

---

# M-learner:A Flexible And Powerful Framework To Study Heterogeneous Treatment Effect In Mediation Model

---

**Xingyu Li**

Department of Biostatistics, The University of Texas MD Anderson Cancer Center

**Qing Liu**

Center for Design and Analysis, Amgen

**Tony Jiang**

Center for Design and Analysis, Amgen

**Hong Amy Xia**

Center for Design and Analysis, Amgen

**Brian P. Hobbs**

Telperian

**Peng Wei**

Department of Biostatistics, The University of Texas MD Anderson Cancer Center

## Abstract

We propose a novel method, termed the M-learner, for estimating heterogeneous indirect and total treatment effects and identifying relevant subgroups within a mediation framework. The procedure comprises four key steps. First, we compute individual-level conditional average indirect/total treatment effect Second, we construct a distance matrix based on pairwise differences. Third, we apply t-SNE to project this matrix into a low-dimensional Euclidean space, followed by K-means clustering to identify subgroup structures. Finally, we calibrate and refine the clusters using a threshold-based procedure to determine the optimal configuration. To the best of our knowledge, this is the first approach specifically designed to capture treatment effect heterogeneity in the presence of mediation. Experimental results validate the robustness and effectiveness of the proposed framework. Application to the real-world Jobs II dataset highlights the broad adaptability and potential applicability of our method. Code is available at <https://anonymous.4open.science/r/M-learner-C4BB>.

## 1 Introduction

Randomized clinical trials are often costly and time-consuming, with significant delays between treatment administration and the observation of final outcomes such as the response variable  $Y$ . However, mediators can serve as early indicators of treatment efficacy. For example, in colorectal cancer trials, cessation of tumor growth may act as a proxy for positive treatment response. This enables interim evaluation of treatment efficacy and the potential to adapt treatment strategies mid-trial based on changes in the mediator. In practice, treatment effects may be mediated only in a subset of patients. Furthermore, the chosen mediator may not be valid, or treatment effects may be homogeneous (i.e., uniformly effective or ineffective across individuals). Existing approaches—such as the T-learner Künzel et al. [2019] and Causal Random Forest Athey and Imbens [2016] have focused on heterogeneous treatment effect estimation but do not account for mediation mechanisms.

For heterogeneous total treatment effects, Foster et al. [2011] introduced the Virtual Twins to estimate heterogeneous total treatment effects with RCT data. Several methodological approaches within observational studies enable the estimation of flexible and accurate models of heterogeneous total treatment effects. Shalit et al. [2017], Shi et al. [2019], Johansson et al. [2016], Hassanzadeh and Greiner [2019] use neural networks to estimate heterogeneous total treatment effects. Athey and Imbens [2016], Wager and Athey [2018], Athey and Wager [2019] use random forests to estimate heterogeneous total treatment effects. Künzel et al. [2019] proposed the Meta-learners which consist of S-learner, T-learner, and X-learner. Nie and Wager [2021] proposed R-learner, Zhang et al. [2022] extended the R-learner to continuous treatment scenario. And other methods have been proposed, such as DR-learner and Lp-R-learner [Kennedy, 2023]. Dwivedi et al. [2020] proposed model selection solution in causal inference.

When a mediator is present, the primary analytical focus is on evaluating the indirect treatment effect on the outcome via the mediator. These methods decompose the average treatment effect into direct treatment effect and indirect effect[Lin et al., 1997, Preacher, 2015, Robins and Greenland, 1992, Petersen et al., 2006, van der Laan and Petersen, 2008, Imai et al., 2010, Tchetgen and Shpitser, 2012, VanderWeele, 2015, Vandenberghe et al., 2017, Dorresteijn et al., 2011, Pearl, 2022, VanderWeele and Vansteelandt, 2009, Vansteelandt and VanderWeele, 2012, Angrist, 2004, Imbens, 2004]. Recently, to address commonly observed intermediate confounders that would be affected by the covariates and then affect both mediators and outcome, multiple methods have been developed to extend the classical mediation analysis[Tchetgen and VanderWeele, 2014, Díaz et al., 2021, 2023, Gilbert et al., 2024]. Ge et al. [2023, 2025], Luo et al. [2025] and Wang and Song [2025] proposed the method to use reinforcement learning to deal with the dynamic mediation analysis. Cheng et al. [2022] use deep learning to estimate causal effects in mediation model. However, these methods do not account for treatment effect heterogeneity in mediation model. In contrast, our approach is the first to incorporate heterogeneity of treatment effects within a mediation framework. This enables the identification of subgroups characterized by distinct indirect treatment effects and provides a principled way to evaluate the effectiveness of the mediator and causal decision making.

To bridge these gaps, we propose a flexible and powerful method, termed M-learner, designed to capture heterogeneity in treatment effects mediated by a mediator, we aim to estimate the Conditional Average Total Treatment Effect (CATTE) and the Conditional Average Indirect Treatment Effect (CAITE) from randomized clinical trial (RCT) data, and to identify subpopulations that benefit from treatment by examining heterogeneity in these effects.

Our approach facilitates the identification of subgroups exhibiting differential response patterns through mediated pathways, thereby improving the interpretability and clinical relevance of treatment effect estimation.

The key contributions of this paper are:

1. We propose M-learner, a flexible and powerful method for estimating the conditional indirect treatment effect, which operates independently of any specific model architecture.
2. We introduce a novel clustering approach for subgroup identification based on treatment effect heterogeneity. By leveraging differences in treatment effects, our method reformulates the unsupervised task of discovering benefiting subgroups into a supervised learning problem.
3. We propose a novel calibration framework to assess the effectiveness of mediators and evaluate the reliability of identified subgroups.

## 2 Models

### 2.1 Model Formulation

We assume the superpopulation or distribution  $\mathcal{P}$  from which a realization of  $N$  independent random variables is given as the training data. That is,

$$(Y_i(0, M_i(0)), Y_i(1, M_i(0)), Y_i(0, M_i(1)), Y_i(1, M_i(1)), X_i, W_i, M_i(1), M_i(0)) \sim \mathcal{P},$$

where  $X_i \in \mathbb{R}^d$  is a d-dimensional covariate or feature vector,  $W_i \in \{0, 1\}$  is the treatment assignment indicator,  $Y_i(0, M_i(0)) \in \mathbb{R}$  is the potential outcome of unit  $i$  when  $i$  is assigned to control

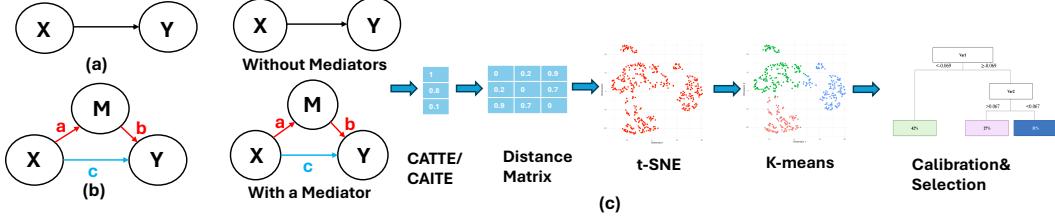


Figure 1: (a)Directed Acyclic Graph between the covariates  $X$  and response  $Y$ , (b)Directed Acyclic Graph between the mediator  $M$ , covariates  $X$  and response  $Y$ ,(c)Pipeline of the M-learner.

group, and the mediator  $M$  is assigned to control group,  $Y_i(0, M_i(1)) \in \mathbb{R}$  is the potential outcome of unit  $i$  when  $i$  is assigned to control group, and the mediator  $M$  is assigned to treatment group,  $Y_i(1, M_i(0)) \in \mathbb{R}$  is the potential outcome of unit  $i$  when  $i$  is assigned to treatment group, while the mediator  $M$  is assigned to control group,  $Y_i(1, M_i(1)) \in \mathbb{R}$  is the potential outcome of unit  $i$  when  $i$  is assigned to treatment group, and the mediator  $M$  is assigned to treatment group. What's more, this paper follows the four assumptions specified in the VanderWeele and Vansteelandt [2009] regarding the interrelations among covariate  $X$ , outcome  $Y$ , and the mediator  $M$ . When a mediator  $M$  is present, the relationships among  $X$ ,  $Y$ , and  $M$  follow the structure shown in the Directed Acyclic Graph (DAG) in Figure 1 (b). The treatment effect can be transmitted through two distinct pathways. If the treatment effect is transmitted along path  $a \rightarrow b$  via the mediator  $M$ , this is referred to as the indirect treatment effect (ITE). If the treatment effect is transmitted directly to  $Y$  along path  $c$ , we refer to it as the direct treatment effect (DTE). The total treatment effect (TTE) is the combined influence of both the direct and indirect pathways. These effects are formally defined as follows,

$$\begin{aligned} ITE &= Y(1, M(1)) - Y(0, M(1)), DTE = Y(0, M(1)) - Y(0, M(0)), \\ TTE &= Y(1, M(1)) - Y(0, M(0)) = ITE + DTE. \end{aligned}$$

When the mediator  $M$  is absent, the relationship between  $X$  and  $Y$  corresponds to the DAG in Figure 1 (a). In this case, the treatment effect is transmitted to  $Y$  through a single pathway, representing the total treatment effect only.

For a new unit  $i$  with covariate vector  $x_i$ , to decide whether to give the treatment via total treatment effect, we wish to estimate the individual total treatment effect (ITTE) of each  $i$ ,  $D_i$ , which is defined as  $D_i := Y_i(1, M(1)) - Y_i(0, M(0))$ . However, we can not observe  $D_i$  for any unit, and  $D_i$  is not identifiable without strong assumptions. Instead, Künzel et al. [2019] proposed that we can estimate the CATTE function, which is defined as  $\tau^{TOT}(x) = \mathbb{E}[D|X = x] = \mathbb{E}[Y(1, M(1)) - Y(0, M(0))|X = x]$ , the best estimator for the CATEE is also the best estimator for the ITTE in terms of the mean squared error (MSE). Let  $\hat{\tau}_i^{TOT}$  be an estimator of  $D_i$  and decompose the MSE at  $x_i$ ,

$$\mathbb{E}[(D_i - \hat{\tau}_i^{TOT})^2 | X_i = x_i] = \mathbb{E}[(D_i - \tau^{TOT}(x_i))^2 | X_i = x_i] + \mathbb{E}[(\tau^{TOT}(x_i) - \hat{\tau}_i^{TOT})^2]. \quad (1)$$

Because we can not affect the first term in 1, the estimator that minimizes the MSE for the ITTE of  $i$  also minimizes the MSE for the CATTE at  $x_i$ . To decide whether to give the treatment via indirect treatment effect, we wish to estimate the individual indirect treatment effect (IITE) of each  $i$ , where the definition of IITE is  $Y_i(1, M(1)) - Y_i(1, M(0))|X$ . Analogous to Expression 1, the CAITE can also be regarded as the optimal estimator of the IITE under the MSE criterion, where the definition of CAITE is  $\tau^{ITE} := \mathbb{E}[Y(1, M(1)) - Y(1, M(0))|X]$ . With the estimation of CATTE and CAITE, we can use the estimation to study the subgroups of TTE and ITE. For subgroups via TTE, it can be defined as  $\mathbb{E}[Y(1, M(1)) - Y(0, M(0))] \in \mathcal{U}_i^{TTE}|X$ ,  $i = 1, 2, 3, \dots$ , where  $\mathcal{U}_i^{TTE}$  is the set, the number of the sets are unknown, and when there is no heterogeneity, the number of the set is 1. For subgroups via ITE, it can be defined as  $\mathbb{E}[Y(1, M(1)) - Y(1, M(0))] \in \mathcal{U}_i^{ITE}|X$ ,  $i = 1, 2, 3, \dots$ , where  $\mathcal{U}_i^{ITE}$  is the set, the number of the sets are unknown, and when there is no heterogeneity, the number of the set is 1.

We propose a machine learning-based approach to separately estimate the CATTE and the CAITE, with the goal of identifying subgroups characterized by differences in both total and indirect treatment effects. This study addresses several key research questions: How can treatment effects be accurately estimated? How can meaningful and interpretable subgroups be identified based on treatment effect

heterogeneity? Are the resulting subgroups statistically valid and practically useful? Finally, to what extent is the mediator effective—that is, does a significant portion of the treatment effect operate through the mediator?

## 2.2 M-learner

We propose an algorithm called M-learner<sup>1</sup> shown in Figure 1 (c) to address the aforementioned problem. This algorithm consists of three main steps:

1. Estimating individual treatment effect (either the total treatment effect or the indirect treatment effect),
2. Measuring the difference in treatment effects between pairs of samples (referred to as the “distance”),
3. Projecting the distance matrix into a two-dimensional Euclidean space using t-SNE, and clustering the projected data use K-means,
4. Selecting the optimal clustering result and calibration.

When studying treatment effects within models involving mediators, we typically distinguish between the total treatment effect, which does not account for the mediator, and the indirect treatment effect transmitted through the mediator. In the following sections, we systematically discuss methods applicable to different scenarios and propose corresponding solutions.

In the absence of the mediator, the first step of the algorithm estimates the ITTE. At this stage, the T-learner is employed to estimate the CATTE. First, the control response function,  $g_0(x) = \mathbb{E}[Y(0)|X = x]$ , is estimated by a base learner, which could be any supervised learning estimator using the observations in the control group,  $\{(X_i, Y_i)\}_{W_i=0}$  and denoting the estimator by  $\hat{g}_0$ . Second, the treatment response function,  $g_1(x) = \mathbb{E}[Y(1)|X = x]$ , is estimated with a base learner, using the treated observations  $\{(X_i, Y_i)\}_{W_i=1}$  and denoting the estimator by  $\hat{g}_1$ . The T-learner is then obtained as  $\hat{\tau}^{tot}(x) = \hat{g}_1(x) - \hat{g}_0(x)$ .

In the presence of a mediator, the first step of the algorithm estimates the IITE. we use M-learner proposed in our paper to estimate CAITE, which takes three steps. First, estimate the treatment mediator function,  $g_1^M(x) = \mathbb{E}[M(1)|X = x]$ , with a base learner, using the treated observations  $\{(M_i, Y_i)\}_{W_i=1}$  and denoting the estimator by  $\hat{g}_1^M$ . Second, estimate the control mediator function,  $g_0^M(x) = \mathbb{E}[M(0)|X = x]$ , with a base learner, using the treated observations  $\{(X_i, Y_i)\}_{W_i=0}$  and denoting the estimator by  $\hat{g}_0^M$ . Third, fit the treatment response function,  $g_1^Y(x, m) = \mathbb{E}[Y(1, M(1))|X = x, M = m]$ , with a base learner, using the treated observations  $\{(X_i, M_i, Y_i)\}_{W_i=1}$  and denoting the estimator by  $\hat{g}_1^Y$ . So, the M-learner is then obtained as  $\hat{\tau}^{ITE}(x) = \hat{g}_1^Y(x, \hat{g}_1^M(x)) - \hat{g}_1^Y(x, \hat{g}_0^M(x))$ . Now, We have estimated the CATTE and CAITE, how to get the subgroups of different units? We propose a new method to transform the estimation of treatment effects to the clustering.

As the algorithms used to identify subgroups are identical for both CATTE and CAITE scenarios, we henceforth refer to them collectively as treatment effects (TE). The estimated TE for each unit  $i$  is denoted by  $\hat{\tau}_i$ .

First, we evaluate the treatment effect difference between each pair of units  $i$  and  $j$ , which we refer to as the distance. It is defined by  $dis(i, j)$ , The distance metric can be defined as Euclidean distance, Manhattan distance, or other formulations. In the following analysis, we consistently adopt Euclidean distance to compute the distance between units  $i$  and  $j$  which means  $dis(i, j) = (\hat{\tau}_i - \hat{\tau}_j)^2$ . Consequently, an  $n \times n$  distance matrix is obtained. This matrix is then projected into a two-dimensional Euclidean space using t-SNE[Van der Maaten and Hinton, 2008], where each unit  $i$  is assigned a coordinate. Subsequently, K-means clustering is performed on the projected points in the Euclidean space. The range of cluster numbers (from 2 to  $k$ ) for K-means clustering must be specified in advance, based on the sample size and the number of covariates; an excessively large  $k$  can result in substantial overfitting, based on our experience, we recommend  $k = \lfloor \sqrt{d} \rfloor + 2$ , where  $\lfloor \rfloor$  represents rounding down to the nearest integer. Then, a decision tree is employed to model the

---

<sup>1</sup>We refer to the method as the M-learner, as it is specifically designed to operate within the framework of mediation models.

clustering results obtained for each predefined number of clusters, with the objective of mapping the unknown categories to interpretable categorical information. For each leaf of the decision tree, we refer to it as a subtype. We use  $p_{leaf}$  to select the final, unique subtype grouping from different decision tree results. The definition of  $p_{leaf}$  is as follows,

$$M = \beta_1 \text{leaf} + \beta_2 W, \quad (2)$$

$$M = \beta_3 \text{leaf} + \beta_4 W + \beta_5 \text{leaf} * W, \quad (3)$$

the likelihood functions of 2 and 3 are  $L_0$  and  $L_1$ , respectively. Then  $2(\log L_1 - \log L_0)$  follows a chi-squared distribution, where the degrees of freedom correspond to the number of decision tree leaves minus 1. Based on this, we calculate  $p_{leaf}$ . The decision tree result with the minimal  $p_{leaf}$  is chosen as the final subtype classification. When no mediating variable  $M$  is involved, the corresponding expression can be reformulated by substituting  $M$  with the outcome variable  $Y$ .

### 3 Simulation

In real-world datasets, ground truth causal treatment effects are rarely directly observable. Consequently, empirical evaluation of causal inference methods often relies on synthetic data. For such evaluations to yield meaningful conclusions, the synthetic data must closely reflect real-world characteristics. In all experiments, we employed Random Forests(RF) and XGBoost(XGB) as base learners[Breiman, 2001, Chen and Guestrin, 2016]. Unless otherwise stated, all experiments in this section are conducted with a sample size of 1000 and 10 covariates.

#### 3.1 Subgroup analysis without mediators

To identify subgroups that benefit from treatment through total treatment effects, we design four scenarios that reflect varying real-world heterogeneity structures. Specifically, the scenarios represent (i) Simple heterogeneity, (ii) Complex heterogeneity, (iii) Global , and (iv) Null setting. Treatment effect heterogeneity exists in the simple and complex scenarios, whereas it is absent in the global and null scenarios. These controlled simulations allow us to systematically evaluate the performance of our method across different levels and patterns of treatment effect variation. Further experimental details are available in the Appendix A.5.1.

Table 1 presents the simulation results of the unmediated model based on 100 experimental replications, using two different base learners, and reports the frequency with which covariates  $X_1$  and  $X_2$  are included in the identified final subtypes. These results are obtained through a calibration-based procedure. Specifically, we determine a threshold under the Null scenario such that the Type I error rate is controlled at 10%, and apply this threshold to assess the validity of subtype groupings in other scenarios. The empirical cumulative distribution functions (ECDF) of the p-values under different scenarios are illustrated in Figure 2 (a) and (b), the figures elucidates why the  $p_{leaf}$  is instrumental in discerning non-heterogeneous scenarios. In addition, the distribution of the number of covariates contained in the final decision trees is shown in the Appendix A.3 Table A.5.

Table 1: The table summarizes the correct covariates in profiles in each scenario.  $X_1/X_2$ :Final profile contain  $X_1/X_2$ ,  $X_1\&X_2$ : final profile contain both  $X_1$  and  $X_2$ ,  $X_1$  and  $X_2$  are the two variables associated with treatment effect heterogeneity. Each value denotes the count of occurrences across 100 simulations.

| Base Learner | Random Forest |       |       | XGBoost    |       |       |
|--------------|---------------|-------|-------|------------|-------|-------|
|              | Covariates    | $X_1$ | $X_2$ | $X_1\&X_2$ | $X_1$ | $X_2$ |
| Simple       | 100           | 100   | 100   | 100        | 100   | 100   |
| Complex      | 68            | 63    | 57    | 70         | 70    | 68    |
| Global       | 0             | 2     | 0     | 0          | 2     | 0     |
| Null         | 4             | 4     | 2     | 1          | 4     | 1     |

These results indicate that when the underlying heterogeneity is relatively simple, both RF and XGB perform well in identifying heterogeneous treatment groups. As the heterogeneity structure increases in complexity, the performance of the proposed method deteriorates. However, as the complexity of the heterogeneity structure increases, XGB consistently outperforms RF. Moreover, our

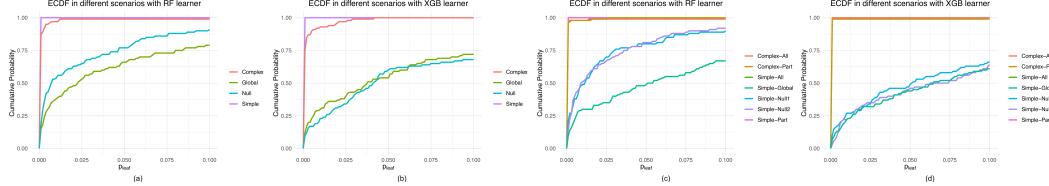


Figure 2: Empirical cumulative distribution functions (ECDF) of  $p_{leaf}$  under different scenarios, using RF and XGB as base learners. These results illustrate the sensitivity of each base learner to underlying treatment effect heterogeneity across varying levels of complexity. (a) results for RF without a mediator, (b) results for XGB without a mediator, (c) results for RF with a mediator, (d) results for XGB with a mediator.

proposed method not only effectively identifies heterogeneous treatment groups but also maintains a low Type I error rate, underscoring its robustness and reliability across a variety of conditions. We compared our algorithm with K-Means, and the results show that K-Means lacks the ability to identify heterogeneous groups ( Appendix A.4 provides detailed procedures and results of the comparative algorithms can be found in Appendix A.4 Table A.12,A.13 and Figure A.9).

### 3.2 Subgroup analysis with a mediator

To identify subgroups that benefit from the treatment effects through the mediator, we design seven scenarios that reflect varying real-world heterogeneity structures . Specifically, the scenarios represent (i) simple heterogeneity, all treatment effect via mediator (Simple-All); (ii) simple heterogeneity, part treatment effect via mediator (Simple-Part); (iii) complex heterogeneity, all treatment effect via mediator (Complex-Part); (iv) complex heterogeneity, part treatment effect via mediator (Complex-Part); (v) simple no heterogeneity, no treatment effect via mediator (Simple-Null1); (vi) simple no heterogeneity, no treatment effect via mediator (Simple-Null2); (vii) simple no heterogeneity, all units benefit from the treatment and all treatment effect via mediator (Simple-Global); further experimental details are available in the Appendix A.5.2.

Table 2 presents the results from 100 replicated simulations of the mediation model, using two different base learners, and reports the frequency with which covariates  $X_1$  and  $X_2$  are included in the identified final subtypes. Table 3 presents a comparison between the subtypes identified by the M-learner and the true heterogeneous regions, based on sample size and mediation proportion, across four heterogeneity scenarios. As shown in Table 3, the subtype regions identified by the M-learner are smaller than the true heterogeneous regions; however, the estimated mediation proportions closely approximate those of the true regions. These results suggest that the M-learner is effective in identifying subtype regions based on the indirect treatment effect (for the detailed computation methods of Table 3, please refer to the Appendix A.5.2). The distribution of the number of covariates contained in the final decision trees is shown in the Appendix A.3 Table A.6. In contrast to the unmediated setting, we calibrate the threshold based on Scenario Simple-Null2 by controlling the Type I error at 10%, and then apply this threshold to assess the validity of subtype identification across other scenarios.

Table 2: The table summarizes the correct covariates in profiles in each scenario.  $X_1/X_2$ :Final profile contain  $X_1/X_2$ ,  $X_1 \& X_2$ : final profile contain both  $X_1$  and  $X_2$ ,  $X_1$  and  $X_2$  are the two variables associated with treatment effect heterogeneity. Each value denotes the count of occurrences across 100 simulations.

| Covariates    | Random Forest |       |              | XGBoost |       |              |
|---------------|---------------|-------|--------------|---------|-------|--------------|
|               | $X_1$         | $X_2$ | $X_1 \& X_2$ | $X_1$   | $X_2$ | $X_1 \& X_2$ |
| Simple-All    | 99            | 99    | 98           | 100     | 100   | 100          |
| Simple-Part   | 100           | 99    | 99           | 100     | 100   | 100          |
| Complex-All   | 84            | 81    | 69           | 98      | 99    | 97           |
| Complex-Part  | 78            | 72    | 54           | 95      | 97    | 93           |
| Simple-Null1  | 9             | 6     | 4            | 8       | 4     | 2            |
| Simple-Null2  | 4             | 4     | 1            | 2       | 1     | 0            |
| Simple-Global | 2             | 2     | 1            | 3       | 5     | 1            |

Table 3: Comparison of true and M-learner–estimated mediation proportions and sample sizes within heterogeneous treatment effect regions across scenarios. Standard deviations are shown in parentheses.

| Scenario     | Ground Truth |            | Random Forest |            | XGBoost |            |
|--------------|--------------|------------|---------------|------------|---------|------------|
|              | N            | Med Prop   | N             | Med Prop   | N       | Med Prop   |
| Simple-All   | 251(13)      | 1.37(0.09) | 214(46)       | 1.33(0.10) | 229(50) | 1.33(0.11) |
| Simple-Part  | 251(13)      | 0.76(0.03) | 219(47)       | 0.78(0.06) | 221(39) | 0.78(0.07) |
| Complex-All  | 251(13)      | 1.54(0.30) | 221(89)       | 1.29(0.19) | 192(64) | 1.36(0.22) |
| Complex-Part | 251(13)      | 0.85(0.08) | 216(84)       | 0.91(0.12) | 187(95) | 0.84(0.13) |

In scenarios with a simple underlying structure, both learners are capable of effectively identifying subgroups reflecting treatment effects transmitted through the mediator. However, under more complex settings, XGBoost consistently outperforms Random Forest. Specifically, the M-learner framework with XGBoost demonstrates strong capacity to detect indirect treatment effects and to recover meaningful subgroups accordingly. It also robustly rejects cases where the indirect effect is zero, indicating an ineffective mediator. Simple-Null1 and Simple-Null2 both assume a null indirect effect, but differ in whether the mediator influences the outcome. In Simple-Null1, the treatment does not affect the mediator, while in Simple-Null2, the mediator has no association with the outcome, implying it is not a true mediator. Our proposed method successfully identifies both types of scenarios using either base learner. For non-heterogeneous settings, we simulate only simple scenarios, while in heterogeneous cases, we consider both simple and complex structures to reflect realistic uncertainties regarding the complexity of heterogeneity. Simple scenarios serve as a baseline to filter out non-heterogeneous structures. Overall, the experimental results strongly support our hypotheses: across varying degrees of complexity, the proposed approach reliably determines the effectiveness of the mediator  $M$ , estimates the mediation effect, and identifies the corresponding subgroups. Figure 2 (c) and (d) show the ECDFs of p-values under different scenarios. The figure elucidates why the  $p_{leaf}$  is instrumental in discerning non-heterogeneous scenarios and in assessing the efficacy of a mediator. We compared our algorithm with K-Means, and the results show that K-Means lacks the ability to identify heterogeneous groups in the mediation model(the Appendix A.4 provides detailed procedures and results of the comparative algorithms can be found in Appendix Table A.14,A.15 ,A.16 and Figure A.9).

### 3.3 Effect of sample size on M-learner performance and visualization of the M-learner

To assess the performance of the M-learner under varying sample sizes, we additionally compare the results to those obtained with a sample size of 500, while holding all other conditions constant. In Appendix A.3, the Table A.7-A.11 and the Figure A.4-A.8 show the detailed results. The results indicate that, in the absence of mediators, the M-learner is capable of accurately identifying heterogeneous treatment regions in the simple scenario, even with a reduced sample size. However, its performance deteriorates substantially in the complex scenario, suggesting that the identification of treatment effect heterogeneity under more complex conditions requires larger sample sizes. Notably, the M-learner continues to perform well in detecting non-heterogeneous scenarios, even when the available sample size is limited.

In the presence of a mediator, the X-learner remains effective in identifying heterogeneous treatment regions mediated by the mediator, even with a reduced sample size. In the complex scenario, the performance of the RF-based learner declines significantly, whereas the performance degradation of the XGB-based learner is relatively limited. This suggests that, under limited sample sizes, XGB serves as a more robust base learner. The M-learner also demonstrates stable performance; however, accurately estimating complex relationships among  $M$ ,  $X$ , and  $Y$  still requires a sufficiently large sample size to reliably identify heterogeneous treatment effects.

Comparisons of ECDFs across varying sample sizes and scenarios reveal that, in the simple scenario, the ECDFs for sample sizes of 1000 and 500 exhibit minimal divergence. In contrast, more pronounced differences are observed in the complex scenario. Notably, an ECDF curve positioned closer to the upper-left corner reflects stronger treatment effect heterogeneity, whereas a curve nearer to the lower-right corner indicates weaker heterogeneity. A comprehensive comparison across different learners, scenarios, and sample sizes is presented in Appendix Figure A.5-A.8.

The visualization results of the M-learner are provided in Appendix A.5 for reference.

## 4 Real Data Application

In this section, we apply the proposed method to analyze the JOBS II real dataset, collected from a randomized of a job training intervention on unemployed workers. The dataset can be downloaded from R package "mediation"[Tingley et al., 2014]. The JOBS II study is a well-known randomized controlled trial conducted in the United States, designed to evaluate the effectiveness of a job search intervention program for unemployed individuals. The study enrolled 899 participants, who were randomly assigned either to a job training intervention group or to a control group. The dataset includes a rich set of covariates, such as demographic information and psychological measures (e.g., self-efficacy, depression). In follow-up interviews, the outcome—a continuous measure of depressive symptoms was assessed. The mediator,  $M$ , is a continuous measure representing job search self-efficacy[Price et al., 1992, Vinokur et al., 2000, Vuori and Silvonen, 2005, Cheng et al., 2022].

We use age, sex, education, prior occupational status, and the level of economic hardship experienced by participants as covariates. For educational background, we classify participants into three categories: individuals who did not complete high school or whose highest degree is a high school diploma are grouped together; individuals who attended some college but did not obtain a bachelor's degree form the second group; and those who earned a bachelor's degree or higher are classified into the third group. Other covariates are kept in their original form.

We first conducted a preliminary analysis of the JOBS II data using the mediation package[Imai et al., 2010]. Table 4 shows the detailed analysis. Job search self-efficacy has been widely employed as a mediator in the literature[Vinokur et al., 1995, 2000]. However, it exhibits relatively large p-value.

Applying our proposed method to the JOBS II data, we first analyzed the heterogeneity based solely on the TTE. We then applied the mediation analysis separately to each subtype to estimate the TTE, ITE, mediation proportions, and corresponding p-values. Figure 3, specifically the first decision tree, identifies four subtypes. Subtype 1 and 2 are associated with occupation and the level of economic hardship. Subtype 1 includes individuals from occupation categories A, B, D, and F, as well as those with either higher or lower levels of education. Subtype 2 consists of individuals from the same occupation categories (A, B, D, F) but with moderate levels of education. According to the results in Table 4, subtype 1 shows significant TTE, ITE, and mediation proportions. In this subtype, receiving job search assistance (treatment) significantly improves the mood, with  $M$  identified as a significant mediator. In contrast, subtype 2 exhibits a weaker treatment effect, suggesting that job search assistance may have a limited impact on this subtype, with  $M$  showing no significant mediating role. Subtype 3 and 4 are also associated with occupation and economic hardship. Subtype 3 is characterized by individuals in occupation categories C, E, and G, as well as those experiencing greater economic difficulties, while subtype 4 includes individuals from the same occupation categories (C, E, G) but with relatively better economic conditions. In Subtype 3, both TTE and ITE are observed, though the p-values are relatively large, which may be attributable to the smaller sample size. For subtype 4, the TTE is significant, but the ITE is not, along with a lower mediation proportion. Interestingly, when not receiving job search assistance, the mood of unemployed individuals in this group tends to improve.

Employing a mediation analysis approach of M-learner, as depicted in the second decision tree of Figure 3, yielded two distinct subtypes. In contrast to findings derived without considering a mediator, these subtypes exhibited a broader scope and were exclusively associated with occupational factors. Specifically, subtype 1 in this analysis represents the aggregation of subtypes 1 and 2 from the initial, non-mediated analysis, while subtype 2 comprises the combined subtypes 3 and 4. As presented in Table 4, subtype 1 demonstrated significant TTE and ITE, accompanied by a substantial proportion of mediation and a reduced p-value. Conversely, subtype 2 exhibited smaller and statistically non-significant TTE and ITE.

Table 4: Summary of the total treatment effect, indirect treatment effect, mediation proportion, and corresponding p-values, calculated using methods from the "mediation" package Imai et al. [2010] and the proposed approach. Med prop denotes mediation proportion:=indirect treatment effect/ total treatment effect. TTE denotes total treatment effect The value in parentheses is the p-value. ITE denotes indirect treatment effect. The value in parentheses is the p-value. N denotes sample size.

| Method                      | Subtype  | N   | TTE           | ITE           | Med Prop     |
|-----------------------------|----------|-----|---------------|---------------|--------------|
| "Mediation"                 | NA       | 899 | -0.063(0.29)  | -0.015(0.21)  | 24.0%(0.29)  |
| M-learner,<br>no mediator   | Subtype1 | 344 | -0.243(0.004) | -0.057(0.010) | 23.6%(0.014) |
|                             | Subtype2 | 190 | 0.003(0.95)   | 0.005(0.77)   | 160%(0.81)   |
|                             | Subtype3 | 129 | -0.10(0.42)   | -0.013(0.76)  | 12.9%(0.73)  |
|                             | Subtype4 | 236 | 0.145(0.03)   | 0.009(0.47)   | 6.1%(0.48)   |
| M-learner,<br>with mediator | Subtype1 | 534 | -0.15(0.016)  | -0.030(0.010) | 20.0%(0.052) |
|                             | Subtype2 | 365 | 0.065(0.30)   | 0.005(0.75)   | 8.0%(0.74)   |

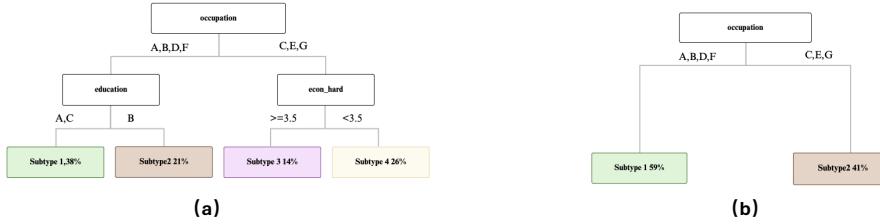


Figure 3: Subtype results identified by the M-learner on the Jobs II dataset. (a) It shows subtype results obtained without mediators, (b) it shows subtype results obtained with the mediator. In each decision tree, each leaf represents a subtype. econ\_hard refers to the level of economic hardship. In occupation, A denotes "clericalkindred",B denotes "laborersservice works",C denotes "operativeskindred works",D denotes "sales workers",E denotes "craftsmenforemenkindred",F denotes "managerial", G denotes "professionals".In education, A denotes individuals who did not complete high school or whose highest degree is a high school diploma, B denotes individuals who attended some college but did not obtain a bachelor's degree, C denotes those who earned a bachelor's degree or higher.

These results suggest that the mediator exhibits heterogeneity across different subtypes and serves as an effective mediator only in certain groups. Regardless of whether the mediator is included, our method yielded similar subgroup structures, with the key difference being the further subdivision of certain occupational groups when considering the mediator. Using this approach, in future government-initiated randomized clinical trials, it would be possible to monitor changes in the mediator within specific subtypes to quickly assess whether a new intervention is effective for those groups. Trials could be stopped early for ineffective interventions in particular groups, allowing better, more targeted treatments to be administered. This strategy can significantly reduce government costs while enabling timely adjustments to intervention strategies, thereby minimizing the negative impact of ineffective treatments on participants.

## 5 Discussion

This article introduced a novel method that is, to our knowledge, the first to examine treatment effect heterogeneity transmitted through the mediator. In addition to estimating mediated heterogeneity, our approach enables data-driven subgroup identification based on distinct mediator patterns. This represents a significant methodological advancement, as it allows researchers to evaluate the informativeness of mediators and to detect heterogeneous treatment effects that are revealed through them. The proposed framework is highly flexible and can accommodate a variety of base learners, including random forests, XGBoost, and neural networks, depending on the data characteristics and specific application context. Moreover, the method is broadly applicable to both mediated and unmediated settings. In the absence of a mediator, it enables estimation of the CATTE and facilitates the detection of heterogeneity and subgroup structures driven by the TTE. When a mediator is present, the method

estimates the CAITE and identifies heterogeneity and subgroup structures that arise through the mediator. By capturing treatment effect heterogeneity from both total and mediated perspectives, the framework offers a comprehensive understanding of the complex interplay among treatment, covariates, mediators, and outcomes.

Importantly, our method permits the inference of individual treatment responsiveness without requiring observation of the final outcome variable  $Y$ . This finding holds substantial implications across disciplines such as economics, psychology, medicine and sociology. In practical applications, such as in technology companies, mediator behavior can guide personalized interventions. In healthcare, particularly in pharmaceutical settings, the method can assist clinicians in adapting treatment strategies, thereby contributing to the development of precision medicine.

A potential direction for future research is to extend our method to survival analysis settings, which could enable the pharmaceutical industry to predefine subgroups based on mediating variables, ultimately enhancing the success rate of drug development.

## References

- J. D. Angrist. Treatment effect heterogeneity in theory and practice. *The economic journal*, 114(494):C52–C83, 2004.
- S. Athey and G. Imbens. Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, 113(27):7353–7360, 2016.
- S. Athey and S. Wager. Estimating treatment effects with causal forests: An application. *Observational studies*, 5(2):37–51, 2019.
- L. Breiman. Random forests. *Machine learning*, 45:5–32, 2001.
- T. Chen and C. Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 785–794, 2016.
- L. Cheng, R. Guo, and H. Liu. Causal mediation analysis with hidden confounders. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*, pages 113–122, 2022.
- I. Díaz, N. S. Hejazi, K. E. Rudolph, and M. J. van Der Laan. Nonparametric efficient causal mediation with intermediate confounders. *Biometrika*, 108(3):627–641, 2021.
- I. Díaz, N. Williams, and K. E. Rudolph. Efficient and flexible mediation analysis with time-varying mediators, treatments, and confounders. *Journal of Causal Inference*, 11(1):20220077, 2023.
- J. A. Dorresteijn, F. L. Visseren, P. M. Ridker, A. M. Wassink, N. P. Paynter, E. W. Steyerberg, Y. van der Graaf, and N. R. Cook. Estimating treatment effects for individual patients based on the results of randomised clinical trials. *Bmj*, 343, 2011.
- R. Dwivedi, Y. S. Tan, B. Park, M. Wei, K. Horgan, D. Madigan, and B. Yu. Stable discovery of interpretable subgroups via calibration in causal studies. *International Statistical Review*, 88: S135–S178, 2020.
- J. C. Foster, J. M. Taylor, and S. J. Ruberg. Subgroup identification from randomized clinical trial data. *Statistics in medicine*, 30(24):2867–2880, 2011.
- L. Ge, J. Wang, C. Shi, Z. Wu, and R. Song. A reinforcement learning framework for dynamic mediation analysis. In *International Conference on Machine Learning*, pages 11050–11097. PMLR, 2023.
- L. Ge, H. Cai, R. Wan, Y. Xu, and R. Song. A review of causal decision making. *arXiv preprint arXiv:2502.16156*, 2025.
- B. Gilbert, K. L. Hoffman, N. Williams, K. E. Rudolph, E. J. Schenck, and I. Díaz. Identification and estimation of mediational effects of longitudinal modified treatment policies. *arXiv preprint arXiv:2403.09928*, 2024.

- N. Hassanzadeh and R. Greiner. Learning disentangled representations for counterfactual regression. In *International Conference on Learning Representations*, 2019.
- K. Imai, L. Keele, and D. Tingley. A general approach to causal mediation analysis. *Psychological methods*, 15(4):309, 2010.
- G. W. Imbens. Nonparametric estimation of average treatment effects under exogeneity: A review. *Review of Economics and statistics*, 86(1):4–29, 2004.
- F. Johansson, U. Shalit, and D. Sontag. Learning representations for counterfactual inference. In *International conference on machine learning*, pages 3020–3029. PMLR, 2016.
- E. H. Kennedy. Towards optimal doubly robust estimation of heterogeneous causal effects. *Electronic Journal of Statistics*, 17(2):3008–3049, 2023.
- S. R. Künzel, J. S. Sekhon, P. J. Bickel, and B. Yu. Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the national academy of sciences*, 116(10):4156–4165, 2019.
- D. Lin, T. Fleming, and V. De Gruttola. Estimating the proportion of treatment effect explained by a surrogate marker. *Statistics in medicine*, 16(13):1515–1527, 1997.
- L. Luo, C. Shi, J. Wang, Z. Wu, and L. Li. Multivariate dynamic mediation analysis under a reinforcement learning framework. *The Annals of Statistics*, 53(1):400–425, 2025.
- X. Nie and S. Wager. Quasi-oracle estimation of heterogeneous treatment effects. *Biometrika*, 108(2):299–319, 2021.
- J. Pearl. Direct and indirect effects. In *Probabilistic and causal inference: the works of Judea Pearl*, pages 373–392. 2022.
- M. L. Petersen, S. E. Sinisi, and M. J. van der Laan. Estimation of direct causal effects. *Epidemiology*, 17(3):276–284, 2006.
- K. J. Preacher. Advances in mediation analysis: A survey and synthesis of new developments. *Annual review of psychology*, 66(1):825–852, 2015.
- R. H. Price, M. Van Ryn, and A. D. Vinokur. Impact of a preventive job search intervention on the likelihood of depression among the unemployed. *Journal of health and social behavior*, pages 158–167, 1992.
- J. M. Robins and S. Greenland. Identifiability and exchangeability for direct and indirect effects. *Epidemiology*, 3(2):143–155, 1992.
- U. Shalit, F. D. Johansson, and D. Sontag. Estimating individual treatment effect: generalization bounds and algorithms. In *International conference on machine learning*, pages 3076–3085. PMLR, 2017.
- C. Shi, D. Blei, and V. Veitch. Adapting neural networks for the estimation of treatment effects. *Advances in neural information processing systems*, 32, 2019.
- E. J. Tchetgen and I. Shpitser. Semiparametric theory for causal mediation analysis: efficiency bounds, multiple robustness, and sensitivity analysis. *Annals of statistics*, 40(3):1816, 2012.
- E. J. Tchetgen and T. J. VanderWeele. Identification of natural direct effects when a confounder of the mediator is directly affected by exposure. *Epidemiology*, 25(2):282–291, 2014.
- D. Tingley, T. Yamamoto, K. Hirose, L. Keele, and K. Imai. Mediation: R package for causal mediation analysis. *Journal of statistical software*, 59:1–38, 2014.
- M. J. van der Laan and M. L. Petersen. Direct effect models. *The international journal of biostatistics*, 4(1), 2008.
- L. Van der Maaten and G. Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

- S. Vandenberghe, S. Vansteelandt, and T. Loeys. Boosting the precision of mediation analyses of randomised experiments through covariate adjustment. *Statistics in medicine*, 36(6):939–957, 2017.
- T. VanderWeele. *Explanation in causal inference: methods for mediation and interaction*. Oxford University Press, 2015.
- T. VanderWeele and S. Vansteelandt. Conceptual issues concerning mediation, interventions and composition. *Statistics and its Interface*, 2:457–468, 2009.
- S. Vansteelandt and T. J. VanderWeele. Natural direct and indirect effects on the exposed: effect decomposition under weaker assumptions. *Biometrics*, 68(4):1019–1027, 2012.
- A. D. Vinokur, R. H. Price, and Y. Schul. Impact of the jobs intervention on unemployed workers varying in risk for depression. *American journal of community psychology*, 23(1):39–74, 1995.
- A. D. Vinokur, Y. Schul, J. Vuori, and R. H. Price. Two years after a job loss: long-term impact of the jobs program on reemployment and mental health. *Journal of occupational health psychology*, 5(1):32, 2000.
- J. Vuori and J. Silvonen. The benefits of a preventive job search program on re-employment and mental health at 2-year follow-up. *Journal of Occupational and Organizational Psychology*, 78(1): 43–52, 2005.
- S. Wager and S. Athey. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242, 2018.
- J. Wang and R. Song. Dynamic causal structure discovery and causal effect estimation. *arXiv preprint arXiv:2501.06534*, 2025.
- Y. Zhang, D. Kong, and S. Yang. Towards r-learner of conditional average treatment effects with a continuous treatment: T-identification, estimation, and inference. *arXiv preprint arXiv:2208.00872*, 2022.

## A Technical Appendices and Supplementary Material

### A.1 Setting in no mediator model

Across all experimental designs in this section, we fix the sample size at 1,000, with 10 covariates generated for each unit. Subjects were randomly assigned to treatment and control groups in a 1 : 1 ratio. We specify two underlying functions:  $\eta(x)$ , representing the conditional mean outcome, and  $\kappa(x)$ , capturing the conditional treatment effect. Both functions are defined for units under treatment ( $w = 1$ ) and control ( $w = 0$ ). For simple scenario,

$$Y_i(w) = \eta(X_i) + \frac{1}{2}(2w - 1) \cdot \kappa(X_i) + b + \epsilon_i, \quad (4)$$

for complex scenario,

$$Y_i(w) = \frac{1}{1 + \exp(\eta(X_i) + \frac{1}{2}(2w - 1) \cdot \kappa(X_i) + b + \epsilon_i)}, \quad (5)$$

where  $\epsilon_i \sim \mathcal{N}(0, 0.01)$ , and the  $X_i$  are independent of  $\epsilon_i$  and one another, and  $X_i \sim \mathcal{N}(0, 1)$ ,  $b$  is the intercept. The four scenarios designs follow:

1. Simple scenario in (4), existing heterogeneity:  $\eta(x) = \frac{1}{2}(x_1 + x_2) + x_3 + x_4, \kappa(x) = \sum_{i=1}^2 \mathbb{I}(x_i > 0) \cdot x_i, b = 1$ .
2. Complex scenario in (5), existing heterogeneity:  $\eta(x) = \frac{1}{2}(x_1 + x_2) + x_3 + x_4, \kappa(x) = \sum_{i=1}^2 \mathbb{I}(x_i > 0) \cdot x_i, b = 1$ .
3. Simple scenario in (4), no heterogeneity, all units benefit from treatment(Global):  $\eta(x) = x_3 + x_4, \kappa(x) = 1, b = 1$ .
4. Simple scenario in (4), no heterogeneity, no units benefit from treatment(NULL):  $\eta(x) = x_3 + x_4, \kappa(x) = 0, b = 1$ .

For the RF model, the number of trees is set to 2000, with all other parameters kept at their default values. For XGBoost, the number of boosting rounds is set to 100, while all remaining parameters were left at their default settings. In all simulation experiments, the range of cluster numbers was predefined as 2 to 5.

In this paper, all experiments were conducted on a MacBook Pro equipped with an M3 Max CPU and 36GB of RAM. The software environment includes R version 4.3.3, with the randomForestSRC package version 3.2.3, xgboost version 1.7.9.1 and rpart version 4.1.23.

### A.2 Setting in mediator model

Across all experimental designs in mediator model, we fixed the sample size at 1,000, with 10 covariates generated for each unit. Subjects were randomly assigned to treatment and control groups in a 1 : 1 ratio. We specified four underlying functions:  $\eta_1(x)$ , representing the conditional mean outcome in mediator covariates model,  $\eta_2(x)$ , representing the conditional mean outcome in covariates reponse model, and  $\kappa_1(x)$ , capturing the conditional treatment effect in mediator covariates model, and  $\kappa_2(x)$ , capturing the conditional treatment effect in covariates reponse model. All functions were defined for units under treatment ( $w = 1$ ) and control ( $w = 0$ ), for simple scenario,

$$\begin{aligned} M_i(w) &= \eta_1(X_i) + \frac{1}{2}(2w - 1) \cdot \kappa_1(X_i) + b_1 + \epsilon_{1i}, \\ Y_i(w) &= \eta_2(X_i) + \frac{1}{2}(2w - 1) \cdot \kappa_2(X_i) + b_2 + c \cdot M_i(w) + \epsilon_{2i}. \end{aligned} \quad (6)$$

For complex scenario,

$$\begin{aligned} M_i(w) &= \eta_1(X_i) + \frac{1}{2}(2w - 1) \cdot \kappa_1(X_i) + b_1 + \epsilon_{1i}, \\ Y_i(w) &= \frac{1}{1 + \exp(\eta_2(X_i) + \frac{1}{2}(2w - 1) \cdot \kappa_2(X_i) + b_2 + c \cdot M_i(w) + \epsilon_{2i})}, \end{aligned} \quad (7)$$

where  $\epsilon_{1i} \sim \mathcal{N}(0, 0.01)$ ,  $\epsilon_{2i} \sim \mathcal{N}(0, 0.01)$ , and the  $X_i$  are independent of  $\epsilon_{1i}$ ,  $\epsilon_{2i}$  and one another, and  $X_i \sim \mathcal{N}(0, 1)$ ,  $b_1$  and  $b_2$  are the intercept terms,  $c$  is the coefficients of mediator. The seven scenarios designs follow:

1. Simple scenario in (6), existing heterogeneity, all treatment effects via mediator:  $\eta_1(x) = \frac{1}{2}(x_1 + x_2) + x_3 + x_4, \kappa_1(x) = \sum_{i=1}^2 \mathbb{I}(x_i > 0) \cdot x_i, b_1 = 0, \eta_2(x) = \frac{1}{2}(x_3 + x_4), \kappa_2(x) = 0, b_2 = 1, c = 1.$
2. Simple scenario in (6), existing heterogeneity, part of the treatment effect via mediator:  $\eta_1(x) = \frac{1}{2}(x_1 + x_2) + x_3 + x_4, \kappa_1(x) = \sum_{i=1}^2 \mathbb{I}(x_i > 0) \cdot x_i, b_1 = 0, c = 1, \eta_2(x) = \frac{1}{2}(x_3 + x_4), \kappa_2(x) = \sum_{i=1}^2 \mathbb{I}(x_i > 0) \cdot x_i, b_2 = 1, c = 1.$
3. Complex scenario in (7), existing heterogeneity, all treatment effects via mediator:  $\eta_1(x) = \frac{1}{2}(x_1 + x_2) + x_3 + x_4, \kappa_1(x) = \sum_{i=1}^2 \mathbb{I}(x_i > 0) \cdot x_i, b_1 = 0, \eta_2(x) = \frac{1}{2}(x_3 + x_4), \kappa_2(x) = 0, b_2 = 1, c = 1.$
4. Complex scenario in (7), existing heterogeneity, part of the treatment effect via mediator:  $\eta_1(x) = \frac{1}{2}(x_1 + x_2) + x_3 + x_4, \kappa_1(x) = \sum_{i=1}^2 \mathbb{I}(x_i > 0) \cdot x_i, b_1 = 0, c = 1, \eta_2(x) = \frac{1}{2}(x_3 + x_4), \kappa_2(x) = \sum_{i=1}^2 \mathbb{I}(x_i > 0) \cdot x_i, b_2 = 1, c = 1.$
5. Simple scenario in (6), existing heterogeneity, 0% treatment effects via mediator (NULL 1):  $\eta_1(x) = \frac{1}{2}(x_1 + x_2) + x_3 + x_4, \kappa_1(x) = 0, b_1 = 0, \eta_2(x) = \frac{1}{2}(x_3 + x_4), \kappa_2(x) = \sum_{i=1}^2 \mathbb{I}(x_i > 0) \cdot x_i, b_2 = 1, c = 1.$
6. Simple scenario in (6), existing heterogeneity,  $M$  is not mediator, All treatment effects are directly transmitted to  $Y$  (NULL 2):  $\eta_1(x) = \frac{1}{2}(x_1 + x_2) + x_3 + x_4, \kappa_1(x) = 0, b_1 = 0, \eta_2(x) = \frac{1}{2}(x_3 + x_4), \kappa_2(x) = \sum_{i=1}^2 \mathbb{I}(x_i > 0) \cdot x_i, b_2 = 1, c = 0.$
7. Simple scenario in (6), no heterogeneity, all units benefit from the treatment and all treatment effects via mediator (Global):  $\eta_1(x) = \frac{1}{2}(x_1 + x_2) + x_3 + x_4, \kappa_1(x) = 1, b_1 = 0, \eta_2(x) = \frac{1}{2}(x_3 + x_4), \kappa_2(x) = 0, b_2 = 1, c = 1.$

For the RF model, the number of trees is set to 2000, with all other parameters kept at their default values. For XGBoost, the number of boosting rounds is set to 100, while all remaining parameters are left at their default settings. In all simulation experiments, the range of cluster numbers is predefined as 2 to 5.

In the analysis presented in Table 3, the ground truth heterogeneous region is defined by samples satisfying the conditions  $X_1 > 0$  and  $X_2 > 0$ . The mediation proportion for this region is calculated using the R package "mediation". To evaluate the mediation proportion and sample size corresponding to the heterogeneous region identified by the M-learner, we adopt the following procedure: among the final subtype regions generated by the decision tree, we assess the mediation effect within each region and select the region exhibiting the most statistically significant mediation effect. The sample size and mediation effect of this selected region are then recorded.

### A.3 Additional simulation results for M-learner

Table A.5: The table summarizes the distribution of the number of variables included in the selected decision trees across 100 simulation runs. Each value represents the frequency with which a specific number of variables was selected over the course of the experiments. The performance of Random Forest and XGBoost as base learners was systematically evaluated under four distinct experimental scenarios. Selecting zero covariates in a given experimental run indicates that the decision tree's classification was not supported by the calibration process and was therefore rejected.

| Base Learner         | Random Forest |   |    |    |    | XGBoost |   |    |    |    |
|----------------------|---------------|---|----|----|----|---------|---|----|----|----|
|                      | 0             | 1 | 2  | 3  | 4  | 0       | 1 | 2  | 3  | 4  |
| Number of covariates | 0             | 0 | 29 | 52 | 19 | 0       | 0 | 57 | 28 | 15 |
| Simple               | 0             | 0 | 29 | 52 | 19 | 0       | 0 | 57 | 28 | 15 |
| Complex              | 25            | 1 | 9  | 34 | 31 | 28      | 1 | 7  | 30 | 34 |
| Global               | 92            | 0 | 5  | 2  | 1  | 98      | 0 | 2  | 0  | 0  |
| Null                 | 90            | 1 | 3  | 5  | 1  | 90      | 3 | 3  | 3  | 1  |

Table A.6: The table summarizes the distribution of the number of variables included in the selected decision trees across 100 simulation runs. Each value represents the frequency with which a specific number of variables was selected over the course of the experiments. The performance of Random Forest and XGBoost as base learners was systematically evaluated under seven distinct experimental scenarios. Selecting zero covariates in a given experimental run indicates that the decision tree's classification was not supported by the calibration process and was therefore rejected.

| Base Learner  | Random Forest |   |    |    |    |          | XGBoost |   |    |    |    |          |
|---------------|---------------|---|----|----|----|----------|---------|---|----|----|----|----------|
|               | 0             | 1 | 2  | 3  | 4  | $\geq 5$ | 0       | 1 | 2  | 3  | 4  | $\geq 5$ |
| Simple-All    | 0             | 0 | 35 | 47 | 17 | 1        | 0       | 0 | 58 | 33 | 9  | 0        |
| Simple-Part   | 0             | 0 | 34 | 51 | 14 | 1        | 0       | 0 | 54 | 40 | 6  | 0        |
| Complex-All   | 4             | 1 | 10 | 36 | 48 | 1        | 0       | 1 | 23 | 43 | 33 | 0        |
| Complex-Part  | 4             | 1 | 8  | 51 | 36 | 0        | 1       | 1 | 14 | 49 | 35 | 0        |
| Simple-Null1  | 84            | 1 | 6  | 7  | 2  | 0        | 83      | 1 | 8  | 6  | 1  | 1        |
| Simple-Null2  | 90            | 1 | 4  | 4  | 1  | 0        | 90      | 5 | 3  | 2  | 0  | 0        |
| Simple-Global | 94            | 0 | 2  | 3  | 1  | 0        | 87      | 1 | 6  | 4  | 2  | 0        |

Table A.7: The table summarizes the correct covariates in profiles in each scenario when sample size is 500 in no mediators model.  $X_1/X_2$ :Final profile contain  $X_1/X_2$ ,  $X_1 \& X_2$ : final profile contain both  $X_1$  and  $X_2$ ,  $X_1$  and  $X_2$  are the two variables associated with treatment effect heterogeneity. Each value denotes the count of occurrences across 100 simulations.

| Base Learner | Random Forest |       |       | XGBoost      |       |       |              |
|--------------|---------------|-------|-------|--------------|-------|-------|--------------|
|              | Covariates    | $X_1$ | $X_2$ | $X_1 \& X_2$ | $X_1$ | $X_2$ | $X_1 \& X_2$ |
| Simple       | 95            | 94    | 93    | 96           | 96    | 96    | 95           |
| Complex      | 26            | 24    | 19    | 33           | 25    | 25    | 19           |
| Global       | 4             | 3     | 1     | 0            | 0     | 0     | 0            |
| Null         | 5             | 4     | 2     | 4            | 3     | 3     | 1            |

Table A.8: The table summarizes the distribution of the number of variables included in the selected decision trees across 100 simulation runs when there is no mediator and sample size is 500. Each value represents the frequency with which a specific number of variables was selected over the course of the experiments. The performance of Random Forest and XGBoost as base learners was systematically evaluated under four distinct experimental scenarios. Selecting zero covariates in a given experimental run indicates that the decision tree's classification was not supported by the calibration process and was therefore rejected.

| Base Learner | Random Forest        |   |    |    |    | XGBoost  |   |    |    |    |
|--------------|----------------------|---|----|----|----|----------|---|----|----|----|
|              | Number of covariates | 0 | 1  | 2  | 3  | $\geq 4$ | 0 | 1  | 2  | 3  |
| Simple       | 4                    | 1 | 42 | 37 | 16 | 3        | 1 | 70 | 21 | 5  |
| Complex      | 67                   | 3 | 9  | 11 | 10 | 61       | 3 | 7  | 18 | 11 |
| Global       | 91                   | 0 | 3  | 5  | 2  | 99       | 0 | 1  | 0  | 0  |
| Null         | 90                   | 2 | 2  | 4  | 5  | 90       | 2 | 3  | 3  | 2  |

Table A.9: The table summarizes the correct covariates in profiles in each scenario when sample size is 500.  $X_1/X_2$ :Final profile contain  $X_1/X_2$ ,  $X_1\&X_2$ : final profile contain both  $X_1$  and  $X_2$ ,  $X_1$  and  $X_2$  are the two variables associated with treatment effect heterogeneity. Each value denotes the count of occurrences across 100 simulations.

| Base Learner  | Random Forest |       |       | XGBoost    |       |       |            |
|---------------|---------------|-------|-------|------------|-------|-------|------------|
|               | Covariates    | $X_1$ | $X_2$ | $X_1\&X_2$ | $X_1$ | $X_2$ | $X_1\&X_2$ |
| Simple-All    | 93            | 90    | 89    | 92         | 94    | 91    |            |
| Simple-Part   | 97            | 96    | 96    | 92         | 93    | 92    |            |
| Complex-All   | 46            | 39    | 23    | 86         | 83    | 76    |            |
| Complex-Part  | 38            | 33    | 13    | 76         | 69    | 57    |            |
| Simple-Null1  | 9             | 5     | 3     | 0          | 1     | 0     |            |
| Simple-Null2  | 6             | 2     | 1     | 2          | 2     | 1     |            |
| Simple-Global | 2             | 0     | 0     | 0          | 2     | 0     |            |

Table A.10: The table summarizes the distribution of the number of variables included in the selected decision trees across 100 simulation runs when the sample size is 500. Each value represents the frequency with which a specific number of variables was selected over the course of the experiments. The performance of Random Forest and XGBoost as base learners was systematically evaluated under seven distinct experimental scenarios. Selecting zero covariates in a given experimental run indicates that the decision tree's classification was not supported by the calibration process and was therefore rejected.

| Base Learner  | Random Forest        |   |    |    |    |   | XGBoost  |    |    |    |    |   |          |
|---------------|----------------------|---|----|----|----|---|----------|----|----|----|----|---|----------|
|               | Number of covariates | 0 | 1  | 2  | 3  | 4 | $\geq 5$ | 0  | 1  | 2  | 3  | 4 | $\geq 5$ |
| Simple-All    | 6                    | 3 | 42 | 35 | 14 | 0 | 5        | 4  | 65 | 23 | 3  | 0 |          |
| Simple-Part   | 3                    | 1 | 46 | 39 | 11 | 0 | 7        | 1  | 67 | 18 | 7  | 0 |          |
| Complex-All   | 37                   | 6 | 16 | 23 | 13 | 0 | 7        | 10 | 25 | 48 | 9  | 1 |          |
| Complex-Part  | 42                   | 5 | 12 | 29 | 12 | 0 | 12       | 12 | 30 | 31 | 15 | 0 |          |
| Simple-Null1  | 86                   | 2 | 6  | 3  | 3  | 0 | 96       | 1  | 3  | 0  | 0  | 0 |          |
| Simple-Null2  | 90                   | 2 | 6  | 2  | 0  | 0 | 96       | 1  | 1  | 2  | 0  | 0 |          |
| Simple-Global | 94                   | 0 | 1  | 4  | 1  | 0 | 97       | 0  | 0  | 2  | 1  | 0 |          |

Table A.11: Comparison of true and M-learner–estimated mediation proportions and sample sizes within heterogeneous treatment effect regions across scenarios when sample size is 500. Standard deviations are shown in parentheses. The mediation proportion is calculated by R package "mediation". "For each set of results, the subtype exhibiting the lowest p-value for the mediation proportion is designated as the region of interest. Within this region, we estimate the mediation proportion, total treatment effect, and indirect treatment effect.

| Scenario     | True    |            | Random Forest |            | XGBoost |            |
|--------------|---------|------------|---------------|------------|---------|------------|
|              | N       | Med Prop   | N             | Med Prop   | N       | Med Prop   |
| Simple-All   | 124(10) | 1.38(0.14) | 111(28)       | 1.27(0.13) | 114(30) | 1.29(0.15) |
| Simple-Part  | 124(10) | 0.76(0.04) | 113(25)       | 0.78(0.06) | 113(23) | 0.77(0.04) |
| Complex-All  | 124(10) | 1.64(0.69) | 140(66)       | 1.29(0.19) | 116(55) | 1.25(0.23) |
| Complex-Part | 124(10) | 0.87(0.12) | 134(59)       | 0.91(0.12) | 136(61) | 0.93(0.18) |

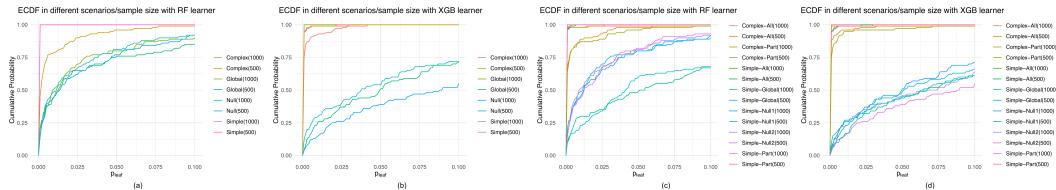


Figure A.4: Comparison of  $p_{leaf}$  ECDFs using RF and XGB as base learners across different scenarios with sample sizes of 1000 and 500.(a)results for RF without a mediator, (b) results for XGB without a mediator, (c)results for RF with a mediator (d) results for XGB with a mediator.

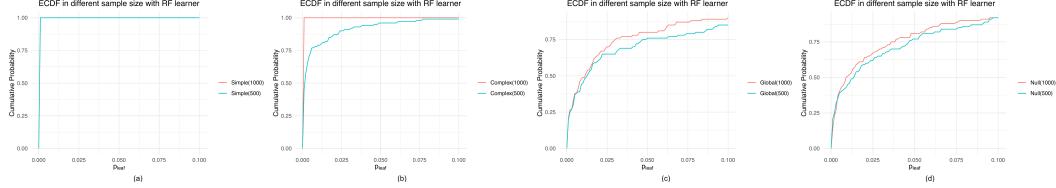


Figure A.5: Comparison of  $p_{leaf}$  ECDFs under four scenarios without a mediator, using RF as base learner with sample sizes of 1000 and 500. (a) results for Simple scenario, (b) results for Complex scenario, (c) results for Global scenario (d) results for Null scenario.

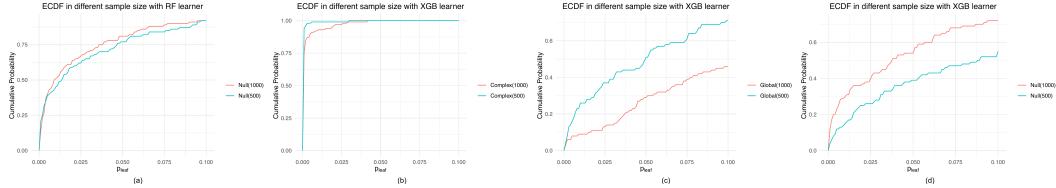


Figure A.6: Comparison of  $p_{leaf}$  ECDFs under four scenarios without a mediator, using XGB as base learner with sample sizes of 1000 and 500. (a) results for Simple scenario, (b) results for Complex scenario, (c) results for Global scenario (d) results for Null scenario.

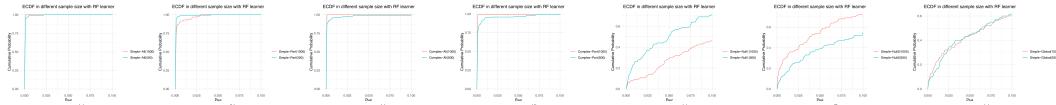


Figure A.7: Comparison of  $p_{leaf}$  ECDFs under seven scenarios with a mediator, using RF as base learner with sample sizes of 1000 and 500. (a) results for Simple-All scenario, (b) results for Simple-Part scenario, (c) results for Complex-All scenario, (d) results for Complex-Part scenario, (e) results for Simple-Null1 scenario, (f) results for Simple-Null2 scenario, (g) results for Simple-Global scenario.

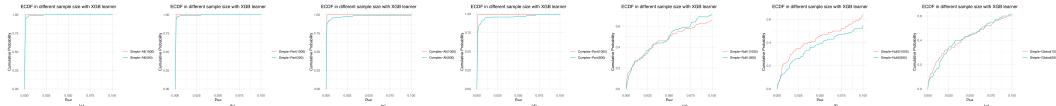


Figure A.8: Comparison of  $p_{leaf}$  ECDFs under seven scenarios with a mediator, using XGB as base learner with sample sizes of 1000 and 500. (a) results for Simple-All scenario, (b) results for Simple-Part scenario, (c) results for Complex-All scenario, (d) results for Complex-Part scenario, (e) results for Simple-Null1 scenario, (f) results for Simple-Null2 scenario, (g) results for Simple-Global scenario.

#### A.4 K-Means

For the benchmark method K-means, clustering is performed directly on the data, with the number of clusters preset to range from 2 to 5. The  $p_{leaf}$  is then estimated using the same approach as in the M-learner framework, followed by a calibration procedure.

When model without mediators, the Appendix Table A.12 and A.13 indicate the K-Means fails to effectively identify heterogeneous treatment subtypes and also struggles to distinguish non-heterogeneous scenarios.

When model with a mediator, the Appendix Table A.14, A.15 and A.16 also indicate the K-Means fails to effectively identify effective subtypes and also struggles to distinguish non-heterogeneous scenarios. K-means completely fails to identify the Simple-Global scenario, which lacks treatment heterogeneity.

The Appendix Figure A.9 shows the ECDFs of K-means method.

Table A.12: The table summarizes the correct covariates in profiles in each scenario for K-means method in no mediators model.  $X_1/X_2$ :Final profile contain  $X_1/X_2$ ,  $X_1\&X_2$ : final profile contain both  $X_1$  and  $X_2$ , $X_1$  and  $X_2$  are the two variables associated with treatment effect heterogeneity. Each value denotes the count of occurrences across 100 simulations.

| Covariates | $X_1$ | $X_2$ | $X_1\&X_2$ |
|------------|-------|-------|------------|
| Simple     | 46    | 33    | 17         |
| Complex    | 22    | 10    | 4          |
| Global     | 65    | 48    | 24         |
| Null       | 3     | 2     | 1          |

Table A.13: The table summarizes the distribution of the number of variables included in the selected decision trees across 100 simulation runs. Each value represents the frequency with which a specific number of variables was selected over the course of the experiments. The performance of K-means was systematically evaluated under four distinct experimental scenarios. Selecting zero covariates in a given experimental run indicates that the decision tree's classification was not supported by the calibration process and was therefore rejected.

| Number of covariates | 0  | 1 | 2  | 3  | 4  | $\geq 5$ |
|----------------------|----|---|----|----|----|----------|
| Simple               | 33 | 6 | 18 | 19 | 15 | 9        |
| Complex              | 65 | 2 | 14 | 10 | 7  | 2        |
| Global               | 7  | 4 | 19 | 32 | 27 | 11       |
| Null                 | 90 | 0 | 2  | 3  | 4  | 1        |

Table A.14: The table summarizes the correct covariates in profiles in each scenario for K-means method in the mediator model.  $X_1/X_2$ :Final profile contain  $X_1/X_2$ ,  $X_1\&X_2$ : final profile contain both  $X_1$  and  $X_2$ , $X_1$  and  $X_2$  are the two variables associated with treatment effect heterogeneity. Each value denotes the count of occurrences across 100 simulations.

| Covariates    | $X_1$ | $X_2$ | $X_1\&X_2$ |
|---------------|-------|-------|------------|
| Simple-All    | 53    | 36    | 17         |
| Simple-Part   | 57    | 38    | 18         |
| Complex-All   | 35    | 24    | 11         |
| Complex-Part  | 47    | 33    | 15         |
| Simple-Null1  | 5     | 4     | 1          |
| Simple-Null2  | 1     | 5     | 1          |
| Simple-Global | 37    | 26    | 11         |

Table A.15: The table summarizes the distribution of the number of variables included in the selected decision trees across 100 simulation runs. Each value represents the frequency with which a specific number of variables was selected over the course of the experiments. The performance of K-means was systematically evaluated under seven distinct experimental scenarios. Selecting zero covariates in a given experimental run indicates that the decision tree's classification was not supported by the calibration process and was therefore rejected.

| Number of covariates | 0  | 1 | 2  | 3  | 4  | $\geq 5$ |
|----------------------|----|---|----|----|----|----------|
| Simple-All           | 12 | 6 | 23 | 25 | 22 | 12       |
| Simple-Part          | 4  | 8 | 25 | 29 | 22 | 12       |
| Complex-All          | 43 | 3 | 14 | 18 | 13 | 9        |
| Complex-Part         | 20 | 6 | 23 | 23 | 16 | 12       |
| Simple-Null1         | 87 | 0 | 3  | 3  | 5  | 2        |
| Simple-Null2         | 91 | 0 | 1  | 3  | 4  | 1        |
| Simple-Global        | 0  | 5 | 20 | 31 | 25 | 9        |

Table A.16: Mediation proportions and sample sizes within heterogeneous treatment effect regions across scenarios with K-means method. Standard deviations are shown in parentheses. The mediation proportion is calculated by R package "mediation". "For each set of results, the subtype exhibiting the lowest p-value for the mediation proportion is designated as the region of interest. Within this region, we estimate the mediation proportion, total treatment effect, and indirect treatment effect.

| Scenario     | True    |            | K-means  |            |
|--------------|---------|------------|----------|------------|
|              | N       | Med Prop   | N        | Med Prop   |
| Simple-All   | 251(13) | 1.37(0.09) | 264(121) | 1.21(0.11) |
| Simple-Part  | 251(13) | 0.76(0.03) | 278(127) | 0.80(0.05) |
| Complex-All  | 251(13) | 1.54(0.30) | 265(118) | 1.30(0.23) |
| Complex-Part | 251(13) | 0.85(0.08) | 284(122) | 1.03(0.14) |

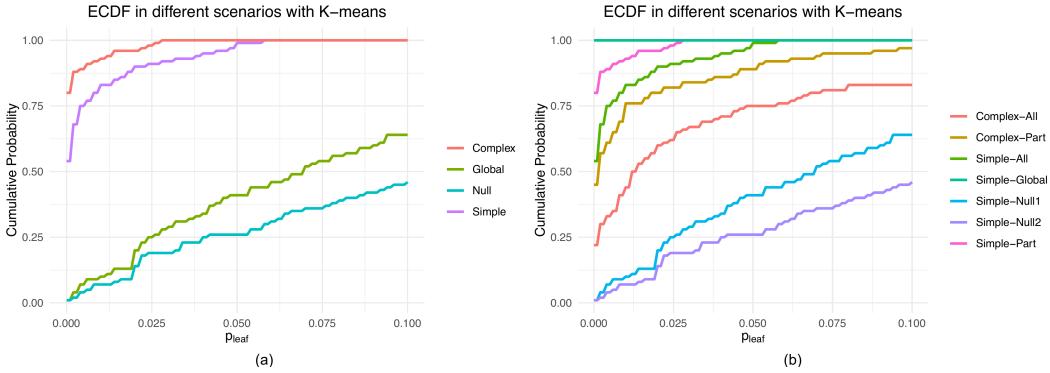


Figure A.9: Empirical cumulative distribution functions (ECDF) of  $p_{leaf}$  under four different scenarios, using K-means. These results illustrate the sensitivity of each base learner to underlying treatment effect heterogeneity across varying levels of complexity.(a)results for K-means without a mediator, (b) results for K-means with a mediator.

## A.5 Visualization

In this section, we visualize an intermediate step of the M-learner algorithm—specifically, the estimation of treatment effects—which provides insights into the underlying mechanisms contributing to its effectiveness.

### A.5.1 Without mediators

When model without mediators, we visualize the estimation of CATTE  $\hat{\tau}^{tot}(x) = \hat{g}_1(x) - \hat{g}_0(x)$ .

For each scenario , we estimate  $\hat{g}_0(x)$  and  $\hat{g}_1(x)$  using randomized clinical trial data. To facilitate visualization, we construct a synthetic grid of covariates where Cov1 and Cov2 vary from  $-1.5$  to  $1.5$  in increments of  $0.05$ , while all other covariates are drawn from a standard normal distribution. The estimated functions are then applied to this grid to compute the CATTE for each unit. We repeat this process across 100 simulated experiments and compute the average CATTE at each grid point. The resulting surface is interpolated using the R package "akima" and visualized to illustrate the spatial patterns in treatment heterogeneity. We visualize the results under a sample size of 1000 for each of the four scenarios—Simple, Complex, Global, and Null—using Random Forest and XGBoost as base learners. The visualizations are presented in Appendix Figure A.10 - A.13.

According to the predefined settings of each scenario in , the heterogeneous region in the Simple scenario exhibits positive CATTE values, while the Complex scenario presents negative CATTE values within its heterogeneous region. The Global scenario demonstrates uniformly positive CATTE across the entire covariate space, whereas the Null scenario yields CATTE values consistently close to zero. Our visualizations effectively capture these patterns, providing a clear and accurate reflection of treatment effect heterogeneity across scenarios.

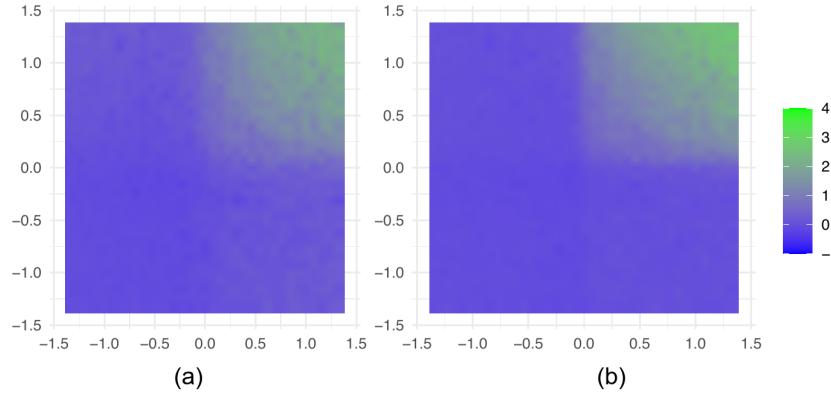


Figure A.10: Visualization of CATTE in the Simple scenario: each panel plots Cov1 (x-axis, -1.4 to 1.4) and Cov2 (y-axis, -1.4 to 1.4), with color representing the estimated CATTE magnitude.(a) visualization of RF learner result, (b) visualization of XGB learner result.

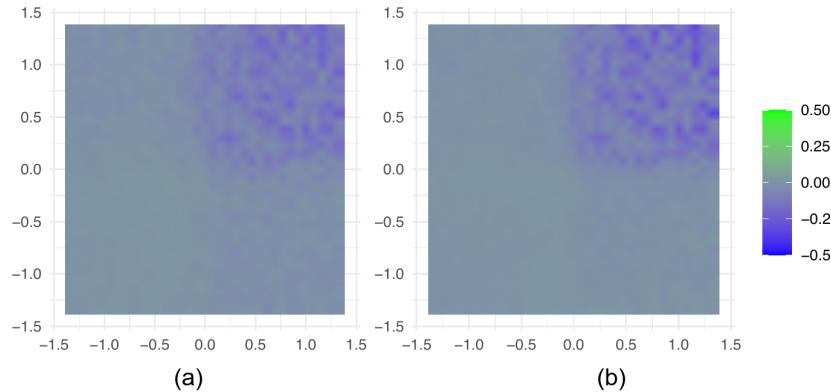


Figure A.11: Visualization of CATTE in the Complex scenario: each panel plots Cov1 (x-axis, -1.4 to 1.4) and Cov2 (y-axis, -1.4 to 1.4), with color representing the estimated CATTE magnitude. (a) visualization of RF learner result, (b) visualization of XGB learner result.

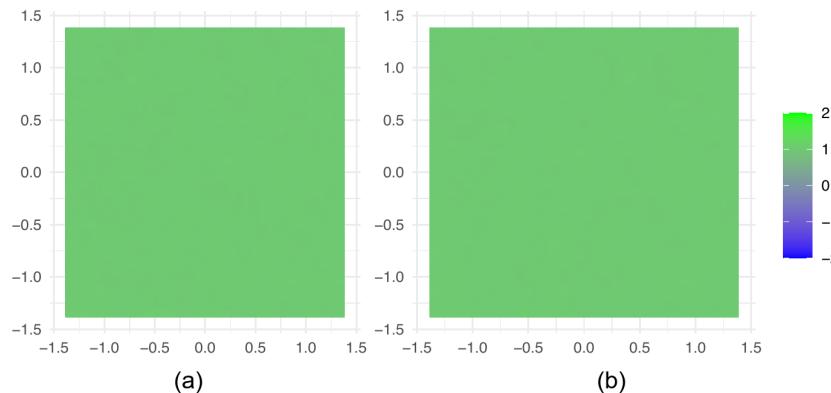


Figure A.12: Visualization of CATTE in the Global scenario:each panel plots Cov1 (x-axis, -1.4 to 1.4) and Cov2 (y-axis, -1.4 to 1.4), with color representing the estimated CATTE magnitude.(a) visualization of RF learner result, (b) visualization of XGB learner result.

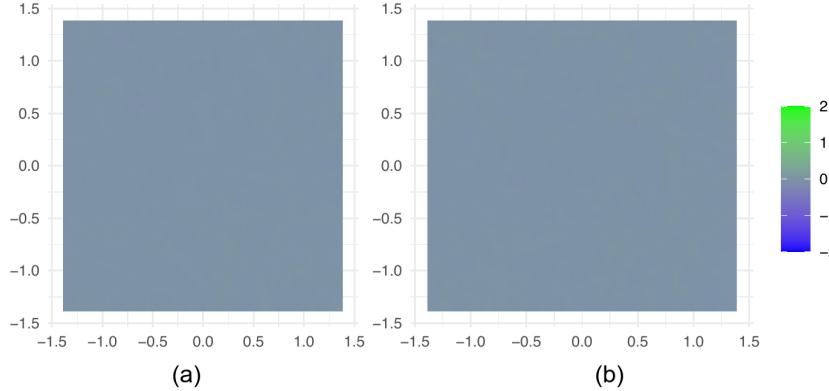


Figure A.13: Visualization of CATTE in the Null scenario:each panel plots Cov1 (x-axis, −1.4 to 1.4) and Cov2 (y-axis, −1.4 to 1.4), with color representing the estimated CATTE magnitude.(a) visualization of RF learner result, (b) visualization of XGB learner result.

### A.5.2 With a mediator

When model with a mediator, we visualize the estimation of CAITE  $\hat{\tau}^{ITE}(x) = \hat{g}_1^Y(x, \hat{g}_1^M(x)) - \hat{g}_1^Y(x, \hat{g}_0^M(x))$ .

For each scenario , we estimate  $\hat{g}_1^Y, \hat{g}_1^M(x)$  and  $\hat{g}_0^M(x)$  using randomized clinical trial data. To facilitate visualization, we construct a synthetic grid of covariates where Cov1 and Cov2 vary from −1.5 to 1.5 in increments of 0.05, while all other covariates are drawn from a standard normal distribution. The estimated functions are then applied to this grid to compute the CAITE for each unit. We repeat this process across 100 simulated experiments and compute the average CAITE at each grid point. The resulting surface is interpolated using the R package "akima" and visualized to illustrate the spatial patterns in treatment heterogeneity. We visualize the results under a sample size of 1000 for each of the seven scenarios—Simple-All, Simple-Part, Complex-All, Complex-Part, Simple-Null1, Simple-Null2 and Simple-Global—using Random Forest and XGBoost as base learners. The visualizations are presented in Appendix Figure A.14 - A.20.

According to the predefined settings of each scenario in , the heterogeneous region in the Simple-All and Simple-Part scenarios exhibit positive CAITE values, while the Complex-All and Complex-Part scenarios present negative CAITE values within its heterogeneous region. The Simple-Global scenario demonstrates uniformly positive CAITE across the entire covariate space, whereas the Simpe-Null1 and Simple-Null2 scenarios yield CAITE values consistently close to zero. Our visualizations effectively capture these patterns, providing a clear and accurate reflection of treatment effect heterogeneity across scenarios.

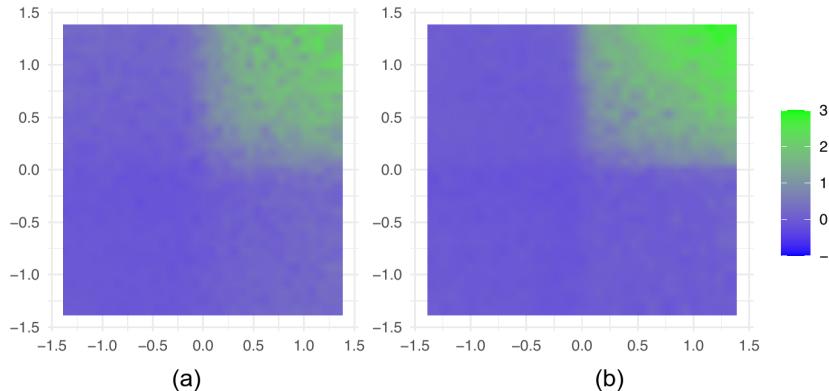


Figure A.14: Visualization of CAITE in the Simple-All scenario:each panel plots Cov1 (x-axis, −1.4 to 1.4) and Cov2 (y-axis, −1.4 to 1.4), with color representing the estimated CAITE magnitude.(a) visualization of RF learner result, (b) visualization of XGB learner result.

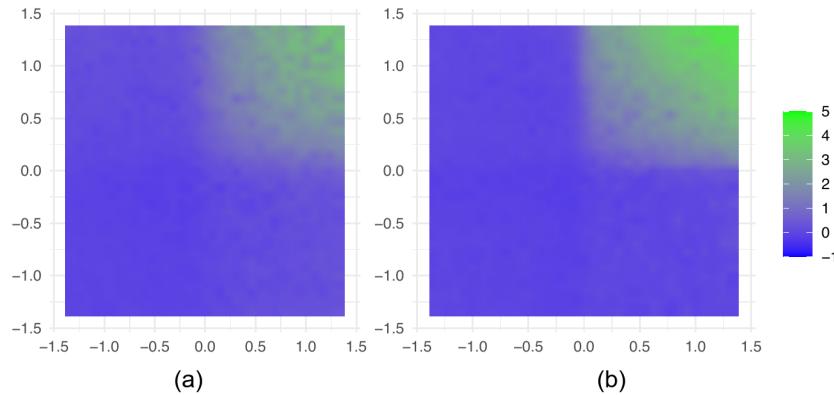


Figure A.15: Visualization of CAITE in the Simple-Part scenario:each panel plots Cov1 (x-axis, −1.4 to 1.4) and Cov2 (y-axis, −1.4 to 1.4), with color representing the estimated CAITE magnitude.(a) visualization of RF learner result, (b) visualization of XGB learner result.

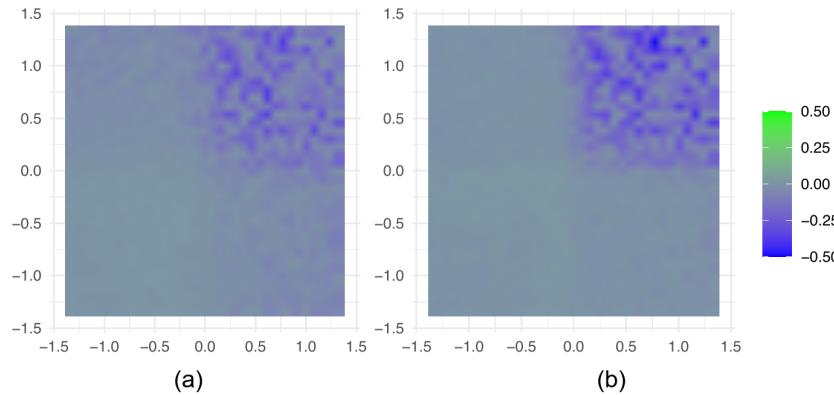


Figure A.16: Visualization of CAITE in the Complex-All scenario:each panel plots Cov1 (x-axis, −1.4 to 1.4) and Cov2 (y-axis, −1.4 to 1.4), with color representing the estimated CAITE magnitude.(a) visualization of RF learner result, (b) visualization of XGB learner result.

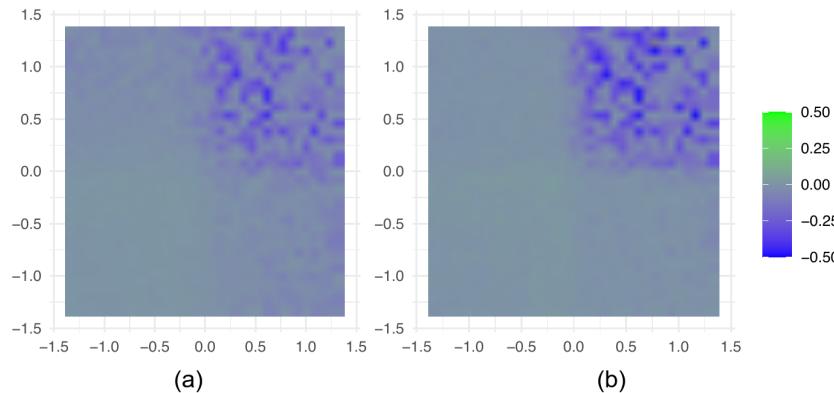


Figure A.17: Visualization of CAITE in the Complex-Part scenario:each panel plots Cov1 (x-axis, −1.4 to 1.4) and Cov2 (y-axis, −1.4 to 1.4), with color representing the estimated CAITE magnitude.(a) visualization of RF learner result, (b) visualization of XGB learner result.

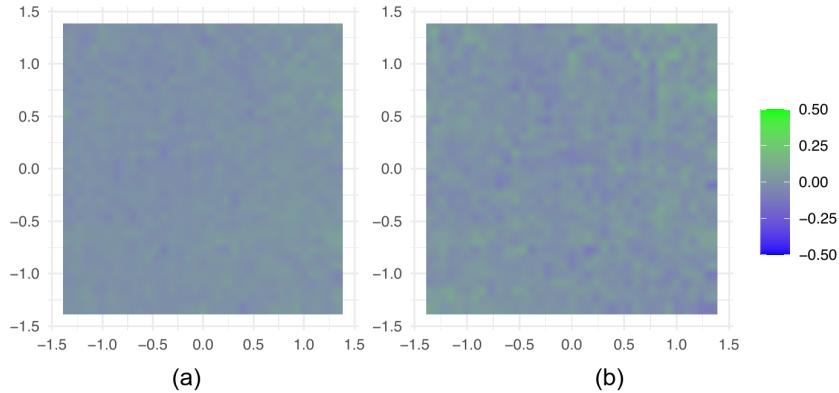


Figure A.18: Visualization of CAITE in the Simple-Null1 scenario:each panel plots Cov1 (x-axis, -1.4 to 1.4) and Cov2 (y-axis, -1.4 to 1.4), with color representing the estimated CAITE magnitude.(a) visualization of RF learner result, (b) visualization of XGB learner result.

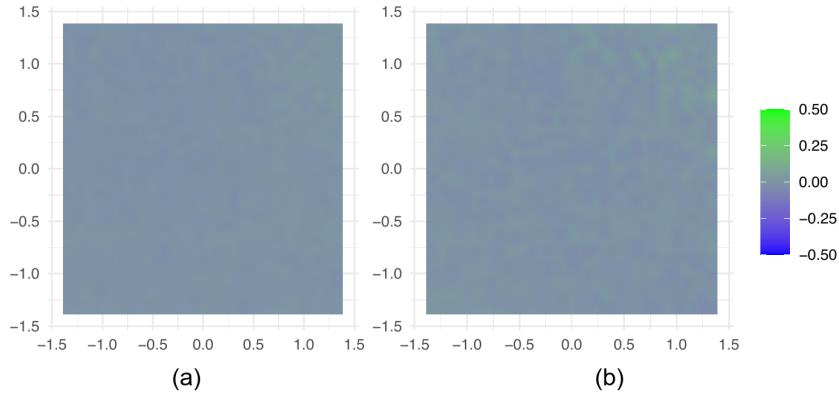


Figure A.19: Visualization of CAITE in the Simple-Null2 scenario:each panel plots Cov1 (x-axis, -1.4 to 1.4) and Cov2 (y-axis, -1.4 to 1.4), with color representing the estimated CAITE magnitude.(a) visualization of RF learner result, (b) visualization of XGB learner result.

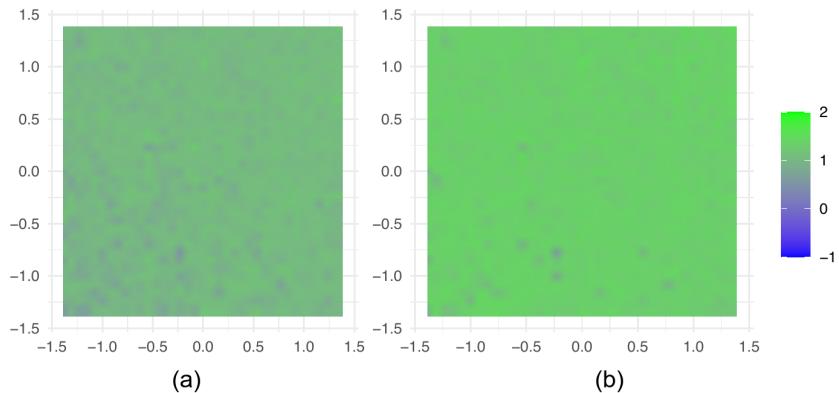


Figure A.20: Visualization of CAITE in the Simple-Global scenario:each panel plots Cov1 (x-axis, -1.4 to 1.4) and Cov2 (y-axis, -1.4 to 1.4), with color representing the estimated CAITE magnitude.(a) visualization of RF learner result, (b) visualization of XGB learner result.