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

Xiaoqing Tan [1]   Chung-Chou H. Chang [1]   Ling Zhou [2]   Lu Tang [1]

Abstract

Accurately estimating personalized treatment effects within a study site (e.g., a hospital) has been challenging due to limited sample size. Furthermore, privacy considerations and lack of resources prevent a site from leveraging subject-level data from other sites. We propose a tree-based model averaging approach to improve the estimation accuracy of conditional average treatment effects (CATE) at a target site by leveraging models derived from other potentially heterogeneous sites, without them sharing subject-level 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. Specifically, under distributed data networks, our framework provides an interpretable tree-based ensemble of CATE estimators that joins models across study sites, while actively modeling the heterogeneity in data sources through site partitioning. The performance of this approach is demonstrated by a real-world study of the causal effects of oxygen therapy on hospital survival rate and backed up by comprehensive simulation results.

1. Introduction

Estimating individualized treatment effects has been a hot topic because of its wide applications, ranging from personalized medicine, policy research, to customized marketing advertisement. 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 (Athey & Imbens, 2016; Wager & Athey, 2018; Hahn et al., 2020; Künzel et al., 2019; Nie & Wager, 2021). 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 (Reynolds et al., 2020; Cohen et al., 2020; Berger et al., 2015). Distributed research networks have been established in many large scale studies (Fleurence et al., 2014; Hripcsak et al., 2015; Platt et al., 2018; Donohue et al., 2021). 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 statistical precision. This concern is mostly noticeable in estimating treatment effects where sample size requirement is high yet observations are typically limited. Furthermore, information exchange between data sites is often highly restricted due to privacy, feasibility, or other concerns, prohibiting centralized analyses that pool data from multiple sources (Maro et al., 2009; Brown et al., 2010; Toh et al., 2011; Raghupathi & Raghupathi, 2014; DeShazo & Hoffman, 2015; Donahue et al., 2018; Dayan et al., 2021). One way to tackle this challenge is through model averaging (Raftery et al., 1997), where multiple research sites collectively contribute to the tasks of statistical modeling without sharing sensitive subject-level data. Although this idea has existed in supervised learning problems (Dai & Zhang, 2011; McMahan et al., 2017), to our best knowledge, there are no established model averaging approach and theoretical results on estimating CATE in a distributed environment. The extension is non-trivial because CATE is unobserved in nature, as opposed to prediction problems where labels are given.

This paper focuses on improving the prediction accuracy of CATE concerning a target site by leveraging models derived from other sites where *transportability* (to be formally defined in Section 3.1, Pearl & Bareinboim, 2011; Stuart et al., 2011; Pearl & Bareinboim, 2014; Bareinboim & Pearl, 2016; Buchanan et al., 2018; Dahabreh et al., 2019) may not hold. Specifically, there may exist heterogeneity in treatment effects. In the context of our multi-hospital example,

---

[1]University of Pittsburgh, Pittsburgh, PA, USA [2]Southwestern University of Finance and Economics, Chengdu, China. Correspondence to: Lu Tang <lutang@pitt.edu>.

these are: 1) **local heterogeneity**: within a hospital, patients with different characteristics may have different treatment effects. This is the traditional notion of CATE; and 2) **global heterogeneity**: where the same patient may experience different treatment effects at different hospitals. The second type of heterogeneity is driven by site-level confounding, and hampers the transportability of models across hospital sites. We also note that these two types of heterogeneity may interact with each other in the sense that transportability is dependent on patient characteristics, which we will address.

We propose a model averaging framework that uses a flexible tree-based weighting scheme to combine learned models from sites that takes into account heterogeneity. The contribution of each learned model to the target site depends on subject characteristics. This is achieved by applying tree splittings (Breiman et al., 1984) at both the site and the subject levels. For example, effects of a treatment in two hospitals may be similar for female patients but not for male, suggesting us to consider borrowing information across sites only on selective subgroups. Our approach extends the classic model averaging framework (Raftery et al., 1997; Wasserman, 2000; Hansen, 2007; Yang, 2001) by allowing data-adaptive weights, which are interpretable in a sense that they can be used to lend credibility to transportability. For example, in the case of extreme heterogeneity where other sites merely contribute to the target, the weights can be used as a diagnostic tool to inform the decision against borrowing information.

Main contributions. 1) We propose a model averaging scheme with interpretable weights that are adaptive to both local and global heterogeneity via tree-splitting dedicated to improving CATE estimation under distributed data networks. 2) We generalize model averaging techniques to study the transportability of causal inference. Causal assumptions with practical implications are explored to warrant the use of our approach. 3) We provide an extensive empirical evaluation of the proposed approach with a concrete real-data example on how to apply the method in practice. 4) Compared to other distributed learning methods, the proposed framework enables causal analysis without sharing subject-level data, is easy to implement, offers ease of operations, and minimizes infrastructure, which facilitates practical collaboration within research networks.

2. Related Work

There are two types of construct of a distributed database (Breitbart et al., 1986): *homogeneous* versus *heterogeneous*. For homogeneous data sources, data across sites are random samples of the global population. Recent modeling approaches (Lin & Zeng, 2010; Lee et al., 2017; McMahan et al., 2017; Battey et al., 2018; Jordan et al., 2019; Tang et al., 2020; Wang et al., 2021) all assume samples are

randomly partitioned, which guarantees identical data distribution across sites. The goal of these works is to improve overall prediction by averaging results from homogeneous sample divisions. The classic random effects meta-analysis (see, e.g., Whitehead (2002); Sutton et al. (2000); Borenstein et al. (2011) describes heterogeneity using modeling assumptions, but its focus mostly is still on global patterns.

Heterogeneous models. In practice, however, there is often too much global 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 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 (Ke et al., 2015; Tang & Song, 2016; Smith et al., 2017; Ma & Huang, 2017; Wang et al., 2020; Tang et al., 2021; Tang & Song, 2021). Most of these require the pooling of subject-level data. The second class of approaches falls in the *model averaging* framework (Raftery et al., 1997) 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 $f^*(\boldsymbol{x}) = \sum_{k=1}^{K} \omega_k f_k(\boldsymbol{x})$ with weights ω_k to balance the contribution of each model and $\sum_k \omega_k = 1$. It provides an immediate interpretation of usefulness of each data source. When the weights are proportional to the prediction performance of f_k on site 1, for example,

$$\omega_k = \frac{\exp\{-\sum_{i \in \mathcal{I}_1}(f_k(\boldsymbol{x}_i) - y_i)^2\}}{\sum_{\ell=1}^{K} \exp\{-\sum_{i \in \mathcal{I}_1}(f_\ell(\boldsymbol{x}_i) - y_i)^2\}},$$

with y_i being the observed outcome of subject i in site 1, indexed by $\mathcal{I}_1$, the method is termed as the exponential weighted model averaging (EWMA). Several variations of ω_k can be found in Yang (2001); Dai & Zhang (2011); Yao et al. (2018); Dai et al. (2018). In general, separate samples are used to obtain the estimates of ω_k's and f_k's, respectively.

Here we focus on the literature review of model averaging. We note that our framework is also related to federated learning (McMahan et al., 2017). But the latter often involves iterative updating rather than a one-shot procedure, and could be hard to apply to nonautomated distributed research networks. Besides, it has been developed mainly to estimate a global prediction model by leveraging distributed data, and is not designed to target any specific site. We further discuss these approaches and other related research topics and their distinctions with model averaging in Appendix A.

Transportability. In causal inference, there is a lot of inter-

est in identifying subgroups with enhanced treatment effects, targeting at the feasibility of customizing estimates for individuals (Athey & Imbens, 2016; Wager & Athey, 2018; Hahn et al., 2020; Künzel et al., 2019; Nie & Wager, 2021). 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 sought after. Pearl & Bareinboim (2011; 2014); Bareinboim & Pearl (2016) introduced the notion of transportability to warrant causal inference models be generalized to a new population. The issue of generalizability is common in practice due to the non-representative sampling of participants in randomized controlled trials (Cook et al., 2002; Druckman et al., 2011; Allcott, 2015; Stuart et al., 2015; Egami & Hartman, 2020). Progress on bridging the findings from an experimental study with observational data can be found in, e.g., Stuart et al. (2015); Kern et al. (2016); Stuart et al. (2018); Ackerman et al. (2019); Yang et al. (2020); Harton et al. (2021). See Tipton & Olsen (2018); Colnet et al. (2020); Degtiar & Rose (2021) and references therein for a comprehensive review. However, most methods require fully centralized data. In contrast, we leverage the distributed nature of model averaging to derive an integrative CATE estimator.

3. A Tree-based Model Averaging Framework

We first formally define the conditional average treatment effect (CATE). 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 (Neyman, 1923; Rubin, 1974), 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 potential outcomes under treatment arms $Z = 1$ and $Z = 0$, respectively. The expected difference of the potential outcomes is dependent on subject features $\boldsymbol{X}$. By the causal consistency assumption, the observed outcome is $Y = ZY^{(Z=1)} + (1 - Z)Y^{(Z=0)}$.

Now suppose the distributed data network consists of K sites, each with sample size of n_k. Site k contains data $\mathcal{D}_k = \{y_i, z_i, \boldsymbol{x}_i\}_{i \in \mathcal{I}_k}$, where $\mathcal{I}_k$ denotes its index set. Its CATE function is 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.

3.1. Causal Assumptions

To ensure information can be properly borrowed across sites, we first impose the following idealistic assumptions, and then present relaxed version of Assumption 2. Let S be the site indicator taking values in $\mathcal{S} = \{1, \ldots, K\}$ such that $S_i = k$ if $i \in \mathcal{I}_k$.

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

Assumption 2 (Transportability):

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

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

Assumption 1 ensures treatment effects are unconfounded within sites so that $\tau_k(\boldsymbol{x})$ can be consistently identified. It holds by design when data are randomized controlled trials or when treatment assignment depends on $\boldsymbol{X}$. By this assumption, we have $\tau_k(\boldsymbol{x}) = E[Y|\boldsymbol{X} = \boldsymbol{x}, S = k, Z = 1] - E[Y|\boldsymbol{X} = \boldsymbol{x}, S = k, Z = 0]$. The 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 Stuart et al. (2011), Buchanan et al. (2018) and Yang et al. (2020) for similar consideration. This assumption may not be satisfied due to 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. Explicitly, we consider a relaxed Assumption 2a to hold for a subset of sites that contains site 1.

Assumption 2a (Partial Transportability):

$$\{Y^{(Z=0)}, Y^{(Z=1)}\} \perp S_1|\boldsymbol{X}.$$

Here, 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 specific subjects. In a special case in Section 4 where $\mathcal{S}_1 = \{1\}$, bias may be introduced to by model averaging. However, our approach is still able to exploits the bias and variance trade off to improve estimation. 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. The former ensures a balance of covariates between site 1 population and the population of other sites. Violation of either one may result in extrapolation and introduce unwanted bias to the ensemble estimates for site 1. This assumption is also used, e.g., in Stuart et al. (2011).

3.2. Model Ensemble

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

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

where τ^* is the weighted model averaging estimator. The weight functions $\omega_k(\boldsymbol{x})$'s are not only site-specific, but also
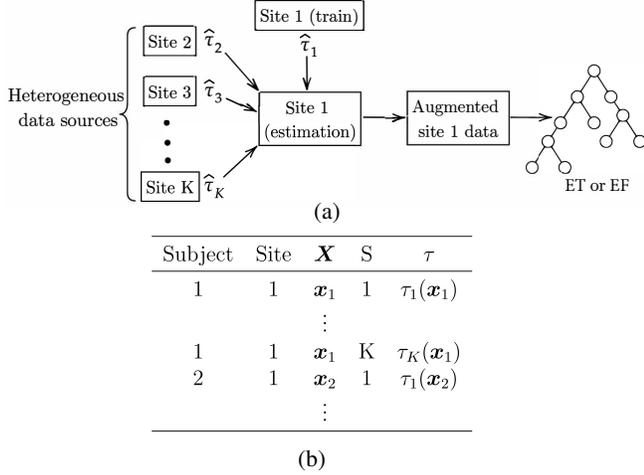
Figure 1. (a) Schema of the proposed algorithm. (b) Illustration of the augmented data constructed from the estimation set of site 1.

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 $\boldsymbol{x}$ are 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. We will describe the approaches to estimate $\widehat{\tau}_k$ in Section 3.3.

A two-stage model averaging approach is proposed. We first split $\mathcal{D}_1$, the data in the target site, into a training set and an estimation set indexed by $\{i \in \mathcal{I}_1^{(1)}\}$ and $\{i \in \mathcal{I}_1^{(2)}\}$, respectively. *1) Local stage:* Obtain $\widehat{\tau}_1$ from subjects in $\mathcal{I}_1^{(1)}$. Obtain $\widehat{\tau}_k$ from local subjects in $\mathcal{I}_k$, $k = 2, \ldots, K$. These $\{\widehat{\tau}_k\}_{k=1}^{K}$ are then passed to site 1 to get K predicted treatment effects for each subject in $\mathcal{I}_1^{(2)}$, resulting in an augmented data set as shown in Figure 1(b). *2) Ensemble stage:* A tree-based ensemble model is trained on the augmented data by either an ensemble tree (ET) or an ensemble random forest (EF), with the predicted treatment effects from the previous stage, i.e., $\widehat{\tau}_k(\boldsymbol{x}_i)$ as the *outcome*. The site indicator S of which local model is used as well as the subject features $\boldsymbol{x}_i$ are fed into the ensemble model as *predictors*. The resulting model will be used to compute our proposed model averaging estimator. Figure 1(a) illustrates a conceptual diagram of the proposed model averaging framework and structure of the augmented data. Note the idea of data augmentation has been used in, e.g., computer vision (Perez & Wang, 2017), statistical computing (van Dyk & Meng, 2001), and imbalanced classification (Chawla et al., 2002). Here the technique is being used to construct weights for model averaging, which will be discussed in the following paragraph. Algorithm 1 provides an algorithmic overview. Our method has been implemented as an R package ifedtree available on GitHub (https://github.com/ellenxtan/ifedtree).

Algorithm 1 Tree-based model averaging for heterogeneous data sources

---

for $k = 1$ **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.
end for
for $i \in \mathcal{I}_1^{(2)}$ **do**
        {Loop through subjects in site 1 estimation set.}
  for $k = 1$ **to** K **do**
           {Loop through K local models.}
    Predict $\widehat{\tau}_k(\boldsymbol{x}_i)$ using local model k.
    $D_{i,k} = [\boldsymbol{x}_i, k, \widehat{\tau}_k(\boldsymbol{x}_i)]$.
  end for
end for
Create augmented site 1 data $\mathfrak{D}_{aug,1}$ by concatenating $D_{i,k}$ vectors.
$\widehat{\mathcal{T}}_{\text{EF}}(\boldsymbol{x}, s) = \text{ENSEMBLEFOREST}(\mathfrak{D}_{aug,1})$
        {Or ENSEMBLETREE when $B = 1$.}

Construction of weights. 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 $\mathfrak{D}_{aug,1} = \{\boldsymbol{x}_i, k, \widehat{\tau}_k(\boldsymbol{x}_i)\}_{i \in \mathcal{I}_1^{(2)}, k \in \mathcal{S}}$, for subjects in $\mathcal{I}_1^{(2)}$. The illustration of this augmented site 1 data is given in Figure 1(b). An ensemble is then trained on this data by either a 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 subject-level 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

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

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

$$
\begin{aligned}
\widehat{\mathcal{T}}_{\mathrm{EF}}(\boldsymbol{x}, s) &= \frac{1}{B} \sum_{b=1}^B \widehat{\mathcal{T}}^{(b)}(\boldsymbol{x}, s) \\
&= \sum_{i \in \mathcal{I}_1^{(2)}} \sum_{k=1}^K \lambda_{i,k}(\boldsymbol{x}, s) \widehat{\tau}_k(\boldsymbol{x}_i), \quad (3)
\end{aligned}
$$

$$
\text{where } \lambda_{i,k}(\boldsymbol{x}, s) = \frac{1}{B} \sum_{b=1}^B \frac{\mathbb{1}\{(\boldsymbol{x}_i, k) \in \mathcal{L}_b(\boldsymbol{x}, s)\}}{|\mathcal{L}_b(\boldsymbol{x}, s)|}.
$$

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. We then obtain the model averaging estimates defined in (1) by fixing $s = 1$ such that $\widehat{\tau}_{\mathrm{ET}}^*(\boldsymbol{x}) = \widehat{\mathcal{T}}_{\mathrm{ET}}(\boldsymbol{x}, s = 1)$ or $\widehat{\tau}_{\mathrm{EF}}^*(\boldsymbol{x}) = \widehat{\mathcal{T}}_{\mathrm{EF}}(\boldsymbol{x}, s = 1)$. 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

$$
\widehat{\tau}_{\mathrm{ET}}^*(\boldsymbol{x}) = \widehat{\mathcal{T}}_{\mathrm{ET}}(\boldsymbol{x}, 1) = \sum_{k=1}^K \widehat{\omega}_k(\boldsymbol{x}) \widehat{\tau}_k(\boldsymbol{x}),
$$

$$
\text{where } \widehat{\omega}_k(\boldsymbol{x}) = \sum_{i \in \mathcal{I}_1^{(2)}} \frac{\mathbb{1}\{(\boldsymbol{x}_i, k) \in \mathcal{L}(\boldsymbol{x}, 1)\}}{|\mathcal{L}(\boldsymbol{x}, 1)|};
$$

$$
\widehat{\tau}_{\mathrm{EF}}^*(\boldsymbol{x}) = \widehat{\mathcal{T}}_{\mathrm{EF}}(\boldsymbol{x}, 1) = \sum_{k=1}^K \widehat{\omega}_k(\boldsymbol{x}) \widehat{\tau}_k(\boldsymbol{x}),
$$

$$
\text{where } \widehat{\omega}_k(\boldsymbol{x}) = \sum_{i \in \mathcal{I}_1^{(2)}} \lambda_{i,k}(\boldsymbol{x}, 1).
$$

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 $B = 2,000$ throughout the paper. Tree and forest estimates are obtained by R packages `rpart` and `grf`, respectively.

Interpretability of weights. The choice of tree-based models naturally results in such kernel weighting $w_k(\boldsymbol{x})$ (Athey et al., 2019), which are not accessible by other ensemble techniques. Such explicit and interpretable weight functions could deliver meaningful rationales for data integration. For example, under scenarios where there exists extreme global heterogeneity (as shown in Section 4 when c is large), $w_k(\boldsymbol{x})$ can be used as a diagnostic tool to decide which external data sources should be co-used. Weights close to 0 inform against model transportability, and they are adaptive to subject-level features $\boldsymbol{x}$ so that decisions can be made based on the subpopulations of interest.

3.3. 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 (Athey & Imbens, 2016; Wager & Athey, 2018; Hahn et al., 2020; Künzel et al., 2019; Nie & Wager, 2021). As an example, we consider using the causal tree (CT) (Athey & Imbens, 2016) to estimate the local model at each site. CT is a non-linear learner that *(i)* allows different types of outcome such as discrete and continuous, and can be applied to a broad range of real data scenarios; *(ii)* can manage hundreds of features and high order interactions by construction; *(iii)* can be applied to both experimental studies and observational studies by propensity score weighting or doubly robust methods. CT is implemented in the R package `causalTree`. We also explore another estimating option for local models in Appendix C.

3.4. Asymptotic Properties

We provide consistency guarantee of the proposed estimator $\widehat{\mathcal{T}}_{\mathrm{EF}}$ for the true target τ_1. Assuming point-wise consistent local estimators are used for $\{\tau_k\}_{k=1}^K$, EF with subsampling procedure described in Appendix B is consistent.

Theorem 3.1. *Suppose the subsample used to build each tree in an ensemble forest is drawn from different subjects of the augmented data and the following conditions hold:*

(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 $\mathbb{E}[\mathcal{T}|\boldsymbol{X} = \boldsymbol{x}, S = 1]$ is Lipschitz-continuous.

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

Then $\widehat{\mathcal{T}}_{EF}(\boldsymbol{x}, 1) \xrightarrow{p} \tau_1(\boldsymbol{x})$, for all $\boldsymbol{x}$, as $\min_k n_k \to \infty$. Hence, $\widehat{\tau}_{EF}^*(\boldsymbol{x}) \xrightarrow{p} \tau_1(\boldsymbol{x})$.

The conditions and a proof of Theorem 3.1 is given in Appendix B. To demonstrate the consistency properties of our methods, we add in Appendix C oracle versions of ET and

EF estimators, denoted as ET-oracle and EF-oracle, which use the ground truth of local models $\{\tau_k\}_{k=1}^K$ in estimating $\{\widehat{\omega}_k\}_{k=1}^K$. This removes the uncertainty in local models. The remaining uncertainty only results from the estimation of the ensemble weights, and we see both oracle estimators achieve minimal MSE. Section 4 gives a detailed evaluation of the finite sample performance.

4. Simulation Study

Monte Carlo simulations are conducted to assess the proposed methods. We specify $m(\boldsymbol{x}, k)$ as the conditional outcome surface and $\tau(\boldsymbol{x}, k)$ as the conditional treatment effect for individuals with features $\boldsymbol{x}$ in site k. The treatment propensity is specified as $e(\boldsymbol{x}) = \Pr(Z = 1 | \boldsymbol{X} = \boldsymbol{x})$. The potential outcomes can be written as $Y_i = m(\boldsymbol{X}_i, S_i) + \{Z_i - e(\boldsymbol{X}_i)\}\tau(\boldsymbol{X}_i, S_i) + \epsilon_i$, following notations in Robinson (1988); Athey & Imbens (2016); Wager & Athey (2018); Nie & Wager (2021). 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 treatment effect 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 global heterogeneity due to site-level confounding, controlled by a scaling factor c, and $\epsilon_i \sim N(0, 1)$. Features follow $\boldsymbol{X}_i \sim N(\boldsymbol{0}, \boldsymbol{I}_D)$, where $D = 5$, and are independent of ϵ_i. The simulation setting within each site (with k fixed) is motivated by designs in Athey & Imbens (2016). Features in τ are determinants of treatment effect while those in m but not in τ are prognostic only. The data are generated under a distributed data networks. We assume there are $K = 20$ sites in total, each with a sample size $n = 500$. In our main exposition, we consider an experimental study design where treatment propensity is $e(\boldsymbol{x}) = 0.5$, i.e., individuals are randomly assigned to treatment and control. Variations of the settings above are discussed, with results presented in Appendix C.

Global heterogeneity: discrete and continuous. Two types for global heterogeneity are considered by the choice of U_k. 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 $\mathcal{G}_1 = \{1, 3, \ldots, K - 1\}$; $\mathcal{G}_2 = \{2, 4 \ldots, K\}$ such that $U_{k \in \mathcal{G}_1} = 0$ and $U_{k \in \mathcal{G}_2} = 1$. 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. For ***continuous grouping***, we consider $U_k \sim Unif[0, 1]$. We vary the scales of the global heterogeneity under the discrete and continuous cases, respectively, with c taking values $c \in \{0, 0.6, 1, 2\}$. A $c = 0$ implies all data sources are homogeneous. In other words, Assumption 2 is satisfied when $c = 0$ but not when $c > 0$.

Table 1. MSE ratios of EF over LOC. As n increases, model averaging becomes more powerful due to better estimation of τ_k, and is more pronounced when c is small.

		$c = 0$	$c = 0.6$	$c = 1$	$c = 2$
Discrete grouping	$n = 100$	0.57	0.59	0.61	0.59
	$n = 500$	0.12	0.17	0.17	0.16
	$n = 1000$	0.07	0.12	0.12	0.13
Continuous grouping	$n = 100$	0.54	0.59	0.63	0.69
	$n = 500$	0.11	0.24	0.31	0.34
	$n = 1000$	0.08	0.17	0.21	0.26

Compared estimators and evaluation. The proposed approaches ET and EF are compared with several competing methods. **LOC:** A local CT estimator that does not utilize external information. It is trained on $\mathcal{I}_1$ only, combining training and estimation sets. **MA:** A naive model averaging method with weights $\omega_k^{\mathrm{MA}} = 1/k$. This approach assumes models are homogeneous. **EWMA:** We consider a modified version of EWMA that can be used for CATE. We obtain an approximation of $\tau_1(\boldsymbol{x})$ by fitting another local model using the estimation set of site 1, denoted by $\widetilde{\tau}_1(\boldsymbol{x})$. Its weights are given by

$$\omega_k^{\mathrm{EWMA}} = \frac{\exp\{-\sum_{i \in \mathcal{I}_1^{(2)}}(\widehat{\tau}_k(\boldsymbol{x}_i) - \widetilde{\tau}_1(\boldsymbol{x}_i))^2\}}{\sum_{\ell=1}^K \exp\{-\sum_{i \in \mathcal{I}_1^{(2)}}(\widehat{\tau}_\ell(\boldsymbol{x}_i) - \widetilde{\tau}_1(\boldsymbol{x}_i))^2\}}.$$

STACK: A stacking ensemble, which is a linear ensemble of predictions of several models (Breiman, 1996). 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 mean squared error (MSE) of these methods over an independent testing set of sample size $n_{te} = 2000$ from site 1. $\mathrm{MSE}(\widehat{\tau}) = n_{te}^{-1} \sum_{i=1}^{n_{te}} \{\widehat{\tau}(\boldsymbol{x}_i) - \tau_1(\boldsymbol{x}_i)\}^2$. Each simulation scenario is repeated for 1000 times. Experiments are performed on a 6-core Intel Xeon CPU E5-2620 v3 2.40GHz equipped with 64GB RAM.

Estimation performance. Figure 2 shows the performance of the proposed estimators and the competing estimators, using LOC as the benchmark. The proposed ET and EF show the best performance in terms of the mean and variation of MSE among other estimators when $c > 0$, and comparable to equal weighting MA when $c = 0$. Although, a forest is more stable than a tree in practice, both ET and EF give similar results because the true model is relatively simple and can be accurately estimated by a single ensemble tree under the given sample size.

Although asymptotically consistency, under finite sample, bias exists in local models and leads to biased model averaging estimates. While explicit quantification of bias and variance remains challenging due to extra uncertainty carried forward from the local estimates, we demonstrated that
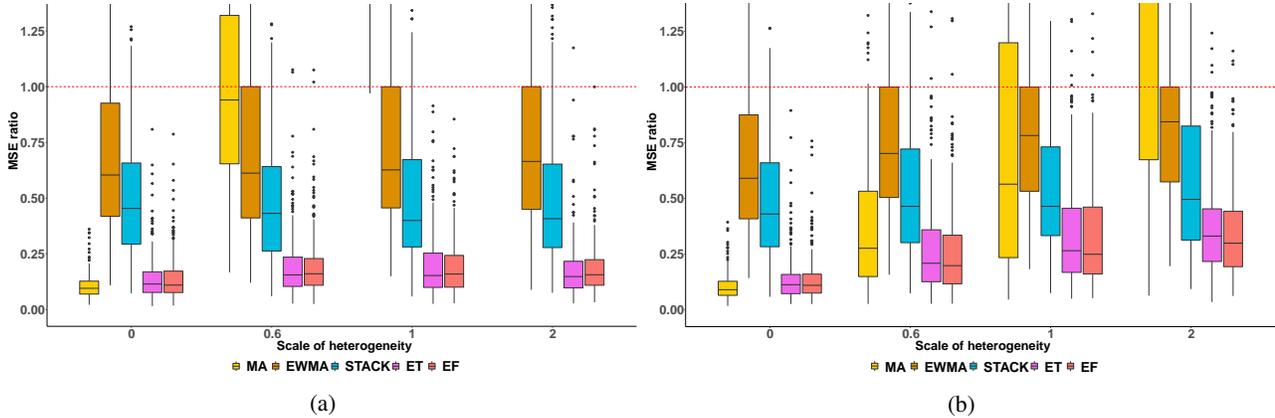
Figure 2. Box plots of MSE ratios of CATE estimators, respectively, over LOC, for **(a) discrete grouping** and **(b) continuous grouping** across site. Different colors imply different estimators, and x-axis, i.e., the value of c, differentiates the scale of global heterogeneity. The red dotted line denotes an MSE ratio of 1. MA performance is truncated due to large MSE ratios. The proposed ET and EF achieve smaller MSE ratios compared to standard model averaging or ensemble methods and are robust to heterogeneity across settings.

the proposed estimators can improve upon the local models under small sample size via Table 1. It shows the MSE ratio of EF over LOC as a measure of gain resulting from model averaging by varying $n = 100, 500, 1000$. The decrease in MSE ratio as n increases, regardless of the choice of c, is consistent with our asymptotic results in Theorem 3.1. This is due to a bias-and-variance trade-off in the ensemble that ensures a small MSE, which remains smaller than that in LOC despite varying n. It also shows our method is robust to the existence of local uncertainty.

Visualization of information borrowing. Figure 3 visualizes the proposed ET and EF. In (a) and (d), the site indicator and X_1 appear as splitting variables in the ETs, which is consistent with the data generation process. The estimated treatment effect (b) and (e) reveals the pattern of transportability across sites and with respect to X_1. Panels (c) and (f) plot the model averaging weights in EFs over X_1. Site 1 has a relatively large contribution to the weighted estimator while models from other sites have different contributions at different values of X_1 depending on their similarity in $\tau(\boldsymbol{x}, k)$ to that in site 1. Corresponding ET and EF show consistent patterns.

Additional simulations. The detailed results of these additional simulations are included in Appendix C. *1) Connection to supervised learning.* The uniqueness of averaging $\tau_k(\boldsymbol{x})$ as opposed to supervised learning that averages prediction models $f_k(\boldsymbol{x})$ is that the outcome of $f_k(\boldsymbol{x})$ is immediately available. In our case, an additional estimation step is needed to construct the model averaging weights. We provide a comparison among estimators that utilize the ground truth $\{\tau_k(\boldsymbol{x})\}_{k=1}^K$ (denoted as "-oracle") when computing ensemble weights. This mimics the case of supervised learning where weights are based on observed outcomes. Oracle methods achieve smaller MSE ratios; the pattern is consistent with Table 1. *2) Simulation under observational stud-*

ies. We also consider the treatment generation mechanism under an observational design. Specifically, the propensity is given as $e(\boldsymbol{x}) = \text{expit}(0.6x_1)$. We consider both a correctly specified propensity model using a logistic regression of Z on X_1 and a misspecified propensity model with a logistic regression of Z on all $\boldsymbol{X}$. In general, the proposed estimators obtain the best performance with similar results as in Figure 2. With the correctly specified propensity score model, the local estimator is consistent in estimating $\tau_k(\boldsymbol{x})$, the proposed framework is valid. When the propensity model is misspecified, extra uncertainty is carried forward from the local estimates, but the proposed estimators can still improve upon LOC. This is due to a bias-and-variance trade-off that leads to small MSE, which remains smaller than the local models. *3) Covariate dimensions.* Besides $D = 5$, we consider other choices of covariate dimension including $D = 20, 50$. With a higher dimension, the MSE ratio between the proposed estimates and LOC estimates increases but the same pattern across methods persists. *4) Unequal sample size at each site.* In the distributed date network, different sites may 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. We consider a simulation setting where site 1 has a sample size of $n_1 = 500$ while other site $n_2, \ldots, n_K$ has a sample size of 200. Results show that the MSE ratio between the proposed estimates and LOC estimates increases compared to the scenario where the sample size in all sites are 500. However, the proposed estimators still enjoy the most robust performance. This also shows our method is robust to the existence of local uncertainty. *5) Different local estimators.* We stress that other consistent estimators could be used as the local model. Options such as causal forest (Wager & Athey, 2018) are explored varying the sample size at local sites. Similar performance is observed as in Figure 2. *6) Further comparisons to non-*
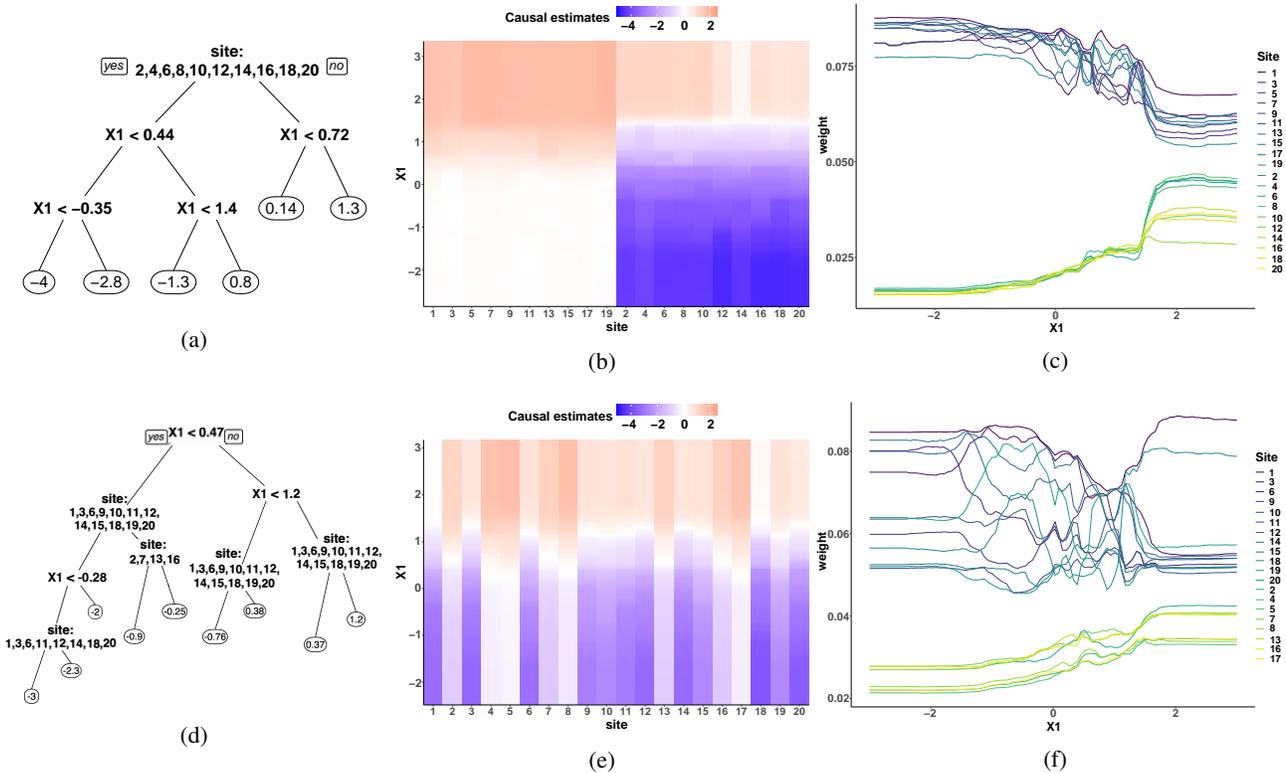
Figure 3. Visualization of simulation results under **discrete grouping (a,b,c)** and **continuous grouping (d,e,f)** when $c = 1$. (a) and (d) visualize the proposed ETs where the site indicator and X_1 are selected as splitting variables, which is consistent with the underlying data generation process. (b) and (e) show the predicted treatment effects of the proposed EFs varying X_1 in each site, marginalized over all other features. (b) is arranged according to the true grouping, odd sites versus even sites. The plot recovers the pattern of local and global heterogeneity. (c) and (f) plot the interpretable model averaging weights in EFs over X_1. The weights of site 1 have a relatively large contribution to the weighted estimator while models from other sites have different contributions for different X_1 depending on their similarity in $\tau(\boldsymbol{x}, k)$ to that in site 1. Corresponding ET and EF show consistent patterns and recover the true grouping.

adaptive ensemble. Here we provide a brief discussion of the implications of the proposed method and how it differs from non-adaptive methods such as stacking. Although unrealistic, when the true weights are non-adaptive, the performance may be similar. Plus, our learned weights can be used to examine adaptivity, as shown in Figure 3(c,f) and Figure 4(c). Stacking is shown to be more robust than non-adaptive model averaging in case of model misspecification. See discussion in Clarke (2003). Our additional simulation results show that in case of a large global heterogeneity, as c increases, the heterogeneity across sites gets larger, reducing the influence of important covariates on heterogeneity, hence the weights become more non-adaptive. However, the proposed methods still enjoy a comparable performance to STACK, which further indicates the robustness of the proposed methods.

5. Example: a Multi-Hospital Data Network

Application with contextual insights is provided based on an analysis of the eICU Collaborative Research Database, a multi-hospital database published by Philips Healthcare

(Pollard et al., 2018). The analysis is motivated by a recent retrospective study that there is a higher survival rate when SpO_2 is maintained at 94-98% among patients requiring oxygen therapy (van den Boom et al., 2020), not "the higher the better". We use the same data extraction code to create our data. We consider SpO_2 within this range as treatment ($Z = 1$) and outside of this range as control ($Z = 0$). A total of 7,022 patients from 20 hospitals, each with at least 50 patients in each treatment arm, are included with a randomly selected target (hospital 1). Hospital-level summary information is provided in Appendix D. Patient-level features include age, BMI, sex, Sequential Organ Failure Assessment (SOFA) score, and duration of oxygen therapy. The outcome is hospital survival ($Y = 1$) or death ($Y = 0$).

Figure 4 visualizes the performance of EF-based estimated effect of oxygen therapy setting on in-hospital survival. CT is used as the local model with propensity score modeled by a logistic regression. Figure 4(a) shows the propensity score-weighted **average survival** for those whose received treatment is consistent with the estimated decision. Specifi-

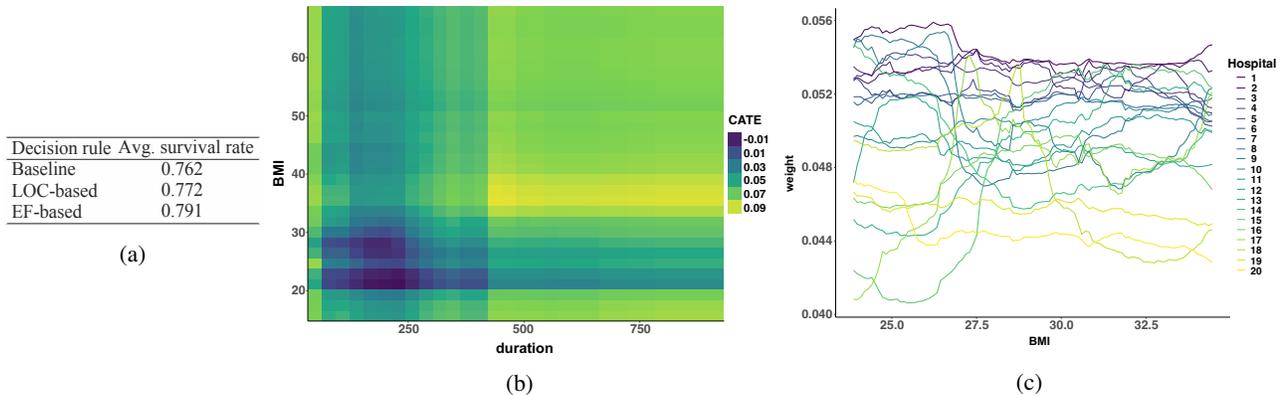| Decision rule	Avg. survival rate
Baseline	0.762
LOC-based	0.772
EF-based	0.791

(a)

(b)

(c)

Figure 4. Application to estimating treatment effects of oxygen therapy on survival. (a) Expected survival of treatment decision following different estimators. The proposed EF shows the largest gain in improving survival rate, more promising than LOC and baseline. (b) Estimated treatment effects varying duration and BMI, two important features in the fitted EF. Patients with a BMI around 35, and a duration above 400 benefited the most. (c) Visualization of data-adaptive weights in the estimated EF varying BMI. Hospitals with a larger bed capacity tend to contribute more, the data of which might be more similar to hospital 1.

cally, the expected reward is given by

$$\frac{\sum_i Y_i 1(Z_i = Z_i^{est})/\pi(Z_i, \boldsymbol{X}_i)}{\sum_i 1(Z_i = Z_i^{est})/\pi(Z_i, \boldsymbol{X}_i)},$$

where $Z_i^{est} = 1(\hat{\tau} > 0)$ denotes the estimated treatment rule and $\pi(Z_i, \boldsymbol{X}_i)$ is the probability of receiving the actual treatment. We provide expected reward for the 1) observed treatment assignment (baseline), 2) LOC-based rule, and 3) EF-based rule. The treatment rule based on our method can increase mean survival by 3% points compared to baseline, and is more promising than LOC.

In the fitted EF, the hospital indicator is the most important, explaining about 50% of the decrease in training error. Figure 4(b) shows the estimated CATE varying two important features, BMI and oxygen therapy duration. Patients with BMI around 36 and duration above 400 show the most benefit from oxygen therapy in the target SpO$_2$ range. Patients with BMI between 20 and 30 and duration around 200 may not benefit from such alteration. Figure 4(c) visualizes the data-adaptive weights $\omega_k(\boldsymbol{x})$ in the fitted EF with respect to BMI for different models, while holding other variables constant. The weights of hospital 1 are quite stable while models from other sites may have different contribution to the weighted estimator for different values of BMI. Judging from hospital information in Appendix D, hospitals with a larger bed capacity tend to be similar to hospital 1, and are shown to provide larger contributions.

In this distributed research network, different hospitals have a different sample size. For sensitivity analysis, we consider a weighting strategy to adjust for the sample size of site k. Results show similar patterns as in Figure 4. Detailed results are provided in Appendix D. The real-data access is provided in Appendix E.

6. Discussion

We have proposed an efficient and interpretable tree-based model averaging framework for enhancing treatment effect estimation at a target site by borrowing information from potentially heterogeneous data sources. We generalize standard model averaging scheme in a data-adaptive way such that the generated weights depend on subject-level features. This work makes multi-site collaborations and especially treatment effect estimation more practical by avoiding the need to share subject-level data. Our approach extends beyond causal inference to estimating a general $f(\boldsymbol{x})$ from heterogeneous data.

Unlike in classic model averaging where prediction performance can be assessed against observed outcomes or labels, treatment effects are not directly observed. While our approach is guaranteed to be consistent under randomized studies, the weights are estimated based on expected treatment effects, hence relying on Assumption 1 (unconfoundedness) to hold. It may be a strong assumption in observational studies with unmeasured confounding.

Acknowledgments

References

Ackerman, B., Schmid, I., Rudolph, K. E., Seamans, M. J., Susukida, R., Mojtabai, R., and Stuart, E. A. Implementing statistical methods for generalizing randomized trial findings to a target population. *Addictive Behaviors*, 94:

124–132, 2019.

Allcott, H. Site selection bias in program evaluation. *The Quarterly Journal of Economics*, 130(3):1117–1165, 2015.

Athey, S. and Imbens, G. Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, 113(27):7353–7360, 2016.

Athey, S., Tibshirani, J., and Wager, S. Generalized random forests. *Annals of Statistics*, 47(2):1148–1178, 2019.

Bareinboim, E. and Pearl, J. Causal inference and the data-fusion problem. *Proceedings of the National Academy of Sciences*, 113(27):7345–7352, 2016.

Battey, H., Fan, J., Liu, H., Lu, J., and Zhu, Z. Distributed testing and estimation under sparse high dimensional models. *Annals of Statistics*, 46(3):1352–1382, 2018.

Berger, M. L., Lipset, C., Gutteridge, A., Axelsen, K., Subedi, P., and Madigan, D. Optimizing the leveraging of real-world data to improve the development and use of medicines. *Value in Health*, 18(1):127–130, 2015.

Borenstein, M., Hedges, L. V., Higgins, J. P., and Rothstein, H. R. *Introduction to Meta-Analysis*. John Wiley & Sons, 2011.

Breiman, L. Stacked regressions. *Machine Learning*, 24(1): 49–64, 1996.

Breiman, L., Friedman, J., Stone, C. J., and Olshen, R. A. *Classification and Regression Trees*. CRC Press, 1984.

Breitbart, Y., Olson, P. L., and Thompson, G. R. Database integration in a distributed heterogeneous database system. In *1986 IEEE Second International Conference on Data Engineering*, pp. 301–310. IEEE, 1986.

Brown, J. S., Holmes, J. H., Shah, K., Hall, K., Lazarus, R., and Platt, R. Distributed health data networks: a practical and preferred approach to multi-institutional evaluations of comparative effectiveness, safety, and quality of care. *Medical Care*, 48(6):S45–S51, 2010.

Buchanan, A. L., Hudgens, M. G., Cole, S. R., Mollan, K. R., Sax, P. E., Daar, E. S., Adimora, A. A., Eron, J. J., and Mugavero, M. J. 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.

Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. Smote: synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16:321–357, 2002.

Cho, Y. J., Wang, J., Chiruvolu, T., and Joshi, G. Personalized federated learning for heterogeneous clients with clustered knowledge transfer. *arXiv preprint arXiv:2109.08119*, 2021.

Clarke, B. Comparing bayes model averaging and stacking when model approximation error cannot be ignored. *Journal of Machine Learning Research*, 4(Oct):683–712, 2003.

Cohen, J. A., Trojano, M., Mowry, E. M., Uitdehaag, B. M., Reingold, S. C., and Marrie, R. A. Leveraging real-world data to investigate multiple sclerosis disease behavior, prognosis, and treatment. *Multiple Sclerosis Journal*, 26 (1):23–37, 2020.

Colnet, B., Mayer, I., Chen, G., Dieng, A., Li, R., Varoquaux, G., Vert, J.-P., Josse, J., and Yang, S. Causal inference methods for combining randomized trials and observational studies: a review. *arXiv preprint arXiv:2011.08047*, 2020.

Cook, T. D., Campbell, D. T., and Shadish, W. *Experimental and quasi-experimental designs for generalized causal inference*. Houghton Mifflin Boston, MA, 2002.

Dahabreh, I. J., Robertson, S. E., Tchetgen, E. J., Stuart, E. A., and Hernán, M. A. Generalizing causal inferences from individuals in randomized trials to all trial-eligible individuals. *Biometrics*, 75(2):685–694, 2019.

Dai, D. and Zhang, T. Greedy model averaging. *Advances in Neural Information Processing Systems*, 24:1242–1250, 2011.

Dai, D., Han, L., Yang, T., and Zhang, T. Bayesian model averaging with exponentiated least squares loss. *IEEE Transactions on Information Theory*, 64(5):3331–3345, 2018.

Dayan, I., Roth, H. R., Zhong, A., Harouni, A., Gentili, A., Abidin, A. Z., Liu, A., Costa, A. B., Wood, B. J., Tsai, C.-S., et al. Federated learning for predicting clinical outcomes in patients with covid-19. *Nature Medicine*, 27 (10):1735–1743, 2021.

Degtiar, I. and Rose, S. A review of generalizability and transportability. *arXiv preprint arXiv:2102.11904*, 2021.

DeShazo, J. P. and Hoffman, M. A. A comparison of a multistate inpatient ehr database to the hcup nationwide inpatient sample. *BMC Health Services Research*, 15(1): 1–8, 2015.

Donahue, M., Bouhaddou, O., Hsing, N., Turner, T., Crandall, G., Nelson, J., and Nebeker, J. Veterans health information exchange: successes and challenges of nationwide interoperability. In *AMIA Annual Symposium*

Proceedings, volume 2018, pp. 385. American Medical Informatics Association, 2018.

Donohue, J. M., Jarlenski, M. P., Kim, J. Y., Tang, L., Ahrens, K., Allen, L., Austin, A., Barnes, A. J., Burns, M., Chang, C.-C. H., 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.

Druckman, J. N., Greene, D. P., Kuklinski, J. H., and Lupia, A. *Cambridge Handbook of Experimental Political Science*. Cambridge University Press, 2011.

Egami, N. and Hartman, E. Elements of external validity: Framework, design, and analysis. *Design, and Analysis (June 30, 2020)*, 2020.

Fallah, A., Mokhtari, A., and Ozdaglar, A. Personalized federated learning with theoretical guarantees: A model-agnostic meta-learning approach. *Advances in Neural Information Processing Systems*, 33:3557–3568, 2020.

Fleurence, R. L., Curtis, L. H., Califf, R. M., Platt, R., Selby, J. V., and Brown, J. S. Launching PCORnet, a national patient-centered clinical research network. *Journal of the American Medical Informatics Association*, 21(4): 578–582, 2014.

Hahn, P. R., Murray, J. S., and Carvalho, C. M. Bayesian regression tree models for causal inference: Regularization, confounding, and heterogeneous effects (with discussion). *Bayesian Analysis*, 15(3):965–1056, 2020.

Hansen, B. E. Least squares model averaging. *Econometrica*, 75(4):1175–1189, 2007.

Hard, A., Rao, K., Mathews, R., Ramaswamy, S., Beaufays, F., Augenstein, S., Eichner, H., Kiddon, C., and Ramage, D. Federated learning for mobile keyboard prediction. *arXiv preprint arXiv:1811.03604*, 2018.

Harton, J., Segal, B., Mamtani, R., Mitra, N., and Hubbard, R. Combining real-world and randomized control trial data using data-adaptive weighting via the on-trial score. *arXiv preprint arXiv:2108.08756*, 2021.

Hripcsak, G., Duke, J. D., Shah, N. H., Reich, C. G., Huser, V., Schuemie, M. J., Suchard, M. A., Park, R. W., Wong, I. C. K., Rijnbeek, P. R., et al. Observational health data sciences and informatics (ohdsi): opportunities for observational researchers. *Studies in Health Technology and Informatics*, 216:574, 2015.

Jordan, M. I., Lee, J. D., and Yang, Y. Communication-efficient distributed statistical inference. *Journal of the American Statistical Association*, 114(526):668–681, 2019.

Ju, C., Combs, M., Lendle, S. D., Franklin, J. M., Wyss, R., Schneeweiss, S., and van der Laan, M. J. Propensity score prediction for electronic healthcare databases using super learner and high-dimensional propensity score methods. *Journal of Applied Statistics*, 46(12):2216–2236, 2019.

Kairouz, P., McMahan, H. B., Avent, B., Bellet, A., Bennis, M., Bhagoji, A. N., Bonawitz, K., Charles, Z., Cormode, G., Cummings, R., et al. Advances and open problems in federated learning. *arXiv preprint arXiv:1912.04977*, 2019.

Ke, Z. T., Fan, J., and Wu, Y. Homogeneity pursuit. *Journal of the American Statistical Association*, 110(509):175–194, 2015.

Kern, H. L., Stuart, E. A., Hill, J., and Green, D. P. Assessing methods for generalizing experimental impact estimates to target populations. *Journal of Research on Educational Effectiveness*, 9(1):103–127, 2016.

Künzel, S. R., Sekhon, J. S., Bickel, P. J., and Yu, B. Meta-learners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the National Academy of Sciences*, 116(10):4156–4165, 2019.

Lee, J. D., Liu, Q., Sun, Y., and Taylor, J. E. Communication-efficient sparse regression. *The Journal of Machine Learning Research*, 18(1):115–144, 2017.

Li, T., Sahu, A. K., Talwalkar, A., and Smith, V. Federated learning: Challenges, methods, and future directions. *IEEE Signal Processing Magazine*, 37(3):50–60, 2020.

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

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

Maro, J. C., Platt, R., Holmes, J. H., Strom, B. L., Hennessy, S., Lazarus, R., and Brown, J. S. Design of a national distributed health data network. *Annals of Internal Medicine*, 151(5):341–344, 2009.

Masoudnia, S. and Ebrahimpour, R. Mixture of experts: a literature survey. *Artificial Intelligence Review*, 42(2): 275–293, 2014.

McMahan, B., Moore, E., Ramage, D., Hampson, S., and y Arcas, B. A. Communication-efficient learning of deep networks from decentralized data. In *Artificial Intelligence and Statistics*, pp. 1273–1282. PMLR, 2017.

Neyman, J. 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.

Nie, X. and Wager, S. Quasi-oracle estimation of heterogeneous treatment effects. *Biometrika*, 108(2):299–319, 2021.

Pearl, J. and Bareinboim, E. Transportability of causal and statistical relations: A formal approach. In *Twenty-fifth AAAI Conference on Artificial Intelligence*, 2011.

Pearl, J. and Bareinboim, E. External validity: From do-calculus to transportability across populations. *Statistical Science*, 29(4):579–595, 2014.

Perez, L. and Wang, J. The effectiveness of data augmentation in image classification using deep learning. *arXiv preprint arXiv:1712.04621*, 2017.

Pirracchio, R., Petersen, M. L., and van der Laan, M. Improving propensity score estimators' robustness to model misspecification using super learner. *American Journal of Epidemiology*, 181(2):108–119, 2015.

Platt, R., Brown, J. S., Robb, M., McClellan, M., Ball, R., Nguyen, M. D., and Sherman, R. E. The FDA sentinel initiative—an evolving national resource. *New England Journal of Medicine*, 379(22):2091–2093, 2018.

Pollard, T. J., Johnson, A. E., Raffa, J. D., Celi, L. A., Mark, R. G., and Badawi, O. The eICU collaborative research database, a freely available multi-center database for critical care research. *Scientific Data*, 5:180178, 2018.

Polley, E. C. and van der Laan, M. J. Super learner in prediction. *U.C. Berkeley Division of Biostatistics Working Paper Series*, 2010.

Raftery, A. E., Madigan, D., and Hoeting, J. A. Bayesian model averaging for linear regression models. *Journal of the American Statistical Association*, 92(437):179–191, 1997.

Raghupathi, W. and Raghupathi, V. Big data analytics in healthcare: promise and potential. *Health Information Science and Systems*, 2(1):1–10, 2014.

Reynolds, M., Christian, J., Mack, C., Hall, N., and Dreyer, N. Leveraging real-world data for covid-19 research: challenges and opportunities. *Journal of Precision Medicine*, 6:1–6, 2020.

Riley, R. D., Higgins, J. P., and Deeks, J. J. Interpretation of random effects meta-analyses. *BMJ*, 342, 2011.

Robinson, P. M. Root-n-consistent semiparametric regression. *Econometrica: Journal of the Econometric Society*, pp. 931–954, 1988.

Röver, C. and Friede, T. Dynamically borrowing strength from another study through shrinkage estimation. *Statistical Methods in Medical Research*, 29(1):293–308, 2020.

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

Shortreed, S. M., Cook, A. J., Coley, R. Y., Bobb, J. F., and Nelson, J. C. Challenges and opportunities for using big health care data to advance medical science and public health. *American Journal of Epidemiology*, 188(5):851–861, 2019.

Smith, V., Chiang, C.-K., Sanjabi, M., and Talwalkar, A. S. Federated multi-task learning. *Advances in Neural Information Processing Systems*, 30, 2017.

Stuart, E. A., Cole, S. R., Bradshaw, C. P., and Leaf, P. J. 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.

Stuart, E. A., Bradshaw, C. P., and Leaf, P. J. Assessing the generalizability of randomized trial results to target populations. *Prevention Science*, 16(3):475–485, 2015.

Stuart, E. A., Ackerman, B., and Westreich, D. Generalizability of randomized trial results to target populations: design and analysis possibilities. *Research on Social Work Practice*, 28(5):532–537, 2018.

Sutton, A. J., Abrams, K. R., Jones, D. R., Jones, D. R., Sheldon, T. A., and Song, F. *Methods for Meta-Analysis in Medical Research*. Wiley Chichester, 2000.

Tan, X., Yang, S., Ye, W., Faries, D. E., Lipkovich, I., and Kadziola, Z. When doubly robust methods meet machine learning for estimating treatment effects from real-world data: A comparative study. *arXiv preprint arXiv:2204.10969*, 2022.

Tang, L. and Song, P. X. Fused lasso approach in regression coefficients clustering: learning parameter heterogeneity in data integration. *The Journal of Machine Learning Research*, 17(1):3915–3937, 2016.

Tang, L. and Song, P. X.-K. Poststratification fusion learning in longitudinal data analysis. *Biometrics*, 77(3):914–928, 2021.

Tang, L., Zhou, L., and Song, P. X.-K. Distributed simultaneous inference in generalized linear models via confidence distribution. *Journal of Multivariate Analysis*, 176: 104567, 2020.

Tang, X., Xue, F., and Qu, A. Individualized multidirectional variable selection. *Journal of the American Statistical Association*, 116(535):1280–1296, 2021.

Tipton, E. and Olsen, R. B. A review of statistical methods for generalizing from evaluations of educational interventions. *Educational Researcher*, 47(8):516–524, 2018.

Toh, S., Platt, R., Steiner, J., and Brown, J. Comparative-effectiveness research in distributed health data networks. *Clinical Pharmacology & Therapeutics*, 90(6):883–887, 2011.

van den Boom, W., Hoy, M., Sankaran, J., Liu, M., Chahed, H., Feng, M., and See, K. C. The search for optimal oxygen saturation targets in critically ill patients: Observational data from large ICU databases. *Chest*, 157(3): 566–573, 2020.

van der Laan, M. J., Polley, E. C., and Hubbard, A. E. Super learner. *Statistical Applications in Genetics and Molecular Biology*, 6(1), 2007.

van Dyk, D. A. and Meng, X.-L. The art of data augmentation. *Journal of Computational and Graphical Statistics*, 10(1):1–50, 2001.

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

Wang, Y., Jang, B., and Hero, A. The sylvester graphical lasso (syglasso). In *International Conference on Artificial Intelligence and Statistics*, pp. 1943–1953. PMLR, 2020.

Wang, Y., Klein, N., Morley, S., Jordanova, V., Henderson, M., Biswas, A., and Lawrence, E. Tributarypca: Distributed, streaming pca for in situ dimension reduction with application to space weather simulations. In *2021 7th International Workshop on Data Analysis and Reduction for Big Scientific Data (DRBSD-7)*, pp. 33–39. IEEE, 2021.

Wasserman, L. Bayesian model selection and model averaging. *Journal of Mathematical Psychology*, 44(1):92–107, 2000.

Whitehead, A. *Meta-Analysis of Controlled Clinical Trials*. John Wiley & Sons, 2002.

Wyss, R., Schneeweiss, S., van der Laan, M., Lendle, S. D., Ju, C., and Franklin, J. M. Using super learner prediction modeling to improve high-dimensional propensity score estimation. *Epidemiology*, 29(1):96–106, 2018.

Yang, Q., Liu, Y., Chen, T., and Tong, Y. Federated machine learning: concept and applications. *ACM Transactions on Intelligent Systems and Technology*, 10(2):1–19, 2019.

Yang, S., Zeng, D., and Wang, X. Elastic integrative analysis of randomized trial and real-world data for treatment heterogeneity estimation. *arXiv preprint arXiv:2005.10579*, 2020.

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

Yao, Y., Vehtari, A., Simpson, D., and Gelman, A. Using stacking to average bayesian predictive distributions (with discussion). *Bayesian Analysis*, 13(3):917–1007, 2018.

Zhao, Y., Li, M., Lai, L., Suda, N., Civin, D., and Chandra, V. Federated learning with non-iid data. *arXiv preprint arXiv:1806.00582*, 2018.

A. Related Topics and Distinctions

In Section 2, we focused on the literature review of model averaging for ease of exposition, because the most innovated part of our method is motivated directly from this class of work. Here we clarify the main differences among model averaging, meta-analysis, federated learning, as well as super learner.

Model averaging: a convex averaging of models via model-specific weights (Raftery et al., 1997; Yang, 2001; Dai & Zhang, 2011; Yao et al., 2018; Dai et al., 2018). The extension of the weights from scalars to functions provides the best motivation for our approach.

Meta-analysis: classic in the way that it describes the site-level heterogeneity using modeling assumptions (Whitehead, 2002; Sutton et al., 2000), rather than a more data-driven approach such as tree models. It can be either frequentist or Bayesian, the latter of which tends to be more useful under limited sample sizes. However, its main interest is typically the overall effect rather than the site-level heterogeneity, which is usually modeled by a nuisance parameter (Borenstein et al., 2011; Riley et al., 2011; Röver & Friede, 2020).

Federated learning: originated from the field of computer science (McMahan et al., 2017), federated learning is a collaborative learning procedure that ensures data privacy by exchanging model parameters only. Federated learning methods often involves iterative updating (Fallah et al., 2020; Cho et al., 2021; Smith et al., 2017; Yang et al., 2019), rather than a one-shot procedure, which could be hard to apply to nonautomated distributed research networks. It has been developed mainly to estimate a global prediction model by leveraging distributed data (Li et al., 2020; Kairouz et al., 2019; Zhao et al., 2018; Hard et al., 2018), and is not designed to target any specific site.

Super learner: an ensemble of multiple statistical and machine learning models (van der Laan et al., 2007). It learns an optimal weighted average of those candidate models by minimizing the cross-validated risk, and assigns higher weights to more accurate models (Polley & van der Laan, 2010). The final prediction on an independent testing data is the weighted combination of the predictions of those models. Super learner has been showed empirically to improve treatment effect estimation via the modeling of propensity score in observational studies (Pirracchio et al., 2015; Wyss et al., 2018; Shortreed et al., 2019; Ju et al., 2019; Tan et al., 2022).

Mixture of experts: an ensemble learning technique that decomposes a task into multiple subtasks with domain knowledge, followed by using multiple expert models to handle each subtask. A gating model is then used to decide which expert to use to make future prediction (Masoudnia & Ebrahimpour, 2014). It differs from other ensemble methods typically in that often only a few experts will be selected for predictions, rather than combining results from all experts (Masoudnia & Ebrahimpour, 2014).

B. Proof of Theorem 3.1

The proof of Theorem 3.1 closely follows arguments given in Wager & Athey (2018). 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 $m < n_1$, 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 Athey & Imbens (2016) 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 (Athey & Imbens, 2016). Following Definitions 1-5 and Theorem 3.1 in Wager & Athey (2018), the proposed estimator $\widehat{\mathcal{T}}_{\text{EF}}(\boldsymbol{x}, 1)$ is a consistent estimator of the true treatment effect function $\tau_1(\boldsymbol{x})$ for site 1.

C. Full Simulation Results

Connection to supervised learning. Similar to ET-oracle and EF-oracle 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-oracle and its weight is given by

$$\omega_k^{\text{EWMA-oracle}} = \frac{\exp\{-\sum_{i \in \mathcal{I}_1^{(2)}} (\widehat{\tau}_k(\boldsymbol{x}_i) - \tau_1(\boldsymbol{x}_i))^2\}}{\sum_{\ell=1}^K \exp\{-\sum_{i \in \mathcal{I}_1^{(2)}} (\widehat{\tau}_\ell(\boldsymbol{x}_i) - \tau_1(\boldsymbol{x}_i))^2\}}.$$

Similarly, the corresponding linear stacking approach, denoted as STACK-oracle, 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. We present simulation results using CT as the local model and the sample size at local sites to be $n = 500$. Figure 5 presents the performance of the proposed estimators along with other competing estimators. Each series of boxes corresponds to a different strength of global heterogeneity c. Table 2 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-oracle achieves minimal MSE for low and moderate degrees of heterogeneity while EF-oracle 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-oracle 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-oracle performs better than STACK in all settings. STACK-oracle, with ground truth $\tau_1(\boldsymbol{x})$ available, outperforms ET and EF when there exists a moderate to high level of heterogeneity across sites.

Various sample sizes in local sites. We provide detailed simulation results varying n $(100, 500, 1000)$ with CT as the local model. Figure 6 and Figure 7 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-oracle and EF-oracle achieve close-to-zero MSE with very small spreads in some settings. Figure 8 shows plots of the bias and MSE of EF-oracle varying sample size at each site ($n = 100, 500, 1000$). As the sample size increases, both bias and MSE of EF-oracle reduce to zero. Consistency of EF-oracle 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.

Simulations under observational studies. We also consider the treatment generation mechanism under an observational design. Specifically, the propensity is given as $e(\boldsymbol{x}) = \text{expit}(0.6x_1)$. We consider both a correctly specified propensity model using a logistic regression of Z on X_1 and a misspecified propensity model with a logistic regression of Z on all $\boldsymbol{X}$. Figure 15 and Figure 16 show box plots of simulation results. In general, the proposed estimators obtain the best performance with similar results are obtained as in the Figure 2. With the correctly specified propensity score model, the local estimator is consistent in estimating $\tau_k(\boldsymbol{x})$, the proposed framework is valid. When the propensity model misspecified, extra uncertainty is carried forward from the local estimates, but the proposed estimators can improve upon the local models. This is due to a bias-and-variance trade-off that guarantees small MSE in prediction, which remains smaller than those from local estimators.

Covariate dimensions We consider various choices of covariate dimensions besides $D = 5$. Specifically, we also try $D = 20$ and $D = 50$. Figure 9 and Figure 10 show box plots of simulation results. With a higher dimension of variables, the MSE ratio between the proposed estimates and LOC estimates increases than that in the scenario with a small dimension.

Unequal sample size at each site. In the distributed date network, different sites may 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. We consider a simulation setting where site 1 has a sample size of $n_1 = 500$ while other site $n_2, \ldots, n_K$ has a sample size of 200. Figure 11 shows box plots of simulation results. Results show that the MSE ratio between the proposed estimates and LOC estimates increases compared to the scenario where the sample size in all sites are 500. However, the proposed estimators still enjoy the most robust performance via bias-and-variance trade-off. This also shows our method is robust to the existence of local uncertainty.

Different local estimators. We explore another option for the local model using the causal forest (CF) (Wager & Athey,

2018) varying the sample size at local sites. A causal forest is a stochastic averaging of multiple causal trees (Athey & Imbens, 2016), 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 (Wager & Athey, 2018). CF is implemented in the R packages `grf`. Figure 12, Figure 13, and Figure 14 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 $\tau_1(\boldsymbol{x})$.

Further comparisons to non-adaptive ensemble. We provide simulation results to compare the proposed methods to the non-adaptive method STACK. Consider the following setting where the heterogeneity is continuous and nonlinear: $\tau(\boldsymbol{x}, k) = \mathbb{1}\{x_1 > 0\} \cdot x_1 + (x_1 - 3) \cdot (U_k)^c$, with $U_k \sim Unif[0, 3]$, $\boldsymbol{X}_i \sim N(\boldsymbol{0}, \boldsymbol{I}_5)$, and $c = (1, 2, 3, 4)$. As c increases, the heterogeneity across sites gets larger, reducing the influence of x_1 on heterogeneity, hence the weights become more non-adaptive. For $c = (1, 2, 3, 4)$, the one-SD ranges of MSE ratios of EF over STACK are [0.73,0.82], [0.86,0.87], [0.99,1.04], [0.87,1.07], respectively. When c is relatively small, the proposed EF has a smaller MSE compared to STACK. As c increases, the performance of EF is similar to that of STACK, in the case of a large global heterogeneity. This further indicates the robustness of the proposed methods.

D. Additional Results for Data Application

In real-life applications, hospitals may have different sample sizes n_k that may affect the accuracy of the estimation of τ_k. Table 3 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 $\eta_k(\boldsymbol{x}) = K n_k \{\sum_{j=1}^{K} n_j\}^{-1}$.

Figure 17 visualizes the performance of oxygen therapy on hospital survival with the weighting strategy adopted. CT is used as the local model with propensity score modeled by a logistic regression. Figure 17(a) shows the propensity score-weighted average survival for those whose received treatment is consistent with the estimated decision. Treatment rule based on our method can increase survival by 4%, more promising than the EF estimates without the weighting strategy and the LOC and the baseline. The weighting strategy takes account into the unequal sample size among the hospital network, and assign weights based on precision of local estimates.

In the fitted EF, the hospital indicator remains the most important, explaining about 48% of the decrease in training error. Figure 17(b) shows the estimated CATEs varying two important features, BMI and oxygen therapy duration. Patients with BMI between 36 and 40 and duration above 400 show the most benefit from oxygen therapy in the target SpO$_2$ range. Patients with BMI between 20 and 30 and duration between 100 and 400 may not benefit from such alteration. The treatment estimates are similar to that in Figure 4(b) Figure 17(c) visualizes the proposed model averaging scheme with data-adaptive weights $\omega_k(\boldsymbol{x})$ in the fitted EF with respect to BMI for different models, while holding other variables constant. The weights of hospital 1 are quite stable while models from other sites may have different contribution to the weighted estimator for different values of BMI. Similar to Figure 4(c), hospitals with a larger bed capacity tend to be similar to hospital 1, and are shown to provide larger contributions. In general, the weighting strategy helps further improve the expected survival rate. The patterns in each subfigure are similar to Figure 4, 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.

E. Real Data Access

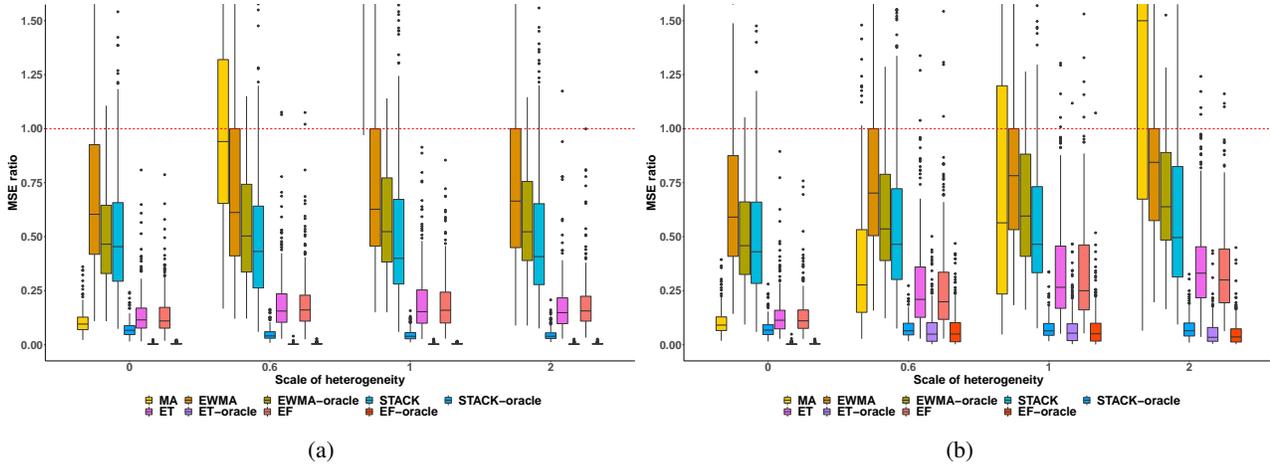Although the eICU data used in our application example cannot be shared subject to the data use agreement, access can be individually requested at https://eicu-crd.mit.edu/gettingstarted/access/.
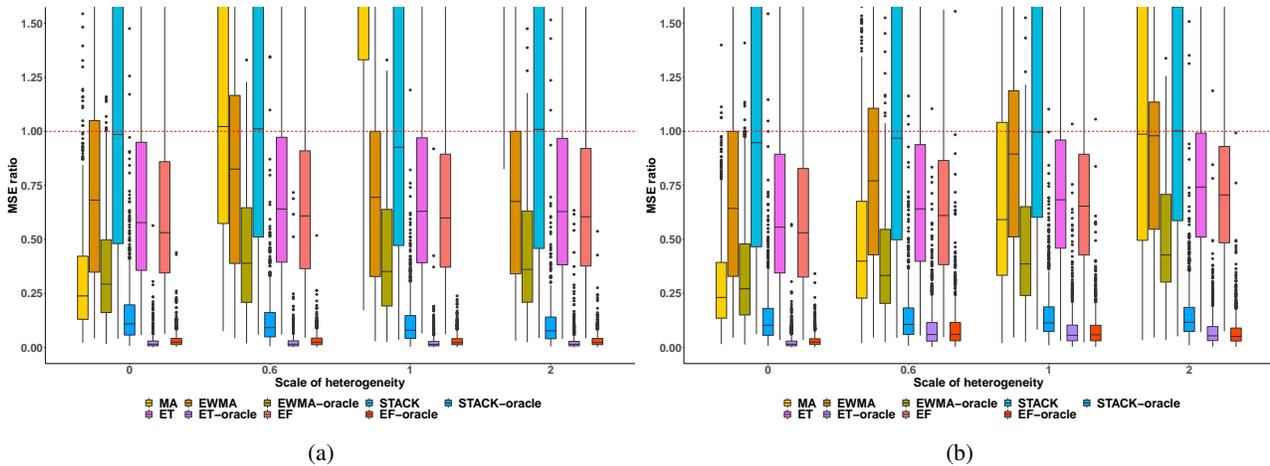
Figure 5. Box plots of the MSE ratios of CATE estimators, respectively, over LOC (**CT**) and a sample size of **500** at each site for **(a) discrete grouping** and **(b) continuous grouping** across site, respectively, varying scale of global heterogeneity. Estimators ending with "-oracle" makes use of ground truth treatment effects. Different colors imply different estimators, and x-axis, i.e., the value of c, differentiates the scale of global heterogeneity. The red dotted line denotes an MSE ratio of 1. MA performance is truncated due to large MSE ratios. The proposed ET and EF achieve competitive performance compared to standard model averaging or ensemble methods and are robust to heterogeneity across settings. Note that ET-oracle and EF-oracle achieve close-to-zero MSE ratios with very small spreads in some settings.

Figure 6. Box plots of the MSE ratios of CATE estimators, respectively, over LOC (**CT**) and a sample size of **100** at each site for **(a) discrete grouping** and **(b) continuous grouping** across site, respectively, varying scale of global heterogeneity. Estimators ending with "-oracle" makes use of ground truth treatment effects. Different colors imply different estimators, and x-axis, i.e., the value of c, differentiates the scale of global heterogeneity. The red dotted line denotes an MSE ratio of 1. MA performance is truncated due to large MSE ratios. The proposed ET and EF achieve competitive performance compared to standard model averaging or ensemble methods and are robust to heterogeneity across settings. Note that ET-oracle and EF-oracle achieve close-to-zero MSE ratios with very small spreads in some settings.

Table 2. Simulation results for ratio between MSE of the estimator and MSE of LOC (**CT**) with a sample size of **500** at each site. A smaller number indicates larger improvement over the local model. Estimators ending with "-oracle" 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 $\tau_1(\boldsymbol{x})$.

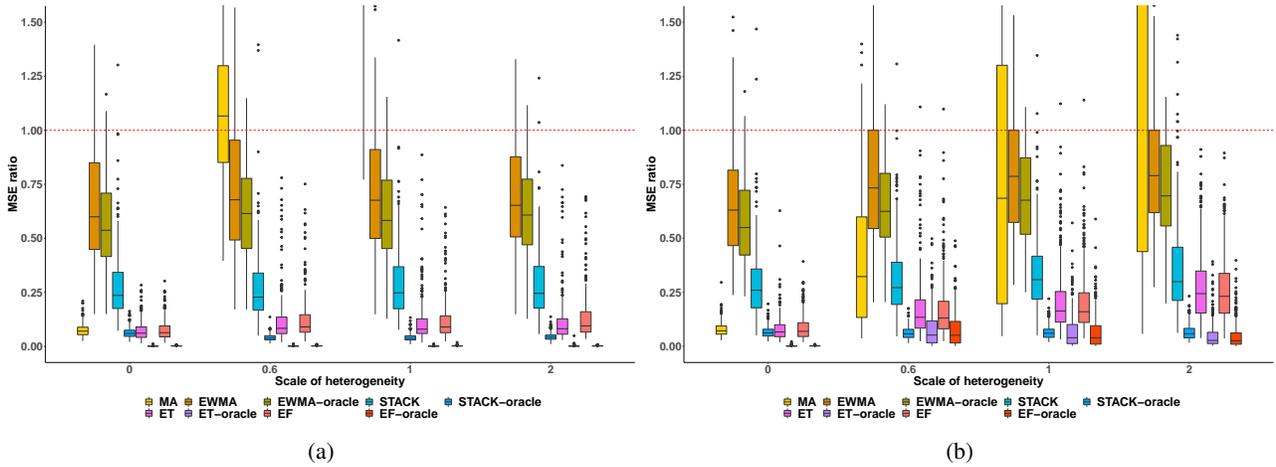| Estimator	Discrete grouping				Continuous grouping			
	$c = 0$	$c = 0.2$	$c = 0.6$	$c = 1$	$c = 0$	$c = 0.2$	$c = 0.6$	$c = 1$
Ratio of average of MSEs over 1000 replicates								
MA	0.09	0.91	2.4	9.87	0.08	0.32	0.65	1.78
EWMA	0.57	0.62	0.61	0.62	0.56	0.65	0.7	0.77
EWMA-oracle	0.42	0.5	0.49	0.5	0.42	0.49	0.53	0.59
STACK	0.44	0.45	0.44	0.45	0.45	0.45	0.48	0.54
STACK-oracle	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-oracle	<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-oracle	<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-oracle	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-oracle	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-oracle	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-oracle	0.03	0.03	0.03	0.03	0.03	0.06	0.07	0.08

(a) (b)

Figure 7. Box plots of the MSE ratios of CATE estimators, respectively, over LOC (**CT**) and a sample size of **1000** at each site for **(a) discrete grouping** and **(b) continuous grouping** across site, respectively, varying scale of global heterogeneity. Estimators ending with "-oracle" makes use of ground truth treatment effects. Different colors imply different estimators, and x-axis, i.e., the value of c, differentiates the scale of global heterogeneity. The red dotted line denotes an MSE ratio of 1. MA performance is truncated due to large MSE ratios. The proposed ET and EF achieve competitive performance compared to standard model averaging or ensemble methods and are robust to heterogeneity across settings. Note that ET-oracle and EF-oracle achieve close-to-zero MSE ratios with very small spreads in some settings.
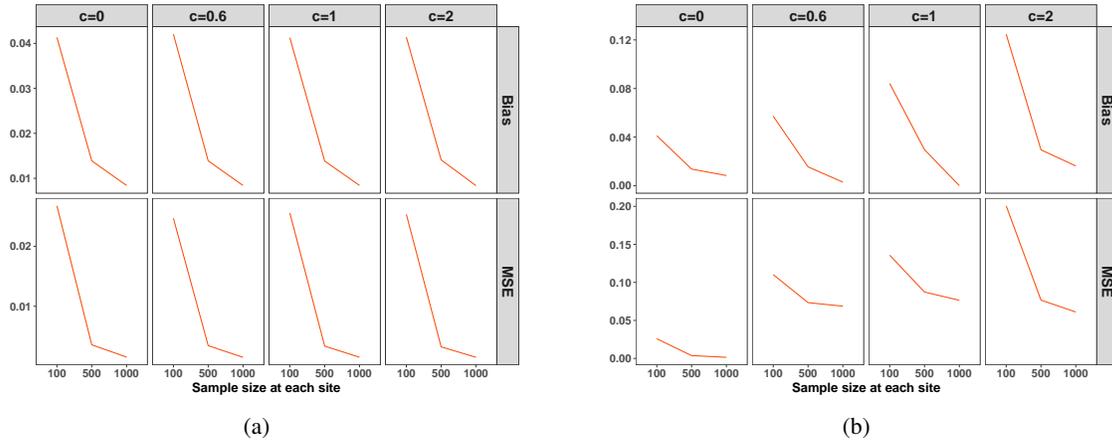
Figure 8. Plots of the bias and MSE of **EF-oracle** varying sample site at each site for **(a) discrete grouping** and **(b) continuous grouping** across site, varying scale of global heterogeneity. Both bias and MSE reduces to zero as the sample size increases.
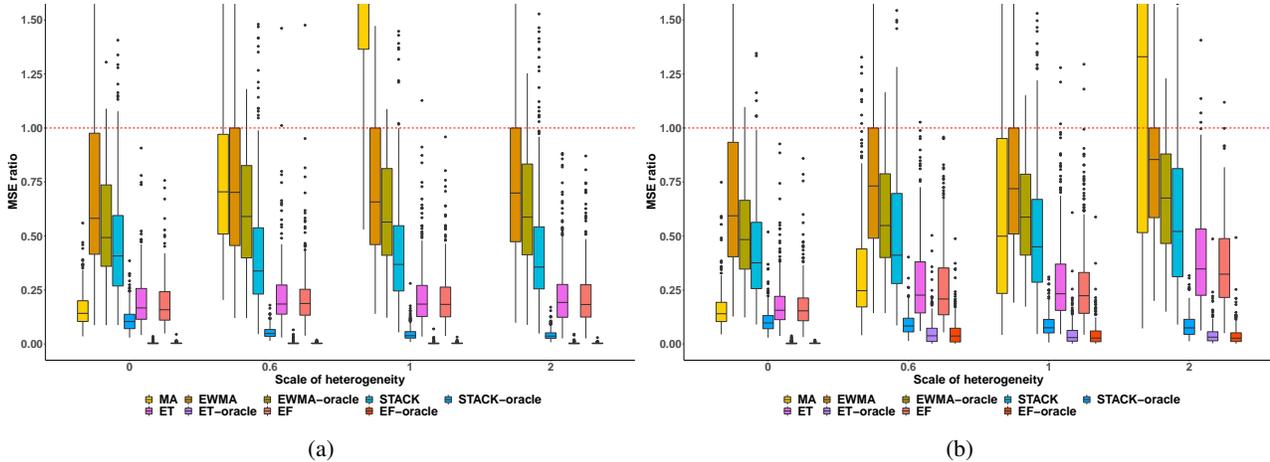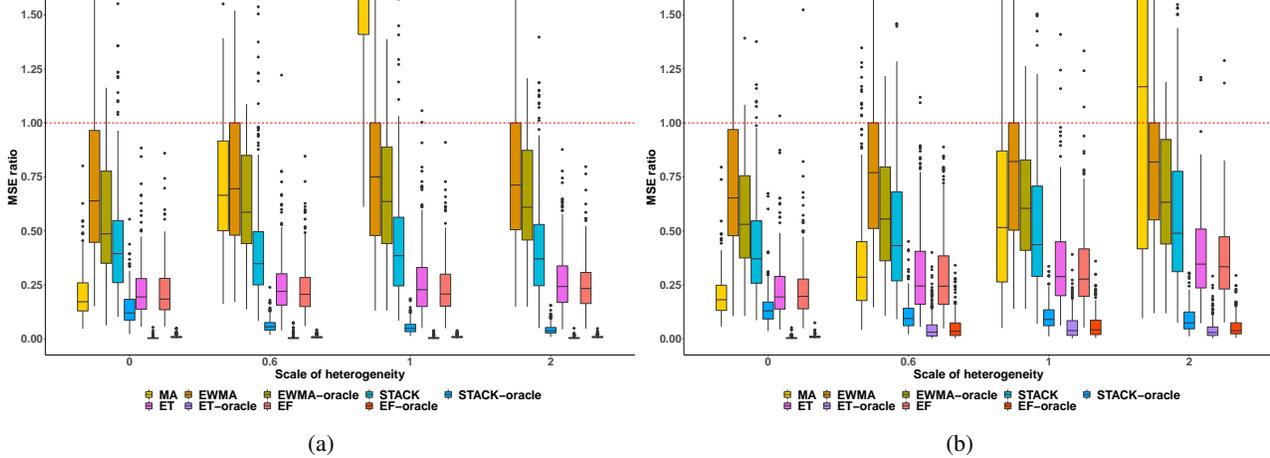

Figure 9. Box plots of the MSE ratios of CATE estimators, respectively, over LOC (**CT**) and a sample size of **500** at each site, and covariate dimension of **20** for **(a) discrete grouping** and **(b) continuous grouping** across site, respectively, varying scale of global heterogeneity. Estimators ending with "-oracle" makes use of ground truth treatment effects. Different colors imply different estimators, and x-axis, i.e., the value of c, differentiates the scale of global heterogeneity. The red dotted line denotes an MSE ratio of 1. MA performance is truncated due to large MSE ratios. The proposed ET and EF achieve competitive performance compared to standard model averaging or ensemble methods and are robust to heterogeneity across settings. Note that ET-oracle and EF-oracle achieve close-to-zero MSE ratios with very small spreads in some settings.
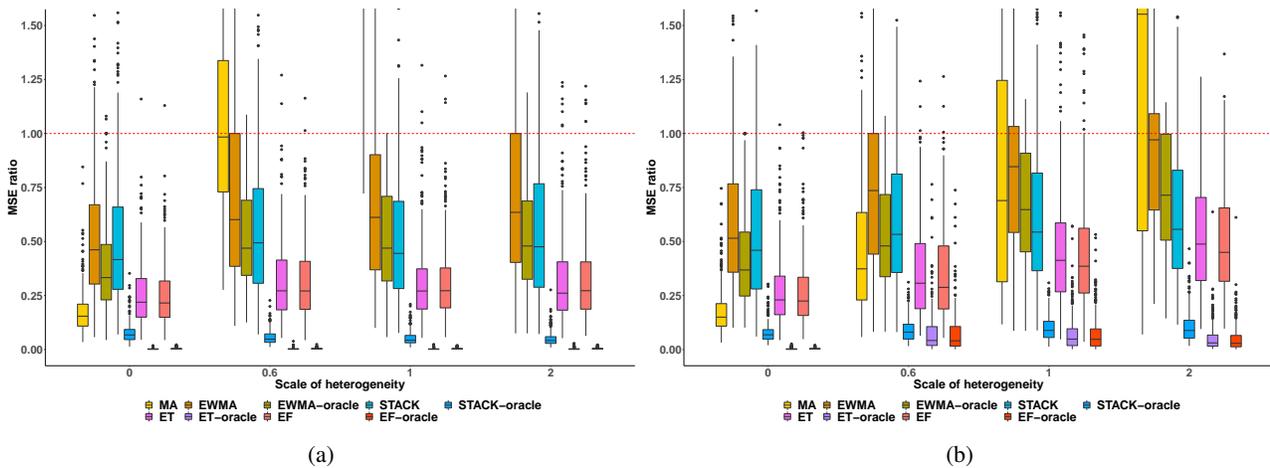
Figure 10. Box plots of the MSE ratios of CATE estimators, respectively, over LOC (**CT**) and a sample size of **500** at each site, and covariate dimension of **50** for **(a) discrete grouping** and **(b) continuous grouping** across site, respectively, varying scale of global heterogeneity. Estimators ending with "-oracle" makes use of ground truth treatment effects. Different colors imply different estimators, and x-axis, i.e., the value of c, differentiates the scale of global heterogeneity. The red dotted line denotes an MSE ratio of 1. MA performance is truncated due to large MSE ratios. The proposed ET and EF achieve competitive performance compared to standard model averaging or ensemble methods and are robust to heterogeneity across settings. Note that ET-oracle and EF-oracle achieve close-to-zero MSE ratios with very small spreads in some settings.
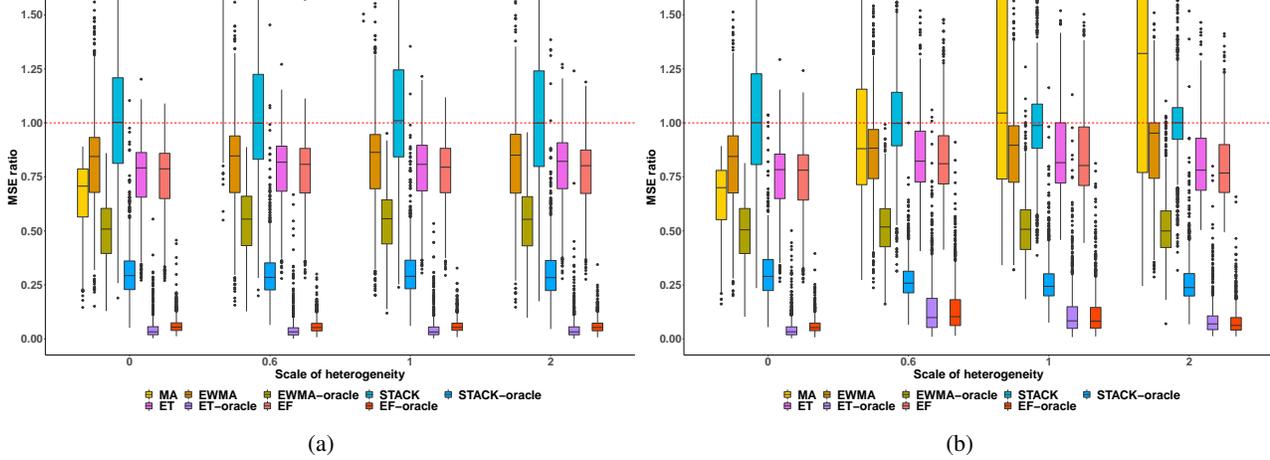

Figure 11. Box plots of the MSE ratios of CATE estimators, respectively, over LOC (**CT**) and a sample size of **500** at site 1, and a sample size of **200** at other sites for **(a) discrete grouping** and **(b) continuous grouping** across site, respectively, varying scale of global heterogeneity. Estimators ending with "-oracle" makes use of ground truth treatment effects. Different colors imply different estimators, and x-axis, i.e., the value of c, differentiates the scale of global heterogeneity. The red dotted line denotes an MSE ratio of 1. MA performance is truncated due to large MSE ratios. The proposed ET and EF achieve competitive performance compared to standard model averaging or ensemble methods and are robust to heterogeneity across settings. Note that ET-oracle and EF-oracle achieve close-to-zero MSE ratios with very small spreads in some settings.

Figure 12. Box plots of the MSE ratios of CATE estimators, respectively, over LOC (**CF**) and a sample size of **100** at each site for **(a) discrete grouping** and **(b) continuous grouping** across site, respectively, varying scale of global heterogeneity. Estimators ending with "-oracle" makes use of ground truth treatment effects. Different colors imply different estimators, and x-axis, i.e., the value of c, differentiates the scale of global heterogeneity. The red dotted line denotes an MSE ratio of 1. MA performance is truncated due to large MSE ratios. The proposed ET and EF achieve competitive performance compared to standard model averaging or ensemble methods and are robust to heterogeneity across settings. Note that ET-oracle and EF-oracle achieve close-to-zero MSE ratios with very small spreads in some settings.
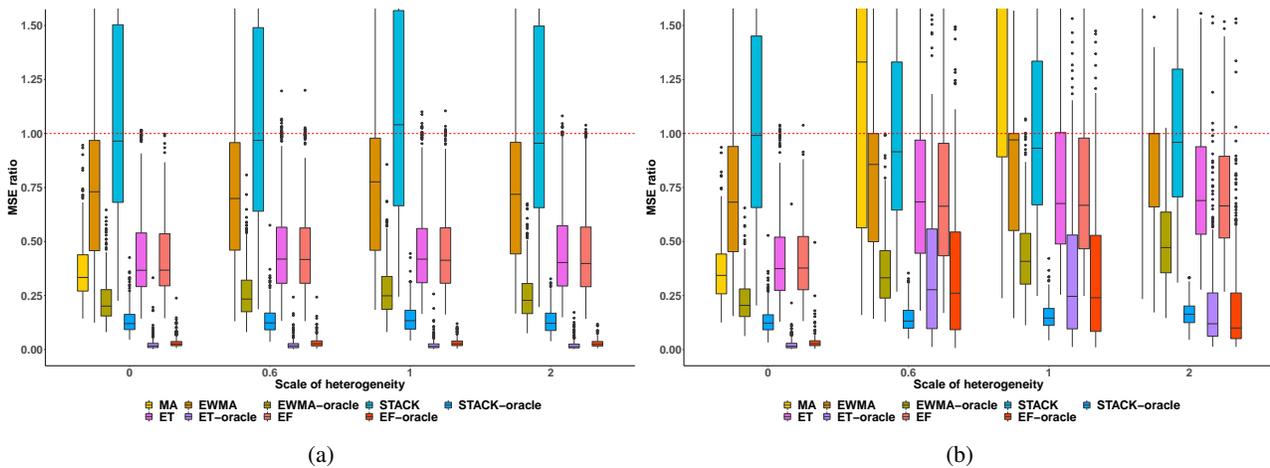

Figure 13. Box plots of the MSE ratios of CATE estimators, respectively, over LOC (**CF**) and a sample size of **500** at each site for **(a) discrete grouping** and **(b) continuous grouping** across site, respectively, varying scale of global heterogeneity. Estimators ending with "-oracle" makes use of ground truth treatment effects. Different colors imply different estimators, and x-axis, i.e., the value of c, differentiates the scale of global heterogeneity. The red dotted line denotes an MSE ratio of 1. MA performance is truncated due to large MSE ratios. The proposed ET and EF achieve competitive performance compared to standard model averaging or ensemble methods and are robust to heterogeneity across settings. Note that ET-oracle and EF-oracle achieve close-to-zero MSE ratios with very small spreads in some settings.
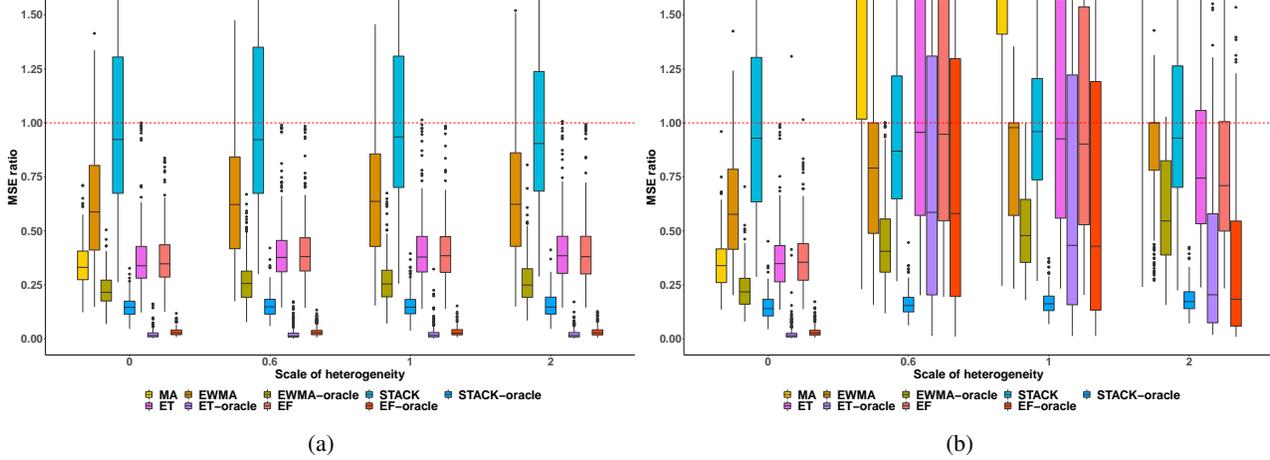
Figure 14. Box plots of the MSE ratios of CATE estimators, respectively, over LOC (**CF**) and a sample size of **1000** at each site for **(a) discrete grouping** and **(b) continuous grouping** across site, respectively, varying scale of global heterogeneity. Estimators ending with "-oracle" makes use of ground truth treatment effects. Different colors imply different estimators, and x-axis, i.e., the value of c, differentiates the scale of global heterogeneity. The red dotted line denotes an MSE ratio of 1. MA performance is truncated due to large MSE ratios. The proposed ET and EF achieve competitive performance compared to standard model averaging or ensemble methods and are robust to heterogeneity across settings. Note that ET-oracle and EF-oracle achieve close-to-zero MSE ratios with very small spreads in some settings.
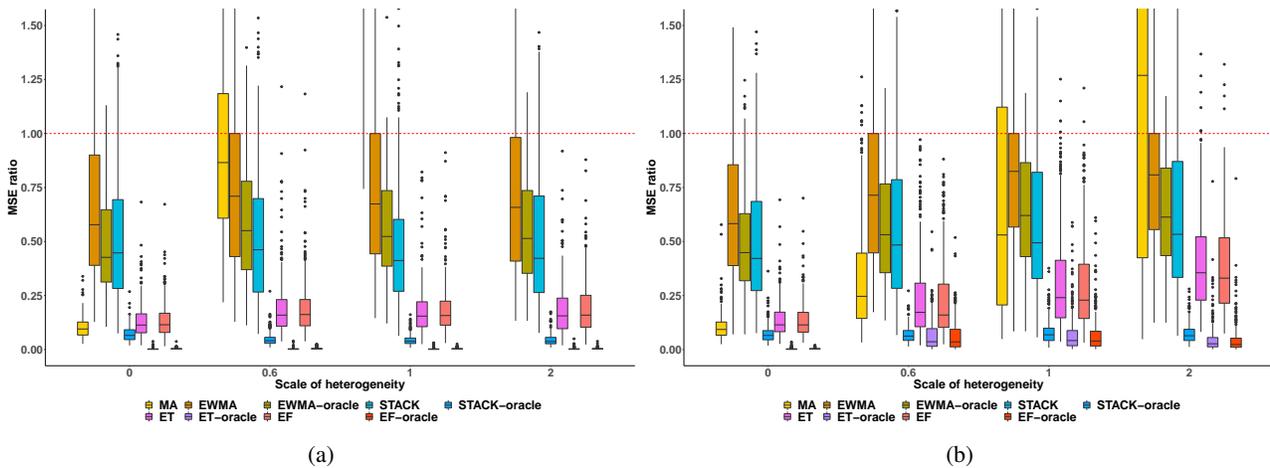

Figure 15. Box plots of the MSE ratios of CATE estimators, respectively, over LOC (**CT**) and a sample size of **500** at each site under **observational design with a correctly specified propensity score model** for **(a) discrete grouping** and **(b) continuous grouping** across site, respectively, varying scale of global heterogeneity. Estimators ending with "-oracle" makes use of ground truth treatment effects. Different colors imply different estimators, and x-axis, i.e., the value of c, differentiates the scale of global heterogeneity. The red dotted line denotes an MSE ratio of 1. MA performance is truncated due to large MSE ratios. The proposed ET and EF achieve competitive performance compared to standard model averaging or ensemble methods and are robust to heterogeneity across settings. Note that ET-oracle and EF-oracle achieve close-to-zero MSE ratios with very small spreads in some settings.
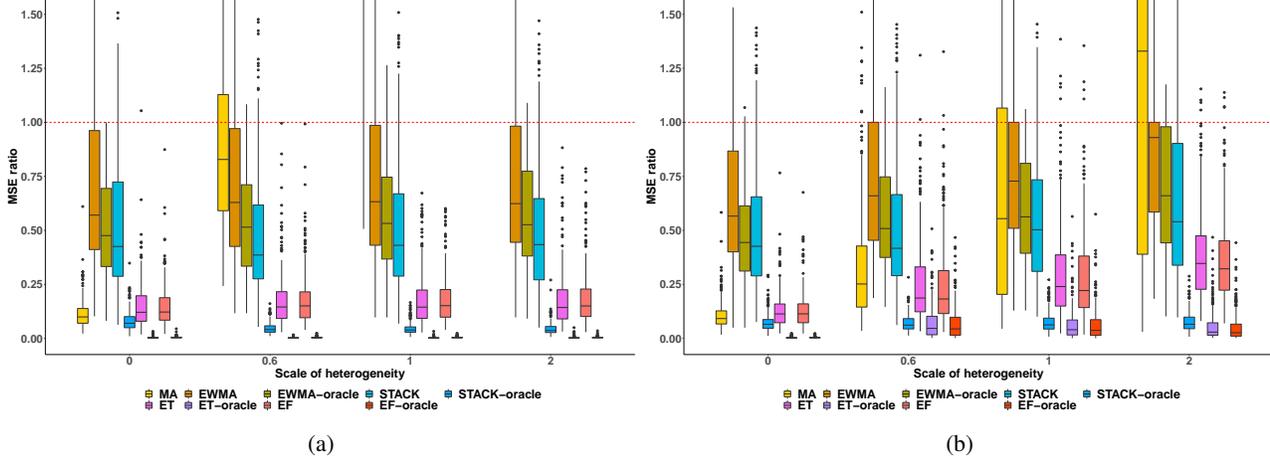
| | (a) | | | (b) |

Figure 16. Box plots of the MSE ratios of CATE estimators, respectively, over LOC (**CT**) and a sample size of **500** at each site under **observational design with a misspecified propensity score model** for **(a) discrete grouping** and **(b) continuous grouping** across site, respectively, varying scale of global heterogeneity. Estimators ending with "-oracle" makes use of ground truth treatment effects. Different colors imply different estimators, and x-axis, i.e., the value of c, differentiates the scale of global heterogeneity. The red dotted line denotes an MSE ratio of 1. MA performance is truncated due to large MSE ratios. The proposed ET and EF achieve competitive performance compared to standard model averaging or ensemble methods and are robust to heterogeneity across settings. Note that ET-oracle and EF-oracle achieve close-to-zero MSE ratios with very small spreads in some settings.

Table 3. Hospital-level information of our analysis cohort in eICU database. Hospitals are relabeled according to their average contribution to the estimation task at hospital 1, the target site.

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

Decision rule	Avg. survival rate
Baseline	0.762
LOC-based	0.772
EF-based	0.801

(a)
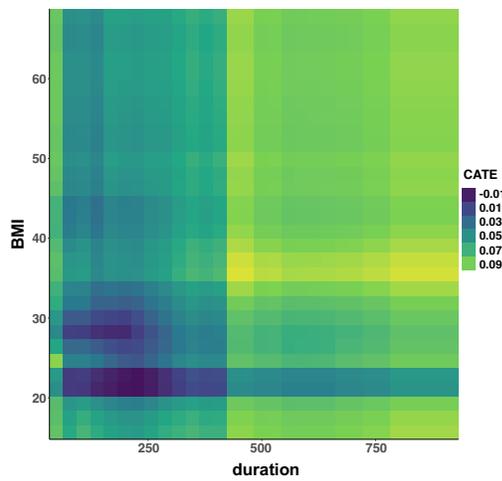
(b)
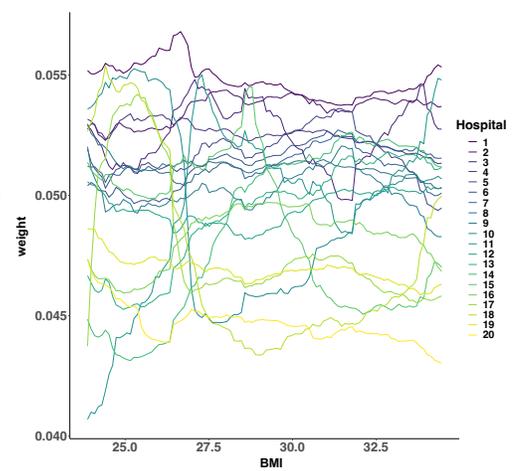
(c)

Figure 17. Application to estimating treatment effects of oxygen therapy on survival with a **sample size weighting strategy**. (a) Expected survival of treatment decision following different estimators. EF shows the largest gain in improving survival rate, more promising than the LOC and the baseline. (b) Estimated treatment effects varying duration and BMI, two important features in the fitted EF. Patients with a BMI around 35, and a duration above 400 benefited the most. (c) Visualization of data-adaptive weights in EF varying BMI. Hospitals with a larger bed capacity tend to contribute more, the data of which might be more similar to hospital 1.