A Tree-based Federated Learning Approach for Personalized Treatment Effect Estimation from Heterogeneous Data Sources

Xiaoqing Tan
Department of Biostatistics, University of Pittsburgh, Pittsburgh, PA 15261
e-mail: xit31@pitt.edu

and

Chung-Chou H. Chang
Department of Medicine and Department of Biostatistics, University of Pittsburgh, Pittsburgh, PA 15261
e-mail: changj@pitt.edu

and

Lu Tang
Department of Biostatistics, University of Pittsburgh, Pittsburgh, PA 15261
e-mail: lutang@pitt.edu

Summary: Federated learning is an appealing framework for analyzing sensitive data from distributed health data networks due to its protection of data privacy. Under this framework, data partners at local sites collaboratively build an analytical model under the orchestration of a coordinating site, while keeping the data decentralized. However, existing federated learning methods mainly assume data across sites are homogeneous samples of the global population, hence failing to properly account for the extra variability across sites in estimation and inference. Drawing on a multi-hospital electronic health records network, we develop an efficient and interpretable tree-based ensemble of personalized treatment effect estimators to join results across hospital sites, while actively modeling for the heterogeneity in data sources through site partitioning. The efficacy of our method is demonstrated by a study of causal effects of oxygen saturation on hospital mortality and backed up by comprehensive numerical results.

Key words: Conditional average treatment effect; Data integration; Heterogeneous data sources; Individualized treatment rule; Meta-analysis.

1. Introduction

Data integration approaches have received wide attention in recent years because distributed data networks are becoming a commonplace of modern data architecture. A distributed data network allows scalable storage of huge data sets without requiring all data to be physically stored in one central location. Compared to the classic centralized storage scheme, such a distributed data storage framework offers a “cloud” solution to allow more efficient management and maintenance of data in a collaborative environment. Analytically, new statistical and machine learning methods have been proposed to handle such data. They focus on restricting the amount of data exchanged between data sites in order to be communication efficient and to respect the privacy of proprietary data, while suffering certain loss of statistical efficiency (Chen et al., 2020).

There are two main types of construct of a distributed data: randomized versus nonrandomized. In the randomized case, collected data are actively randomized and allocated to local sites (Martalo et al., 2011), rendering a homogeneous and identically distributed data network. On the other hand, when a central randomization is not affordable, the nonrandomized case considers local data that are heterogeneously distributed across the network. We study the more challenging latter case. Specifically, we draw our focus on distributed health data networks (DHDNs) where sensitive patient-level data are collected and stored locally among sites, and propose an analytical framework to integratively analyze such heterogeneous data, with the goal of improving statistical estimation and prediction. DHDNs consist of mainly observational health data where there is no control over how data are distributed. Well known examples include FDA’s Sentinel (Platt et al., 2018) initiative for monitoring product safety and PCORI’s PCORnet (Fleurence et al., 2014) initiative for clinical research, both are DHDNs that operate on a national scale.

The analytical counterpart of distributed data is federated learning (Konecný et al., 2016; Yang et al., 2019). Federated learning enlists local sites to collaboratively build an analyti-
In this paper, we propose a tree-based federated learning approach to account for data heterogeneity in DHDNs, with the goal of improving estimation and prediction of personalized treatment effects of site 1 data and revealing potential heterogeneous structure among sites. Our approach generalizes federated learning to handle heterogeneously distributed data. To bridge a gap between existing methods that assumes homogeneity across sites and real data scenarios that are mostly heterogeneous, we model site-level heterogeneity through site-wise partitioning. The coordinating site in the DHDN does not naively combine models obtained from all sites, but assesses the heterogeneity across sites so that only models from sites that are more similar are merged. In the paper, similarity across sites is defined based on the CATEs conditioning on subgroups, where the subgroups are defined by covariates in the population of interest. For example, the treatment effects in two sites may be similar for female patients, but quite different for male patients, and requires us to consider borrowing information across sites only on selective subgroups. We first build localized models for the CATE estimation within each site. These localized models are then passed to a coordinating site to assess the heterogeneity across sites based on an augmented data constructed only from data in the coordinating site. A final tree-based model is then fit on the augmented data, with treatment effect as the outcome on patient characteristics and site indicator. The conceptual diagram is illustrated in Figure 2, and the rationale will be further explained in Section 3.

We organize the paper as follows. A short problem motivation will be given in Section 2 with a brief introduction of the DHDN under consideration. We describe the rationale of the proposed method in details in Section 3 and estimators for comparison in Section 4. The performance of proposed method are demonstrated in simulation experiments and a data application in Section 5 and Section 6, respectively. We highlight the key points of our work and its future extensions in Section 7.

2. Data Motivation

Electronic health records (EHR) collected in intensive care units (ICUs) possess a wealth of information on health conditions, treatment, and outcomes of critically ill patients. In ICU patients with cardiorespiratory complications, blood oxygen saturation should be given extra attention. Although low blood oxygen levels are generally considered to be harmful, recent studies have shown that a high level of blood oxygenation could further jeopardize the health of patients. Asfar et al. [2017] Geoghegan et al. [2018]. The British Thoracic Society recommends oxygen therapy of 94-98% pulse oximetry-derived oxygen saturation (SpO2) might optimize survival for most severely ill patients O’driscoll et al. [2017] Choudhury et al. [2018]. An earlier retrospective study found that the lowest mortality was observed at a SpO2 between 94% and
98% among patients requiring oxygen therapy [van den Boom et al., 2020]. This study considered a mixed-effects model with a random intercept to mitigate biases due to differences across hospitals, but failed to account for such biases when optimal treatment rules are drawn. So far, the optimal oxygenation target is still unclear and more studies of oxygen therapy are needed, especially for critically ill ICU patients (de Graaff et al., 2011; Itagaki et al., 2015; Suzuki et al., 2013). Driven by such motivation, we draw on an observational EHR data and a comparable cohort of patients with oxygen therapy from the eICU Collaborative Research Database (eICU-CRD), a multi-hospital ICU database made available by Philips Healthcare (Pollard et al., 2018). Our goal is to investigate the CATE of SpO2 within the range of 94% to 98% during oxygen therapy on hospital mortality.

Different from earlier works that operated on a large pooled observational cohort, we assume individual-level data from different hospitals cannot be merged, hence the setting where federated learning is required. From a method development perspective, the convenience of using the open eICU-CRD data is that data of all sites are accessible, making it easy to operate for validation purposes. Additionally, heterogeneity across hospitals is prevalent (Sheikhalishahi et al., 2020). Unobserved factors could result in differences in hospital infrastructures, healthcare coverage affected by economic levels, or cultural factors such as trust in medical systems (Brookhart et al., 2010), among others, all of which can be related to both treatment and outcomes. Therefore, they act as unobserved confounders. By viewing hospitals as sites in a distributed health data network, our goal is to define “closeness” between hospitals and selectively borrow information from “neighbors” to increase the power of CATE estimation. We hypothesize that the unobserved factors are correlated with sites in a way that we can adjust for potential unobserved factors through a subgrouping on the sites. For example, patient comorbidity is typically not available for ICU patients upon admission, yet it is related to hospital types, which is associated with both the treatment and the outcome, and should be adjusted for in CATE estimation. In this case, the site indicator may partially reveal comorbidity differences if patients seek treatments according to their respective health conditions and corresponding hospital reputation, etc. Hence, it allows us to adjust for such unobserved confounding through subgrouping of sites. Our analysis shows there is strong heterogeneity across sites compared to the heterogeneity within sites. See Section 5 for the full details of this application.

3. Tree-based Federated Learning

3.1 Rationale

Given patient feature \(\mathbf{x}\), we define the similarity between two sites, \(A\) and \(B\), directly based on the target of inference, in this case a CATE estimate, \(\tau_A(\mathbf{x})\) and \(\tau_B(\mathbf{x})\). This is more appealing than a similarity based on model coefficients, say, \(\beta_A\) and \(\beta_B\) for some \(\tau_k(\mathbf{x}) = \mathbf{x}^T \beta_k, k \in \{A, B\}\), which are indirect to the target of inference. For example, in a recent approach by Shen et al. (2020), fusion is based on the confidence distributions of model parameters \(\beta_k\) in the regression setting. What further distinguishes our work from earlier work is that we define the similarity of sites to be varying with \(\mathbf{x}\). Such an adaptive distance offers the flexibility to refine the inference targets by selectively borrowing information from other sites, depending on the patient characteristics.

The proposed tree-based federated learning framework consists of an ensemble tree model fitted across the entire DHDN, based on the localized models fit in each site. A tree model can be either a single tree or a forest [Breiman, 2001]. For personalized treatment effect estimation, localized models are constructed at each site to directly estimate CATE functions that are potentially different across sites. These estimated CATE functions are then passed to the coordinating site, say the first site, to estimate an ensemble tree or forest model that appropriately accounts for across-site heterogeneity based on data in site 1.

Using federated learning can be intuitively viewed as constructing a super learner that is more powerful and stable than individual learners [Zhang and Singer, 2010]. Existing tree ensemble methods such as bagging, random forest [Breiman, 2001], or boosting tree [Friedman, 2001] consider learners that are built from randomly sampled data sets, the main goal of which is to reduce variance and improve estimation. On the other hand, we consider a tree-based ensemble that, in addition, accounts for the between-site heterogeneity when localized learners are built on nonrandomly sampled data sets. The main goal of the constructed model ensemble is to improve personalized treatment effect estimation and prediction within sites in the existing DHDN, but not to generalize the conclusion to data outside of the network.

3.2 Federated Learning of Personalized Treatment Effects

Let \(Y\) denote the outcome of interest, \(Z\) denote a binary treatment indicator, and \(X\) denote patient characteristics. Correspondingly, let \(y\), \(z\) and \(\mathbf{x}\) denote realizations of the random variables. For ease of exposition, we present treatment effects through the potential outcome framework [Neyman, 1923; Rubin, 1974]. The classic CATE is defined as the difference in mean for the counterfactual outcomes between two treatment groups for individuals with characteristics \(X\), that is, \(\tau(\mathbf{x}) = E[\text{Y}(\text{Z}=1) - \text{Y}(\text{Z}=0)]|X = \mathbf{x}\), where \(\text{Y}(\text{Z}=1)\) and \(\text{Y}(\text{Z}=0)\) are the counterfactual outcomes under treated \(Z = 1\) and control \(Z = 0\), respectively. In the case of heterogeneous DHDN, we obtain from site \(k\) the CATE function \(\tau_k(\mathbf{x}) = E_k[\text{Y}(\text{Z}=1) - \text{Y}(\text{Z}=0)]|X = \mathbf{x}\), where the expectation is taken over the population in that site, for \(k = 1, \ldots, K\) and \(K\) is the total number of sites. We also denote the sample size of site \(k\) as \(n_k\), and the total sample size as \(\sum_{k=1}^{K} n_k = N\). Based on a modified unconfoundedness assumption we will introduce, we have that \(\tau_k(\mathbf{x}) = E_k[\text{Y}(\text{Z}=1) - \text{Y}(\text{Z}=0)]|X = \mathbf{x}, S = k\), where we introduce \(S\) as the site indicator which can take values \(1, \ldots, K\). Consider the data setup \(D_i = \{U_i, S_i, X_i, Z_i, Y_i\}, i = 1, \ldots, N\), where for subject \(i\), \(U_i\) is the unknown underlying grouping, \(S_i\) is the study site that the subject comes from, \(X_i\) is the observed feature vector of the subject, \(Z_i\) is the treatment assignment, and \(Y_i\) is the observed outcome.

Assumptions. Heterogeneity of treatment effect can be due to unobserved confounding, posing a challenge to the classic assumptions. Given \(\mathbf{x}\), \(\tau_k(\mathbf{x})\) obtained from site \(k\) may not be the same as \(\tau_{k'}(\mathbf{x})\) from another site \(k' \neq k\). Thus,
for subject $i$, the standard unconfoundedness assumption $\{Y_i^{(2=0)}, Y_i^{(2=1)}\} \perp Z_i | X_i$, [Rosenbaum and Rubin 1983] is not sufficient to guarantee estimation consistency. To proceed with heterogeneity across sites, we require the treatment assignment $Z_i$ to be independent of the potential outcomes for $Y_i$ conditional on both $X_i$ and an unobserved confounder $U_i$:

$$\{Y_i^{(2=0)}, Y_i^{(2=1)}\} \perp Z_i | X_i, U_i,$$

and that given additional information on site, $S_i$, treatment assignment is randomly assigned and independent of the unobserved factor $U_i$ that causes heterogeneity:

$$\{Y_i^{(2=0)}, Y_i^{(2=1)}, U_i\} \perp Z_i | X_i, S_i.$$

This means that we can take observations with similar $x$ among comparable sites from $\{1, \ldots, K\}$, and assume that this group comes from a randomized experiment. It also means that site similarity can be used to infer $U_i$, hence can be used to adjust for unobserved confounding. We refer to this as the modified unconfoundedness assumption. This is similar to the commonly known subgroup analysis, except that the subgroups are dependent on not only $X$, but also $S$. In other words, the main source of unobserved heterogeneity is reflected in the differences between sites, not within sites. Comparable sites are those that are more similar among all sites in a DHDN. In addition, we need standard positivity assumptions:

$$0 < P(S_i = k | X_i) < 1, k = 1, \ldots, K,$$

and

$$0 < P(Z_i = 1 | X_i, S_i) < 1,$$

for all $X_i$ and $S_i$ such that all subjects are possible to be observed in all sites, and all subjects in all sites are possible to be receiving either arm of treatment.

**Local model: estimation of $\tau_k(x)$ at each site.**

Due to privacy considerations and lack of resources, for most of the time a site can only rely on its own data for estimating treatment effects, without leveraging data from other sites. Recently published methods in directly estimating individualized treatment effects include causal tree [Athey and Imbens 2016], causal forest [Wager and Athey 2018], Bayesian regression tree models [Hahn et al. 2020], and deep neural networks [Farrell et al. 2021], etc. On the other hand, meta-learners such as T-learner, S-learner, X-learner [Künzel et al. 2019], and R-learner [Nie and Wager 2020] that indirectly estimate treatment effects with one or more models of the observed outcome have become increasingly popular.

In our paper, we consider the following methods in estimating $\tau_k(x)$ within a given site $k$: causal tree (CT), causal forest (CF), and X-learner with Bayesian additive regression trees as base learners (XL). Despite the many choices of methods to estimate the local causal effect models, we opt for non-linear tree-based learners due to the following reasons: (i) they allow different types of outcome such as discrete and continuous outcome, hence can be applied to a broad range of real data scenarios; (ii) they can easily handle a large number of covariates and high order interactions by construction. A causal forest is a stochastic averaging of multiple causal trees [Athey and Imbens 2016]. In each tree, the mean squared error (MSE) of treatment effect $\tau$ is used to select the feature and cutoff point in each split. Causal trees can be applied to observational studies by applying methods such as propensity score weighting to modify estimates within leaves [Athey and Imbens 2016]. Causal forest builds deep trees with small leaves without needing to explicitly estimate the propensity [Wager and Athey 2018]. Propensity score estimates could also be incorporated for improved robustness by fitting a separate regression forest as a sub-routine before fitting a causal forest [Athey et al. 2019]. X-learner first models outcome functions $E[Y^{X=x}|X=x, S=k]$ and $E[Y^{X=x}|X=x, S=k]$ separately to obtain the CATEs among the treated subjects and CATEs on the control subjects with base learners, and then apply propensity-weighting for these two estimates to generate CATE estimate for a new data point $x$ [Künzel et al. 2019]. We choose Bayesian additive regression trees as base learners since they tend to be robust to hyperparameter tuning and small data sets [Künzel et al. 2019]. The causal tree, causal forest, and X-learner can be implemented with the R packages causalTree, grf, and hte, respectively.

**Remark 1:** Note that despite our choice of the localized learners in estimating $\tau_k(x)$, the ensemble framework can be applied to any general estimator of $\tau_k(x)$.

**Federated model: $\{\hat{\tau}_k(x)\}_{k=1}^K$ ensemble using site 1 data.**

Aggregation has been a central idea behind model ensemble, especially for trees. A key advantage of an aggregation scheme is to reduce variance and smooth sharp decision boundaries [Bühlmann and Yu 2002]. See for example Breiman (2001), which considers aggregating a forest of $B$ trees, where each is denoted as $\hat{\tau}_k(x)$. The aggregated estimate is a direct average, $\frac{1}{B}\sum_{b=1}^{B}\hat{\tau}_b(x)$. It requires all $B$ trees use random samples of the original sample. The random forest uses the same aggregation on random subsamples, but also selects a random subsets of features to construct each tree. Taking a turn from the above idea, we consider an ensemble of $\{\hat{\tau}_k(x)\}_{k=1}^K$, which are estimated from a nonoverlapping data partitions that are not random. To model heterogeneity, we include a categorical site indicator when building an ensemble, so that it can actively adjust for heterogeneity across sites. Since individual-level data cannot be pooled, we adopt federated learning and construct the ensemble using only data from a single coordinating site, say, site 1. The $K$ tree-based models from local sites, $\{\hat{\tau}_k(x)\}_{k=1}^K$, are passed into the first site to get $K$ treatment effect estimates for each subject in site 1, resulting in a new augmented site 1 data:

$$D_{aug,1} = \{D_{1,k} = [X_i, k, \hat{\tau}_k(X_i)], i : S_i = 1, k = 1, \ldots, K\}.$$

Using this data, an ensemble can be trained by either a regression tree or a random forest [Breiman et al. 1984; Zhang and Singer 2010], with those estimated treatment effects and CATE as the outcome. A site indicator of which localized model is used along with the patient characteristics are fed into the federated model as covariates.

In the ensemble, the estimated treatment effect can be represented as $\tilde{\tau}(x, s)$, which depends on both $x$ and site $s$. Let $L(x, s)$ denote the leaf node corresponding to $x$ and $s$. The ensemble tree (ET) estimate based on the augmented
site 1 data can be derived by

\[
\hat{\tau}_{ET}(x, s) = \frac{1}{\sum_{(i,k): S_i = 1, (X_i, k) \in \mathcal{L}(x, s)} \tau_k(X_i)} \sum_{(i,k): S_i = 1, (X_i, k) \in \mathcal{L}(x, s)} \tau_k(X_i).
\]

This represents the average difference between \( Y | Z = 1 \) and \( Y | Z = 0 \) at the leaf node \( \mathcal{L}(x, s) \). Intuitively, observations with similar characteristics \((x, x')\) and from similar sites \((s, s')\) are more likely to fall in the same leaf node in the ensemble tree, i.e., \((x, s) \in \mathcal{L}(x', s')\) or \((x', s') \in \mathcal{L}(x, s)\).

Therefore, our estimator attempts to borrow information from neighbors in the space of \( X \) and \( S \). Due to our assumptions stated above, we can adjust for potential unobserved confounding factor \( U \) by adjusting for \( S \) through the partitioning over sites. The splits of the tree are based on minimizing in-sample MSE of \( \tau \) within each leaf and pruning by cross-validation over choices of the complexity parameter. Since a single ensemble tree model could be unstable, we can also rely on the forest version of ensemble tree to reduce variance and smooth the partitioning boundaries. An ensemble forest (EF) can be constructed by aggregating many ensemble trees to smooth the boundaries between sites while still benefit from information borrowing across sites. The EF estimate is

\[
\hat{\tau}_{EF}(x, s) = \frac{1}{B} \sum_{b=1}^{B} \tau_b(x, s),
\]

where the form of \( \hat{\tau}_b(x, s) \) closely follows \([1]\) but is based on a subsample of \( \mathcal{D}_{aug, 1} \). The total number of trees in a forest is set at \( B = 2000 \) throughout the paper. The above ensemble approaches could be implemented with R packages \texttt{rpart} and \texttt{ranger}. The diagram for our proposed algorithm is shown in Figure 2 where site 1 serves as the coordinating site. Once the federated model is obtained, we may broadcast it back to local sites for deployment or further validation.

We provide an implementation of our proposed learning strategy with ensemble tree and ensemble forest in Algorithm 1. In Section 4, we further differentiate the above ensemble methods into two variations depending on whether the training data are used for both tree building and tree estimation.

**Remark 2:** With focus on just a single site, say site \( s \), the proposed ensemble estimators \( \hat{\tau}_{ET}(x, s) \) and \( \hat{\tau}_{EF}(x, s) \) can be regarded as improved versions of the local estimator \( \hat{\tau}_s(x) \) through leveraging information from other sites. How much of information being shared is dependent on the ensemble tree construct, with a hard neighborhood assignment in a tree where two sites are either neighbors or not neighbors, and a soft neighborhood assignment in a forest where sites can be neighbors with varying “distances”. Note that the neighborhood definition is dependent on \( x \).

### 3.3 Connection with Existing Federated Learning Methods

The most straightforward way to do data integration without pooling individual-level data is direct aggregation, that is, taking the average of the mean estimates from the \( K \) localized models of each site \([Lenzerini, 2002]\). It falls in the federated learning category, however, heterogeneity across sites is being ignored. Some other federated learning methods require multiple rounds of communication, which may be prohibitive in collaborative studies. See for example \([Jordan et al., 2019]\).

Another popular approach is the meta-analysis by combining summary statistics from each site \([Borenstein et al., 2011]\). Unfortunately, this only provides population average conclusions, and cannot provide answers to subpopulation-driven questions, which could be crucial to personalized medicine.

Random effects models may also fail to converge when random effects occur at treatment interaction or higher-order terms \([Borenstein et al., 2011]\), as in the case when CATE is of interest. Especially, a random effects model becomes numerically unstable when the number of random effects is large.
Grueber et al. (2011). Also, it is a model-based method that requires the normality assumption and might not be flexible compared to nonparametric methods when the heterogeneity structure is unknown. Other feasible approaches for heterogeneous DHDN involve integrating likelihood functions (Tang and Song, 2016) and integrating decision models (Qiu, 2018, Chapter 4), but they require pooling individual-level data.

The conditional treatment effect is of great interest to precision health and personalized medicine. By conditioning, average treatment effects can be different for different population subgroups. Usually, studies are only powered for average treatment effects, but not conditional treatment effects. Inference for conditional treatment effect often suffers from low power because of sample attrition when constructing study cohorts (Assmann et al., 2000). And it is often unclear which subgroup is of interest. This type of question can nicely leverage the benefit of data integration, by pooling information together. Borrowing information from other similar studies can increase power and reduce variance. Our proposed federated learning framework could be generalized to any methods of estimating localized treatment effects beyond tree-based models that are currently considered.

4. Federated Estimators for Heterogeneous Data Sources

We present two versions of our proposed federated methods using adaptive estimation and honest estimation respectively with regards to sample partitioning as in Athey and Imbens (2016) and Athey and Wager (2018). Adaptive approaches use the same training data for both model selection (in our case, building partitions of the feature space) as well as estimation based on a model structure. This approach could lead to biases due to the fact that the selected model could be affected by potential spurious correlations between covariates and outcomes. Fortunately, as the sample size grows, biases disappear and give rise to a small estimation error. Alternatively, so-called 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 of the tree. Although the honest version enjoys unbiased estimation, it could suffer from low precision as it only uses half of the training sample in each step of estimation in case of finite sample size (Athey and Imbens, 2016).

The discussion in Section 3.2 develops our estimators for the ensemble regression tree and the ensemble regression forest model. Using CT, CF, XL as localized models separately, we denote the corresponding ensemble tree estimator as CT-, CF-, XL-ET, and the corresponding ensemble forest estimator as CT-, CF-, XL-meta, respectively. We consider several other methods for comparison purposes. A baseline approach for federated learning is simply a direct aggregation. Specifically, treatment effect estimates from sites are directly combined to form the final estimates. We denote the aggregated estimator based on different localized models to be CT-, CF-, XL-agg, respectively, which is derived as

\[
\hat{\tau}_{agg}(x) = \frac{1}{K} \sum_{k=1}^{K} \hat{\tau}_k(x),
\]

where \(\hat{\tau}_k(x)\) is the local estimator from site \(k\). Note that heterogeneity across sites is ignored. Another alternative method is based on a random effects meta-analysis model where site-specific variations in \(\tau(x)\), i.e., \(u_{k,x} = \tau_k(x) - \tau(x), k = 1, \ldots, K\), are considered as random following a normal distribution, \(u_{1,x}, \ldots, u_{K,x} \sim N(0, \sigma^2(x))\), at a given \(x\). Localized models from sites are passed to site 1 to get treatment effect estimates and the corresponding standard errors from all models for the site 1 subjects. Those estimates are then combined by a random effects meta-analysis with a random site effect. We denote the meta-analysis estimator based on different localized models to be CT-, CF-, XL-meta, respectively.

In contrast to federated learning approaches that assume individual-level data cannot be shared across sites, we also consider the hypothetical analysis when centralized data are available taking advantage of simulation study. Following Athey and Wager (2019), we build a centralized model using cluster-robust causal forest for comparison as well. This type of non-parametric random effects modeling is flexible in that we do not need to specify the distribution of random site effects. Different from the random forest that directly draw subsamples of observations, cluster-robust causal forest first draws a subsample of clusters and then draws a subsample of observations randomly from each cluster (Athey and Wager, 2019). We denote this centralized approach as clustCF and it could be implemented with the R package grf.

A summary of the characteristics of our proposed federated models and other methods on estimating CATEs for heterogeneous data sources is provided in Table 1.

5. Simulation Study

A simulation study is conducted to assess the performance of the proposed method. We assume there are \(K = 20\) sites in total, each with a sample size \(n = 100\). We specify \(m(x, k)\) for the mean effect of individuals in site \(k\) with features \(x\), and \(\tau(x, k)\) for the treatment effect. We design settings with marginal treatment probability \(P = 0.5\). The potential outcomes can be written into

\[
Y_i(z) = m(X_i, S_i) + 1/2 \cdot (2z - 1) \cdot \tau(X_i, S_i) + \epsilon_i,
\]

where \(z = 0, 1\) and \(\epsilon_i \sim N(0,1)\). Denote the number of features as \(D\). Features \(X_i\) are assumed to be independent of \(\epsilon_i\), and \(X_i \sim N(0, I)\). Assume there are two potential underlying groupings among the \(K\) sites (group 1: \(k = 1, 3, \ldots, K - 1\); group 2: \(k = 2, 4, \ldots, K\)), we have the below four simulation designs.

(1) Null treatment effect design:

\[D = 3; \quad m(x, k) = 0; \quad \tau(x, k) = -4 \cdot 1\{k \text{ mod } 2 = 0\}.\]

(2) Complex non-linear design:

\[D = 4; \quad m(x, k) = 0; \quad \tau(x, k) = \epsilon(x_1) \cdot \epsilon(x_2) - 4 \cdot 1\{k \text{ mod } 2 = 0\} + x_1 \cdot 1\{k \text{ mod } 2 = 0\}; \quad \epsilon(x) = \frac{1}{1 + e^{-x^2/17}}.\]

(3) Simple piece-wise linear design:

\[D = 5; \quad m(x, k) = \frac{1}{2} x_1 + \sum_{d=2}^{4} x_d - 4 \cdot 1\{k \text{ mod } 2 = 0\} + x_1 \cdot 1\{k \text{ mod } 2 = 0\}.\]
Table 1: Comparison of characteristics of our proposed federated models with other methods on analyzing heterogeneous data sources.

| Proposed federated models | Not require pooling data from all sites | Borrow information from other sites | Consider heterogeneity across sites | Reveal heterogeneity site patterns | Focus on each individual site |
|---------------------------|----------------------------------------|-------------------------------------|-----------------------------------|----------------------------------|------------------------------|
| Local model               | ✓                                      | ×                                   | ×                                 | ×                                | ✓                             |
| Causal forest             | ×                                      | ✓                                   | √                                 | ×                                | √                             |
| Meta-analysis             | √                                      | √                                   | √                                 | √                                | ×                             |
| Mean aggregation          | √                                      | √                                   | ×                                 | √                                | ×                             |

\[ \tau(x, k) = \mathbb{1}\{ x_1 > 0 \} \cdot x_1 - 4 \cdot \mathbb{1}\{ k \text{ mod } 2 = 0 \} + x_1 \cdot \mathbb{1}(k \text{ mod } 2 = 0). \]

(4) Complex piece-wise linear design:
\[
D = 8; \quad m(x, k) = \frac{1}{2} \sum_{d=1}^{2} x_d + \sum_{d=3}^{5} x_d - 4 \cdot \mathbb{1}(k \text{ mod } 2 = 0) + x_1 \cdot \mathbb{1}(k \text{ mod } 2 = 0);
\]
\[
\tau(x, k) = \sum_{d=1}^{2} \mathbb{1}\{ x_d > 0 \} \cdot x_d - 4 \cdot \mathbb{1}(k \text{ mod } 2 = 0) + x_1 \cdot \mathbb{1}(k \text{ mod } 2 = 0).
\]

Similar to designs in Athey and Imbens (2016), Wager and Athey (2018), and Kunzel et al. (2019), in each design, not all covariates affect the outcome or the treatment effect. Those are so-called “noise” covariates. Covariates in \( \tau(x, k) \) are predictive markers while covariates in \( m(x, k) \) but not in \( \tau(x, k) \) are prognostic only. Design 1 presents a null situation where all covariates are noise and none are predictive of treatment effects. We consider complex linear and non-linear scenarios in design 2, 3, and 4, where sites from the same underlying groupings are designed to have similar treatment effects and mean effects, while sites from different underlying groupings tend to have various treatment effects and mean effects. Here, odd sites and even sites form two distinct groups.

We compare the proposed federated methods with localized model and centralized model as well as methods discussed in Section 4 in terms of bias and MSE of the estimated treatment effects of site 1 data. Specifically, we report the empirical bias and MSE over an independent test samples of a sample size \( n \) from site 1 where
\[
\text{Bias}(\hat{\tau}) = \frac{1}{n} \sum_{i=1}^{n} \hat{\tau}(x_i, k = 1) - \tau(x_i, k = 1)
\]
\[
\text{MSE}(\hat{\tau}) = \frac{1}{n} \sum_{i=1}^{n} (\hat{\tau}(x_i, k = 1) - \tau(x_i, k = 1))^2
\]

Among the choices of localized models, XL has the overall best performance in all designs. Hence, we present simulation results that compare multiple methods on estimating CATEs with XL as the localized model in Table 2 because using XL as localized models achieve the best performance among all localized models. The proposed methods achieve lowest MSEs for adaptive approaches in nearly all designs, where, in contrast, honest approaches have comparatively larger MSEs due to training sample splitting. Among all the methods, the proposed federated models (CT-, CF-, XL-ET and CT-, CF-, XL-EF) enjoy the best performance, outperforming the aggregation approach, meta-analysis, and even the centralized robust approach with an overall 3-4 times smaller MSE when the sample size of each site is limited. CT-, CF-, XL-EF have the lowest MSE among all methods in the three designs. Comparing ET to EF, EF has an even smaller bias and MSE, which is expected because forest models tend to be more stable and accurate than tree models as a result of aggregation. When the sample size of each site is small, the proposed methods tend to have greater advantages than fitting local models stratified by sites (localCF), i.e., treating all sites to be heterogeneous but now allowing information sharing. We compare different localized models and their corresponding federated models in Table 3.

Figure 3 visualizes the results of design 2 using adaptive XL-ET and XL-EF method, respectively. Subfigure 3a plots the results of an ensemble tree in design 2. The site indicator appears to be the root of the tree, recovering the fact that the odd sites belong to one underlying subgroup, while the even sites belong to another subgroup. Features \( X_1 \) and \( X_2 \) are the other two variables appearing on the tree, which is consistent with the fact that only \( X_1 \) and \( X_2 \) are the only two patient characteristics used to simulate the treatment effect \( \tau(x, k) \) in design 2. Feature importance of EF is measured by the total decrease in node impurity defined by the variance reduction using MSE that results from splits over a feature, averaged over all trees in the forest. We scale each feature importance by their sum to obtain the corresponding relative importance. In the generated forest, the site indicator also appears to be the most important, with a relative importance of 77.2%. Features \( X_1 \) and \( X_2 \) take up 15% and 5.83%, respectively. Other covariates take up less than 1%. Figure 3b shows the treatment effect estimates in the ensemble forest for different values of \( X_1 \) in each site, different values of \( X_2 \) in each site, and different values of \( X_1 \) and \( X_2 \), respectively. The color of each grid is the estimated average causal estimates among subjects with a certain value of the covariate of interest in a certain site, with values of other covariates fixed at their expected value. The model is able to recover the underlying groupings across sites indicated by the distinct colors for odd sites and even sites. There is also treatment effect heterogeneity between subjects within each site as indicated by the color gradients of grids within sites, which is consistent with the ensemble tree results and also the ground truth.
Table 2: Simulation results comparing multiple methods with X-learner as the localized model.

| Design | Method | Honest estimation | Adaptive estimation |
|--------|--------|-------------------|---------------------|
|        |        | Bias x 100 | MSE | Bias x 100 | MSE |
| 1      | XL-EF  | 0.254 | 2.5E-04 | 0.324 | 3.0E-04 |
|        | XL-ET  | 0.626 | 4.4E-05 | 0.494 | 2.8E-05 |
|        | localXL| -     | -     | 3.05  | 1.7E-03 |
|        | clustCF| -87.3 | 0.349  | -52.8 | 0.336  |
|        | XL-meta| -     | -     | -200  | 3.997  |
|        | XL-agg | -     | -     | -200  | 3.985  |
| 2      | XL-EF  | -17   | 0.677  | -16   | 0.629  |
|        | XL-ET  | -32.2 | 1.333  | -33.7 | 1.337  |
|        | localXL| -     | -     | -17.7 | 0.868  |
|        | clustCF| -65.6 | 1.337  | -72.7 | 1.694  |
|        | XL-meta| -     | -     | -207  | 4.896  |
|        | XL-agg | -     | -     | -208  | 5.572  |
| 3      | XL-EF  | 0.577 | 0.078  | -2.38 | 0.087  |
|        | XL-ET  | 10.7  | 0.365  | 11.8  | 0.369  |
|        | localXL| -     | -     | -5.03 | 0.208  |
|        | clustCF| -76.6 | 1.319  | -19.5 | 3.156  |
|        | XL-meta| -     | -     | -167  | 2.973  |
|        | XL-agg | -     | -     | -179  | 3.586  |
| 4      | XL-EF  | 12.1  | 0.232  | 11.5  | 0.252  |
|        | XL-ET  | 11.4  | 0.438  | 12.6  | 0.438  |
|        | localXL| -     | -     | 28.1  | 0.526  |
|        | clustCF| -44   | 0.732  | -71.2 | 3.059  |
|        | XL-meta| -     | -     | -160  | 2.992  |
|        | XL-agg | -     | -     | -172  | 3.625  |

Table 3: Simulation results comparing multiple localized models and their corresponding federated models.

| Design | Method | Honest estimation | Adaptive estimation |
|--------|--------|-------------------|---------------------|
|        |        | Bias x 100 | MSE | Bias x 100 | MSE |
| 1      | XL-EF  | 0.254 | 2.5E-04 | 0.324 | 3.0E-04 |
|        | XL-ET  | 0.626 | 4.4E-05 | 0.494 | 2.8E-05 |
|        | localXL| -     | -     | 3.05  | 1.7E-03 |
|        | clustCF| -87.3 | 0.349  | -52.8 | 0.336  |
|        | XL-meta| -     | -     | -200  | 3.997  |
|        | XL-agg | -     | -     | -200  | 3.985  |
| 2      | XL-EF  | -17   | 0.677  | -16   | 0.629  |
|        | XL-ET  | -32.2 | 1.333  | -33.7 | 1.337  |
|        | localXL| -     | -     | -17.7 | 0.868  |
|        | clustCF| -65.6 | 1.337  | -72.7 | 1.694  |
|        | XL-meta| -     | -     | -207  | 4.896  |
|        | XL-agg | -     | -     | -208  | 5.572  |
| 3      | XL-EF  | 0.577 | 0.078  | -2.38 | 0.087  |
|        | XL-ET  | 10.7  | 0.365  | 11.8  | 0.369  |
|        | localXL| -     | -     | -5.03 | 0.208  |
|        | clustCF| -76.6 | 1.319  | -19.5 | 3.156  |
|        | XL-meta| -     | -     | -167  | 2.973  |
|        | XL-agg | -     | -     | -179  | 3.586  |
| 4      | XL-EF  | 12.1  | 0.232  | 11.5  | 0.252  |
|        | XL-ET  | 11.4  | 0.438  | 12.6  | 0.438  |
|        | localXL| -     | -     | 28.1  | 0.526  |
|        | clustCF| -44   | 0.732  | -71.2 | 3.059  |
|        | XL-meta| -     | -     | -160  | 2.992  |
|        | XL-agg | -     | -     | -172  | 3.625  |

To assist with interpretation, we calculate the best linear projector proposed by Cameron and Miller [2015] MacKinnon and White [1985] Semenova and Chernozhukov [2017] for the estimated treatment effects from XL-EF of site 1 data in design 2 by projecting $\hat{\tau}(x, 1)$ to the closest linear function of covariates. Specifically, a doubly robust augmented inverse propensity weighted score that incorporates information of both treatment and outcome of site 1 is calculated for each subject in site 1. Then a linear model is used to regress those doubly robust scores on the covariates. The coefficients of the covariates are 0.42, 0.44, 0.07, -0.11, respectively with $X_1$ and $X_2$ highly significant (corresponding p-values < 0.001) and $X_3$ and $X_4$ insignificant at the significance level of 0.05. We may interpret this as one unit increase in $X_1$, the treatment effect in site 1 increases a unit of 0.42, holding other covariates constant. These linear approximation results are consistent with the $\hat{\tau}(x, k)$ in design 2 where only $X_1$ and $X_2$ are relevant in estimating the treatment effects.

6. Data Application: eICU-CRD Distributed Hospital EHR Data Network

We apply our method to the public eICU-CRD, a multi-hospital electronic health records network. We are interested in estimating the causal effects of oxygen saturation SpO2 within the range of 94% to 98% on hospital mortality among critically ill patients with respiratory diseases and with at least 48-hour of oxygen therapy. A total of 12,626 ICU patients from 127 sites are obtained based on this criterion van den Boom et al. [2020]. We consider the treated arm to be the median SpO2 within the range of 94% to 98% and the control arm to be the median SpO2 outside of that range.

Our final cohort for analysis consists of 7,022 patients from 20 hospitals, each with at least 50 patients on each treatment arm. We use the same covariates as in van den Boom et al. [2020], which are age, body mass index (BMI), sex, Sequential Organ Failure Assessment (SOFA) score, and duration of oxygen therapy. We applied adaptive XL-ET and XL-EF methods to the cohort, where an XL was first fit in each site to get localized models. Those models were then passed into an ensemble regression tree and an ensemble random forest, respectively.

The site indicator appears to be the most important variable in the ensemble forest, with a relative importance taking up 58.8%. The relative importance of BMI, age, oxygen therapy duration, gender, and SOFA score are 9.37%, 9.36%, 8.03%, 7.31%, 7.1%, respectively. Figure 3 plots the estimation of treatment effects for different values of covariates in each site, respectively, with subfigures ranked by relative importance of the covariates. There are strong between-site heterogeneity followed by some within-site heterogeneity indicated by the different color gradients of the grids between and within sites. For better interpretation, the best linear projection of the estimated treatment effects of the coordinating site from XL-EF is mapped into the patient characteristics space. The coefficients from the doubly robust linear model fit are $0.002, -0.001, 0.0002, 0.13, 0.016$ for BMI, age, oxygen therapy duration, gender, and SOFA score, respectively, with none of these covariates significant under the significance level of 0.05. Similar linear projection of using localXL are conducted for the coordinating site. The coefficients of this model fit are similar to that of linear approximation using the ensemble method, with none of the patient characteristics significant. These results indicate strong between-site heterogeneity while the within-site heterogeneity is marginal. The random forest results share some similarity to that in the ensemble tree but not exactly the same. In practice, we suggest fitting an
ensemble forest because it yield a more accurate and stable estimation.

In this application, different hospitals have different sample sizes \( n_k \). Those with a smaller sample size may not be representative of the population, leading to an uneven level of precision for local causal estimates \( \tau_k(x) \). To account for different sample sizes at each hospital, we consider a basic weighting strategy where we add weights to \( \tau_k(x) \) in the ET and EF adjusting for the sample of site \( k \). The weights are defined as

\[
    w_k(x) = \frac{K n_k}{\sum_{j=1}^{K} n_j}
\]

Our results show similar patterns as in Figure 4 as well as similar coefficients of the best linear projection of mapping the estimated CATEs to patient characteristic space. However, this simplistic weighting scheme fails to consider the possibly uneven distribution of treatment proportion and patient characteristics \( x \) within a site and is needed to be improved. Also, to assess the sensitivity of our methods over each individual site, we estimate the causal effects of \( \text{SpO}_2 \) using all other sites as the coordinating site, respectively.

Results give similar plots as that in Figure 4 assuring that the ensemble is robust over the choice of coordinating site. Overall, our results show evidence that there is moderate causal effects of oxygen saturation \( \text{SpO}_2 \) being in the range of 94% to 98% on hospital mortality. However, the causal effects could vary across sites. This could provide insights for identifying the optimal range of \( \text{SpO}_2 \) that could help to reduce hospital mortality in clinical practice.

7. Discussion

In this paper, we have proposed an efficient and interpretable tree-based ensemble framework for personalized treatment effect estimation in the setting of distributed data where individual-level data cannot be pooled, an increasingly common practice in the era of Big Data. The core of our methods is to break the tradition of viewing of data integration as being a yes or no decision, where practitioners may refute integrating data sets altogether when there is a high level of heterogeneity across sites. Instead, the newly proposed approaches explore the similarity and dissimilarity of the...
Figure 4: Visualization of Causal effects estimation of oxygen saturation on hospital mortality using adaptive XL-EF method. The interpretation of is similar to that in Figure 3. Estimation of causal effects for different values of covariates in each site, respectively. The site indicator is the most important variable in the ensemble forest, with importance taking up 58.8%. These subfigures are ranked by importance in the forest. The sites are arranged by estimated average treatment effect within site. There are apparent within-site heterogeneity and across-site heterogeneity indicated by the different color gradients of the grids within and across sites.

targeted treatment effect between sites to yield an optimal information sharing scheme for selectively improving the treatment effect of interest, all through powerful yet highly accessible tree-based algorithms.

Throughout the paper, we focus on the problems of estimation and prediction of the conditional average treatment effects by leveraging information from heterogeneous data sources. The construction of confidence intervals for conditional treatment effects in the ensemble model remains a challenging problem due to the dependence on the localized estimates. Biases may be introduced from the localized estimator, leading to inconsistent estimates in the federated estimator. Our empirical experiments using the infinitesimal jackknife for random forests (Efron, 2014; Wager et al., 2014; Wager and Athey, 2018) show that our federated estimators fail to achieve a satisfactory performance in terms of nominal coverage of a confidence interval. On the one hand, the consistency and asymptotic properties of the localized estimator is based on a large enough sample size \( n_k \) in the local site, which introduce inaccurate standard errors when \( n_k \) is small. The consistency of federated estimators depends on the perfect subrouping of site indicators from the localized level. When \( n_k \) is large enough, localized estimators such as honest causal trees (Athey and Imbens, 2016) and causal forests (Wager and Athey, 2018) obtain consistent estimators and valid confidence intervals of treatment effects for each individual site, making the task of federated learning less meaningful in practice. On the other hand, on the ensemble stage our federated estimators adaptively place splits based on the internal data structure of site 1 data. Subsample splitting mechanism that is previously favorable to achieve honesty would change the distribution of \( \mathbf{x} \) on the augmented site 1 data and fail to work well anymore. However, by borrowing information from multiple sites, variances are greatly reduced, resulting a small MSE in terms of the final estimated treatment effect.

We also stress that despite the function of interest in this paper being the conditional treatment effect \( \tau(\mathbf{x}) \), a more general \( f(\mathbf{x}) \) still applies and may be of interest to other research problems where data are heterogeneous. For example, in a longitudinal study of outcome-time association \( y = f(t) \) where there is a large amount of between-individual heterogeneity, in which case individuals will be treated as sites, or in a meta-analysis of genome-wide association studies where auxiliary individual-level data from one of the sites are available.

In real life applications, sites in the distributed data networks may have uneven sample sizes, posing a challenge as in different precision in local estimates. Improvements to the weighting strategy are needed to develop considering the treatment proportion as well as covariate distributions in order to obtain an accurate estimate with the federated estimator. Another future direction along the line of modeling for data heterogeneity in distributed data networks is to extend our method for the consideration of missing variables, in the sense that not all variables are available in all sites. Additional assumptions may be needed but it will undoubtedly broaden the application of federated learning approaches.

References
Asfar, P., Schortgen, F., Boisramé-Helms, J., Charpentier, J., Guérot, E., Megarbane, B., Grimaldi, D., Grelon, F., Anguel, N., Lasocki, S., et al. (2017). Hyperoxia and hypertonic saline in patients with septic shock (hypers2s): a two-by-two factorial, multicentre, randomised, clinical trial. The Lancet Respiratory Medicine 5, 180–190.
Assmann, S. F., Pocock, S. J., Enos, L. E., and Kasten, L. E. (2000). Subgroup analysis and other (mis) uses of baseline data in clinical trials. The Lancet 355, 1064–1069.
Athey, S., and Imbens, G. (2016). Recursive partitioning for heterogeneous causal effects. Proceedings of the National Academy of Sciences 113, 7353–7360.
Athey, S., Tibshirani, J., Wager, S., et al. (2019). Generalized random forests. The Annals of Statistics 47, 1148–1178.
Athey, S. and Wager, S. (2019). Estimating treatment effects with causal forests: An application. arXiv preprint arXiv:1902.07409.
Battey, H., Fan, J., Liu, H., Lu, J., and Zhu, Z. (2018).
Distributed testing and estimation under sparse high dimensional models. *Annals of Statistics* 46, 1352–1382.

Borenstein, M., Hedges, L. V., Higgins, J. P., and Rothstein, H. R. (2011). *Introduction to meta-analysis*. John Wiley & Sons.

Breiman, L. (2001). Random forests. *Machine Learning* 45, 5–32.

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

Brookhart, M. A., Stürmer, T., Glynn, R. J., Rassen, J., and Schneeweiss, S. (2010). Confounding control in healthcare database research: challenges and potential approaches. *Medical Care* 48, S114–S120.

Bühlmann, P. and Yu, B. (2002). Analyzing bagging. *The Annals of Statistics* 30, 927–961.

Cameron, A. C. and Miller, D. L. (2015). A practitioner's guide to cluster-robust inference. *Journal of human resources* 50, 317–372.

Chen, D.-G., Liu, D., Min, X., and Zhang, H. (2020). Relative challenges and solutions. *Journal of Evolutionary Biology* 33, 1189–1232.

Fleurence, R. L., Curtis, L. H., Califf, R. M., Platt, R., Selby, J. V., and Miller, D. L. (2015). A practitioner's guide to cluster-robust inference. *Journal of the American Statistical Association* 110, 991–1007.

Farrell, M. H., Liang, T., and Misra, S. (2021). Deep neural networks for estimation and inference. *Econometrica* 89, 181–213.

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

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of Statistics* 28, 1189–1232.

Geoghegan, P., Keane, S., and Martin-Loeches, I. (2018). Change is in the air: dying to breathe oxygen in acute respiratory distress syndrome? *Journal of Thoracic Disease* 10, S2133.

Gruber, C., Nakagawa, S., Laws, R., and Jamieson, I. (2011). Multimodel inference in ecology and evolution: challenges and solutions. *Journal of Evolutionary Biology* 24, 699–711.

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

Itagaki, T., Nakano, Y., Okuda, N., Izawa, M., Onodera, M., Imanaka, H., and Nishimura, M. (2015). Hyperoxemia in mechanically ventilated, critically ill subjects: incidence and related factors. *Respiratory care* 60, 335–340.

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

Konečný, J., McMahan, H. B., Yu, F. X., Richtárik, P., Suresh, A. T., and Bacon, D. (2016). Federated learning: strategies for improving communication efficiency. *arXiv preprint arXiv:1610.05492*.

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

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

Lenzerini, M. (2002). Data integration: A theoretical perspective. In *Proceedings of the Twenty-first ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, pages 233–246.

Li, X., Fireman, B. H., Curtis, J. R., Arterburn, D. E., Fisher, D. P., Moyneur, É., Gallagher, M., Raebel, M. A., Newell, W. B., Lagreid, L., et al. (2019). Validity of privacy-protecting analytical methods that use only aggregate-level information to conduct multivariable-adjusted analysis in distributed data networks. *American Journal of Epidemiology* 188, 709–723.

MacKinnon, J. G. and White, H. (1985). Some heteroskedasticity-consistent covariance matrix estimators with improved finite sample properties. *Journal of econometrics* 29, 305–325.

Martalò, M., Picone, M., Amoretti, M., Ferrari, G., and Rabeli, R. (2011). Randomized network coding in distributed storage systems with layered overlay. In *2011 Information Theory and Applications Workshop*, pages 1–7. IEEE.

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

Nie, X. and Wager, S. (2020). Quasi-oracle estimation of heterogeneous treatment effects. *arXiv preprint arXiv:1712.04912*.

O’Driscoll, B., Howard, L., Earis, J., and Mak, V. (2017). Bts guideline for oxygen use in adults in healthcare and emergency settings. *Thorax* 72, ii1–ii90.

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

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

Qiu, X. (2018). *Statistical Learning Methods for Personalized Medicine*. PhD thesis, Columbia University.

Riley, R. D., Higgins, J. P., and Deeks, J. J. (2011). Interpretation of random effects meta-analyses. *BMJ* 342, e2564.

Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika* 70, 41–55.
Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology* **66**, 688.

Semenova, V. and Chernozhukov, V. (2017). Debiasing machine learning of conditional average treatment effects and other causal functions. *arXiv preprint arXiv:1702.06240*.

Sheikhalishahi, S., Balaraman, V., and Osmani, V. (2020). Benchmarking machine learning models on multi-centre eicu critical care dataset. *PloS ONE* **15**, e0235424.

Shen, J., Liu, R. Y., and Xie, M.-g. (2020). i fusion: Individualized fusion learning. *Journal of the American Statistical Association* **115**, 1251–1267.

Shvachko, K., Kuang, H., Radia, S., and Chansler, R. (2010). The hadoop distributed file system. In *2010 IEEE 26th symposium on Mass Storage Systems and Technologies (MSST)*, pages 1–10. IEEE.

Suzuki, S., Eastwood, G. M., Peck, L., Glassford, N. J., and Bellomo, R. (2013). Current oxygen management in mechanically ventilated patients: a prospective observational cohort study. *Journal of Critical Care* **28**, 647–654.

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

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

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

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

Wager, S., Hastie, T., and Efron, B. (2014). Confidence intervals for random forests: The jackknife and the infinitesimal jackknife. *The Journal of Machine Learning Research* **15**, 1625–1651.

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

Zhang, H. and Singer, B. H. (2010). *Recursive Partitioning and Applications*. Springer Science & Business Media.