_i) + \epsilon_i \tag{14}$$

where $\psi : \mathbb{R}^p \mapsto \mathbb{R}$ is the true effect related to the outcome of the covariates $\boldsymbol{x}_i$ and $\tau(\boldsymbol{x}_i)$ is that of the HTE. The error distribution ϵ_i follows the normal distribution, $N(0, 0.25)$. We generated the training and the test data using the same settings and sample sizes. To compare the performances of the proposed method and the other methods, the simulation was conducted with various factors. The total pattern of the simulation was 2 (Factor 1) × 3 (Factor 2) × 4 (Factor 3) × 4 (Factor 4) = 96, and each pattern was repeated 100 times.

We present the factors of the simulation settings below.

Factor 1: Sample Size
The sample size n was 600 and 1000 to examine the influence of the different number of n.

Factor 2: Covariate Variables
The number of variables p was set 100, 200, and 400 to examine the influence of the number of p.

Factor 3: Patterns of $\psi(\boldsymbol{x})$
$\psi(\boldsymbol{x}_i)$ is a function that expresses the prognostic effects. We set four different settings, as listed in the second column of Table 1. Scenarios 1–4 and 5–8 assumed linear function. Scenarios 9–12 were generated from the nonlinear function using the indicator function, and Scenarios 13–16 were generated by sin function and exponential function.

Factor 4: Patterns of $\tau(\boldsymbol{x})$
$\tau(\boldsymbol{x}_i)$ generates data relevant to the HTE. Hence, we set the other four settings, as shown in the third column from the left in Table 1. Scenarios $1, 5, 9$, and 13 were combinations of linear and quadratic functions. Scenarios $2, 6, 10$, and 14 were indicator functions that assumed quantitative data. Scenarios $3, 7, 11$, and 15 are based on sin function and exponential function. Scenarios $4, 8, 12$, and 16 assumed no treatment effects.

To evaluate performance accuracy, we used three different evaluation indices. The first was the mean squared error (MSE), calculated as

$$MSE = \frac{1}{n} \sum_{i=1}^{n} \left(\tau^*(\boldsymbol{x}_i) - \hat{\tau}(\boldsymbol{x}_i)\right)^2 \tag{15}$$

where $\tau^*(\boldsymbol{x}_i)$ is the true HTE value and $\hat{\tau}(\boldsymbol{x}_i)$ is the estimated HTE value. The second evaluation index is the relative bias (RBias) against the true HTE, given as

$$RBias = \frac{1}{n} \sum_{i=1}^{n} \frac{\left(\tau^*(\boldsymbol{x}_i) - \hat{\tau}(\boldsymbol{x}_i)\right)}{\tau^*(\boldsymbol{x}_i)}$$

The third evaluation index is Spearman's correlation coefficient between the true and estimated HTE values. In practical situations, subgroups are detected by subjects ordered based on estimated HTE[23]; therefore, we added this evaluation. The evaluation indices were calculated by test data that were not used in the model fitting process.

We compared the proposed method to four different tree-based approach methods: causal forest,[3] BART,[11] causal multivariate adaptive regression spline (Causal MARS),[9] and pollinated transformed outcome (PTO) forest.[9] In the simulation, we used Rstudio version 1.2.5033. We used the R packages grf[24] for causal forest, bartCause[11] for BART, and causalLearning[25] for the Causal MARS and PTO forest. In the parameter settings of each compared method, causal forest

Table 1. Scenarios of the simulation. $\psi(x)$ is a function related to the prognostic effect, and $\tau(x)$ is the one related to the heterogeneous treatment effect (HTE).

Scenario	$\psi(x)$	$\tau(x)$
1	x_1x_2	$2(x_2) + x_3^2 + x_5x_6 + x_8^2$
2	x_1x_2	$2 + 0.3I(x_4 > -3) - 4I(x_5 > 0) + 0.7I(x_7 < 1)$
3	x_1x_2	$3\sin(x_1x_5)^2 + 5\exp(x_8 + x_3)$
4	x_1x_2	0
5	$x_1 + x_3 - x_5$	$2(x_2) + x_3^2 + x_5x_6 + x_8^2$
6	$x_1 + x_3 - x_5$	$2 + 0.3I(x_4 > -3) - 4I(x_5 > 0) + 0.7I(x_7 < 1)$
7	$x_1 + x_3 - x_5$	$3\sin(x_1x_5)^2 + 5\exp(x_8 + x_3)$
8	$x_1 + x_3 - x_5$	0
9	$0.5I(x_1 > -1) - 1.4I(x_3 > 0)$	$2(x_2) + x_3^2 + x_5x_6 + x_8^2$
10	$0.5I(x_1 > -1) - 1.4I(x_3 > 0)$	$2 + 0.3I(x_4 > -3) - 4I(x_5 > 0) + 0.7I(x_7 < 1)$
11	$0.5I(x_1 > -1) - 1.4I(x_3 > 0)$	$3\sin(x_1x_5)^2 + 5\exp(x_8 + x_3)$
12	$0.5I(x_1 > -1) - 1.4I(x_3 > 0)$	0
13	$3\sin(x_4 + x_5)^2 - 0.2\exp(x_7)$	$2(x_2) + x_3^2 + x_5x_6 + x_8^2$
14	$3\sin(x_4 + x_5)^2 - 0.2\exp(x_7)$	$2 + 0.3I(x_4 > -3) - 4I(x_5 > 0) + 0.7I(x_7 < 1)$
15	$3\sin(x_4 + x_5)^2 - 0.2\exp(x_7)$	$3\sin(x_1x_5)^2 + 5\exp(x_8 + x_3)$
16	$3\sin(x_4 + x_5)^2 - 0.2\exp(x_7)$	0

was set to default, apart from "tune.parameters," which was set to "all." Parameter "method.trt" in BART was set to "none." The parameter settings in Causal MARS and PTO forest were set to default as well.

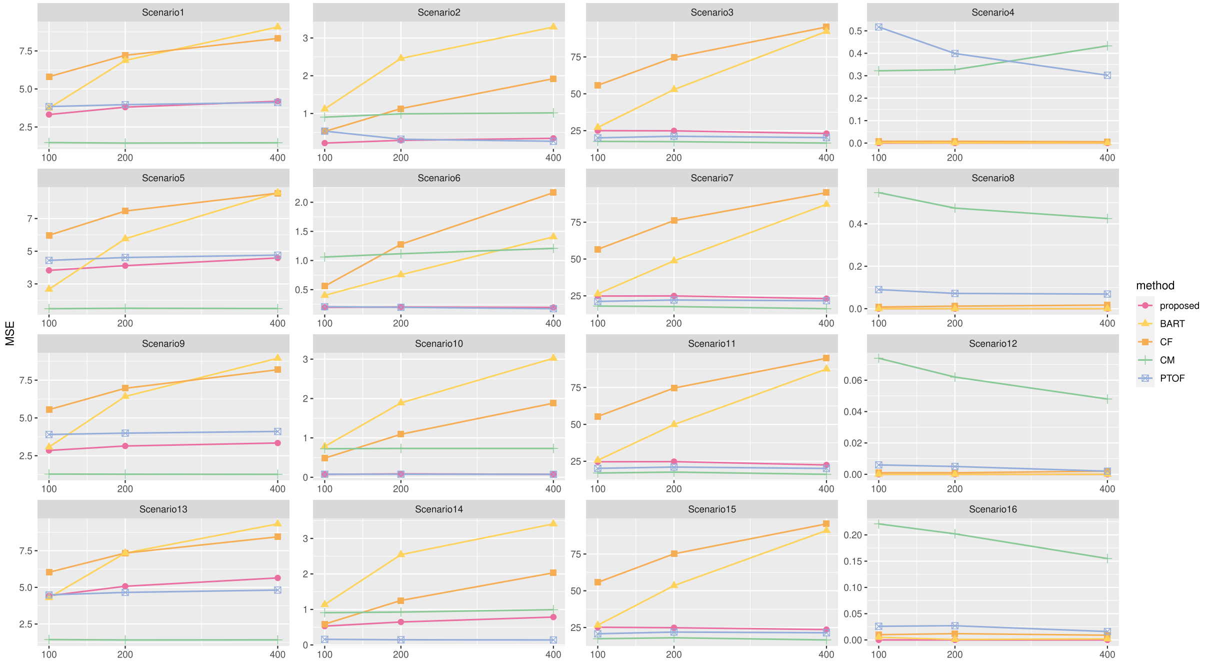

Figure 1. Plots of mean squared error (MSE) for $n = 600$. The horizontal axis is the number of covariate variables, and the vertical axis is the MSE.

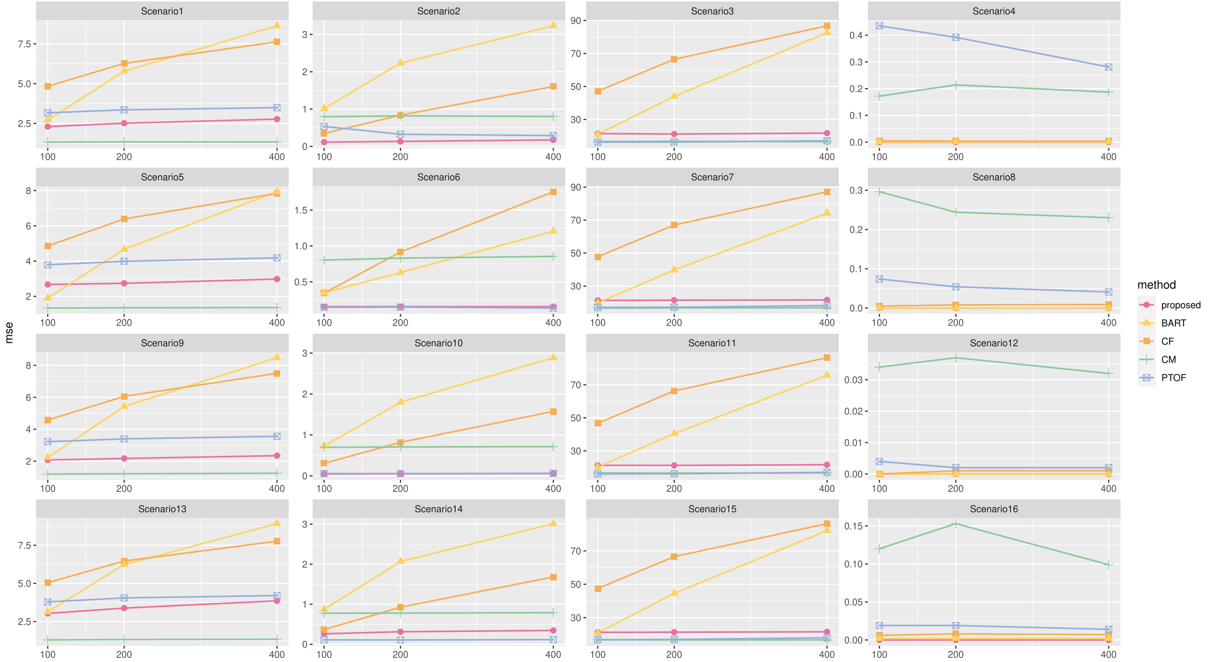

Figure 2. Plots of mean squared error (MSE) for $n = 1000$. The horizontal axis is the number of covariate variables, and the vertical axis is the MSE.

4.2 Simulation results

The results are presented in Figures 1 to 6. First, we explain the results of MSE. Figure 1 presents the MSE for $n = 600$ and Figure 2 presents the MSE for $n = 1000$. The horizontal axis represents the number of p, and the vertical axis represents the MSE values. Both figures are drawn by method and plotted using the scenario in Table 1. Overall, the results of the proposed method were better, particularly in cases where the settings of the true HTE τ were more complicated nonlinear functions. In many scenarios, the proposed method was stable, regardless of the values of n and p, whereas the MSE of causal forest and BART increased based on the value of p. Next, we consider the results for the scenarios. In Scenarios $1, 4, 9$, and 13 for $n = 600$ (the leftmost column of Figure 1), in which setting of the true HTE was a combination of linear and quadratic functions, the MSE values of Causal MARS were found to be the lowest. The proposed method was superior to the other compared methods, apart from PTO forest with $p = 200$ and 400 in Scenario 13. For Scenarios $2, 6, 10$, and 14 (the second column from the left of Figure 1), whose HTE setting was piecewise constant, the proposed method and PTO forest were superior to the other methods in Scenarios 6 and 10. For $p = 100$ and 200 in Scenario 2, the MSE values of the proposed method were smaller than those of PTO forest, although it was slightly increased at $p = 400$. The MSE of PTO forest was smaller in Scenario 14. In Scenarios $3, 7, 11$, and 15 for $n = 600$ (the second column from the right of Figure 1), whose HTE setting was a combination of sin function and exponential function, the proposed method was not smaller than Causal MARS and PTO forest; however, the differences were rather small compared with those of the causal forest and BART. Regarding the trend of the MSE values, the proposed method, Causal MARS, and PTO forest remained unchanged as the value of p increased. The MSE values of the causal forest and BART increased significantly as the number of variables increased. In Scenarios 4, 8, 12, and 16 (the rightmost column of Figure 1), the MSE values of the proposed method and BART were estimated as 0, and causal forest estimated nearly the true τ value. On the other hand, Causal MARS and PTO forest in Scenarios 4 and 8 estimated the presence of the treatment effect, and the trend depending on p was unstable. In the case of $n = 1000$, Scenarios $1, 4, 9$, and 13 (the leftmost column of Figure 2), Causal MARS was better than the other methods, however, the proposed method was superior to PTO forest in all scenarios. In Scenarios $2, 6, 10$, and 14 with $n = 1000$ (the second column from the left of Figure 2), the proposed method was superior to the other methods in Scenarios 2 and was nearly the same as PTO forest in Scenarios 6 and 10. In Scenario 14, PTO forest was better than the proposed method; however, the difference between these methods was closer than that with $n = 600$ in the same scenario. In all scenarios, the trend of the MSE with respect to the number of p showed a tendency similar to that of the case of $n = 600$. The MSE values of the proposed method, Causal MARS, and PTO forest were not influenced by the number of p, whereas those of causal forest and BART increased.

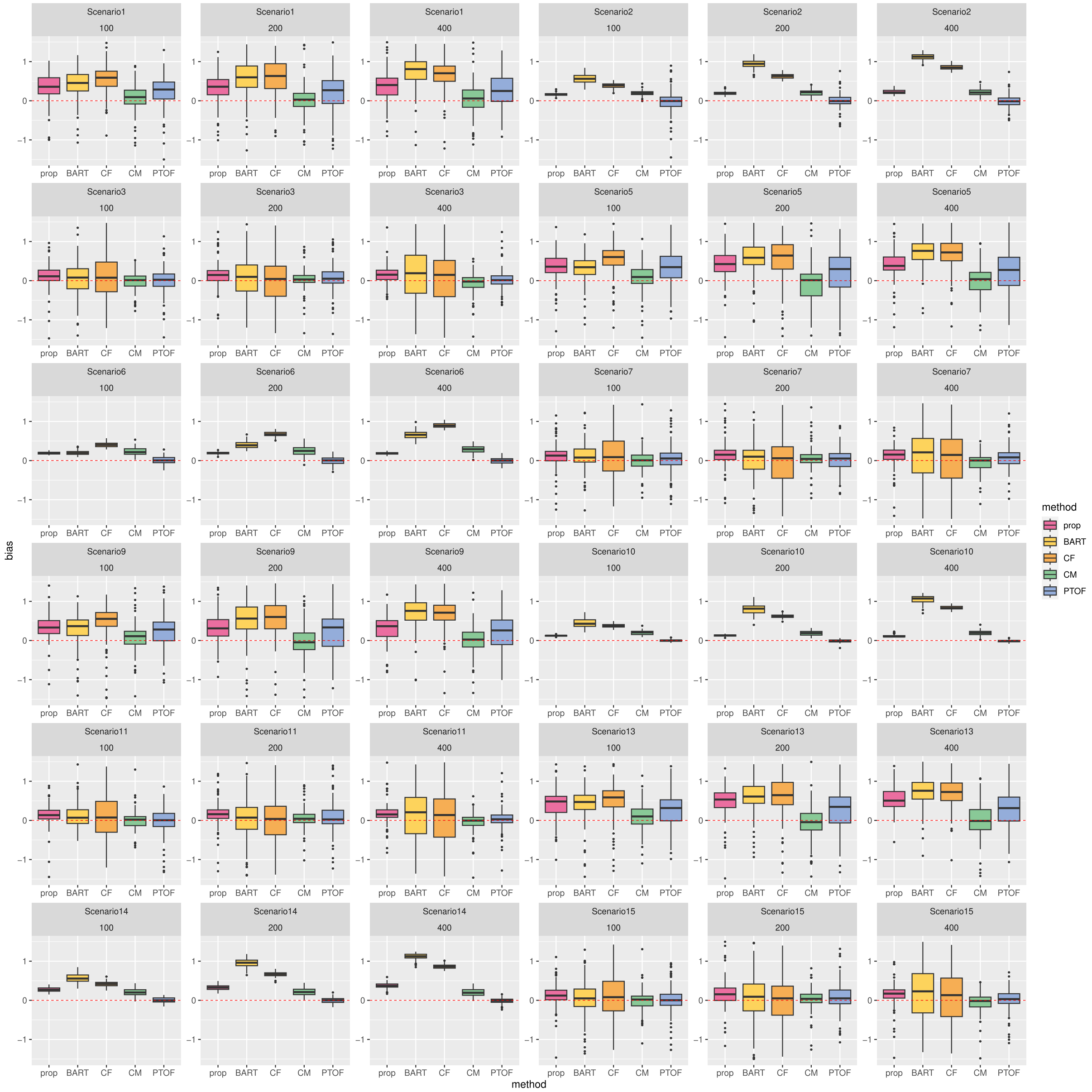

Figure 3. Plots of bias for $n = 600$. Each scenario is depicted by $p = 100, 200$, and 400. The horizontal axis is the method, and the vertical axis is the bias. Scenarios 4, 8, 12, and 16 were excluded due to the bias not being calculated by $\tau(x) = 0$.

Next, we compared the relative bias between τ^* and $\hat{\tau}$, as displayed in Figures 3 and 4. Figure 3 shows the bias of $n = 600$. Each scenario had plots for $p = 100, 200$, and 400. Plots for the bias in Scenarios 4, 8, 12, and 16 were not created because τ^* was set to 0. Almost all results of the median values were positive for all methods. The proposed method, Causal MARS, and PTO forest were mostly stable even when p increased. However, BART and causal forest showed a larger bias as p increased. These results indicate that the proposed method, Causal MARS, and PTO forest were not dependent of the number of p, whereas BART and causal forest increased the bias with the influence of the number of p. Next, we examined the results of each scenario. In Scenarios 1, 5, 9, and 13, the setting of the true HTE was a combination of linear and quadratic functions, the median values of Causal MARS were close to the true τ value, and PTO forest included a bias of 0. In Scenarios 2, 6, 10, and 14, where the HTE was set as a piecewise constant, the median of PTO forest was nearly 0. The proposed method performed better next to PTO forest, however, causal forest and BART were greater than 0.5 at $p = 400$. In Scenarios 3, 7, 11, and 15, the HTE setting was a combination of the sin and exponential functions, the median of almost all methods was close to 0. The results of the proposed method were slightly > 0, whereas the range of the results was narrower than those of causal forest and BART. Figure 4 shows the results for $n = 1000$. The tendency for $n = 1000$ was similar to that of $n = 600$ in all scenarios.

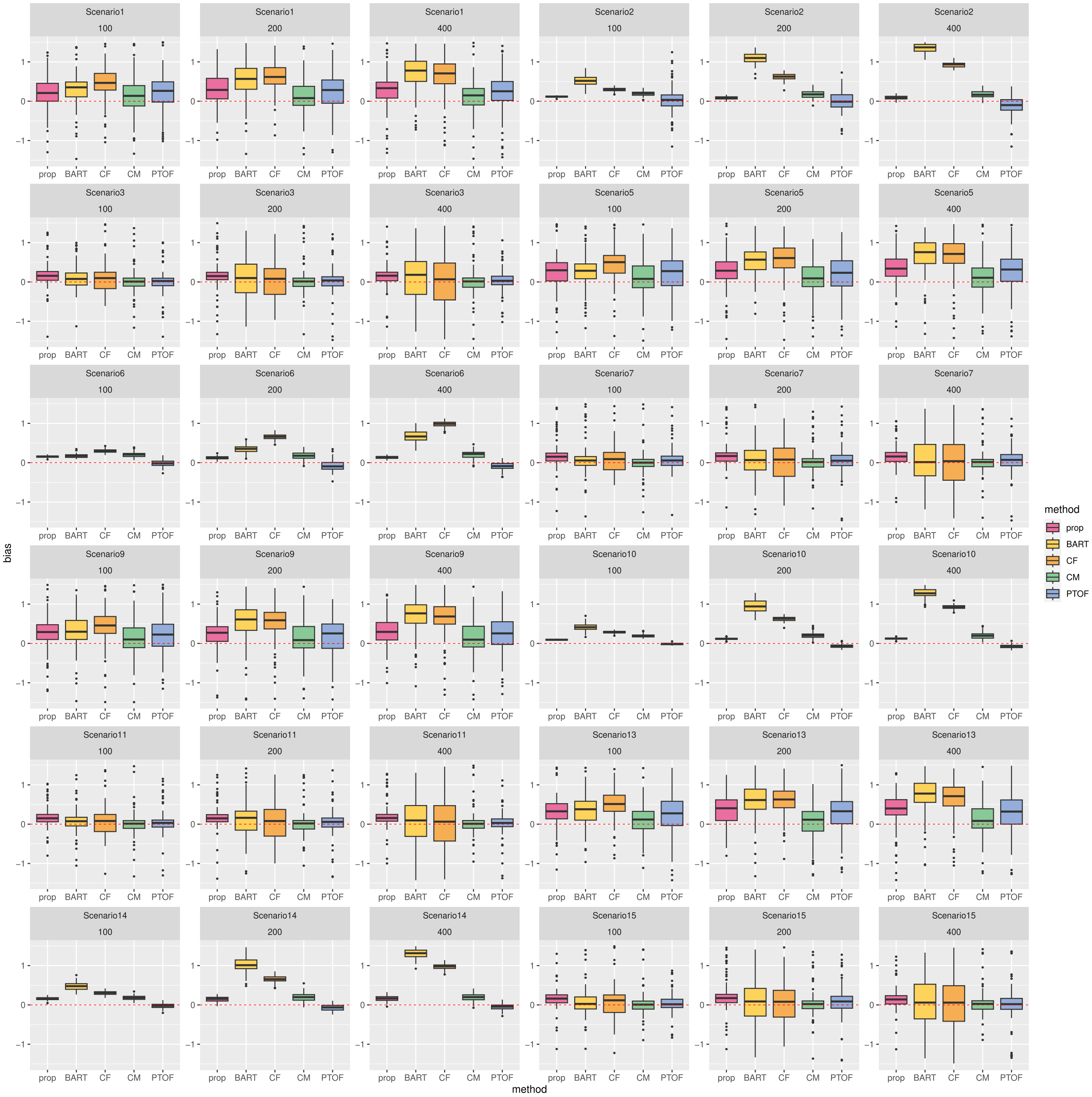

Figure 4. Plots of bias for $n = 1000$. Each scenario is depicted by $p = 100, 200$, and 400. The horizontal axis is the method and the vertical axis is the bias. Scenarios 4, 8, 12, and 16 were excluded due to the bias not being calculated by $\tau(x) = 0$.

We next show the results of the correlations between the estimated treatment effect $\hat{\tau}$ and the true treatment effect τ^*. The results are shown in Figures 5 and 6. Figure 5 presents the results for $n = 600$. The proposed method and the compared methods were stable in almost all scenarios regardless of p, apart from for BART as its correlation was affected by the value of p. Causal MARS had the highest correlation in Scenarios 1, 5, 9, and 13, and the correlation of the proposed method was higher than that of PTO forest. In Scenarios 2, 6, and 14, the proposed method had a higher correlation than the compared methods apart from Scenario 6 when $p = 100$. In Scenarios 3, 7, 11, and 15, the correlation of all methods were close to 0, except for BART at $p = 400$. The results of $n = 1000$ are presented in Figure 6, where the overall correlation was slightly higher than that at $n = 600$; however, the trends of the results for $n = 1000$ were similar to those for $n = 600$.

From these results of MSE, the proposed method exhibited a stable tendency as the number of variables increased, whereas causal forest and BART were affected by the increase in the number of variables. These trends were confirmed by those results of the RBias and correlation coefficients. Moreover, in most scenarios of the true treatment effects that comprise threshold functions, the MSE of the proposed method showed a better performance than the other compared methods and was better or very close to the MSE of Causal MARS and PTO forest for scenarios in which the true treatment

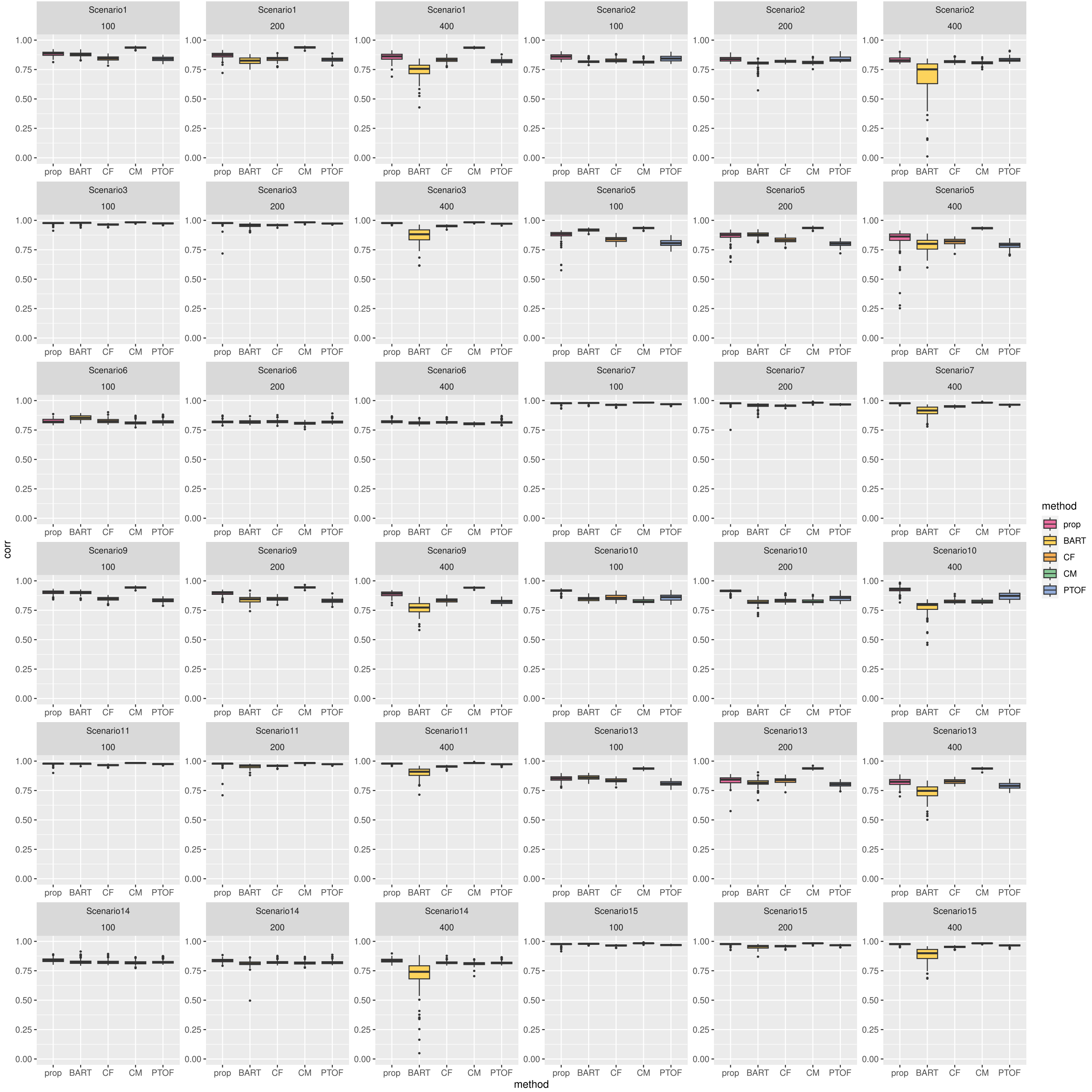

Figure 5. Plots of correlation for $n = 600$. Each scenario is depicted by $p = 100, 200$, and 400. The horizontal axis is the method and the vertical axis is the correlation. Scenarios 4, 8, 12, and 16 were excluded due to the correlation not being calculated by $\tau(x) = 0$.

follows a combination of sin and exponential functions. The tendencies of the results for Rbias and correlation coefficients were similar to those of the MSE. These results confirm the estimation performance of the proposed method with nonlinear structures. However, in the scenarios involving a combination of linear and quadratic functions in the true treatment, Causal MARS performed better than the proposed method, whereas the proposed method yielded better results than the other four compared methods. Causal MARS captured the structure of the linear tendency; however, the estimation performance was inferior for nonlinear structure. In this simulation, we confirmed that the proposed method performed well when the true treatment followed a nonlinear structure, and it maintained an estimation accuracy in the near-linear structure.

5 Real data application

In this section, we demonstrate the usefulness of the proposed method by applying it to actual clinical study data named AIDS Clinical Trials Group Study 175 (ACTG 175)[26] from the package speff2trial[27] in the R software. We also evaluate the HTE estimation by using the same data.

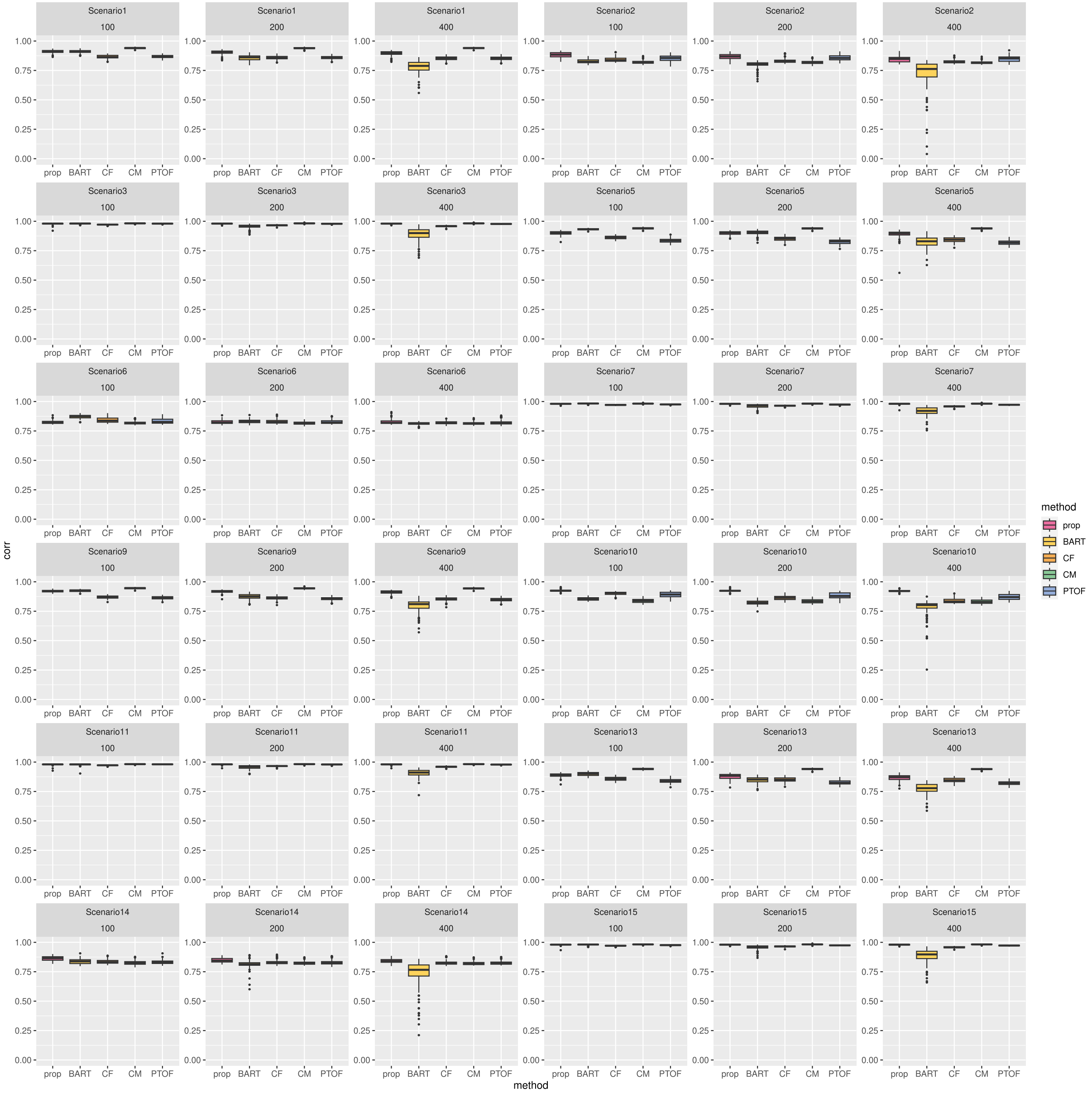

Figure 6. Plots of correlation in $n = 1000$. Each scenario is depicted by $p = 100, 200$, and 400. The horizontal axis is the method and the vertical axis is the correlation. Scenarios 4, 8, 12, and 16 were excluded due to the correlation not being calculated by $\tau(x) = 0$.

5.1 Application of the proposed method to ACTG 175

In this double-blind randomized study, 2139 subjects infected with human immunodeficiency virus type 1 (HIV-1) at 200 to 500 per mL CD4 cell counts were randomly assigned to one of four arms: zidovudine with didanosine, zidovudine with zalcitabine, zidovudine only, or didanosine only. We selected 522 subjects from the zidovudine plus zalcitabine group as the target treatment group and 532 subjects from the zidovudine only group as the control group. The outcome was defined as the difference in CD4 cell counts at 20 ± 5 weeks from their baseline. Table 2 lists the 14 covariates selected. To apply the data, the data of each arm were split into training and test data. The model was fitted using the training data with 527 subjects, and the HTE was predicted for 527 subjects using the test data. For the hyperparameters of the proposed method, the number of trees was set to M =500, the shrinkage rate was set to $\eta = 0.01$, and the mean depth of each tree-based function was set to $\bar{L} = 2$.

Table 2. Selected covariates at baseline in a real data application.

Variable name	Description
age	age in years
wtkg	weight in kg
karnof	Karnofsky score, a scale of 0 – 100
preanti	number of days of antiretroviral therapy previously received
cd40	CD4 cell count at baseline, cells/mm^3
cd80	CD8 cell count at baseline, cells/mm^3
hemo	hemophilia, 0 = no, 1 = yes
homo	homoseuxal activity, 0 = no, 1 = yes
drugs	history of intravenous drug use, 0 = no, 1 = yes
oprior	non-zidovudine antiretroviral therapy prior to initiation of study treatment, 0 = no, 1 = yes
race	0 = white, 1 = non-white
gender	0 = female, 1 = male
str2	antiretroviral history, 0 = naive, 1 = experienced
symptom	symptom indicator, 0 = asymptomatic, 1 = symptomatic

Based on the real data application results, we obtained estimation results and the rules for the proposed method. For the estimation results, we ordered the estimated HTE values in ascending order and divided them into three equal portions: low, middle, and high. If the HTE is properly estimated, the mean of the outcome is expected to be low, middle, and high, corresponding to the low, middle, and high groups, respectively. The procedure used to divide the groups is described below. First, the sample IDs were arranged in ascending order of the estimated HTE. This ordered sample ID list was then divided into three groups: low for 176 subjects, middle for 176 subjects, and high for 175 subjects from the test data. The mean and standard error of the outcome for each treatment group were then calculated for each ordered group. If the HTE is properly estimated, the low group is expected to show the smallest difference in the mean of the outcome between the two treatment groups. Conversely, the high group is expected to show the largest difference in the mean of the outcome. Next, we show the results of the three groups ordered by treatment arm in Figure 7. The bar pairs shown in the left, middle, and right correspond with the results of the small, middle, and high groups, respectively. The green and pink bars represent the target treatment group and the control group, respectively. The differences between the treatment arms mostly increased in the high group, which confirms that the estimated results of the proposed method exhibited a trend of a magnitude of outcome value.

Additionally, we calculated the rule importance and its support to observe the subgroups of the data.[13] An advantage of the RuleFit method is its rule-based interpretability, and the conventional RuleFit method evaluates the importance of the rule and linear terms to the coefficient values. We focused on the rule importance[13] of rule terms related to the HTE values. Support refers to the percentage of subjects who meet the conditions of the rule. The k^*th importance Q_{k^*} for the rules of the proposed method are calculated as follows:

$$Q_{k^*} = \left(|\hat{\beta}^{(a)}_{k^*} - \hat{\beta}^{(c)}_{k^*}| \cdot \sqrt{o_{k^*}(1 - o_{k^*})} \right)$$

where o_{k^*} reflects the support of the rule $r^*_{k^*}$. The support in kth rule is computed as

$$o_{k^*} = \frac{1}{n} \sum_{i=1}^{n} r^*_{k^*}(\boldsymbol{x}_i)$$

In this application, 70 rules were chosen to estimate the HTE. Figure 8 shows the rule importance on the left and its support on the right. On the left side of Figure 8, the rules with high rule-importance values are indicated by pink bars. The pink bar on the right side of Figure 8 indicates support values > 0.7.

Notably, the proposed method can represent the characteristics of the subgroups relevant to the treatment effects as rules. To demonstrate this capability, we depicted a distribution of 70 rules for the estimated HTE and their support values in Figure 9. The vertical axis represents the estimated HTE for each rule, and the horizontal axis represents the support value corresponding to each rule. From this plot, we can observe the overall trend in the results. There was wide variation in the HTE values of each rule. Many of the support values were between 0 and 0.25, while some points showed higher values. In particular, rule # 1, for example, had a high HTE value, whereas its support value was small, indicating that the subgroup that fits rule # 1 did not meet $< 10\%$ of this data. In contrast, the HTE of rule # 63 was $\sim$ 1.23, and its support value

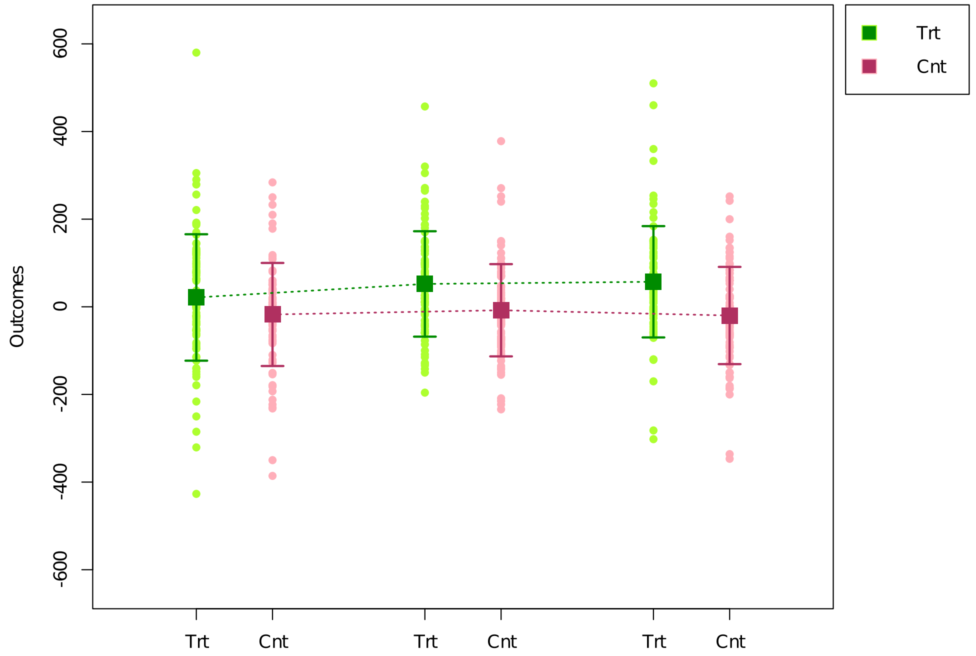

Figure 7. Plots of the distribution of the outcome, mean, and standard error by each arm in the low, middle, and high groups. The vertical axis shows the outcome and the horizontal axis shows the treatment. The green bars and plots represent the treatment group, and the pink bars and plots represent the control group. The leftmost green and pink bars are the low group, whose estimated heterogeneous treatment effect (HTE) values were smaller in order. The middle bars and plots show the middle group, and the rightmost green and pink bars and plots show the high group.

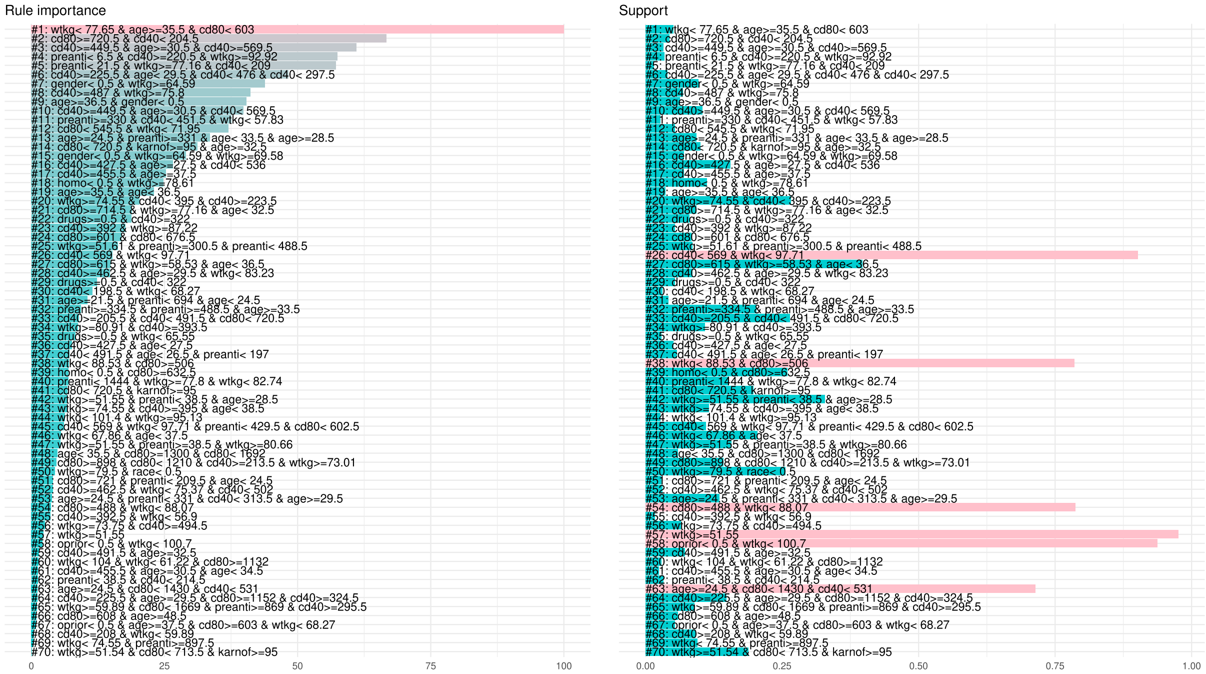

Figure 8. Plots of rule importance and the support of the rule importance. The left bar plot describes rule importance, the horizontal axis shows the importance value of the rule, and the vertical axis shows the rule. The right bar plot is the support of the rules. The horizontal axis describes the support value, which is depicted in pink if the value is > 0.7.

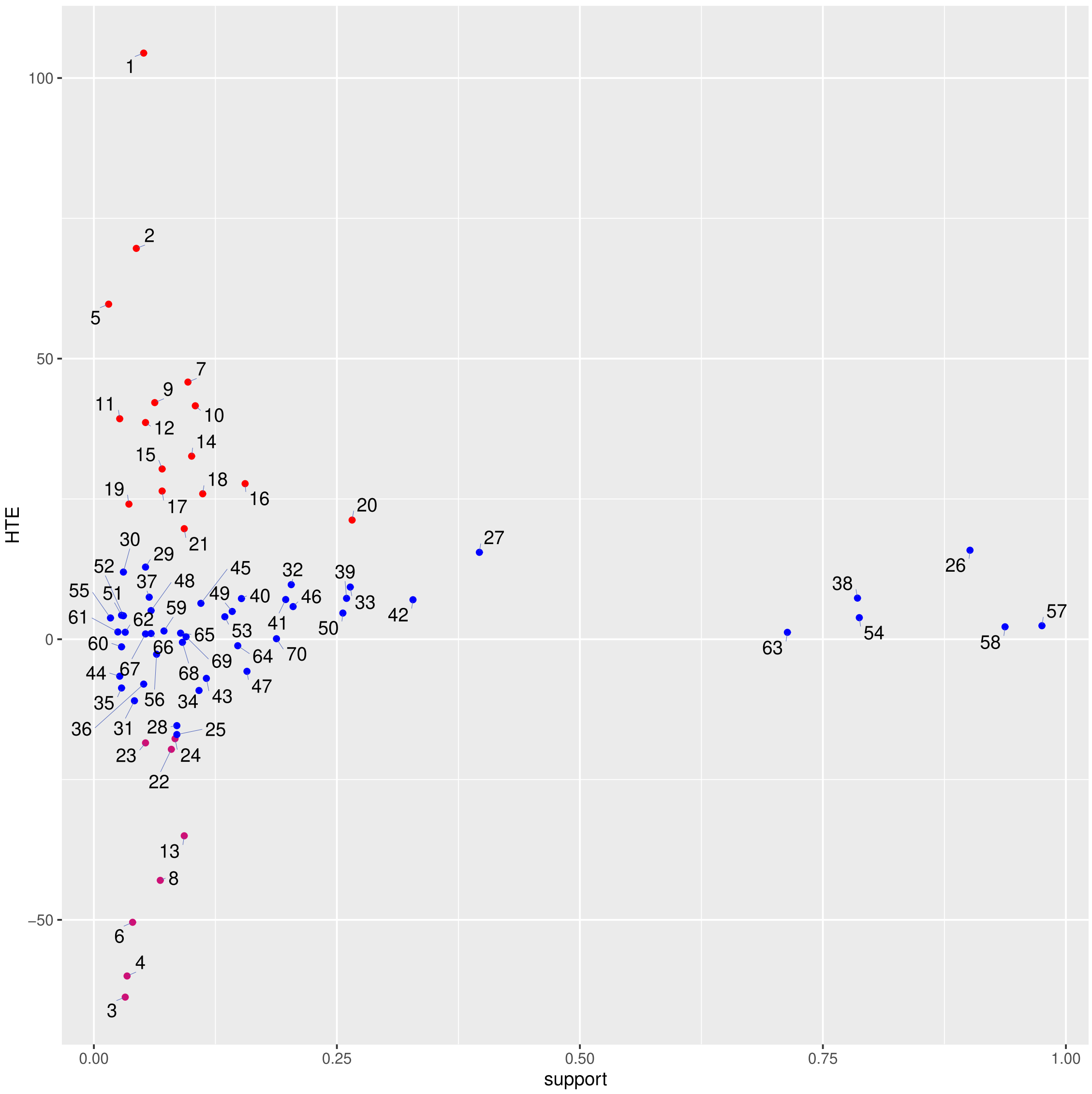

Figure 9. Plot of the distribution of estimated heterogeneous treatment effect (HTE) values of 70 selected rules and their support values. The vertical axis shows the HTE value, and the horizontal axis shows the support value. The number represents the rule number. The red points designate rules whose rule importance is greater than its mean and its HTE ≥ 0. The purple points designate rules whose rule importance is greater than its mean and its HTE < 0.

was ~ 0.71. Although the HTE was not high, 71% of the subjects belonged to this subgroup, indicating that this subgroup was suitable for the subjects in this study. To select the subgroups, we employed which the 16 rules had rule importance more than their mean value, and with HTE value more than 0. The rules depicted by the red points in Figure 9 are listed in Table 3.

The HTE of the selected 16 rules was positive, so the subgroups meeting these rules had more effective treatment with combined zidovudine and didanosine therapy than with zidovudine-only therapy. This result is consistent with the ACTG 175 study results.[26,28] Based on the context of rule “wtkg< 77.65 & age $\geq$ 35.5 & cd80 < 603,” < 77.65 kg and $\geq$ 35.5 years old and < 603 cells/mm^3 in CD8 cell counts benefited from the combination treatment. Figure 10 shows the ATE and the 95% confidence interval (CI) of the selected rules listed in Table 3. The 95% CI for the ATE of 10 selected rules did not include 0, indicating that the treatment group performed better in these rules. The point estimations of the ATE in 9 rules were higher than the overall ATE, and the confidence intervals of rules # 12 and 18 were higher than the overall ATE. This shows that these subgroups benefited from the targeted treatment.

Table 3. Rule importance, the heterogeneous treatment effect (HTE), and its support of the proposed method for a subset of the selected rules.

Rule #	Rule	Rule importance	HTE	Support
1	wtkg $<$ 77.65 & age$\geq$ 35.5 & cd80 $<$ 603	100.00	104.44	0.05
2	cd80 $\geq$ 720.5 & cd40 $<$ 204.5	66.69	69.65	0.04
5	preanti $<$ 21.5 & wtkg $\geq$ 77.16 & cd40$<$ 209	57.17	59.70	0.02
7	gender $<$ 0.5 & wtkg $\geq$ 64.59	43.87	45.82	0.10
9	age $\geq$ 36.5 & gender $<$ 0.5	40.37	42.16	0.06
10	cd40 $\geq$ 449.5 & age$\geq$ 30.5 & cd40$<$ 569.5	39.83	41.60	0.10
11	preanti $\geq$ 330 & cd40 $<$ 451.5 & wtkg $<$ 57.83	37.62	39.29	0.03
12	cd80 $<$ 545.5 & wtkg $<$ 71.95	36.97	38.62	0.05
14	cd80 $<$ 720.5 & karnof $\geq$ 95 & age $\geq$ 32.5	31.24	32.62	0.10
15	gender $<$ 0.5 & wtkg $\geq$ 64.59 & wtkg $\geq$ 69.58	29.05	30.34	0.07
16	cd40 $\geq$ 427.5 & age $\geq$ 27.5 & cd40 $<$ 536	26.55	27.72	0.16
17	cd40 $\geq$ 455.5 & age $\geq$ 37.5	25.28	26.41	0.07
18	homo $<$ 0.5 & wtkg $\geq$ 78.61	24.82	25.92	0.11
19	age $\geq$ 35.5 & age $<$ 36.5	23.05	24.08	0.04
20	wtkg $\geq$ 74.55 & cd40 $<$ 395 & cd40 $\geq$ 223.5	20.33	21.24	0.27
21	cd80 $\geq$ 714.5 & wtkg $\geq$ 77.16 & age $<$ 32.5	18.87	19.70	0.09

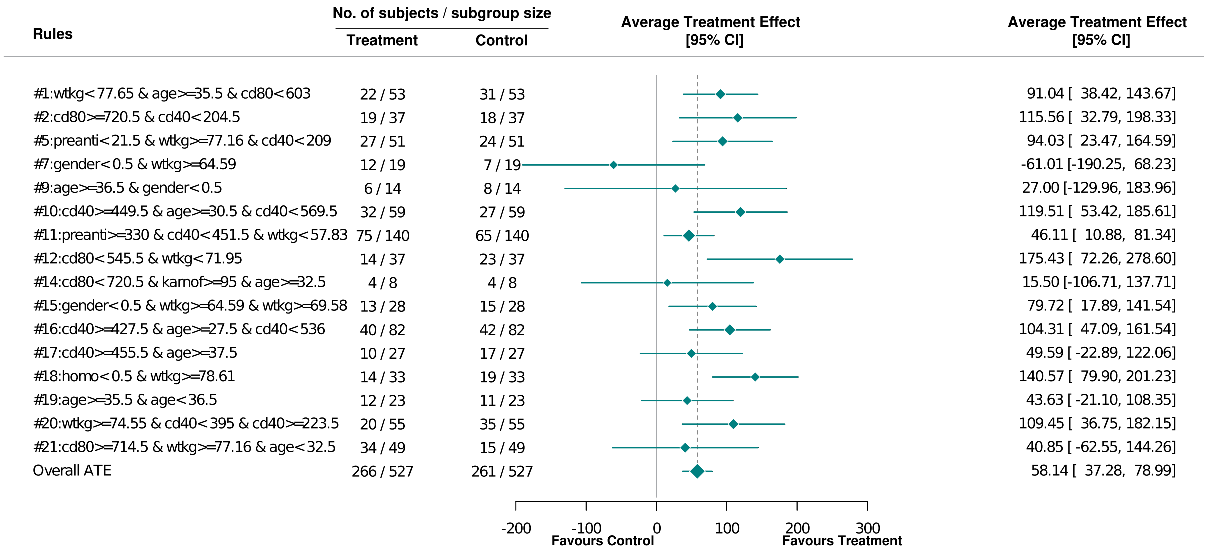

Figure 10. Forest plot of the rules in Table 3. The rules selected in Table 3 are listed in the first column, the subgroup size of each treatment group is shown in the second and third columns, respectively, the average treatment effect of each rule is depicted in the fourth column, and its 95% confidence interval is in the fifth column.

5.2 Comparison between the proposed method and control method

In this subsection, we report the evaluation of predictions using real data. The proposed method is expected to enhance the estimation accuracy through the combination of the prognostic and prescriptive rule terms. Therefore, we compare the prediction accuracy in the real data application between the proposed method and the RuleFit method which contains rule term for only the prognostic effect and linear term. We used the same data in Section 5.1. We conducted 10-fold cross-validation and the mean of root mean squared error (RMSE) was calculated as

$$RMSE = \frac{1}{n^{\dagger}} \sum_{i=1}^{n^{\dagger}} \sqrt{\left(y_i^{*} - \hat{y}_i\right)^2}$$

where y_i^* $(i = 1, 2, \ldots, n^{\dagger})$ is the outcome in the test data and $n^{\dagger}$ is the size of each fold.

Next, we explain the procedure of the evaluation. It was performed based on the following steps:

STEP 1: The data is randomly split into 10 as $(\boldsymbol{y}_w, (\boldsymbol{X}_w, \boldsymbol{z}_w))$ $(w = 1, 2, \dots, 10)$.
STEP 2: Set $w = 1$.
STEP 3: Let the wth split data as the test data and the merged data of $w \neq l$ as the training data.
STEP 4: Apply equation (5) in the proposed method to the training data and estimate all regression parameters of $\hat{\beta}_0, \hat{\beta}_{k^\dagger}, \hat{\alpha}_j, \hat{\beta}^{(a)}_{k^*}$ and $\hat{\beta}^{(c)}_{k^*}$.
STEP 5: With the parameters in **STEP 4**, equation (16) is employed by the proposed method to calculate the prediction value of the test data, and the RMSE is computed. In contrast, the control method used equation (17) for the prediction and calculated the RMSE.

$$
\begin{aligned}
\hat{y}_{proposed} = \hat{\beta}_0 &+ \sum_{k^\dagger=1}^{K^\dagger} \hat{\beta}_{k^\dagger} r^\dagger_{k^\dagger}(\boldsymbol{x}_i) + \sum_{j=1}^{p} \hat{\alpha}_j l_j(\boldsymbol{x}_i) \\
&+ \sum_{k^*=1}^{K^*} \hat{\beta}^{(a)}_{k^*} r^*_{k^*}(\boldsymbol{x}_i) \cdot I(z_i = 1) + \sum_{k^*=1}^{K^*} \hat{\beta}^{(c)}_{k^*} r^*_{k^*}(\boldsymbol{x}_i) \cdot I(z_i = 0)
\end{aligned} \tag{16}
$$

$$
\hat{y}_{control} = \hat{\beta}_0 + \sum_{k^\dagger=1}^{K^\dagger} \hat{\beta}_{k^\dagger} r^\dagger_{k^\dagger}(\boldsymbol{x}_i) + \sum_{j=1}^{p} \hat{\alpha}_j l_j(\boldsymbol{x}_i) \tag{17}
$$

where $\hat{y}_{proposed}$ and $\hat{y}_{control}$ represent the predicted values by the proposed method and control method, respectively, and $\boldsymbol{x}_i$ and z_i are the covariates and the treatment arm of the test data. Here, the parameters of the proposed method, $\hat{\beta}_0, \hat{\beta}_{k^\dagger}, \hat{\alpha}_j, \hat{\beta}^{(a)}_{k^*}$, and $\hat{\beta}^{(c)}_{k^*}$, are obtained from equation (16). The coefficients $\hat{\beta}_0, \hat{\beta}_{k^\dagger}$, and $\hat{\alpha}_j$ and the rule term for prognostic effect and linear term in equation (17) were the same as those of the proposed method in equation (16).
STEP 6: If $w = 10$, this process ends. If $w \neq 10$, return to **STEP 3**.

The mean (standard deviation) of the RMSE of the proposed method was 118.44 (11.86), while that of the rule ensemble comprising the prognostic rule term and linear term was 137.16 (13.93). The proposed method was confirmed to provide reasonable prediction accuracy for this data application by including not only the prognostic rule term, but also the rule term related to the HTE.

6 Discussion and conclusion

We proposed a novel framework based on the RuleFit method to estimate the HTE. Through numerical simulations, we found in many scenarios that compared with other methods, the proposed method estimates with good stability regardless of the number of covariates . Based on Spearman's correlation coefficient results, we confirmed that the proposed method captures the appropriate order of the magnitude of correlation coefficients between the true and predicted treatment effects. In most scenarios, the MSE of the proposed method was better than that of the compared methods, where the true treatment effects were formed from the threshold function. The relative bias in these scenarios was nearly 0 and had a high correlation. In scenarios where the treatment effects consisted of a combination of sin and exponential functions, the MSE values of the proposed method were close to those of Causal MARS and PTO forest. The proposed method maintained a low bias, and its correlation values were nearly the same as those of Causal MARS and PTO forest in these scenarios. Conversely, when the true treatment effect comprised both linear and quadratic functions, the MSE of Causal MARS was better than that of the proposed method. Nonetheless, the proposed method exhibited superior performance over Causal MARS in scenarios in which the true treatment effects included threshold functions or no treatment effects. The results of these simulations confirmed that the proposed method performed equivalently to the compared methods. Focusing on the bias between the differences of $n = 600$ and 1000 in the proposed method, the median of bias in $n = 1000$ was closer to 0 compared to $n = 600$ in $p = 200$ and 400 for Scenario 2 and in all p for Scenarios 13 and 14. With these settings, the values of the bias difference between $n = 1000$ and $n = 600$ were around 0.1.

In contrast, the proposed method generates rules by using the rpart module of the R package, while bias issues in rpart has been raised.[29,30] Therefore, we also compared the proposed method with rpart to the proposed method that generates rules using the method with conditional inference trees (ctree) in partykit.[31] The evaluation indices were the same as those used for the numerical simulation. The overall results of the proposed method with rpart were better than those of the proposed method with ctree. Notably, ctree tends to exhibit lower complexity when generating rules.[19] The generated rules were divided into those associated with the HTE and those associated with the prognostic effect in the proposed method. That is, the number of rules related to HTE was lower than that of the original RuleFit method. As a result, the proposed method with ctree was unable to estimate the HTE in many cases. From additional simulations, the use of `rpart` was confirmed to be acceptable in the proposed method. Furthermore, adaptive group lasso[32] may be a candidate for the regularization of the proposed method. As a result of the comparison of the proposed method with adaptive group lasso, which used `rpart` for rule generation, the proposed method with group lasso had a better performance in many scenarios. Adaptive group lasso estimates less important variables as 0 more often than group lasso. Additionally, the rules for HTE are selected less than the original RuleFit. Therefore, we considered that these may have led to estimating many number of HTE as 0 in the proposed method with adaptive group lasso, similar to the proposed method using ctrees. Therefore, the group lasso was found to be reasonable for regularization based on our simulation. The plots of the additional simulation results are presented as Supplemental Material.

Through its application to real clinical trial data, we confirmed the usefulness of the proposed method in terms of the interpretability of the estimated results using the estimated rules and the validity of the selected rules through 10-fold cross-validation. The selected rules were visually represented by the HTE value, rule importance, and support value, which provided suggestions for the interpretation of subgroup characteristics. Though we confirmed the interpretability of the rules as a model, whether the selected rules can be interpretable in practice requires further consideration, because there may be varying interpretations depending on the different domain perspectives. Discussion with specialists in clinical practice is therefore required.

Here, focusing on metalearners,[33] the proposed method corresponds to S-learner, which uses the rule function as the base function to estimate HTE. Metalearner is a framework in machine learning for causal inference that estimates HTE. Although other learner methods, such as T-learner (where "T" denotes "two"), are formed by two models per treatment group, S-learner, where "S" denotes "single," provides a single model for HTE estimation. The proposed method is based on the S-learner due to the structure of the framework, which uses the rule function as the base function for estimating the HTE. For both S-learner and T-learner, HTE is estimated directly using the predictions of the regression model fitted to the responses. However, because T-learner constructs models separately between each treatment group, it is difficult to estimate the HTE considering that the treatment and control group share common effects.[33] The results of previous numerical simulations[34] showed that S-learner had a better performance than T-learner in some situations. Moreover, when calculating the HTE based on the difference between each treatment group with T-learner, the common effects between the two treatment groups were also estimated separately. This makes it difficult to interpret treatment-specific effects because the estimated treatment effect includes the prognostic effect. S-learner is the only learner that allows the construction of a model with prognostic effect and interaction terms. Additionally, the proposed framework adopted the idea of a shared basis,[9] which is based on T-learner, using the S-learner framework to ensure the comparability of the HTE between the two treatment groups. We therefore incorporated the advantages of S-learner into the proposed method for easy selection of the same rules, considering that the method contains the rule terms of both treatment groups in one model.

As for future works, five things need to be considered. First, further settings should be evaluated in the numerical simulations to clarify the effectiveness of the proposed method. The treatment effect in equation (14) was set symmetrically following previous studies, but this setting should be further examined. Next, although the MSE of the proposed method did not vary depending on the number of p in the present setting, it is necessary to further examine how the estimation results will change with the value of p in other settings. Additionally, the proposed method compared the bias between $n = 600$ and 1000, however, it is necessary to verify what changes in the performance of bias in increasing n. In addition to these simulations, theoretical analysis also needs to be conducted. Fourth, we showed that the proposed method provided rules related to treatment effect. However, in clinical practice, interpreting rules requires knowledge of the relevant domain, so discussion with professionals is required in the interpretation of the selected rules. In the real data applications considered in this study, we found that the results of proposed method were comparable to those of causal forest in terms of variable importance. We have provided the evaluation results as Supplemental Material. However, these findings require confirmation with other real data. Finally, Figure 7 visualizes the HTE evaluation, however, the true value of HTE remains unknown. Addressing this issue should be considered from various perspectives, such as the method proposed by Yadlowsky et al.[35]

Acknowledgements

The authors sincerely thank the two reviewers and the associate editor for their assessments and constructive comments related to improving this article. The authors are also grateful to Benjamin Phillis at the Clinical Study Support Center, Wakayama Medical University Hospital, for his helpful advice in refining the language and expression when improving this article.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship and/or publication of this article.

ORCID iDs

Mayu Hiraishi https://orcid.org/0000-0001-9944-9569
Ke Wan https://orcid.org/0000-0002-1563-7181

Supplemental material

Supplemental material for this article is available online.

References

1. Holland PW. Statistics and causal inference. *J Am Stat Assoc* 1986; **81**: 945–960.
2. Gail M and Simon R. Testing for qualitative interactions between treatment effects and patient subsets. *Biometrics* 1985; **41**: 361–372.
3. Wager S and Athey S. Estimation and inference of heterogeneous treatment effects using random forests. *J Am Stat Assoc* 2018; **113**: 1228–1242.
4. Breiman L, Friedman R, Olshen J, et al. *Classification and regression trees*. New York: Wardsworth, 1984.
5. Su X, Tsai CL, Wang H, et al. Subgroup analysis via recursive partitioning. *J Mach Learn Res* 2009; **10**: 141–158.
6. Athey S and Imbens G. Recursive partitioning for heterogeneous causal effects. *Proce Nat Acad Sci* 2016; **113**: 7353–7360.
7. Breiman L. Random forests. *Mach Learn* 2001; **45**: 5–32.
8. Athey S, Tibshirani J and Wager S. Generalized random forests. *Ann Stat* 2019; **47**: 1148–1178.
9. Powers S, Qian J, Jung K, et al. Some methods for heterogeneous treatment effect estimation in high dimensions. *Stat Med* 2018; **37**: 1767–1787.
10. Chipman HA, George EI and McCulloch RE. Bart: Bayesian additive regression trees. *Ann Appl Stat* 2010; **4**: 266–298.
11. Hill JL. Bayesian nonparametric modeling for causal inference. *J Comput Graph Stat* 2011; **20**: 217–240.
12. Hahn PR, Murray JS and Carvalho CM. Bayesian regression tree models for causal inference: regularization, confounding, and heterogeneous effects (with discussion). *Bayesian Anal* 2020; **15**: 965–1056.
13. Friedman JH and Popescu BE. Predictive learning via rule ensembles. *Ann Appl Stat* 2008; **2**: 916–954.
14. Bargagli-Stoffi FJ, Cadei R, Lee K, et al. Causal rule ensemble: Interpretable discovery and inference of heterogeneous treatment effects. *arXiv:2009.09036v4*, 2023.
15. Yuan M and Lin Y. Model selection and estimation in regression with grouped variables. *J R Stat Soc: Ser B (Statistical Methodology)* 2006; **68**: 49–67.
16. Tibshirani R. Regression shrinkage and selection via the lasso. *J R Stat Soc: Ser B (Methodological)* 1996; **58**: 267–288.
17. Wan K, Tanioka K and Shimokawa T. Rule ensemble method with adaptive group lasso for heterogeneous treatment effect estimation. *Stat Med* 2023; **42**: 3413–3442.
18. Wan K. *Causal-Rule-Ensemble-grfit-*. URL https://github.com/Kwan12321/Causal-Rule-Ensemble-grfit-/blob/main/. GitHub repository. 2023.
19. Fokkema M. Fitting prediction rule ensembles with R package pre. *J Stat Softw* 2020; **92**: 1–30.
20. Friedman JH. Greedy function approximation: a gradient boosting machine. *Ann Stat* 2001; **29**: 1189–1232.
21. Therneau T and Atkinson B. *rpart: Recursive Partitioning and Regression Trees*. URL https://CRAN.R-project.org/package=rpart. R package version 4.1.16., 2022.
22. Breheny P. grpreg: Regularization Paths for Regression Models with Grouped Covariates. R package version 3.4.0., 2021.
23. Tian L, Alizadeh AA, Gentles AJ, et al. A simple method for estimating interactions between a treatment and a large number of covariates. *J Am Stat Assoc* 2014; **109**: 1517–1532.
24. Tibshirani J, Athey S, Sverdrup E, et al. *grf: Generalized random forests*. URL https://CRAN.R-project.org/package=grf. R package version 2.2.1., 2022.
25. Powers S, Qian J, Hastie T, et al. *causalLearning: Methods for heterogeneous treatment effect estimation*. R package version 1.0.0., 2022.
26. Hammer SM, Katzenstein DA, Hughes MD, et al. A trial comparing nucleoside monotherapy with combination therapy in HIV-infected adults with cd4 cell counts from 200 to 500 per cubic millimeter. *New Engl J Med* 1996; **335**: 1081–1090.
27. Juraska M, Gilbert PB, Lu X, et al. speff2trial: Semiparametric efficient estimation for a two-sample treatment effect. URL https://CRAN.R-project.org/package=speff2trial. R package version 1.0.5., 2022.
28. Saravolatz LD, Winslow DL, Collins G, et al. Zidovudine alone or in combination with didanosine or zalcitabine in HIV-infected patients with the acquired immunodeficiency syndrome or fewer than 200 cd4 cells per cubic millimeter. Investigators for the terry beirn community programs for clinical research on aids. *N Engl J Med* 1996; **335**: 1099–1106.
29. Hothorn T, Hornik K and Zeileis A. Unbiased recursive partitioning: a conditional inference framework. *J Comput Graph Stat* 2006; **15**: 651–674.
30. Shih YS and Tsai HW. Variable selection bias in regression trees with constant fits. *Comput Stat Data Anal* 2004; **45**: 595–607.

31. Hothorn T and Zeileis A. *partykit: A modular toolkit for recursive partytioning in R*. URL http://partykit.r-forge.r-project.org/partykit/. *Journal of Machine Learning Research*, 2015, **16**: 3905–3909.
32. Wang H and Leng C. A note on adaptive group lasso. *Comput Stat Data Anal* 2008; **52**: 5277–5286.
33. Künzel SR, Sekhon JS, Bickel PJ, et al. Metalearners for estimating heterogeneous treatment effects using machine learning. *Proce Nat Acad Sci* 2019; **116**: 4156–4165.
34. Nie X and Wager S. Quasi-oracle estimation of heterogeneous treatment effects. *Biometrika* 2020; **108**: 299–319.
35. Yadlowsky S, Fleming S, Shah N, et al. *Evaluating treatment prioritization rules via rank-weighted average treatment effects*. *Arxiv.* arXiv:2111.07966, 2021.