A Tree-based Model Averaging Approach for Personalized Treatment Effect Estimation from Heterogeneous Data Sources

Xiaoqing Tan
University of Pittsburgh
Pittsburgh, PA 15213
xit31@pitt.edu

Chung-Chou Ho Chang
University of Pittsburgh
Pittsburgh, PA 15213
changj@pitt.edu

Lu Tang
University of Pittsburgh
Pittsburgh, PA 15213
lutang@pitt.edu

Abstract

Accurately estimating personalized treatment effects within a single study site has been challenging due to small sample sizes. Here we propose a tree-based model averaging approach to improve the estimation efficiency of conditional average treatment effects concerning the population of a target research site by leveraging models derived from potentially heterogeneous populations of other sites, but without them sharing individual-participant data. To our best knowledge, there is no established model averaging approach for distributed data with a focus on improving the estimation of treatment effects. Under distributed data networks, we develop an efficient and interpretable tree-based ensemble of personalized treatment effect estimators to join results across study sites, while actively modeling for the heterogeneity in data sources through site partitioning. The performance of this approach is demonstrated by a study of the causal effects of oxygen therapy settings on in-hospital mortality and backed up by comprehensive numerical results.

1 Introduction

Estimating individualized treatment effects, ranging from personalized medicine, policy research, to customized marketing advertisement, has been a hot topic. Treatment effects of certain subgroups within the population are often of interest. Recently, there has been an explosion of research devoted to improving estimation and inference of covariate-specific treatment effects, or conditional average treatment effects (CATE) at a target research site [1, 2, 3, 4, 5]. However, due to the limited sample size in a single study, improving the accuracy of the estimation of treatment effects remains challenging.

Leveraging data and models from various research sites to conduct statistical analyses is becoming increasingly popular [6, 7, 8]. Distributed research networks have been established in many large scale studies [9, 10, 11, 12]. A question often being asked is whether additional data or models from other research sites could bring improvement to a local estimation task, especially when a single site does not have enough data to achieve a desired power. This concern is most noticeable in estimating treatment effects where sample size requirement is high yet observations are typically limited. Also, the amount of data exchanged between data sites is restricted due to efficiency and privacy concerns, hence prohibiting data from being pooled into a central location [13]. One way to tackle this challenge is through model averaging [14], where multiple research sites collectively contribute to the tasks of statistical modeling without sharing sensitive data. To our best knowledge, there is no established model averaging approach and result for distributed data with the goal of improving the estimation of CATE.

35th Conference on Neural Information Processing Systems (NeurIPS 2021), Sydney, Australia.

Our paper focuses on improving the prediction efficiency of CATE (to be formally defined) concerning a target site by leveraging models derived from other sites where populations and treatment effects are potentially different compared to the target site, but without them sharing individual-level data. To be specific, we consider two levels of potential heterogeneity in treatment effects. The first is local heterogeneity where patients with different characteristics may have different treatment effects within a hospital. It is also known as CATE, covariate-specific treatment effects, or heterogeneous treatment effects. The second is site-level heterogeneity where the same patient may experience different treatment effects at different hospitals that are driven by site-level confounding. We refer to this distributed data network as heterogeneous data sources and details are discussed in Section 2.

In the paper, we propose a flexible tree-based weighting scheme to combine models from each site that takes into account model heterogeneity, where the contribution of each model to the target site depends on subject characteristics. Tree splitting may occur at both site level and subpopulation level, resulting in a flexible information-sharing scheme that is site and feature-dependent. For example, treatment effects in two hospitals may be similar for female patients but different for males, suggesting us to consider borrowing information across sites only on selective subgroups, i.e., females. This is a more data-adaptive weighting scheme than the global weighting schemes used in classic model averaging [15, 16, 17].

The key contributions of this paper are summarized as follows. *(i)* We propose a model averaging scheme that is adaptive to both model heterogeneity and subject features via tree-splitting. *(ii)* We generalize model averaging techniques to the field of causal inference. Causal assumptions with practical implications are explored to warrant the use of our approach. *(iii)* Compared to other data distributed learning methods, the proposed framework ensures the privacy of individual-participant data, facilitating practical collaboration research within distributed research networks.

2 Model Averaging and Related Works

Data integration approaches have received wide attention in recent years partly due to the increasing availability of distributed research networks [9, 10, 11, 12]. There are two main types of construct of a distributed database [18]: *homogeneous* versus *heterogeneous*. For homogeneous data sources, data across sites are random samples of the global population. Recent works [19, 20, 21, 22, 23, 24] all assume samples are randomly partitioned, which guarantees identical data distribution across sites. The goal of these works is to improve the overall estimation by averaging results from homogeneous sample divisions.

In practice, however, there is too much site-level heterogeneity in a distributed data network to warrant direct aggregation of models obtained from local sites. The focus shifts to improving the estimation of a target site by selectively leveraging information from other data sources. There are two main classes of approaches. The first class [25, 26, 27] is based on comparison of the learned model parameters $\{\widehat{\boldsymbol{\theta}}_1, \ldots, \widehat{\boldsymbol{\theta}}_K\}$ from K different sites where for site k we adopt model $f_k(\boldsymbol{x}) = f(\boldsymbol{x}; \boldsymbol{\theta}_k)$ with subject features $\boldsymbol{x}$ to approximate the outcome of interest Y. Clustering and shrinkage approaches are then used by merging data or models that are similar. Most of these require the pooling of individual-participant data. The second class of approaches falls in the model averaging framework [14] with weights directly associated with the local prediction. Let site 1 be our target site, and the goal is to improve f_1 using a weighted estimator $f^*(\boldsymbol{x}) = \sum_{k=1}^{K} \omega_k f_k(\boldsymbol{x})$, where $\omega_k = \frac{\exp\{-\sum_{i \in \mathcal{I}_1}(f_k(\boldsymbol{x}_i)-y_i)^2\}}{\sum_{\ell=1}^{K}\exp\{-\sum_{i \in \mathcal{I}_1}(f_\ell(\boldsymbol{x}_i)-y_i)^2\}}$, with Y_i the observed outcome of subject i in $\mathcal{I}_1$ (the index set of site 1) and ω_k the weights proportional to the prediction performance of f_k on site 1 (e.g., residual sum of squares), and $\sum_k \omega_k = 1$. The above is termed exponential weighted model averaging (EWMA), one of the classic model averaging approaches. Several variations of ω_k can be found in [15, 16, 17]. In general, separate samples are used to obtain the estimates of ω_k's and f_k's, respectively.

In causal inference, there is a lot of interest in identifying subgroups with enhanced treatment effects, targeting at the feasibility of customizing estimates for individuals [1, 2, 3, 4, 5]. These methods aim to estimate the CATE function $\tau(\boldsymbol{x})$, denoting the difference in potential outcomes between treatment and control, conditional on subject characteristics $\boldsymbol{x}$. To reduce uncertainty in estimation of personalized treatment effects, incorporating additional data or models are sough after [28]. There is some recent progress on bridging the findings from an experimental study with observational data

[29, 30, 31]. However, their methods require fully centralized data. In contrast, we leverage the distributed nature of model averaging to derive an integrative CATE estimator. In this paper, we propose a tree-based model averaging framework, designed to improve treatment effect estimation for subgroups within a target site. We assume a heterogeneously distributed data network as this is the more common yet challenging case in practice. The extension is non-trivial because the output of CATE is unobserved in nature, as compared to a standard $f(\boldsymbol{x})$ whose output Y is readily available. Our model averaging weights not only depend on sites, but also on the subject characteristics.

3 A Tree-based Model Averaging Framework

3.1 Notations and Definitions

We first introduce notations related to our goal of conditional average treatment effect (CATE) estimation. Let Y denote the outcome of interest, $Z \in \{0, 1\}$ denote a binary treatment indicator, and $\boldsymbol{X}$ denote subject features. Correspondingly, let y, z and $\boldsymbol{x}$ denote their realizations. Using the potential outcome framework [32, 33], we define CATE as $\tau(\boldsymbol{x}) = E[Y^{(Z=1)} - Y^{(Z=0)}|\boldsymbol{X} = \boldsymbol{x}]$, where $Y^{(Z=1)}$ and $Y^{(Z=0)}$ are the counterfactual outcomes under treated $Z = 1$ and control $Z = 0$, respectively. In other words, it is the expected difference of the potential outcomes between two treatment groups for individuals with characteristics $\boldsymbol{X}$. By the causal consistency assumption, the observed outcome is $Y = Y^{Z} = ZY^{Z=1} + (1 - Z)Y^{Z=0}$.

Suppose the distributed data network $\mathfrak{D} := \{\mathcal{D}_k\}_{k=1}^{K}$ consists of K sites, each of which has a sample size of n_k. We consider for site k the data setup $\mathcal{D}_k = \{y_i, z_i, \boldsymbol{x}_i\}_{i\in\mathcal{I}_k}$, where $\mathcal{I}_k$ is the index set of site k. The CATE function is hence given by $\tau_k(\boldsymbol{x}) = E_k[Y^{(Z=1)} - Y^{(Z=0)}|\boldsymbol{X} = \boldsymbol{x}]$, where the expectation is taken over the data distribution in site k. Without loss of generality, we assume the goal is to estimate the CATE function in site 1, τ_1.

3.2 Assumptions

To ensure information can be properly borrowed across sites without introducing additional bias, we first impose the following idealistic assumptions, and then discuss relaxations of Assumption 2.

Assumption 1: $\{Y^{(Z=0)}, Y^{(Z=1)}\} \perp Z|\boldsymbol{X}, S$.

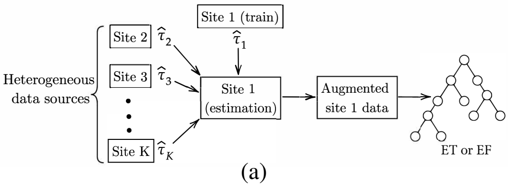
Assumption 2: $\{Y^{(Z=0)}, Y^{(Z=1)}\} \perp S|\boldsymbol{X}$.

Assumption 3: $0 < P(S = 1|\boldsymbol{X}) < 1$ and $0 < P(Z = 1|\boldsymbol{X}, S) < 1 \quad$ for all $\boldsymbol{X}$ and S.

Here S is the site indicator taking values in $\mathcal{S} = \{1, \ldots, K\}$ such that $S_i = k$ if $i \in \mathcal{I}_k$. Assumption 1 ensures treatment effects are unconfounded within sites so that $\tau_k(\boldsymbol{x})$ can be identified. This assumption holds by design when data are collected from randomized controlled trials or when treatment assignment depends only on a subset of $\boldsymbol{X}$. CATE can then be consistently estimated with data in each site, i.e., $\tau_k(\boldsymbol{x}) = E_k[Y^{(Z=1)} - Y^{(Z=0)}|\boldsymbol{X} = \boldsymbol{x}] = E[Y|\boldsymbol{X} = \boldsymbol{x}, S = k, Z = 1] - E[Y|\boldsymbol{X} = \boldsymbol{x}, S = k, Z = 0]$. The second equality directly results from the assumption. Assumption 2 essentially states that the CATE functions are transportable, i.e., $\tau_k(\boldsymbol{x}) = \tau_{k'}(\boldsymbol{x})$ for $k, k' \in \{1, \ldots, K\}$. See also in [34] and [35] for similar consideration. This assumption may not be satisfied due to site level heterogeneity across sites. In other words, site can be a confounder which prevents transporting of CATE functions across sites. Our method allows Assumption 2 to be violated and use model averaging weights to determine transportability. In a special case in Section 4, we consider the assumption to hold for a subset of sites that contains site 1, i.e., $\{Y^{(Z=0)}, Y^{(Z=1)}\} \perp S_1|\boldsymbol{X}$, where S_1 takes values in $\mathcal{S}_1 = \{k : \tau_k(\boldsymbol{x}) = \tau_1(\boldsymbol{x})\}$ and $\{1\} \subset \mathcal{S}_1 \subset \mathcal{S}$. We denote $\mathcal{S}_1$ as the set of transportable sites with regard to site 1. Hence, transportability holds across some sites and some patient types. When the above assumption fails to hold and $\mathcal{S}_1 = \{1\}$, bias may be introduced to site 1 by model averaging. However, our approach is still able to exploits the bias and variance trade off to improve prediction. Assumption 3 ensures that all subjects are possible to be observed in site 1 and all subjects in all sites are possible to receive either arm of treatment.

3.3 Tree-based Adaptive Model Averaging

A two-stage model averaging approach is proposed. We first split the data in the target site (site 1) into a training set and an estimation set. *(i) Local stage:* Obtain the estimated CATE function $\widehat{\tau}_1$ from the training set in site 1. Other sites locally obtain $\{\widehat{\tau}_k\}_{k=2}^K$. These K local models are then passed to site 1 to get K predicted treatment effects for each subject in the estimation set in site 1, resulting in an augmented data set. *(ii) Ensemble stage:* A tree-based ensemble model is trained on the augmented data by either an ensemble regression tree (ET) or an ensemble random forest (EF), with the predicted treatment effects from the previous stage as the outcome. The site indicator of which local model is used as well as the subject characteristics are fed into the ensemble model as predictors. The resulting model will be used to compute our proposed model averaging estimator. Figure 1 illustrates a conceptual diagram of the proposed model averaging framework and structure of the augmented data. Algorithm 1 provides an algorithmic overview. Our method has been implemented as an R package `ifedtree` available on GitHub (`github.com/ellenxtan/ifedtree`, see A.4).



(a)

Subject	Site	$\boldsymbol{X}$	S	τ
1	1	$\boldsymbol{x}_1$	1	$\tau_1(\boldsymbol{x}_1)$
		$\vdots$		
1	1	$\boldsymbol{x}_1$	K	$\tau_K(\boldsymbol{x}_1)$
2	1	$\boldsymbol{x}_2$	1	$\tau_1(\boldsymbol{x}_2)$
		$\vdots$		

(b)

Figure 1: (a) Schema of the proposed algorithm. (b) Illustration of the augmented data constructed from the estimation set of site 1.

Algorithm 1 Tree-based model averaging for heterogeneous data sources

```
for $k \leftarrow 1, 2, \ldots$ to $K$ do                    ▷ Loop through $K$ sites. Can be run in parallel.
    Build a local model using site $k$ data. Site 1 model uses its training set only.
for $i \in \mathcal{I}_1^{(2)}$ do                            ▷ Loop through subjects in site 1 estimation set.
    for $k \leftarrow 1, 2, \ldots$ to $K$ do                          ▷ Loop through $K$ local models.
        Predict $\widehat{\tau}_k(\boldsymbol{x}_i)$ using local model $k$.
        $D_{i,k} \leftarrow [\boldsymbol{x}_i, k, \widehat{\tau}_k(\boldsymbol{x}_i)]$.
Create augmented site 1 data $\mathfrak{D}_{aug,1}$ by concatenating $D_{i,k}$ vectors.
$\widehat{\mathcal{T}}_{\text{EF}}(\boldsymbol{x}, s) \leftarrow$ EnsembleForest($\mathfrak{D}_{aug,1}$)          ▷ Or EnsembleTree when $B = 1$.
```

We consider an adaptive weighting of $\{\tau_1, \ldots, \tau_K\}$ by

$$\tau^*(\boldsymbol{x}) = \textstyle\sum_{k=1}^K \omega_k(\boldsymbol{x})\tau_k(\boldsymbol{x}) \tag{1}$$

where τ^* is the weighted model averaging estimator. The weight functions $\omega_k(\boldsymbol{x})$'s are not only site-specific, but also depend on $\boldsymbol{x}$, and follow $\sum_{k=1}^K \omega_k(\boldsymbol{x}) = 1$. It measures the importance of τ_k in assisting site 1 when subjects with characteristics $\boldsymbol{x}$ is of interest. We rely on each of the sites to derive their respective $\widehat{\tau}_k$ from $\mathcal{D}_k$ so that $\mathcal{D}_1, \ldots, \mathcal{D}_K$ do not need to be pooled. Only the estimated functions $\{\widehat{\tau}_2, \ldots, \widehat{\tau}_K\}$ are passed to site 1. Site 1 also estimates its own $\widehat{\tau}_1$ using half of the samples in $\mathcal{D}_1$, whose indices belong to $i \in \mathcal{I}_1^{(1)}$, the training set. We describe the approaches to estimate $\widehat{\tau}_k$ in Section 3.4. The weight functions are then estimated using the remaining samples in $\mathcal{D}_1$, denoted as $i \in \mathcal{I}_1^{(2)}$, the estimation set. Unlike in classic model averaging where Y is observed, since treatment effects are not directly observed, weights are estimated based on expected treatment effects.

A tree-based ensemble is constructed to estimate the weighting functions $\{\omega_k\}_{k=1}^K$. Heterogeneity across sites is explained by including the site index into an augmented training set when building trees. An intuition of our approach is that sites that are split away from site 1 (by tree nodes) are ignored and the sites that fall into the same leaf node are considered homogeneous to site 1 hence contribute to the estimation of $\tau_1(\boldsymbol{x})$. A splitting by site may occur in any branches of a tree, resulting in an information sharing scheme across sites that is dependent on $\boldsymbol{x}$. We construct the ensemble by first creating an augmented data set for subjects in $\mathcal{I}_1^{(2)}$, the estimation set where $\mathfrak{D}_{aug,1} = \{\boldsymbol{x}_i, k, \widehat{\tau}_k(\boldsymbol{x}_i)\}_{i \in \mathcal{I}_1^{(2)}, k \in \mathcal{S}}$. The illustration of this augmented site 1 data is given in Figure 1b. An ensemble is then trained on this data by either a regression tree or a random forest, with the estimated treatment effects $\widehat{\tau}_k(\boldsymbol{x}_i)$ as the outcome, and a categorical site indicator of which local model is used along with all patient features as predictors, i.e., $(\boldsymbol{x}_i, k)$. We denote the resulting function as $\mathcal{T}(\boldsymbol{x}, s)$ which depends on both $\boldsymbol{x}$ and site s, specifically, $\mathcal{T}_{\text{ET}}(\boldsymbol{x}, s)$ and $\mathcal{T}_{\text{EF}}(\boldsymbol{x}, s)$ for ensemble tree (ET) and ensemble forest (EF), respectively. Let $\mathcal{L}(\boldsymbol{x}, s)$ denote the final partition of the feature space by the tree to which the pair $(\boldsymbol{x}, s)$ belongs. The ET estimate based on the augmented site 1 data can be derived by

$$\begin{aligned}\widehat{\mathcal{T}}_{\text{ET}}(\boldsymbol{x}, s) = & \tfrac{1}{|\{(i,k):(\boldsymbol{x}_i,k)\in\mathcal{L}(\boldsymbol{x},s)\}_{i\in\mathcal{I}_1^{(2)},k\in\mathcal{S}}|} \\ & \sum_{\{(i,k):(\boldsymbol{x}_i,k)\in\mathcal{L}(\boldsymbol{x},s)\}_{i\in\mathcal{I}_1^{(2)},k\in\mathcal{S}}} \widehat{\tau}_k(\boldsymbol{x}_i) \\ = & \textstyle\sum_{i\in\mathcal{I}_1^{(2)}} \sum_{k=1}^{K} \frac{\mathbb{1}\{(\boldsymbol{x}_i,k)\in\mathcal{L}(\boldsymbol{x},s)\}}{|\mathcal{L}(\boldsymbol{x},s)|} \widehat{\tau}_k(\boldsymbol{x}_i).\end{aligned} \tag{2}$$

Intuitively, observations with similar characteristics ($\boldsymbol{x}$ and $\boldsymbol{x}'$) and from similar sites (s and s') are more likely to fall in the same partition region in the ensemble tree, i.e., $(\boldsymbol{x}, s) \in \mathcal{L}(\boldsymbol{x}', s')$ or $(\boldsymbol{x}', s') \in \mathcal{L}(\boldsymbol{x}, s)$. This resembles a non-smooth kernel where weights are $1/|\mathcal{L}(\boldsymbol{x}, s)|$ for observations that are within the neighborhood of $(\boldsymbol{x}, s)$, and 0 otherwise. The estimator borrows information from neighbors in the space of $\boldsymbol{X}$ and S. The splits of the tree are based on minimizing in-sample MSE of $\widehat{\tau}$ within each leaf and pruned by cross-validation over choices of the complexity parameter. Since a single tree is prone to be unstable, in practice, we use random forest to reduce variance and smooth the partitioning boundaries. By aggregating B ET estimates each based on a subsample of the augmented data, $\{\widehat{\mathcal{T}}^{(b)}\}_{b=1}^{B}$, an EF estimate can be constructed by

$$\begin{aligned}\widehat{\mathcal{T}}_{\text{EF}}(\boldsymbol{x}, s) &= \tfrac{1}{B} \textstyle\sum_{b=1}^{B} \widehat{\mathcal{T}}^{(b)}(\boldsymbol{x}, s) \\ &= \textstyle\sum_{i\in\mathcal{I}_1^{(2)}} \sum_{k=1}^{K} \lambda_{i,k}(\boldsymbol{x}, s) \widehat{\tau}_k(\boldsymbol{x}_i), \\ \text{where } \lambda_{i,k}(\boldsymbol{x}, s) &= \tfrac{1}{B} \textstyle\sum_{b=1}^{B} \frac{\mathbb{1}\{(\boldsymbol{x}_i,k)\in\mathcal{L}_b(\boldsymbol{x},s)\}}{|\mathcal{L}_b(\boldsymbol{x},s)|}.\end{aligned} \tag{3}$$

The form of $\widehat{\mathcal{T}}^{(b)}(\boldsymbol{x}, s)$ closely follows (2) but is based on a subsample of $\mathfrak{D}_{aug,1}$. The weights, $\lambda_{i,k}(\boldsymbol{x}, s)$, are similar to that in (2), and can be viewed as kernel weighting that defines an adaptive neighborhood of $\boldsymbol{x}$ and s. Each site can contribute partial information but not all or none. We then obtain the model averaging estimates defined in (1) by fixing $s = 1$ such that $\widehat{\tau}^*_{\text{ET}}(\boldsymbol{x}) = \widehat{\mathcal{T}}_{\text{ET}}(\boldsymbol{x}, s = 1)$ or $\widehat{\tau}^*_{\text{EF}}(\boldsymbol{x}) = \widehat{\mathcal{T}}_{\text{EF}}(\boldsymbol{x}, s = 1)$. The weight functions $\{\omega_k(\boldsymbol{x})\}_{k=1}^{K}$ for $\widehat{\tau}^*(\boldsymbol{x})$ can be immediately obtained from the ET or EF by

$$\begin{aligned}\widehat{\tau}^*_{\text{ET}}(\boldsymbol{x}) &= \widehat{\mathcal{T}}_{\text{ET}}(\boldsymbol{x}, 1) = \textstyle\sum_{k=1}^{K} \widehat{\omega}_k(\boldsymbol{x}) \widehat{\tau}_k(\boldsymbol{x}), \\ \text{where } \widehat{\omega}_k(\boldsymbol{x}) &= \textstyle\sum_{i\in\mathcal{I}_1^{(2)}} \frac{\mathbb{1}\{(\boldsymbol{x}_i,k)\in\mathcal{L}(\boldsymbol{x},1)\}}{|\mathcal{L}(\boldsymbol{x},1)|}; \\ \widehat{\tau}^*_{\text{EF}}(\boldsymbol{x}) &= \widehat{\mathcal{T}}_{\text{EF}}(\boldsymbol{x}, 1) = \textstyle\sum_{k=1}^{K} \widehat{\omega}_k(\boldsymbol{x}) \widehat{\tau}_k(\boldsymbol{x}), \\ \text{where } \widehat{\omega}_k(\boldsymbol{x}) &= \textstyle\sum_{i\in\mathcal{I}_1^{(2)}} \lambda_{i,k}(\boldsymbol{x}, 1).\end{aligned}$$

It can be verified that $\sum_{k=1}^{K} \widehat{\omega}_k(\boldsymbol{x}) = 1$ for all $\boldsymbol{x}$. As our simulations in Section 4 show, $\widehat{\tau}^*$ improves the local functional estimate $\widehat{\tau}_1$. We set $B = 2000$ throughout the paper. Tree and forest estimates are obtained by R packages `rpart` and `grf`, respectively.

3.4 Local Models: obtaining $\widehat{\tau}_k$

Estimate of $\tau_k(\boldsymbol{x})$ at each local site must be obtained separately before the ensemble. Our proposed ensemble framework can be applied to a general estimator of $\tau_k(\boldsymbol{x})$. For each site, the local estimate could be obtained using different methods. Recently, there has been many work dedicated to the estimation of individualized treatment effects [1, 2, 3, 4, 5]. As an example, we consider using the causal tree (CT) [1] to estimate the local model at each site. CT is a non-linear learner that enjoys the following convenience: *(i)* allows different types of outcome such as discrete and continuous outcomes, hence can be applied to a broad range of real data scenarios; *(ii)* can easily handle a large number of features and high order interactions by construction; *(iii)* can be applied to both experimental studies and observational studies. See [1] for a detailed description. CT is implemented in the R package `causalTree`. We explore other estimating options for local models in Appendix A.2.

3.5 Asymptotic Properties

We provide the consistency guarantee of the proposed estimator $\widehat{\mathcal{T}}_{\text{EF}}$ for the true target τ_1. Assuming a consistent local estimator, the EF with subsampling procedure described in Appendix A.1 is consistent.

Theorem 1 *Suppose the subsamples used to build each tree in an ensemble forest are drawn from different subjects in the augmented site 1 data. Under the following conditions:*

(i) Bounded covariates: Features $\boldsymbol{X}_i$ and the site indicator S_i are independent and have a density that is bounded away from 0 and infinity.

(ii) Lipschitz response: the conditional mean function $\mathbb{E}[\mathcal{T}|\boldsymbol{X}=\boldsymbol{x}, S=k]$ is Lipschitz-continuous.

(iii) Honest trees: trees in the random forest use different data for placing splits and estimating leaf-wise responses.

Then $\widehat{\mathcal{T}}_{EF}(\boldsymbol{x}, s) \xrightarrow{p} \tau_s(\boldsymbol{x})$, for all $\boldsymbol{x}$ and s, as $\min_k n_k \to \infty$. Hence, $\widehat{\tau}^*_{EF}(\boldsymbol{x}) \xrightarrow{p} \tau_1(\boldsymbol{x})$.

The detailed description of the conditions and the proof of Theorem 1 is presented in Appendix A.1. With finite sample, bias will always be introduced from the local models, leading to biased model averaging estimates. To better explore the consistency properties of our proposed methods, we add in Section 4 enhanced versions of ET and EF estimators, denoted as ET-cate and EF-cate, which apply ground truth when building the augmented site 1 data so that $\widehat{\tau}_k$ in (2) and (3) is replaced by τ_k. This removes the bias and uncertainty in local model estimation so that bias and uncertainty only results from the estimation of the ensemble model. Empirical results show that ET-cate and EF-cate achieve minimal bias and variance. In other words, the extra bias and uncertainty introduced by model averaging is small.

4 Simulation Study

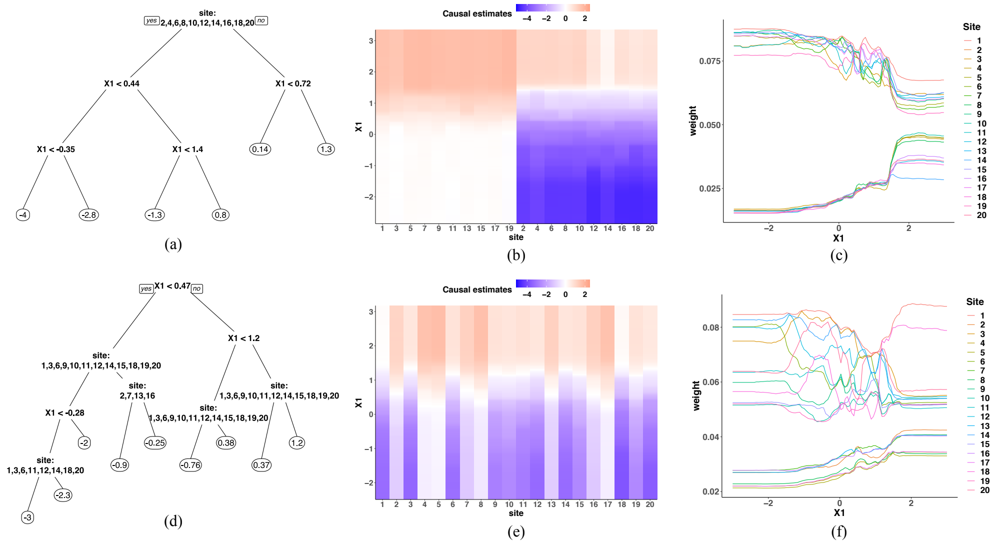


Figure 2: Tree-based model averaging results under **discrete grouping (a,b,c)** and **continuous grouping (d,e,f)** when $c = 1$. (a) and (d) visualize the fitted ETs where the site indicator and X_1 appear as splitting variables. (b) and (e) show the predicted treatment effects of EFs for different values of X_1 in each site, marginalized over all other features. (b) is arranged according to the true grouping, odd sites versus even sites. (c) and (f) plot the model averaging weights in EFs by varying X_1. Corresponding ET and EF show consistent patterns and recover the true grouping.

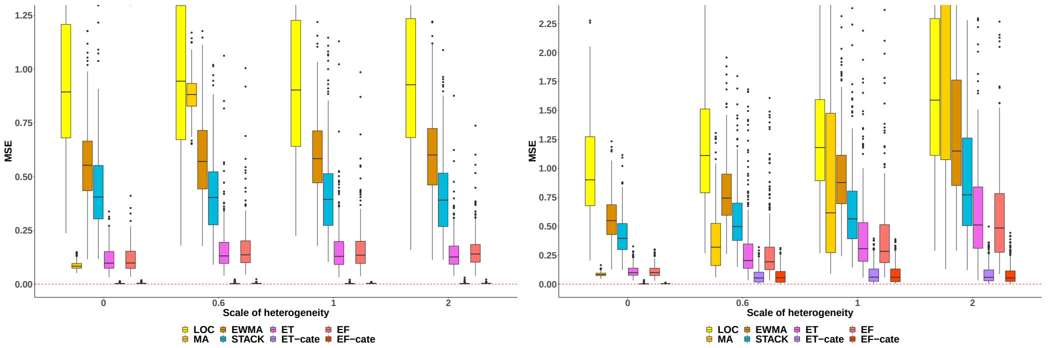
Monte Carlo simulations are conducted to assess the proposed methods. We specify $m(\boldsymbol{x}, k)$ as the conditional mean of outcome and $\tau(\boldsymbol{x}, k)$ as the conditional treatment effect for individuals with features $\boldsymbol{x}$ in site k. The marginal treatment probability is $P = 0.5$. The potential outcomes can be written as $Y_i^{(z)} = m(\boldsymbol{X}_i, S_i) + \frac{1}{2} \cdot (2z-1) \cdot \tau(\boldsymbol{X}_i, S_i) + \epsilon_i$. The mean function is $m(\boldsymbol{x}, k) = \frac{1}{2}x_1 + \sum_{d=2}^{4} x_d + (x_1 - 3) \cdot c \cdot U_k$, and the CATE function is specified as

$$\tau(\boldsymbol{x}, k) = \mathbb{1}\{x_1 > 0\} \cdot x_1 + (x_1 - 3) \cdot c \cdot U_k,$$

where $z = 0, 1$, U_k denotes the site-level heterogeneity, and $\epsilon_i \sim N(0, 1)$. Features $\boldsymbol{X}_i \in \mathbb{R}^5$ are independent of ϵ_i, and $\boldsymbol{X}_i \sim N(\boldsymbol{0}, \boldsymbol{I})$. The simulation setting within each site (with k fixed) is motivated by designs in [1]. Features in τ are predictive markers while those in m but not in τ are prognostic only. Features that do not affect outcomes are noise covariates. The data are generated under a distributed data networks where heterogeneity may exist across sites. We assume there are $K = 20$ sites in total, each with a sample size $n = 500$. Two scenarios for site-level heterogeneity U_k are considered. For **discrete grouping**, we assume there are two underlying groups among the K sites $U_k \sim Bernoulli(0.5)$. Specifically, we assume odd-index sites and even-index sites form two distinct groups $\mathcal{G}_1 = \{k : k \bmod 2 = 1\} = \{1, 3, \ldots, K-1\}$; $\mathcal{G}_2 = \{k : k \bmod 2 = 0\} = \{2, 4 \ldots, K\}$ such that $U_{k \in \mathcal{G}_1} = 0$ and $U_{k \in \mathcal{G}_2} = 1$. The case of **continuous grouping** is considered as well where we assume $U_k \sim Unif[0, 1]$. Sites from similar underlying groupings have similar treatment effects and mean effects, while sites from different underlying groupings have different treatment effects and mean effects. We test different scales of the site-level heterogeneity under the discrete grouping and continuous grouping cases, respectively, with the scale factor denoted as c, taking values in $c \in \{0, 0.6, 1, 2\}$. A scale $c = 0$ implies all data sources are homogeneous. In other words, Assumption 2 is satisfied when $c = 0$ but not when $c > 0$.

The proposed approaches (ET and EF) are compared with several competing methods. The hypothetical cases when the ground truth τ_k's are used when constructing model averaging weights are denoted as ET-cate and EF-cate. We compare with the local estimator CT which does not utilize external information, denoted as LOC. A naive model averaging method is compared, denoted as MA, with weights $\omega_k^{\text{MA}} = 1/k$. This approach assumes models are homogeneous. We also consider a modified version of EWMA that can be used for CATE. Under our setting, the target outcome τ is unobserved. We obtain an approximation of $\tau_1(\boldsymbol{x})$ by fitting another local model using the estimation set of site 1, denoted by $\widetilde{\tau}_1(\boldsymbol{x})$. Essentially, we would like to measure weights that could provide information about the relative degree of similarity between sites so that CATE models closer to site 1 could be borrowed. Its weights are given by $\omega_k^{\text{EWMA}} = \frac{\exp\{-\sum_{i \in \mathcal{I}_1^{(2)}} (\widehat{\tau}_k(\boldsymbol{x}_i) - \widetilde{\tau}_1(\boldsymbol{x}_i))^2\}}{\sum_{\ell=1}^{K} \exp\{-\sum_{i \in \mathcal{I}_1^{(2)}} (\widehat{\tau}_\ell(\boldsymbol{x}_i) - \widetilde{\tau}_1(\boldsymbol{x}_i))^2\}}$. Moreover, we consider the commonly used stacking approach, denoted as STACK, an ensemble method in machine learning that combines predictions of several models. Specifically, we adopt linear stacking, which posits a linear regression of the outcome of interest on the predicted values of multiple models and obtains coefficients (or so-called "weights") of contribution of each model. To our end, we regress $\widetilde{\tau}_1(\boldsymbol{x})$ on the predictions of the estimation set in site 1 from each local model, $\{\widehat{\tau}_1(\boldsymbol{x}), \ldots, \widehat{\tau}_k(\boldsymbol{x})\}$. The stacking weights are not probabilistic hence not directly interpretable. We report the empirical bias and MSE of these methods over an independent testing set of sample size $n_{te} = 2000$ from site 1 where $\text{Bias}(\widehat{\tau}) = \frac{1}{n_{te}} \sum_{i=1}^{n_{te}} \{\widehat{\tau}(\boldsymbol{x}_i) - \tau_1(\boldsymbol{x}_i)\}$ and $\text{MSE}(\widehat{\tau}) = \frac{1}{n_{te}} \sum_{i=1}^{n_{te}} \{\widehat{\tau}(\boldsymbol{x}_i) - \tau_1(\boldsymbol{x}_i)\}^2$. Each simulation scenario is repeated for 1000 times. More information on the simulation code is provided in Appendix A.4. Experiments are performed on a 6-core Intel Xeon CPU E5-2620 v3 2.40GHz equipped with 64GB RAM.

Figure 2 visualizes the proposed tree-based model averaging approaches ET and EF, which use a data-adaptive weighting scheme depending on both model heterogeneity and subject features via tree-splitting. In the individual trees (a) and (d), the site indicator and X_1 appear as splitting variables, which is consistent with the data generation process. Our estimated treatment effect (b) and (e) recover the pattern of heterogeneity and homogeneity across sites and the range of X_1. Model 1 has a relatively larger contribution to the weighted estimator while models from other sites may have different contributions at different values of X_1. Figure 3 presents the box plots based on repeated experiments. Each series of boxes corresponds to a different strength of site-level heterogeneity c. Our proposed estimators ET and EF show the best performance in terms of the mean and variation of MSE among other estimators. Tree and forest have similar MSE due to the fact that the true model is relatively simple to be captured under the given sample size and hence can be already accurately approximated by a single ensemble tree. EF is preferable in practice for better stability. Also note that ET-cate and EF-cate achieve close-to-zero MSE with very small spreads. This is the case when the uncertainty in local model estimation is ignored. Results show that the uncertainty introduced by model averaging is small. Figure 4 shows the ratio of MSE in EF over MSE in LOC as a measure of accuracy gain resulting from model averaging, varying n $(100, 500, 1000)$. As sample size increases, model averaging becomes more powerful due to better estimation of τ_k, and is more pronounced under continuous grouping when c is small. A full comparison among estimators that utilize the ground truth, including EWMA and stacking with weights built on the true τ_k's, is provided in Appendix A.2. Therein, we also explore performance under various sample sizes in local sites as well as different estimating options for local models. Similar patterns are shown and our approach enjoys robust performance.



(a)

(b)

Figure 3: Box plots of the MSE of multiple CATE estimators for **(a) discrete grouping** and **(b) continuous grouping** across site, varying scale of site-level heterogeneity. The proposed methods ET and EF achieve competitive performance compared to standard model averaging or ensemble methods in all settings. Note that ET-cate and EF-cate achieve close-to-zero MSE with very small spreads in some settings.

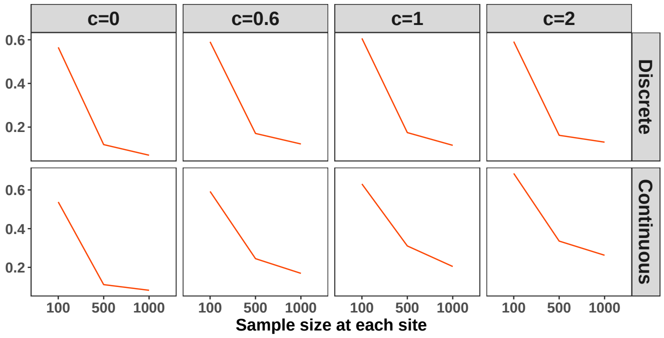
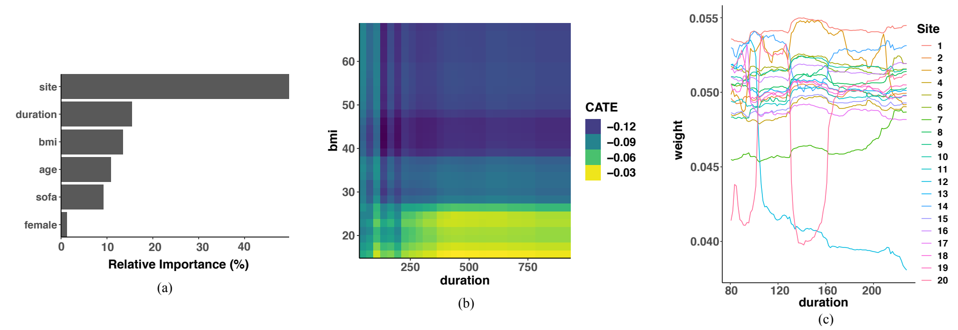


Figure 4: Ratio of MSE between EF and LOC.

5 Data Application: a Distributed Multi-Hospital Data Network

We apply the proposed EF-based model averaging estimator to evaluate the effect of oxygen therapy settings on in-hospital mortality among critically ill patients with respiratory diseases and with at least 48-hour of oxygen therapy. The data of this study are obtained from the eICU Collaborative Research Database (eICU-CRD), a multi-hospital database made available by Philips Healthcare [36]. A recent retrospective study found that the lowest mortality was observed when SpO_2 is in the range of 94-98% among patients requiring oxygen therapy [37]. We consider SpO_2 within this range as the treatment arm ($Z = 1$) and SpO_2 outside of this range as the control arm ($Z = 0$). The final analysis cohort consists of 7,022 patients from 20 hospitals, each with at least 50 patients in each treatment arm. The treatment effects of a randomly selected hospital (hospital site 1) is of interest, and we aim to adopt our proposed estimator to enhance the treatment effect estimation combining models of data from other hospitals. Hospital-level summary information is provided in Appendix A.3. We include the same five covariates as in [37], which are age, body mass index (BMI), sex, Sequential Organ Failure Assessment (SOFA) score, and duration of oxygen therapy. The outcome is the indicator of in-hospital mortality, i.e. $Y = 1$ is death in hospital and $Y = 0$ otherwise.



(a)

(b)

(c)

Figure 5: Application to estimating treatment effects of oxygen therapy on hospital mortality using EF. (a) Variable importance plot in the ensemble forest. The site indicator appears to be the most important variable with the relative importance taking up about 50%, followed by oxygen therapy duration and BMI. (b) Partial dependence plot of estimated treatment effects varying duration and BMI while holding the other covariates constant. (c) Visualization of data-adaptive weights in EF varying duration and site indicator, and varying BMI and site indicator, respectively. The weights of model 1 is stable while models from other sites may have different contributions to the weighted estimator for different values of duration.

Figure 5 visualizes the results of the EF estimated effect of oxygen therapy setting on in-hospital mortality. CT is used as the local model. Subfigure 5a shows the variable importance plot of the fitted EF. The site indicator appears to be most important, explaining about 50% of the decrease in training error. Subfigure 5b shows partial dependence plots of the estimated treatment effects as a function of the two other important features oxygen therapy duration and BMI adjusting for other covariates. A negative treatment effect corresponds to an improvement in survival. Patients of a BMI between 40 and 50, and an oxygen therapy duration about 230 show the most benefit from oxygen therapy at the SpO$_2$ 94-98% range with a lower hospital mortality. Patients with a BMI lower than 30 and a duration greater than 330 do not have a large differential treatment effect. Subfigure 5c visualizes our proposed model averaging scheme with data-adaptive weights $\omega_k(\boldsymbol{x})$ in the fitted EF with respect to oxygen therapy duration for different models, respectively, while holding other covariates constant. The weights of model 1 are quite stable while models from other sites may have different contribution to the weighted estimator for different values of duration. We also provide for comparison a fitted local model for hospital 1 using CT in Appendix A.3. It shares similar patterns as that in Figure 5b while the estimated treatment effect may differ.

In this distributed research network, different hospitals have a different sample size n_k. Those with a smaller sample size may not be representative of their population, leading to an uneven level of precision for local causal estimates. For sensitivity analysis, we consider a weighting strategy to adjust for the sample size of site k. Results show similar patterns as in Figure 5. Detailed results are provided in Appendix A.3. The data access and replication code are provided in Appendix A.4.

6 Discussion

In this paper, we have proposed an efficient and interpretable tree-based model averaging framework for enhancing the estimation of treatment effects for subgroups within a target site by borrowing information from data sources that are potentially heterogeneous. We generalize the standard model averaging scheme so that it is data-adaptive in a sense that the generated weights depend on baseline features. This work contributes to facilitating multi-site collaborations within a distributed research network by providing an analytical framework to leverage models from sites, avoiding the need to pool individual-participant data. Such kind of distributed data scheme is increasingly common in the era of Big Data and we hope to highlight the importance of accounting for data heterogeneity through this work. We break free of traditional thinking of viewing data integration as being a binary (yes/no) decision. Instead, the proposed approaches explore the similarity and dissimilarity of the target treatment effect between sites to yield an optimal information-sharing scheme for selectively improving the estimation of treatment effect of interest. We also stress that despite the function of interest in this paper being the CATEs, our approach extends beyond causal inference to a general $f(\boldsymbol{x})$ which may be of interest in other research problems where data are heterogeneous.

References

[1] Susan Athey and Guido Imbens. Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, 113(27):7353–7360, 2016.

[2] Stefan Wager and Susan Athey. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242, 2018.

[3] P Richard Hahn, Jared S Murray, Carlos M Carvalho, et al. Bayesian regression tree models for causal inference: Regularization, confounding, and heterogeneous effects (with discussion). *Bayesian Analysis*, 15(3):965–1056, 2020.

[4] Sören R Künzel, Jasjeet S Sekhon, Peter J Bickel, and Bin Yu. Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the National Academy of Sciences*, 116(10):4156–4165, 2019.

[5] Xinkun Nie and Stefan Wager. Quasi-oracle estimation of heterogeneous treatment effects. *arXiv preprint arXiv:1712.04912*, 2020.

[6] MW Reynolds, JB Christian, CD Mack, N Hall, and NA Dreyer. Leveraging real-world data for covid-19 research: challenges and opportunities. *J Precision Med*, 6:1–6, 2020.

[7] Jeffrey A Cohen, Maria Trojano, Ellen M Mowry, Bernard MJ Uitdehaag, Stephen C Reingold, and Ruth Ann Marrie. Leveraging real-world data to investigate multiple sclerosis disease behavior, prognosis, and treatment. *Multiple Sclerosis Journal*, 26(1):23–37, 2020.

[8] Marc L Berger, Craig Lipset, Alex Gutteridge, Kirsten Axelsen, Prasun Subedi, and David Madigan. Optimizing the leveraging of real-world data to improve the development and use of medicines. *Value in Health*, 18(1):127–130, 2015.

[9] Rachael L Fleurence, Lesley H Curtis, Robert M Califf, Richard Platt, Joe V Selby, and Jeffrey S Brown. Launching PCORnet, a national patient-centered clinical research network. *Journal of the American Medical Informatics Association*, 21(4):578–582, 2014.

[10] George Hripcsak, Jon D Duke, Nigam H Shah, Christian G Reich, Vojtech Huser, Martijn J Schuemie, Marc A Suchard, Rae Woong Park, Ian Chi Kei Wong, Peter R Rijnbeek, et al. Observational health data sciences and informatics (ohdsi): opportunities for observational researchers. *Studies in Health Technology and Informatics*, 216:574, 2015.

[11] Richard Platt, Jeffrey S Brown, Melissa Robb, Mark McClellan, Robert Ball, Michael D Nguyen, and Rachel E Sherman. The FDA sentinel initiative—an evolving national resource. *New England Journal of Medicine*, 379(22):2091–2093, 2018.

[12] Julie M Donohue, Marian P Jarlenski, Joo Yeon Kim, Lu Tang, Katherine Ahrens, Lindsay Allen, Anna Austin, Andrew J Barnes, Marguerite Burns, Chung-Chou H Chang, et al. Use of medications for treatment of opioid use disorder among us medicaid enrollees in 11 states, 2014-2018. *The Journal of the American Medical Association*, 326(2):154–164, 2021.

[13] Ding-Geng Chen, Dungang Liu, Xiaoyi Min, and Heping Zhang. Relative efficiency of using summary versus individual data in random-effects meta-analysis. *Biometrics*, 2020.

[14] Adrian E Raftery, David Madigan, and Jennifer A Hoeting. Bayesian model averaging for linear regression models. *Journal of the American Statistical Association*, 92(437):179–191, 1997.

[15] Yuhong Yang. Adaptive regression by mixing. *Journal of the American Statistical Association*, 96(454):574–588, 2001.

[16] Dong Dai and Tong Zhang. Greedy model averaging. *Advances in Neural Information Processing Systems*, 24:1242–1250, 2011.

[17] Dong Dai, Lei Han, Ting Yang, and Tong Zhang. Bayesian model averaging with exponentiated least squares loss. *IEEE Transactions on Information Theory*, 64(5):3331–3345, 2018.

[18] Yuri Breitbart, Peter L Olson, and Glenn R Thompson. Database integration in a distributed heterogeneous database system. In *1986 IEEE Second International Conference on Data Engineering*, pages 301–310. IEEE, 1986.

[19] Dan-Yu Lin and Daniel Zeng. On the relative efficiency of using summary statistics versus individual-level data in meta-analysis. *Biometrika*, 97(2):321–332, 2010.

[20] Michael Borenstein, Larry V Hedges, Julian PT Higgins, and Hannah R Rothstein. *Introduction to Meta-Analysis*. John Wiley & Sons, 2011.

[21] Jason D Lee, Qiang Liu, Yuekai Sun, and Jonathan E Taylor. Communication-efficient sparse regression. *The Journal of Machine Learning Research*, 18(1):115–144, 2017.

[22] Heather Battey, Jianqing Fan, Han Liu, Junwei Lu, and Ziwei Zhu. Distributed testing and estimation under sparse high dimensional models. *Annals of Statistics*, 46(3):1352–1382, 2018.

[23] Michael I Jordan, Jason D Lee, and Yun Yang. Communication-efficient distributed statistical inference. *Journal of the American Statistical Association*, 114(526):668–681, 2019.

[24] Lu Tang, Ling Zhou, and Peter X-K Song. Distributed simultaneous inference in generalized linear models via confidence distribution. *Journal of Multivariate Analysis*, 176:104567, 2020.

[25] Zheng Tracy Ke, Jianqing Fan, and Yichao Wu. Homogeneity pursuit. *Journal of the American Statistical Association*, 110(509):175–194, 2015.

[26] Shujie Ma and Jian Huang. A concave pairwise fusion approach to subgroup analysis. *Journal of the American Statistical Association*, 112(517):410–423, 2017.

[27] Xiwei Tang, Fei Xue, and Annie Qu. Individualized multidirectional variable selection. *Journal of the American Statistical Association*, pages 1–17, 2020.

[28] Elias Bareinboim and Judea Pearl. Causal inference and the data-fusion problem. *Proceedings of the National Academy of Sciences*, 113(27):7345–7352, 2016.

[29] Bénédicte Colnet, Imke Mayer, Guanhua Chen, Awa Dieng, Ruohong Li, Gaël Varoquaux, Jean-Philippe Vert, Julie Josse, and Shu Yang. Causal inference methods for combining randomized trials and observational studies: a review. *arXiv preprint arXiv:2011.08047*, 2020.

[30] Shu Yang, Donglin Zeng, and Xiaofei Wang. Elastic integrative analysis of randomized trial and real-world data for treatment heterogeneity estimation. *arXiv preprint arXiv:2005.10579*, 2020.

[31] Joanna Harton, Brian Segal, Ronac Mamtani, Nandita Mitra, and Rebecca Hubbard. Combining real-world and randomized control trial data using data-adaptive weighting via the on-trial score. *arXiv preprint arXiv:2108.08756*, 2021.

[32] Jerzy Neyman. Sur les applications de la thar des probabilities aux experiences agaricales: Essay des principle. Excerpts reprinted (1990) in English. *Statistical Science*, 5:463–472, 1923.

[33] Donald B Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688, 1974.

[34] Elizabeth A Stuart, Stephen R Cole, Catherine P Bradshaw, and Philip J Leaf. The use of propensity scores to assess the generalizability of results from randomized trials. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 174(2):369–386, 2011.

[35] Ashley L Buchanan, Michael G Hudgens, Stephen R Cole, Katie R Mollan, Paul E Sax, Eric S Daar, Adaora A Adimora, Joseph J Eron, and Michael J Mugavero. Generalizing evidence from randomized trials using inverse probability of sampling weights. *Journal of the Royal Statistical Society. Series A,(Statistics in Society)*, 181(4):1193, 2018.

[36] Tom J Pollard, Alistair EW Johnson, Jesse D Raffa, Leo A Celi, Roger G Mark, and Omar Badawi. The eICU collaborative research database, a freely available multi-center database for critical care research. *Scientific Data*, 5:180178, 2018.

[37] Willem van den Boom, Michael Hoy, Jagadish Sankaran, Mengru Liu, Haroun Chahed, Mengling Feng, and Kay Choong See. The search for optimal oxygen saturation targets in critically ill patients: observational data from large icu databases. *Chest*, 157(3):566–573, 2020.

A Appendix

A.1 provides the proof of the consistency of the proposed model averaging estimator stated in Section 3.5.

A.2 provides details on the simulation results in Section 4 varying sample sizes and local models.

A.3 provides sensitivity analysis for the real-data application in Section 5.

A.4 provides details on the access to the application data and the replication code.

A.1 Proof of Theorem 1

The proof of Theorem 1 closely follows arguments given in [2]. Suppose the subsamples for building each tree in an ensemble forest are drawn from different subjects in the augmented site 1 data. Specifically, in one round of EF, we draw m samples from the augmented data, where m is less than the rows in the augmented data, i.e., $m < (n_1 \cdot K)$. By randomly picking m unique subjects from site 1 and then randomly picking a site indicator k out of K sites for each of the m subjects. The resulted m subsamples should not be from the same subject and are hence independent and identically distributed. As long as $s < n_1$, we can ensure that all the subsamples are independent. In practice, when the ratio of n_1/K is relatively large, the probability of obtaining samples from the same subject is small.

Assume that subject features $\boldsymbol{X}_i$ and the site indicator S_i are independent and have a density that is bounded away from 0 and infinity. Suppose moreover that the conditional mean function $\mathbb{E}[\mathcal{T}|\boldsymbol{X} = \boldsymbol{x}, S = k]$ is Lipschitz continuous. We adopt the honesty definition in [1] when building trees in a random forest. Honest approaches separate the training sample into two halves, one half for building the tree model, and another half for estimating treatment effects within the leaves [1]. Following Definitions 1-5 and Theorem 3.1 in [2], the proposed estimator $\widehat{\mathcal{T}}_{\text{EF}}(\boldsymbol{x}, s)$ is a consistent estimator of the true treatment effect function $\tau_s(\boldsymbol{x})$ for any site s.

A.2 Full Simulation Results

Similar to ET-cate and EF-cate whose weights are built on the ground truth CATE functions τ_k's, we also consider for EWMA and STACK under a similar hypothetical setting. Specifically, we assume the true τ_1 is known and use it to compute the weights. This version of EWMA estimator is denoted as EWMA-cate and its weight is given by $\omega_k^{\text{EWMA-cate}} = \frac{\exp\{-\sum_{i\in\mathcal{I}_1^{(2)}}(\widehat{\tau}_k(\boldsymbol{x}_i)-\tau_1(\boldsymbol{x}_i))^2\}}{\sum_{\ell=1}^{K}\exp\{-\sum_{i\in\mathcal{I}_1^{(2)}}(\widehat{\tau}_\ell(\boldsymbol{x}_i)-\tau_1(\boldsymbol{x}_i))^2\}}$. Similarly, the corresponding linear stacking approach, denoted as STACK-cate, regresses the ground truth $\tau_1(\boldsymbol{x})$ on the predictions of the estimation set in site 1 from each local model, $\{\widehat{\tau}_1(\boldsymbol{x}), \ldots, \widehat{\tau}_k(\boldsymbol{x})\}$.

We compare the proposed model averaging estimators with the local estimator, MA, two versions of modified EWMA, as well as two versions of the linear stacking approach. Using CT as the local model, we present simulation results varying the sample size at local sites. The number of replications is 1000 throughout. Figure 6 presents the box plots based on 1000 simulated datasets. Each series of boxes corresponds to a different strength of site-level heterogeneity c. Table 1 reports the ratio between MSE of the estimator and MSE of the local model in terms of average and standard deviation of MSE, respectively, over 1000 replicates. Our proposed estimators ET and EF shows the best performance overall in terms of the mean and variation of MSE among the estimators without using the information of ground truth $\tau_1(\boldsymbol{x})$. Comparing with ET, EF has a slightly smaller MSE when c is large, which is expected because forest models tend to be more stable and accurate than a single tree. ET-cate achieves minimal MSE for low and moderate degrees of heterogeneity while EF-cate has the minimal MSE under all settings. The local estimator (LOC) in general shows the largest MSE compared to other estimators, as it does not leverage information from other sites. By borrowing information from additional sites, variances are greatly reduced, resulting in a small MSE of ensemble estimators. MA that naively adopts the inverse of sample size as weights performs well under low levels of heterogeneity, but suffers from a huge MSE with large variation as c increases. EWMA estimators perform slightly better and are more stable than LOC and MA. EWMA-cate has better performance than EWMA in all settings as the information of true CATE is used for weight construction. STACK estimators performs better than EWMA estimators. Similarly, STACK-cate performs better than STACK in all settings. STACK-cate, with ground truth $\tau_1(\boldsymbol{x})$ available, outperforms ET and EF when there exists a moderate to high level of heterogeneity across sites.

Figure 7 and Figure 8 show box plots of simulation results with a sample size of 100 and 1000, respectively, at each site. Our proposed methods ET and EF show robust performance in all settings. ET-cate and EF-cate achieve close-to-zero MSE with very small spreads in some settings. Figure 9 shows plots of the bias and MSE of EF-cate varying sample size at each site ($n = 100, 500, 1000$). As the sample size increases, both bias and MSE of EF-cate reduce to zero. Consistency of EF-cate can be shown via simulation when perfect estimates are obtained from local models. Meanwhile, our proposed method greatly reduce MSE by selectively borrowing information from multiple sites.

We explore another option for the local model using the causal forest (CF) [2] varying the sample size at local sites. A causal forest is a stochastic averaging of multiple causal trees [1], and hence is more powerful in estimating treatment effects. In each tree of the causal forest, MSE of treatment effect is used to select the feature and cutoff point in each split [2]. CF is implemented in the R packages `grf`. Figure 10, Figure 11, and Figure 12 show box plots of simulation results with a sample size of 100, 500, and 1000, respectively, at each site. Our proposed methods ET and EF show robust performance in all settings regardless of the use of information of the ground truth $\tau_1(\boldsymbol{x})$.

A.3 Additional Results for Data Application

Figure 13 plots a local CT fitted with data in hospital 1. Overall there is a similar pattern as in Figure 5b. Subjects with an oxygen therapy duration smaller than about 330 and a BMI smaller than about 30 do not have a large differential treatment effect.

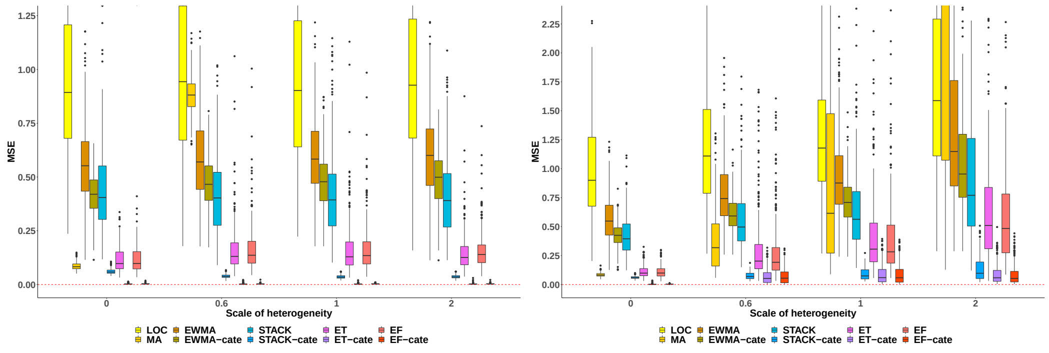
In real-life applications, hospitals may have different sample sizes n_k that may affect the accuracy of the estimation of τ_k. Table 2 shows hospital-level information for the 20 hospitals where the number of patients across sites varies. Information includes the region of the U.S. where the hospital is located, whether it is a teaching hospital, the bed capacity, and the number of patients within the hospital.

Hospitals with a smaller sample size may not be representative of the population, leading to an uneven level of precision for local causal estimates. To account for different sample sizes at each hospital, we consider a basic weighting strategy where we add weights to each observation $\widehat{\tau}_k(\boldsymbol{x})$ in the augmented site 1 data adjusting for the sample size of site k. The weights are defined as $\eta_k(\boldsymbol{x}) = \frac{K n_k}{\sum_{j=1}^{K} n_j}$.

Figure 14 visualizes the results of oxygen therapy on hospital mortality with the basic weighting strategy adopted. The site indicator appears to be most important with a relative importance taking up about 48%, indicating there may exist moderate to high degree of heterogeneity across sites. Two of other important variables are oxygen therapy duration and BMI, taking up about 16% and 13%, respectively. Subfigure 14b shows partial dependence plots of estimated treatment effects as a function of the two most important features oxygen therapy duration and BMI adjusting for other covariates. Similar patterns are observed as in Subfigure 5b where patients of a BMI between 40 and 50, and an oxygen therapy duration about 230 shows a benefit from the oxygen therapy at the SpO$_2$ 94-98% range with a lower hospital mortality. Patients with a BMI lower than 26 and a duration greater than 350 did not benefit much from the oxygen therapy. Subfigure 14c visualizes our proposed model averaging scheme depending on features with data-adaptive weights $\omega_k(\boldsymbol{x})$ in the fitted EF for oxygen therapy duration for different models, respectively, while holding other covariates constant. The patterns are similar as in Subfigure 5c. The weights of model 1 is stable while models from other sites may have different contribution to the weighted estimator for different values of duration. Overall, the patterns in each plot are similar to Figure 5, which indicates the robustness of our proposed estimators. We do stress that improvements to the weighting strategy for different sample sizes at each site are needed. A strategy considering both treatment proportion as well as covariate distributions across sites may further enhance the data-adaptive model averaging estimator.

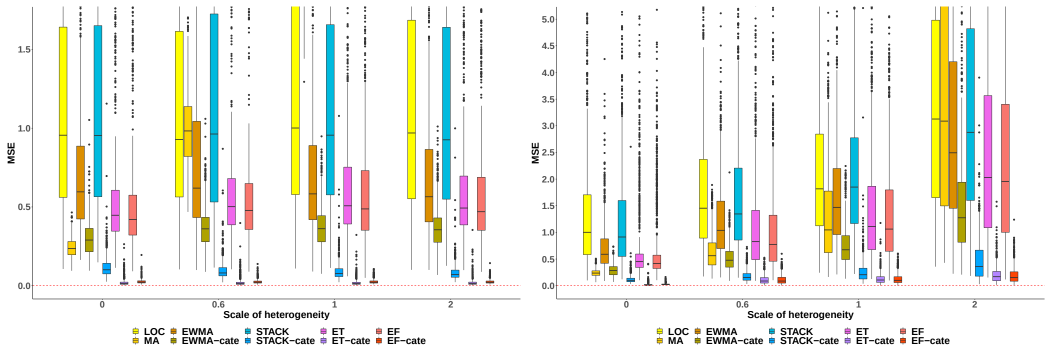
A.4 R Package, Real data, and Replication Code

The proposed method has been implemented in an R package `ifedtree` and has been made available on Github (`github.com/ellenxtan/ifedtree`). Although the eICU-CRD data used in our application example cannot be shared subject to the data use agreement, access can be individually requested at `https://eicu-crd.mit.edu/gettingstarted/access/`. We also provide the R code used to replicate the simulation and data application results at `github.com/ellenxtan/ifedtree`.



(a) (b)

Figure 6: Box plots of the MSE of multiple CATE estimators with **CT** as the local model and a sample size of **500** at each site for **(a) discrete grouping** and **(b) continuous grouping** across site, respectively, varying scale of site-level heterogeneity. Estimators ending with “-cate" makes use of ground truth treatment effects. The proposed methods ET and EF achieve competitive performance compared to standard model averaging or ensemble methods in all settings.

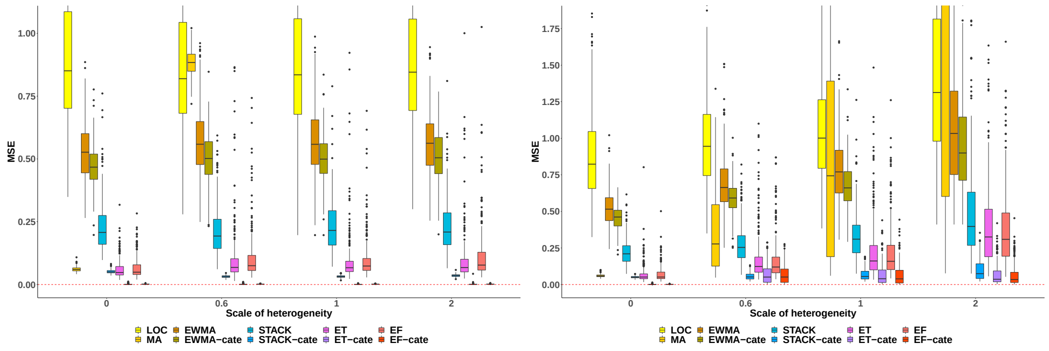


(a) (b)

Figure 7: Box plots of the MSE of multiple CATE estimators with **CT** as the local model and a sample size of **100** at each site for **(a) discrete grouping** and **(b) continuous grouping** across site, respectively, varying scale of site-level heterogeneity. Estimators ending with “-cate" makes use of ground truth treatment effects. The proposed methods ET and EF achieve competitive performance in all settings.

Table 1: Simulation results for ratio between MSE of the estimator and MSE of **CT** (local model) with a sample size of **500** at each site. A smaller number indicates larger improvement over the local model. Estimators ending with "-cate" makes use of ground truth treatment effects. Our proposed methods ET and EF shows robust performance in all settings whether or not using the information of ground truth $\tau_1(\boldsymbol{x})$.

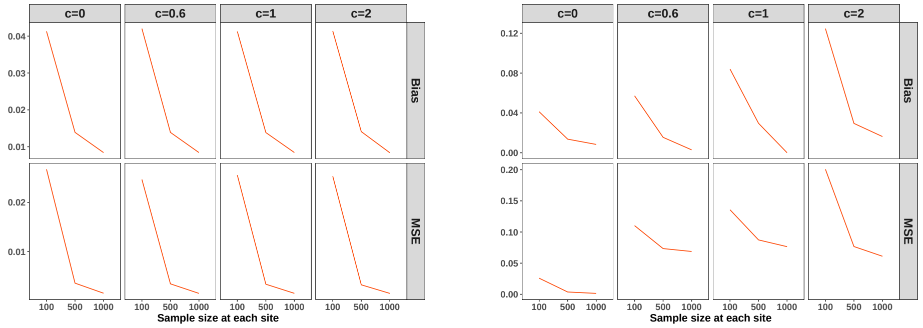
	Discrete grouping				Continuous grouping			
Estimator	$c = 0$	$c = 0.2$	$c = 0.6$	$c = 1$	$c = 0$	$c = 0.2$	$c = 0.6$	$c = 1$
Ratio of average of MSEs over 1000 replicates								
MA	0.09	0.91	2.4	9.87	0.08	0.32	0.65	1.78
EWMA	0.57	0.62	0.61	0.62	0.56	0.65	0.7	0.77
EWMA-cate	0.42	0.5	0.49	0.5	0.42	0.49	0.53	0.59
STACK	0.44	0.45	0.44	0.45	0.45	0.45	0.48	0.54
STACK-cate	0.06	0.04	0.04	0.04	0.06	0.06	0.06	0.07
ET	0.12	0.17	0.16	0.16	0.13	0.24	0.29	0.37
ET-cate	<0.01	<0.01	<0.01	<0.01	<0.01	0.08	0.1	0.07
EF	0.1	0.13	0.13	0.13	0.1	0.19	0.25	0.3
EF-cate	<0.01	<0.01	<0.01	<0.01	<0.01	0.06	0.06	0.05
Ratio of standard deviation of MSEs over 1000 replicates								
MA	0.15	0.35	0.76	3.05	0.14	0.24	0.38	0.81
EWMA	0.61	0.65	0.67	0.66	0.58	0.65	0.69	0.75
EWMA-cate	0.46	0.52	0.54	0.54	0.44	0.52	0.55	0.6
STACK	0.47	0.46	0.47	0.47	0.45	0.49	0.52	0.6
STACK-cate	0.1	0.08	0.08	0.08	0.09	0.11	0.12	0.14
ET	0.18	0.23	0.22	0.22	0.18	0.26	0.32	0.43
ET-cate	0.02	0.03	0.02	0.02	0.02	0.06	0.07	0.07
EF	0.17	0.19	0.19	0.2	0.17	0.23	0.29	0.39
EF-cate	0.03	0.03	0.03	0.03	0.03	0.06	0.07	0.08



(a)

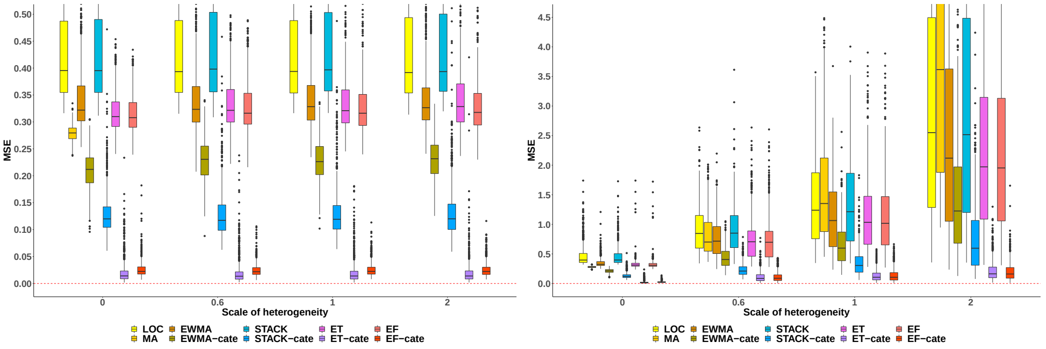
(b)

Figure 8: Box plots of the MSE of multiple CATE estimators with **CT** as the local model and a sample size of **1000** at each site for **(a) discrete grouping** and **(b) continuous grouping** across site, respectively, varying scale of site-level heterogeneity. Estimators ending with "-cate" makes use of ground truth treatment effects. The proposed methods ET and EF achieve competitive performance in all settings.



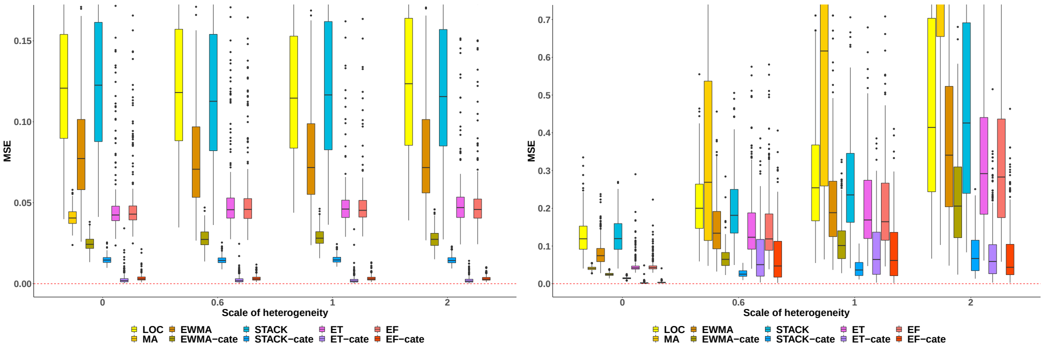
(a) (b)

Figure 9: Plots of the bias and MSE of **EF-cate** varying sample site at each site for **(a) discrete grouping** and **(b) continuous grouping** across site, varying scale of site-level heterogeneity. Both bias and MSE reduces to zero as the sample size increases.



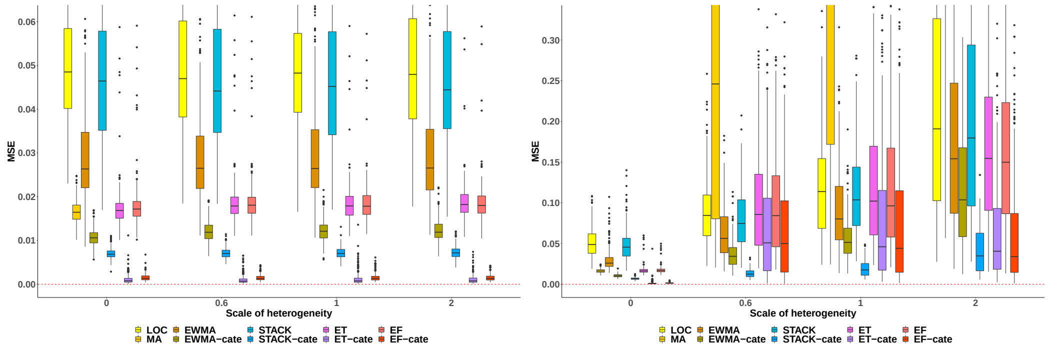
(a) (b)

Figure 10: Box plots of the MSE of multiple CATE estimators with **CF** as the local model and a sample size of **100** at each site for **(a) discrete grouping** and **(b) continuous grouping** across site, respectively, varying scale of site-level heterogeneity. Estimators ending with "-cate" makes use of ground truth treatment effects. The proposed methods ET and EF achieve competitive performance in all settings.



(a) (b)

Figure 11: Box plots of the MSE of multiple CATE estimators with **CF** as the local model and a sample size of **500** at each site for **(a) discrete grouping** and **(b) continuous grouping** across site, respectively, varying scale of site-level heterogeneity. Estimators ending with "-cate" makes use of ground truth treatment effects. The proposed methods ET and EF achieve competitive performance in all settings.



(a) (b)

Figure 12: Box plots of the MSE of multiple CATE estimators with **CF** as the local model and a sample size of **1000** at each site for **(a) discrete grouping** and **(b) continuous grouping** across site, respectively, varying scale of site-level heterogeneity. Estimators ending with “-cate" makes use of ground truth treatment effects. The proposed methods ET and EF achieve competitive performance in all settings.

Table 2: Hospital-level information of our analysis cohort in eICU-CRD database.

Hospital site	Number of patients	Number of control	Number of treated	Bed capacity	Teaching status	Region
1	477	205	272	$\geq$ 500	False	South
2	388	94	294	$\geq$ 500	False	Midwest
3	464	129	335	$\geq$ 500	True	South
4	523	162	361	$\geq$ 500	False	South
5	149	71	78	250 - 499	False	South
6	305	174	131	$\geq$ 500	False	South
7	297	109	188	$\geq$ 500	True	West
8	210	78	132	Unknown	False	Unknown
9	183	52	131	250 - 499	False	West
10	379	161	218	$\geq$ 500	True	Midwest
11	659	165	494	$\geq$ 500	True	Midwest
12	200	55	145	250 - 499	False	South
13	166	64	102	100 - 249	False	Midwest
14	222	58	164	250 - 499	False	South
15	163	58	105	$\geq$ 500	True	Midwest
16	747	185	562	$\geq$ 500	True	Northeast
17	435	240	195	$\geq$ 500	True	South
18	234	70	164	$\geq$ 500	True	Midwest
19	474	229	245	$\geq$ 500	False	South
20	347	109	238	$\geq$ 500	True	Midwest

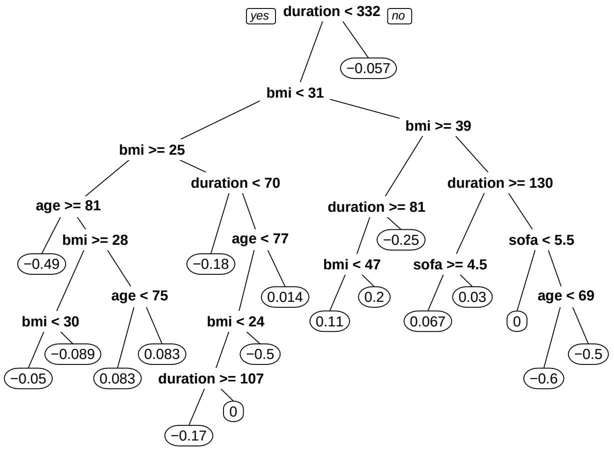
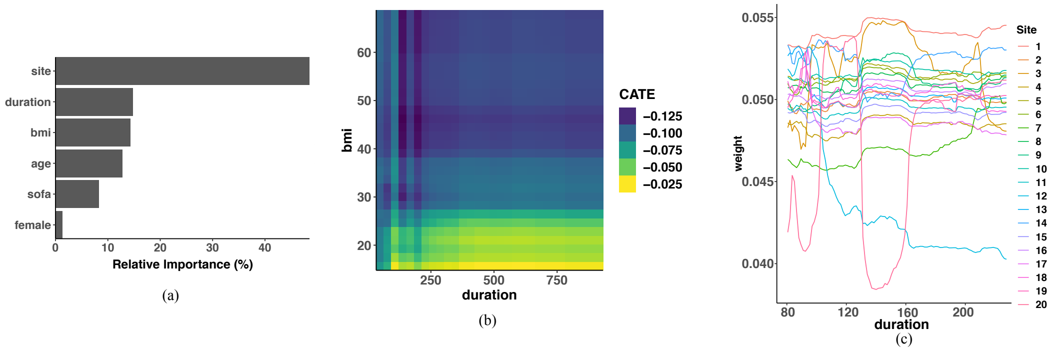


Figure 13: A local CT fitted with data in hospital 1 in the real-data application to estimating treatment effects of oxygen therapy on hospital mortality.



(a)

(b)

(c)

Figure 14: Application to estimating treatment effects of oxygen therapy on hospital mortality with a sample size weighting strategy. (a) Variable importance plot in the ensemble forest. The site indicator appears to be the most important variable with the relative importance taking up about 48%, followed by oxygen therapy duration and BMI. (b) Partial dependence plot of estimated treatment effects varying duration and BMI while holding the other covariates constant. (c) Visualization of data-adaptive weights in EF varying duration and site indicator, and varying BMI and site indicator, respectively.