Testing and Estimating Causal Treatment Effect Heterogeneity in Observational Studies via Revised Deep Semiparametric Regression: A Lung Transplant Case Study

Shuai Yuan
Department of Biostatistics, University of North Carolina at Chapel Hill
shuaiy@ad.unc.edu

Fei Zou*
Department of Biostatistics, University of North Carolina at Chapel Hill

Baiming Zou†
Department of Biostatistics, University of North Carolina at Chapel Hill

Abstract

Lung transplantation programs need to decide when bilateral lung transplantation (BLT) offers meaningful functional benefit over single lung transplantation (SLT). Because donor and recipient characteristics jointly shape outcomes, the BLT-SLT contrast may differ across patients. However, observational registries pose a key statistical challenge: apparent subgroup differences can be artifacts of complex confounding,

*Co-corresponding author: feizou@email.unc.edu
†Corresponding author: bzou@email.unc.edu

while true heterogeneity can be missed or poorly quantified. Using a large national lung transplant registry, we study whether the BLT effect varies across recipients and identify clinically relevant profiles of benefit using post-transplant lung function measured by forced expiratory volume in 1 second (FEV1).

We develop *deepHTL*, an analysis framework that first tests whether treatment effect heterogeneity is supported by the data and then estimates how the BLT-SLT effect changes with patient features when heterogeneity is present. In extensive simulations designed to resemble registry-like confounding, *deepHTL* controls false positives for detecting heterogeneity and yields more accurate individualized effect estimates than common machine learning methods. In the lung transplant cohort, we find strong evidence of heterogeneity in the BLT-SLT effect on FEV1: younger, lower risk recipients with better baseline status show the largest FEV1 gains from BLT, whereas older, higher risk candidates exhibit diminished marginal benefit. These findings provide statistically grounded guidance for patient selection and allocation of scarce donor organs.

Keywords: Causal inference; Heterogeneous treatment effects; Neural networks; Organ allocation; Precision medicine; Score test.

1 Introduction

Individualized decision-making is central to many areas of clinical practice, shifting the focus from the average treatment effect to identifying which patient subgroups a therapy benefits, has no meaningful effect on, or harms. Causal heterogeneous treatment effects (HTE) are essential for this precision medicine paradigm, as they characterize how intervention outcomes vary across patients with diverse demographic profiles, comorbidities, and biological markers. While HTE is well recognized in fields ranging from oncology to infectious disease (Piccart-Gebhart et al., 2005; Reck et al., 2016; RECOVERY Collaborative Group, 2021), the need for rigorous risk stratification is particularly acute in lung transplantation. In this setting, the choice between bilateral lung transplantation (BLT) and single lung transplantation (SLT) involves complex trade-offs that differ markedly across recipients characteristics. Factors such as age, body mass index (BMI), baseline frailty, and comorbidity burden can fundamentally alter the comparative effectiveness of BLT versus SLT on physiological outcomes like Forced Expiratory Volume in one second (FEV1) (Schaffer et al., 2015). Accurately testing and estimating HTE is critical for optimizing individual patient recovery and for effective allocation of scarce donor organs.

Inferring causal HTE from observational data, however, poses substantial analytical and modeling challenges. Unlike randomized trials, treatment assignment in observational registries is driven by complex, often interacting donor and recipient characteristics (e.g., age, sex, blood type), baseline disease severity, and comorbidity burden. This nonrandom assignment induces confounding that, if not rigorously addressed, can substantially bias treatment effect estimates (Rigdon et al., 2018). The problem is particularly acute in covariate-rich settings such as lung transplantation, where many donor and recipient variables may jointly influence both the transplant decision and post transplant outcomes through nonlinear, nonadditive pathways. In these regimes, conventional treatment assignment

mechanism and outcome models can be severely misspecified and fail to capture key confounding structures. While the causal machine learning literature has rapidly expanded to offer flexible estimators for HTE, such as meta-learning strategies (e.g., T-, S-, and X-learners) (Künzel et al., 2019) which wrap off-the-shelf prediction algorithms, and Bayesian causal forests (Hahn et al., 2020; Wager and Athey, 2018) which provide powerful nonparametric modeling, these methods can be sensitive to hyperparameter tuning and prior specification (D'Amour et al., 2021; Curth and van der Schaar, 2021; Lu et al., 2017). Semiparametric regression based on Robinson's transformation (Robinson, 1988) offers a robust alternative by applying Neyman-orthogonal estimating equations to mitigate the bias for estimating the nuisance functions (e.g., the conditional expectation of outcome estimation given observed covariates), a framework formalized by the quasi-oracle R-learner (Nie and Wager, 2021). Nevertheless, the robustness of the R-learner relies heavily on the quality of the nuisance estimators. In finite samples, conventional learners may fail to capture the complex confounding structures or overfit the data. Particularly, when the underlying treatment effect signal is strong, these learners are ineffective to deal with bimodal distributions even when HTE does not exist (Mi et al., 2019). Crucially, errors in these nuisance function estimations can propagate into the HTE estimates, leading to a systematic bias (Chernozhukov et al., 2018).

Furthermore, a fundamental yet often overlooked step is to separate genuine treatment effect heterogeneity from sampling variability. While recent deep learning approaches can outperform traditional learners for complex function approximations, the primary focus of this entire body of work remains on robust point estimation and interval construction of homogeneous treatment effect estimate (Mi et al., 2021). Without a statistically principled global test for the existence of HTE, investigators may overfit to noise, producing spurious subgroup recommendations and needlessly complex treatment rules. Formal hypothesis

testing for the existence of treatment effect heterogeneity is largely overlooked, with existing tests restricted to randomized trials (Chang et al., 2021) or prespecified subgroups (Dai and Stern, 2022). Consequently, there remains a lack of an integrated framework that can simultaneously deliver efficient and robust HTE estimation along with a calibrated global test for its presence in observational studies.

Motivated by these gaps and the critical need for reliable evidence in high-stakes clinical decision-making, we propose a unified framework, *deep Heterogeneous Treatment Learning* (*deepHTL*). This framework integrates three key strategies to robustly test for and estimate causal HTE in observational studies. First, we employ bagged deep neural networks (bagged-DNN) to estimate nuisance functions, ensuring robust and accurate estimations while mitigating stability issue in traditional DNN models under the finite sample setting (Mi et al., 2021; Zou et al., 2025). Second, we estimate heterogeneous treatment effects via a semiparametric regression using the revised nuisance function estimation, a step designed to reduce bias that often undermines standard R-learners. Third, to rigorously assess the existence of treatment effect heterogeneity, we implement a global kernel score test and calculate its significance using the Davies method (Liu et al., 2007; Davies, 1980), complemented by a cross-fitted permutation test to ensure robust and principled inference under finite sample settings.

We evaluate the performance of *deepHTL* through extensive simulation studies, showing improved estimation accuracy relative to standard R-learners and demonstrating that the proposed testing methods maintain nominal Type I error while achieving high power in detecting treatment effect heterogeneity. We then apply *deepHTL* to a large cohort from the United Network for Organ Sharing (UNOS) registry to investigate the comparative effectiveness of BLT vs. SLT on post-transplant lung function, measured by FEV1. The analysis reveals statistically significant and clinically meaningful heterogeneity: younger

recipients with lower baseline risk experience the largest functional gains from BLT, whereas older, higher risk patients exhibit attenuated marginal benefit. These results illustrate how *deepHTL* can support more targeted decision-making and potentially refine allocation criteria for scarce donor organs. Although we focus on lung transplantation, the framework is broadly applicable to other observational settings where individualized treatment effects are of primary interest.

The remainder of the paper is organized as follows. Section 2 describes the detailed framework of *deepHTL*. In Section 3, we report extensive results from simulation studies mimicing various practical confounding and HTE settings. Section 4 presents the UNOS case study application, and we conclude the paper in Section 5 with a brief discussion.

2 Methods

To model potentially complex heterogeneous treatment effects and accommodate intricate confounding structures commonly observed in clinical practice, we consider the following semiparametric regression model:

$$Y_i = \tau(\boldsymbol{X}_i)Z_i + f(\boldsymbol{X}_i) + \varepsilon_i$$

where $Y_i \in \mathbb{R}$ denotes the observed outcome, $Z_i \in \{0, 1\}$ is the binary treatment assignment indicator, and $\boldsymbol{X}_i \in \mathbb{R}^p$ represents the set of vector of pre-treatment covariates for the ith individual ($i = 1, \ldots, n$). The random error term ε_i is assumed to satisfy $\mathbb{E}[\varepsilon_i \mid \boldsymbol{X}_i, Z_i] = 0$ with a finite variance. To accommodate potential complex confounding structures, we leave the nuisance function $f(\boldsymbol{X})$ unspecified. Similarly, the treatment effect of interest, $\tau(\boldsymbol{X})$, is defined as an unspecified function of $\boldsymbol{X}$, allowing for flexible modeling of treatment effect heterogeneity. Under this semiparametric regression framework, we present a set of robust and powerful methods to estimate and test for the existence of treatment effect heterogeneity

as described in the following subsections.

2.1 Deep Heterogeneous Treatment Learning for Estimating HTE

To avoid estimating the unknown nuisance function $f(\boldsymbol{X})$ in the above semiparametric regression model, we employ the Robinson transformation:

$$Y - \mu(\boldsymbol{X}) = \tau(\boldsymbol{X})\{Z - e(\boldsymbol{X})\} + \varepsilon, \tag{2.1}$$

where $\mu(\boldsymbol{X}) = \mathbb{E}[Y \mid \boldsymbol{X}]$ and $e(\boldsymbol{X}) = \mathbb{E}[Z \mid \boldsymbol{X}]$ represent nuisance functions that can be pre-estimated straightforwardly using, for example, the kernel methods. (Robinson, 1988). Alternatively, many standard machine learning methods such as Lasso (Tibshirani, 1996), XGBoost (Chen and Guestrin, 2016), and Kernel Ridge Regression (KRR) (Vovk, 2013), as utilized by Nie and Wager (2021) can be adopted for estimating these nuisance functions. However, they often struggle in the presence of complex, for example, nonlinear and non-additive confounding. In such scenarios, nuisance estimators trained on finite samples using these machine learning models can exhibit systematic bias, and such errors propagate into the estimation of $\tau(\boldsymbol{X})$ (Chernozhukov et al., 2018) since they may not accurately approximate the potential complex associations in these scenarios. While deep neural network (DNN) can universally and accurately approximate complex functions asymptotically (Hornik et al., 1989; Cybenko, 1989), its performance under finite sample settings can be poor where a DNN model may not converge to the global optimal due to random parameter initialization. Bagged-DNN method addresses this challenge by aggregating predictions from multiple bootstrap-trained DNN models while explicitly discarding poorly performed DNN models, thereby stabilizing the fits against the randomness of network initialization. Based on this observation, to obtain more reliable nuisance function estimates, we implement a bagged-DNN model enhanced with a filtering mechanism (Mi et al., 2019, 2021; Zou et al., 2025).

To further mitigate overfitting and induce the orthogonality of nuisance parameter estimation errors, we randomly partition the observation indices $\{1, \ldots, n\}$ into K disjoint folds, denoted by $I_1, \ldots, I_K$. For each fold k, the bagged-DNN models for $\mu(\boldsymbol{X})$ and $e(\boldsymbol{X})$ are trained on the out-of-fold samples to generate predictions exclusively for the held-out observations $i \in I_k$. We then use these out-of-sample nuisance estimates $\hat{\mu}(\boldsymbol{X}_i)$ and $\hat{e}(\boldsymbol{X}_i)$ to estimate the heterogeneous treatment effects by minimizing the squared loss derived from (2.1):

$$\hat{\tau}(\cdot) = \arg\min_{\tau} \frac{1}{n} \sum_{i=1}^{n} \Big\{ \Big(Y_i - \hat{\mu}(\boldsymbol{X}_i) \Big) - \tau(\boldsymbol{X}_i) \Big(Z_i - \hat{e}(\boldsymbol{X}_i) \Big) \Big\}^2, \tag{2.2}$$

and minimization is performed using the bagged-DNN. However, in settings like transplantation where the treatment effect magnitude $|\tau(\boldsymbol{X})|$ can be large, the finite sample accuracy of this estimator can deteriorate. As shown in (2.1), the residual $Y - \mu(\boldsymbol{X})$ comprises $\tau(\boldsymbol{X})\{Z - e(\boldsymbol{X})\}$ and random noise ε. When $|\tau(\boldsymbol{X})|$ is small, the residual is dominated by ε, which typically follows a standard unimodal distribution (e.g., Gaussian) that is efficiently learned by mean-squared-error minimization. Conversely, when $|\tau(\boldsymbol{X})|$ is large, the dominance of $\tau(\boldsymbol{X})\{Z - e(\boldsymbol{X})\}$ induces residual bimodality, which introduces estimation bias in $\hat{\mu}(\boldsymbol{X})$ and propagates to $\hat{\tau}(\boldsymbol{X})$.

To address this, we decompose the treatment effect into a constant average component and a residual heterogeneous component: $\tau(X) = \tau_0 + r(X)$. We reformulate (2.1) with the following model:

$$Y^* - \mathbb{E}[Y^* \mid \boldsymbol{X}] = r(\boldsymbol{X}) \Big\{ Z - e(\boldsymbol{X}) \Big\} + \varepsilon, \tag{2.3}$$

where $Y^* = Y - \tau_0 Z$ denotes the centered outcome and $\mu^*(\boldsymbol{X}) = \mathbb{E}(Y^* \mid \boldsymbol{X})$ is estimated using the bagged-DNN via the same cross-fitting procedure. This centering step effectively removes the large main effect of treatment and strengthens the unimodal residual structure,

allowing $\mu^*(\boldsymbol{X})$ to be estimated more accurately. We estimate the constant average component τ_0 via a computationally efficient double robust estimator:

$$\hat{\tau}_0 = \frac{\sum_{i=1}^n \{Y_i - \hat{\mu}(\boldsymbol{X}_i)\}\{Z_i - \hat{e}(\boldsymbol{X}_i)\}}{\sum_{i=1}^n \{Z_i - \hat{e}(\boldsymbol{X}_i)\}^2}. \tag{2.4}$$

The residual heterogeneous component estimator $\hat{r}(\boldsymbol{X})$ is then obtained by minimizing the squared loss:

$$\hat{r}(\cdot) = \arg\min_r \frac{1}{n}\sum_{i=1}^n \Big\{\Big(Y_i - \hat{\tau}_0 Z_i - \hat{\mu}^*(\boldsymbol{X}_i)\Big) - r(\boldsymbol{X}_i)\Big(Z_i - \hat{e}(\boldsymbol{X}_i)\Big)\Big\}^2. \tag{2.5}$$

Combining the constant and residual components produces the revised estimator of HTE:

$$\hat{\tau}^*(\boldsymbol{X}) = \hat{\tau}_0 + \hat{r}(\boldsymbol{X}). \tag{2.6}$$

A complete HTE estimation procedure based on *deepHTL* is summarized in Algorithm 1.

Algorithm 1 deepHTL: Estimating Heterogeneous Treatment Effects

Input: Observations $\{(Y_i, Z_i, \boldsymbol{X}_i)\}_{i=1}^n$ partitioned into K folds.
for each fold $k = 1, \ldots, K$ **do**
    Estimate nuisance functions $\hat{e}(\boldsymbol{X}_i)$ and $\hat{\mu}(\boldsymbol{X}_i)$ for $i \in I_k$.
end for
Compute the constant effect $\hat{\tau}_0$ using (2.4) and the centered outcome $Y_i^* = Y_i - \hat{\tau}_0 Z_i$.
for each fold $k = 1, \ldots, K$ **do**
    Estimate the centered nuisance function $\hat{\mu}^*(\boldsymbol{X}_i) = \mathbb{E}[Y_i^* \mid \boldsymbol{X}_i]$ for $i \in I_k$.
end for
Estimate residual heterogeneity $\hat{r}(\boldsymbol{X})$ by minimizing the squared loss function in (2.5).
Output: Revised HTE estimate $\hat{\tau}^*(\boldsymbol{X}) = \hat{\tau}_0 + \hat{r}(\boldsymbol{X})$.

Note: All functions are estimated using the bagged-DNN.

2.2 Testing for the Existence of HTE

In clinical practice, it is critically important to test the existence of HTE to avoid deploying unnecessarily complex individualized treatment rules. In observational registries, searching for subgroups without a global protection against Type I error risks identifying "false positive" heterogeneity driven by noise or residual confounding. We therefore test the null hypothesis $H_0 : \mathrm{Var}\{\tau(\boldsymbol{X})\} = 0$, which implies that $\tau(\boldsymbol{X})$ reduces to an unknown constant under the null. Specifically, we propose two procedures for testing the existence of HTE: a computationally efficient analytic method using the kernel score test, and a more robust, albeit computationally intensive, empirical method based on a permutation procedure as described in the following.

Kernel Score Test: Accurately estimating the constant effect is essential for constructing residuals that capture genuine departures from the null hypothesis rather than artifacts of estimation error. Although $\hat{\tau}_0$ in (2.4) provides an initial ("unrevised") estimate, its performance can deteriorate in finite samples when the treatment effect is large in magnitude, i.e., the non-Gaussian term may dominate, leading to inaccurate estimation of $\hat{\mu}(\boldsymbol{X})$. To address this issue while preserving our cross-fitting architecture for the nuisance parameters, we incorporate a bias correction step. Specifically, we compute a global revised estimator, denoted as $\hat{\tau}^*$:

$$\hat{\tau}^* = \hat{\tau}_0 + \frac{\sum_{i=1}^n (Y_i - \hat{\tau}_0 Z_i - \hat{\mu}^*(\boldsymbol{X}_i))(Z_i - \hat{e}(\boldsymbol{X}_i))}{\sum_{i=1}^n (Z_i - \hat{e}(\boldsymbol{X}_i))^2}. \tag{2.7}$$

Using $\hat{\tau}^*$, we derive the raw residual score u_i and its studentized version s_i for each observation $i = 1, \ldots, n$:

$$u_i = \frac{Y_i - \hat{\tau}_0 Z_i - \hat{\mu}^*(\boldsymbol{X}_i)}{Z_i - \hat{e}(\boldsymbol{X}_i)} - (\hat{\tau}^* - \hat{\tau}_0), \quad s_i = \frac{\sqrt{w_i}\, u_i}{\hat{\sigma}},$$

where $w_i = (Z_i - \hat{e}(\boldsymbol{X}_i))^2$ and $\hat{\sigma}^2 \approx n^{-1}\sum_{i=1}^n w_i u_i^2$ estimates the conditional variance $\mathrm{Var}(\varepsilon_i|\boldsymbol{X}_i, Z_i)$. Under the null hypothesis H_0 and mild regularity conditions, using cross-fitted nuisance parameters allows the estimation error to be asymptotically orthogonal to the score (Chernozhukov et al., 2018). Thus, the raw residual admits the asymptotic expansion $u_i \approx \frac{\varepsilon_i}{Z_i - e(\boldsymbol{X}_i)} + o_p(1)$, the leading term of u_i is proportional to the noise. Consequently, the studentized scores $s_i = \frac{\sqrt{w_i}\, u_i}{\hat{\sigma}}$ are marginally asymptotically standard normal.

Let $\boldsymbol{s} = (s_1, \ldots, s_n)^\top$ and $\boldsymbol{a} = (\sqrt{w_1}, \ldots, \sqrt{w_n})^\top$. By definition of the weighted least squares objective used to estimate the constant treatment effect, the resulting residuals satisfy $\boldsymbol{a}^\top \boldsymbol{s} = 0$. This linear constraint inherently induces dependence among the components of $\boldsymbol{s}$. To properly accommodate this structural restriction, we project $\boldsymbol{s}$ onto the orthogonal complement of $\mathrm{span}(\boldsymbol{a})$:

$$\tilde{\boldsymbol{s}} = \boldsymbol{P}\boldsymbol{s}, \qquad \boldsymbol{P} = \boldsymbol{I}_n - \frac{\boldsymbol{a}\boldsymbol{a}^\top}{\boldsymbol{a}^\top \boldsymbol{a}} = \boldsymbol{I}_n - \frac{\sqrt{\boldsymbol{w}}\sqrt{\boldsymbol{w}}^\top}{\sum_{i=1}^n w_i}. \tag{2.8}$$

Here $\boldsymbol{P}$ is symmetric and idempotent ($\boldsymbol{P}^\top = \boldsymbol{P}$, $\boldsymbol{P}^2 = \boldsymbol{P}$), satisfies $\boldsymbol{P}\boldsymbol{a} = \boldsymbol{0}$, and has rank $n-1$ provided $\boldsymbol{a} \neq \boldsymbol{0}$. Following Liu et al. (2007), we construct the quadratic form statistic $Q = \tilde{\boldsymbol{s}}^\top \boldsymbol{K} \tilde{\boldsymbol{s}}$, where $\boldsymbol{K}$ is a symmetric positive semidefinite kernel matrix with entries $K_{ij} = K(\boldsymbol{X}_i, \boldsymbol{X}_j)$ for some Mercer kernel $K(\cdot,\cdot)$ (e.g., Gaussian RBF). Under the null hypothesis and $\boldsymbol{s} \overset{d}{\approx} N(\boldsymbol{0}, \boldsymbol{\Sigma})$, the quadratic form admits a weighted chi-square representation:

$$Q \xrightarrow{d} \sum_{j=1}^{r} \lambda_j \chi^2_{1,j},$$

where $\{\lambda_j\}_{j=1}^r$ are the nonzero eigenvalues of $\boldsymbol{P}\boldsymbol{K}\boldsymbol{P}$ (with rank $r \leq n-1$) and $\chi^2_{1,j}$ are independent central chi-square random variables with 1 degree of freedom. An analytic

p-value is obtained using the Davies method (Davies, 1980):

$$p \;=\; \Pr\left(\sum_{j=1}^{r} \lambda_j \, \chi^2_{1,j} \geq Q\right).$$

Permutation Test: While the kernel score test provides a powerful and computationally efficient analytic framework, its validity relies on asymptotic convergence to a mixture of chi-square distributions. The projection matrix $\boldsymbol{P}$ corrects for the degree of freedom lost in estimating the global constant treatment effect, but it may not fully account for the effective degrees of freedom consumed by the complex DNN nuisance estimators. In challenging finite sample settings (e.g., high dimensionality, high noise, moderate sample sizes), this uncaptured estimation bias may compromise the asymptotic approximation and lead to mild Type I error inflation. To provide a more reliable and robust inference tool for testing existence of HTE, we propose a permutation test. By empirically constructing the null distribution, this approach aims to naturally accommodate the variance structure and control Type I error without relying strictly on asymptotic assumptions.

We construct a test statistic based directly on the out-of-sample prediction loss. Using the same K disjoint folds $I_1, \ldots, I_K$ from the nuisance estimation stage, we cross-fit the treatment effect model in (2.6). For each fold $k \in \{1, \ldots, K\}$, the model is trained exclusively on the remaining $K-1$ folds to generate predictions $\hat{\tau}^*(\boldsymbol{X}_i)$ for the holdout observations $i \in I_k$. The observed mean squared error (MSE) across all K folds is calculated as:

$$L_{\text{obs}} = \frac{1}{n} \sum_{k=1}^{K} \sum_{i \in I_k} \Big\{ \Big(Y_i^* - \hat{\mu}^*(\boldsymbol{X}_i) \Big) - \hat{\tau}^*(\boldsymbol{X}_i) \Big(Z_i - \hat{e}(\boldsymbol{X}_i) \Big) \Big\}^2. \tag{2.9}$$

To generate the null distribution, we perform B random permutations. For each permutation $b = 1, \ldots, B$, we operate within each fold $k \in \{1, \ldots, K\}$ to randomly shuffle

the indices of the holdout predictions within their respective treatment arms ($Z_i \in \{0, 1\}$). Letting $\pi^{(b)}(i)$ denote the permuted index for observation i, the permuted loss is:

$$L_{\text{perm}}^{(b)} = \frac{1}{n} \sum_{k=1}^{K} \sum_{i \in I_k} \Big\{ \Big(Y_i^* - \hat{\mu}^*(\boldsymbol{X}_i) \Big) - \hat{\tau}^*(\boldsymbol{X}_{\pi^{(b)}(i)}) \Big(Z_i - \hat{e}(\boldsymbol{X}_i) \Big) \Big\}^2. \tag{2.10}$$

The empirical p-value is computed as the proportion of permutations where the permuted loss is less than or equal to the observed loss:

$$p_{\text{perm}} = \frac{1}{B+1} \left(1 + \sum_{b=1}^{B} \mathbb{I}\Big\{ L_{\text{perm}}^{(b)} \leq L_{\text{obs}} \Big\} \right).$$

Under the null hypothesis, the covariates $\boldsymbol{X}$ provide no informative signal for treatment heterogeneity. Consequently, breaking the alignment between the observed data and the predicted heterogeneity $\hat{\tau}^*(\boldsymbol{X}_i)$ via permutation does not systematically alter the loss, meaning L_{obs} and $L_{\text{perm}}^{(b)}$ are drawn from the same underlying distribution. Under the alternative hypothesis H_1, shuffling destroys the alignment with the true heterogeneous signal, causing $L_{\text{perm}}^{(b)}$ to be larger than L_{obs}.

This framework applies identically to the unrevised estimator $\hat{\tau}(\boldsymbol{X}_i)$ by cross-fitting the loss function in (2.2). The complete procedure for testing the existence of HTE is summarized in Algorithm 2.

3 Simulation Studies

We evaluate the finite sample performance of *deepHTL* through extensive simulation studies designed to mimic the complex confounding structures commonly encountered in observational biomedical registries. Our primary objectives are to assess (i) the performance of the proposed kernel score and permutation tests under the null hypothesis $H_0 : \text{Var}\{\tau(\boldsymbol{X})\} = 0$, (ii) the power of the test to detect heterogeneity under the alternative

Algorithm 2 deepHTL: Testing for Heterogeneous Treatment Effects

Input: Observations $\{(Y_i, Z_i, \boldsymbol{X}_i)\}_{i=1}^n$, cross-fitting folds K, permutations B.
Kernel Score Test
Partition data into K disjoint folds.
Obtain out-of-fold nuisance estimates $\hat{e}(\boldsymbol{X}_i)$, $\hat{\mu}(\boldsymbol{X}_i)$, and $\hat{\mu}^*(\boldsymbol{X}_i)$ for all i.
Compute the revised constant estimator $\hat{\tau}^*$ via (2.7).
Compute studentized residuals $\boldsymbol{s}$ and the projection matrix $\boldsymbol{P}$ in (2.8).
Calculate test statistic $Q = \boldsymbol{s}^\top \boldsymbol{PKPs}$ and derive p_{davies}.
Permutation Test
Train the target model (2.2) and (2.6) via cross-fitting.
Obtain out-of-fold predictions $\hat{\tau}(\boldsymbol{X}_i)$ for the holdout fold k.
Calculate the observed loss L_{obs} via (2.9).
for $b = 1, \ldots, B$ **do**
    Permute treatment indices within folds and calculate $L_{\text{perm}}^{(b)}$ via (2.10).
end for
Compute $p_{\text{perm}} = \frac{1}{B+1}\left(1 + \sum_{b=1}^{B} \mathbb{I}\left\{L_{\text{perm}}^{(b)} \leq L_{\text{obs}}\right\}\right)$.
Output: p_{davies} and p_{perm}.

Note: All functions are estimated using the bagged-DNN.

hypothesis $H_1 : \text{Var}\{\tau(\boldsymbol{X})\} > 0$, and (iii) the estimation accuracy of the proposed HTE estimator across diverse settings. To this end, we generate data from the following outcome model:

$$Y_i = \tau(\boldsymbol{X}_i)Z_i + f(\boldsymbol{X}_i) + \varepsilon_i, \qquad i = 1, \ldots, n,$$

where $\boldsymbol{X}_i \sim \mathcal{N}_p(\boldsymbol{0}, \mathbf{I}_p)$ denotes the p-dimensional covariate vector (including confounders and nuisance factors), and $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$ is an independent noise term. Binary treatment assignment mechanism follows the following:

$$Z_i \mid \boldsymbol{X}_i \sim \text{Bernoulli}\{e(\boldsymbol{X}_i)\}, \quad e(\boldsymbol{X}) = \text{logit}^{-1}\{0.8\sin(\pi X_1 X_2) + 0.6X_3X_4 + 0.5\tanh(X_5)\}.$$

The structural nuisance function is set to $f(\boldsymbol{X}) = \log(|X_1| + 1) - X_2^2 + \sin(X_3) + 0.5\, X_4 X_5$, and the treatment effect function $\tau(\boldsymbol{X})$ varies by experiment as the following:

$$\tau(\boldsymbol{X}) = \begin{cases} 0, & \text{(Zero effect without HTE)}, \\ 3, & \text{(Large effect without HTE)}, \\ -1 + 2\log\Big(\exp\Big(\sum_{j=1}^{5} X_j\Big) + 1\Big), & \text{(Simple HTE)}, \\ -1 + X_1 X_2 + \cos^2(X_3) + \max(X_4 - X_5, 0) & \text{(Complex HTE)}. \end{cases}$$

We evaluate performance over a grid of sample sizes $n \in \{1000, 2000\}$, covariate dimensions $p \in \{20, 40\}$, and noise levels $\sigma \in \{1, 3\}$, using 1000 independent replicates per configuration.

3.1 Performance on Testing HTE

We first evaluate the performance of the proposed testing procedures under the null hypothesis of homogeneous treatment effects. To probe robustness across effect magnitudes, we consider two homogeneous scenarios: a zero effect ($\tau = 0$) and a large constant effect ($\tau = 3$). For each scenario, we construct test statistics using both the unrevised estimator $\hat{\tau}$ and the revised estimator $\hat{\tau}^*$. We report the empirical Type I error rates across 1,000 replicates at the nominal level $\alpha = 0.05$ for the kernel score test and the permutation test (with $B = 2000$ permutations).

Table 1 shows that the revised estimator delivers uniformly reliable Type I error control across all configurations, whereas the unrevised estimator can break down sharply when the treatment signal is strong relative to noise. Under $\tau = 0$ using the kernel score test, both tests constructed from the revised and unrevised estimators generally produce type I errors that are generally close to nominal, with only modest inflation in the most challenging high-dimensional/high-noise regime ($n = 1000, p = 40, \sigma = 3$). In contrast, under $\tau = 3$, the

unrevised estimator exhibits substantial and systematic inflation whenever the signal-to-noise ratio is high (notably $\sigma = 1$), with rejection rates ranging from 0.15 to 0.30 (e.g., 0.299 at $n = 2000, p = 20, \sigma = 1$), far exceeding the 0.05 target. This inflation is consistent with the mechanism described in Section 2.2: when τ is large, the residual $Y - \mu(\boldsymbol{X})$ is dominated by $\tau\{Z - e(\boldsymbol{X})\}$ rather than ε, producing a pronounced bimodal structure. As a result, the nuisance function estimator $\hat{\mu}(\boldsymbol{X})$ becomes inaccurate, which in turn biases the treatment effect estimate $\hat{\tau}_0$. Moreover, the residual score computed using this unrevised estimate $\hat{\tau}_0$ can deviate substantially from unimodality and Gaussianity, leading to uncontrolled Type I error.

This failure mode is eliminated by the proposed revision. Across every setting in Table 1 including the low noise, large effect regimes where the unrevised method is most anti-conservative, the revised estimator $\hat{\tau}_0^*$ yields near-nominal Type I error (typically 0.04–0.06), demonstrating robustness to sample size, dimensionality, and signal-to-noise ratio. Moreover, the remaining mild inflation in the hardest regimes ($p = 40, \sigma = 3$) attenuates as n increases (to ≈ 0.06 at $n = 2000$), consistent with improved nuisance function estimation in larger samples typical of registry studies (e.g., our UNOS application with $n > 10{,}000$). Furthermore, the permutation test successfully relieves the mild inflation observed in the hardest analytical settings (e.g., reducing the rejection rate from 0.085 to roughly 0.06 at $n = 1000, p = 40, \sigma = 3$). Because the permutation test relies on random shuffling rather than asymptotic approximations, it avoids relying on the asymptotic normality of the residuals, providing tighter empirical control when conditions are challenging.

Although strictly controlling Type I error is paramount to preventing false discoveries, a practical testing procedure must also demonstrate sufficient power to uncover true heterogeneity when it exists. We maintain the data generating process from the previous section but introduce covariate-dependent treatment effects $\tau(\boldsymbol{X})$ under two different

scenarios. **Scenario I** posits a simple, mildly nonlinear effect structure, $\tau(\boldsymbol{X}) = -1 + 2\log\left(\exp\left(\sum_{j=1}^{5} X_j\right) + 1\right)$, while **Scenario II** introduces complex nonlinear interactions, $\tau(\boldsymbol{X}) = -1 + X_1X_2 + \cos^2(X_3) + \max(X_4 - X_5, 0)$. Having established that the unrevised approach fails to control the Type I error rate, we restrict our power evaluations to the properly calibrated revised estimator. For each setting, we report the empirical power, defined as the proportion of replicates where the null hypothesis is rejected at the nominal significance level $\alpha = 0.05$.

In **Scenario I**, the treatment effect function is simple, and the heterogeneity signal is pronounced (average $\mathrm{Var}\{\tau(\boldsymbol{X})\} \approx 6.05$). Consequently, detection is straightforward: both estimators under the two testing frameworks achieve a power of 1.0 across all considered sample sizes and noise levels (results not tabulated). This confirms that the proposed testing methods reliably identify overt heterogeneity without loss of sensitivity.

Scenario II presents a more challenging test bed due to its lower signal variance and complex functional form. Due to its inflated Type I error (Table 1), we omit the unrevised test from our power evaluation. Table 2 reports results exclusively for the proposed revised estimator. Similar to the Type I error analysis, statistical power is sensitive to the data complexity. Since detecting heterogeneity relies on distinguishing systematic variation in $\tau(\boldsymbol{X})$ from random error, increasing the noise level σ or the dimension p naturally attenuates power by diluting the signal-to-noise ratio. The test demonstrates a strong responsiveness to sample size. Under the most challenging scenario ($\sigma = 3, p = 40$), increasing n from 1000 to 2000 more than doubles the analytical power (43.7% to 89.3%). The permutation test achieved a similar power pattern to the kernel score test, confirming that both frameworks reliably detect the heterogeneity signal in larger samples. These results confirm the revised estimator balances rigorous error control with the sensitivity needed for practical observational studies.

Table 1: Performance Comparison on Type I Error Control

τ	Estimator	n	p	σ	Type I Error (Anal.)	Type I Error (Perm.)
$\tau = 0$	Unrevised ($\hat{\tau}_0$)	1000	20	1	0.042	0.049
		1000	20	3	0.056	0.052
		1000	40	1	0.062	0.066
		1000	40	3	0.083	0.061
		2000	20	1	0.045	0.048
		2000	20	3	0.058	0.059
		2000	40	1	0.055	0.061
		2000	40	3	0.060	0.066
	Revised ($\hat{\tau}_0^*$)	1000	20	1	0.044	0.046
		1000	20	3	0.057	0.048
		1000	40	1	0.070	0.065
		1000	40	3	0.086	0.054
		2000	20	1	0.043	0.045
		2000	20	3	0.057	0.064
		2000	40	1	0.053	0.055
		2000	40	3	0.059	0.065
$\tau = 3$	Unrevised ($\hat{\tau}_0$)	**1000**	**20**	**1**	**0.176**	**0.091**
		1000	20	3	0.064	0.049
		1000	**40**	**1**	**0.152**	**0.111**
		1000	40	3	0.086	0.064
		2000	**20**	**1**	**0.299**	**0.090**
		2000	20	3	0.072	0.058
		2000	**40**	**1**	**0.181**	**0.087**
		2000	40	3	0.070	0.063
	Revised ($\hat{\tau}_0^*$)	1000	20	1	0.044	0.049
		1000	20	3	0.057	0.049
		1000	40	1	0.063	0.061
		1000	40	3	0.085	0.059
		2000	20	1	0.043	0.045
		2000	20	3	0.053	0.055
		2000	40	1	0.055	0.060
		2000	40	3	0.060	0.059

Table 2: Empirical Power for Detecting Complex HTE Using the Revised Estimator

n	p	σ	Power (Anal.)	Power (Perm.)
1000	20	1	0.998	1.000
1000	20	3	0.631	0.614
1000	40	1	0.953	0.940
1000	40	3	0.437	0.388
2000	20	1	1.000	1.000
2000	20	3	0.977	0.984
2000	40	1	1.000	1.000
2000	40	3	0.893	0.938

3.2 Performance on Estimating HTE

Having established the validity of the testing procedure, we next evaluate the accuracy of HTE estimation. We maintain the data-generating process and the two heterogeneous scenarios from the previous section and apply the HTE estimator in Section 2.1. For each configuration in the simulation grid, models are trained on a dataset and evaluated on an independent test set of equal size using the logarithmic mean squared error (log-MSE).

To benchmark the proposed *deepHTL* framework, we compare it with three machine learning methods: Least absolute shrinkage and selection operator (*Lasso* (Tibshirani, 1996)), eXtreme Gradient Boosting (*XGBoost* (Chen and Guestrin, 2016)), and Kernel Ridge Regression (*KRR* (Vovk, 2013)). Each learner is used consistently throughout the estimation pipeline to obtain the estimation of the nuisance functions $\mu(\boldsymbol{X})$ and $e(\boldsymbol{X})$, as well as the treatment effect function $\tau(\boldsymbol{X})$. To isolate the contribution of our revision step, we implement two versions of every learner: the unrevised estimator in (2.2) (denoted ***R–learner***, e.g., *R–Lasso*) and the proposed revised estimator in (2.6)(denoted ***Rev–learner***, e.g., *Rev–XGBoost*). Within this comparison framework, *deepHTL* corresponds to the revised estimator implemented with bagged-DNN, while its unrevised counterpart is denoted

as *R–DNN*.

Figure 1 summarizes the predictive performance (log-MSE) across simulation settings. The results highlight a critical interaction between learner capacity and the utility of the revision step. For parametric methods like Lasso, the revised estimator offers negligible benefit, as the linear additive assumption fails to capture the complex nuisance structure $\mu(\boldsymbol{X})$. In contrast, flexible learners (XGBoost, KRR, DNN) exhibit substantial gains in **Scenario II**. For example, the revised estimator reduces the error metric by $0.1 \sim 0.2$ for both XGBoost and KRR. Similarly, *deepHTL* consistently improves, lowering the log-MSE by over 0.19 in complex configurations (e.g., $n = 1000, p = 40, \sigma = 1$). In the smoother **Scenario I**, benefits are more subtle, though *deepHTL* still displays discernible gains at $n = 2000$. Importantly, we observe no performance penalty for the revision step. Despite the theoretical risk of added noise from the multi-step procedure, the method maintains accuracy in simpler settings, confirming the revision step as a safe and robust enhancement.

Overall, *deepHTL* outperforms all competing methods considered across the simulation grid. The largest margins occur at $n = 2000$, confirming that the revision enables high capacity networks to leverage larger sample sizes effectively for resolving complex confounding and heterogeneity. These findings suggest a context aware strategy: while *deepHTL* excels in settings with complex confounding, large magnitude effects, and sufficiently large sample sizes, simpler learners (e.g., *R–KRR*) may suffice in smoother, lower-dimensional problems where computational efficiency is a priority. Returning to the clinical challenge that motivated this methodology, we apply *deepHTL* to a real-world lung transplant dataset in the following section.

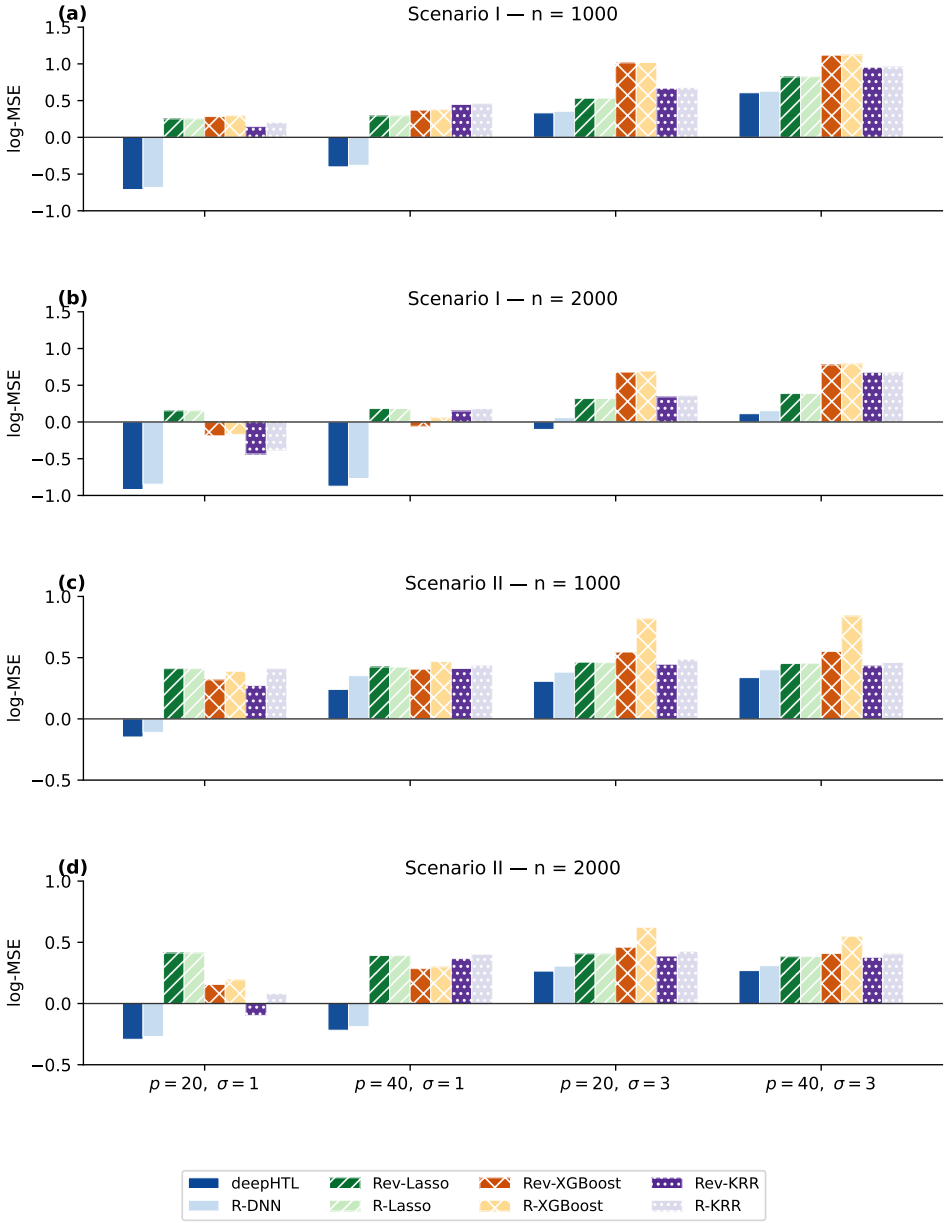

Figure 1: Log-MSE of $\hat{\tau}(X)$ across simulation settings for all HTE estimators.

4 A Lung Transplant Study

To demonstrate the practical utility of our proposed framework *deepHTL*, we apply it to registry data from the United Network for Organ Sharing (UNOS) to address a central question in lung transplantation: whether bilateral lung transplantation confers superior post transplant lung function measured by FEV1 compared with single lung transplantation, and whether any advantage varies across recipients and clinically relevant subgroups. Although BLT is often associated with improved lung function, the extent and clinical drivers of heterogeneity in its benefit remain underexplored (Oliveira et al., 2012; Force et al., 2011).

Our analytic cohort comprises 14,306 adult recipients transplanted between May 2005 and September 2011. The outcome Y is post transplant forced expiratory volume in 1 second (FEV1), a key marker of graft performance and long-term survival risk (Young et al., 2007). Treatment is encoded as $Z \in \{0, 1\}$, with $Z = 1$ indicating BLT ($n = 9{,}758$) and $Z = 0$ indicating SLT ($n = 4{,}548$). The covariate vector $\boldsymbol{X}$ includes baseline donor and recipient characteristics used to adjust for confounding. To select adjustment variables in a data adaptive yet clinically grounded manner, we used the Group Lasso—Doubly Robust (GL-DR) procedure (Koch et al., 2018), which retained 26 clinically relevant predictors (e.g., recipient age, body mass index, and ischemic time) from 29 candidate covariates (Table 3).

We begin the analysis by formally testing for treatment effect heterogeneity using the revised null estimator. The analytical kernel score test is highly significant ($p = 0.002$), and this finding is further corroborated by the permutation test ($p = 0.001$). Together, these results provide strong evidence that the benefit of BLT varies across the recipient population. To further investigate the complexity of the detected heterogeneity, we estimate the effective degrees of freedom, k_{eff}, of the limiting mixture chi-squared distribution (Satterthwaite, 1946). Here, k_{eff} acts as an interpretable proxy for the dimensionality of heterogeneity: values near 1 suggest a sparse pattern dominated by a single covariate direction, whereas larger

Table 3: Key patient and donor characteristics used as confounders by transplant type (BLT vs. SLT). The complete 26 covariates selected are shown in the Table S1.

Covariates	**BLT** (n = 9758)	**SLT** (n = 4548)
Recipient Age (years), Mean (SD)	53.4 (13.5)	63.5 (7.5)
Recipient BMI, Mean (SD)	24.8 (4.7)	26.3 (4.1)
Lung Allocation Score, Mean (SD)	40.8 (13.5)	40.3 (11.1)
Diabetes, n (%)	1999 (20.5%)	836 (18.4%)
Recipient Height (cm), Mean (SD)	170.0 (9.9)	170.6 (9.6)
Ventilator Use, n (%)	465 (4.8%)	71 (1.6%)
6-Minute Walk Distance, Mean (SD)	773.8 (426.1)	794.6 (411.4)
Donor Age (years), Mean (SD)	34.7 (14.0)	34.9 (13.8)
Donor BMI, Mean (SD)	26.2 (5.6)	26.2 (5.3)
CMV Serology Positive, n (%)	6093 (62.4%)	2793 (61.4%)
Donor Diabetes, n (%)	701 (7.2%)	315 (7.0%)
Ischemic Time (hours), Mean (SD)	5.6 (1.76)	4.3 (1.52)

values indicate a more diffuse and multidirectional structure. The estimated $k_{\text{eff}} \approx 137.5$ is substantial, implying that the BLT-SLT contrast varies along many covariate directions and is driven by a complex interplay of recipient characteristics rather than a single dominant clinical factor.

Given the detected heterogeneity, we proceed to estimate $\tau(\boldsymbol{X})$, referred to hereafter as the individual treatment effects (ITE), using the *deepHTL* estimator. As a population level baseline, the estimated ATE is 17.79 (95% CI: $16.84, 18.74$) mL/sec, confirming that BLT confers a significant functional advantage on average. However, the distribution of estimated ITE (Figure 2) exposes substantial variation obscured by this aggregate metric. The effects span a wide range (-20.6 to 59.2 mL/sec) and exhibit a distinct bimodal structure. While the vast majority of patients (93%, $n = 13{,}210$) are predicted to derive functional benefit from BLT, a clinically relevant minority (7%, $n = 986$) are predicted to achieve superior

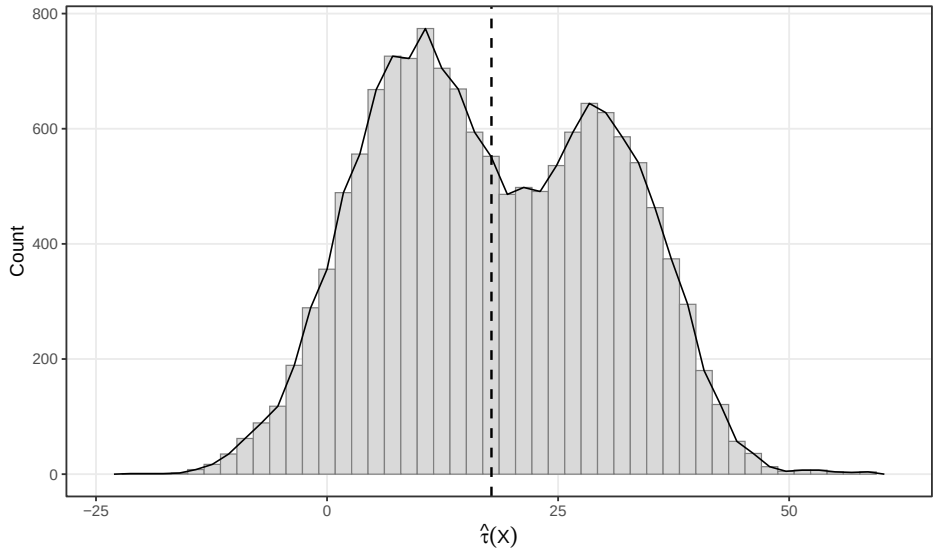

Figure 2: Estimated individualized treatment effects (BLT vs. SLT) in the whole UNOS lung transplant cohort. The dashed line denotes the average treatment effect.

outcomes with SLT. This suggests that a uniform BLT-for-all strategy risks exposing specific patient subgroups to the heightened surgical morbidity of bilateral transplantation without commensurate physiological gain.

To translate the individual estimates into actionable insights, we evaluate Group Average Treatment Effects (GATE) for key subpopulations (Figure 3). The functional advantage of BLT is the largest among patients with high physiological reserve, specifically, younger recipients (<50 years), those with low BMI (<18.5), and those with lower acuity (LAS < 35). Conversely, the benefit diminishes markedly for older recipients ($\geq$75 years) and individuals with high metabolic burden (obese class II/III, BMI ≥ 30). Interestingly, donor characteristics such as ischemic time exert comparatively little influence on the differential benefit. These findings support a precision medicine approach to organ allocation: given the scarcity of donor organs, bilateral transplantation provides the greatest benefit among younger,

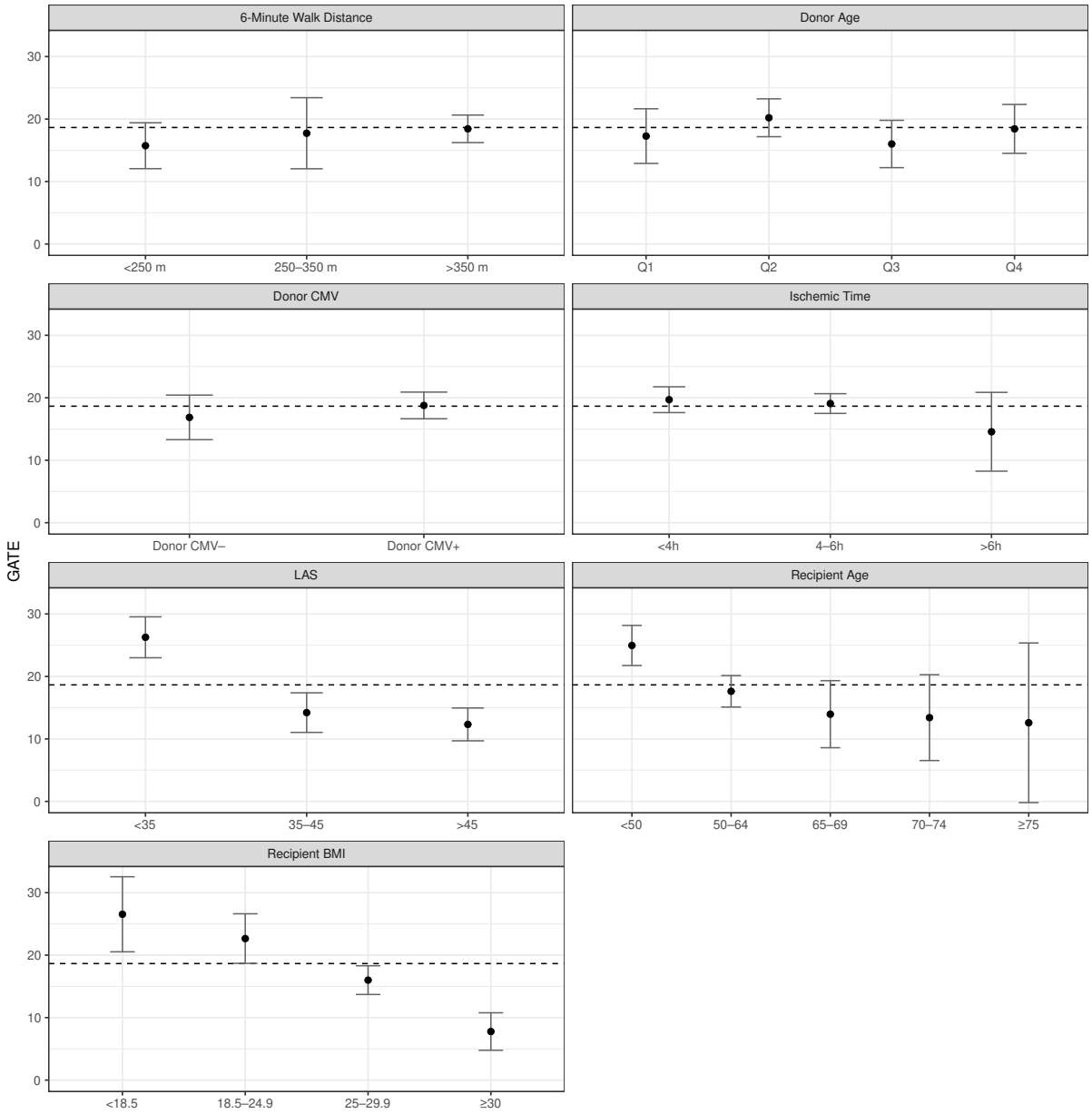

Figure 3: Estimation of GATE (BLT vs. SLT) across prespecified clinical subgroups. The horizontal dashed line denotes the average treatment effect.

low-risk, and lower-BMI candidates, whereas SLT remains a viable, resource-efficient option for older or higher-risk recipients. As a robustness check, an unsupervised k-means clustering ($k = 5$) of the estimated ITE (see Figure S1) recovers clinically meaningful phenotypes that closely mirror these GATE patterns, confirming that the observed heterogeneity is driven by these core physiological characteristics.

Our findings should be interpreted in light of several limitations inherent to observational registry analyses. First, causal inference relies on the assumption of no unmeasured confounding. While we adjusted for 26 key clinical covariates selected via a robust GL-DR procedure, unrecorded factors such as donor's lung quality, center-specific preferences for bilateral procedures, or granular frailty metrics guiding surgical candidate selection could simultaneously influence treatment assignment and FEV1 (Subramanian et al., 2018; Hayes et al., 2022). Second, we assume missing data were ignorable conditional on observed covariates, which may not hold for all variables. Finally, since the study cohort spans 2005 – 2011, generalizability to current practice may be affected by subsequent changes in donor allocation policies and surgical techniques. Despite these potential restrictions, the analysis demonstrates the utility of the proposed framework to uncover actionable heterogeneity in observational data.

5 Discussion

In this paper, we develop a unified framework for rigorous testing and estimation of heterogeneous treatment effects in observational studies with potentially complex confounding. Central to the framework is *deepHTL*, a revised deep semiparametric regression approach that targets both valid inference on the *existence* of heterogeneity and accurate estimation of covariate-dependent causal effects when heterogeneity is present. To gatekeep estimation with a statistically principled check, we propose a kernel score global test and a permutation

test that achieve Type I error control under complex confounding structures regardless of the true homogeneous treatment effect magnitude, while maintaining robust power to detect potential complex heterogeneity. Extensive simulations demonstrate finite sample robustness, and an application to the UNOS lung transplant registry reveals clinically meaningful subgroups that drive differential functional benefit of bilateral versus single lung transplantation. Although motivated by a biomedical case study, the framework is broadly applicable to settings where effect variation must be disentangled from intricate confounding.

Methodologically, *deepHTL* refines and extends existing learners in several directions. Building on the quasi-oracle R-learner of (Nie and Wager, 2021), we introduce a revision step that stabilizes estimation of $\mathbb{E}(Y \mid \boldsymbol{X})$ when treatment effects are not small, preventing the signal-to-noise–driven distortions that can invalidate downstream heterogeneity testing and the accuracy of HTE estimation. For nuisance function estimation, we use a bagged DNN ensemble (Mi et al., 2019, 2021; Zou et al., 2025) to reduce variance while retaining flexibility for complex (e.g., nonlinear and nonadditive) association. Finally, by embedding a global heterogeneity test as a first-stage decision rule, *deepHTL* addresses a critical but often overlooked inferential step and provides a rigorous procedure with controlled Type I error before estimating clinically meaningful individualized effects. It is worth pointing out that though *deepHTL* demonstrates its robustness for testing and estimating HTE under various confounding complexity, the validity relies on the assumption of absence of unmeasured confounding. Extending *deepHTL* to accommodate unmeasured confounding factors is practically important but it is out of scope of this paper.

Acknowledgments

This work was partially supported by the National Institutes of Health R01 (R01LM014407 and 1R01HL173044) grants.

SUPPLEMENTARY MATERIAL

R-package for *deepHTL*: R-package `deepHTL` containing code to perform the hypothesis testing and estimation methods described in the paper. The package is available at `https://github.com/BZou-lab/deepHTL`.

Lung transplant data: Raw registry data are not publicly available due to privacy restrictions but can be accessed through the United Network for Organ Sharing national registry upon approval.

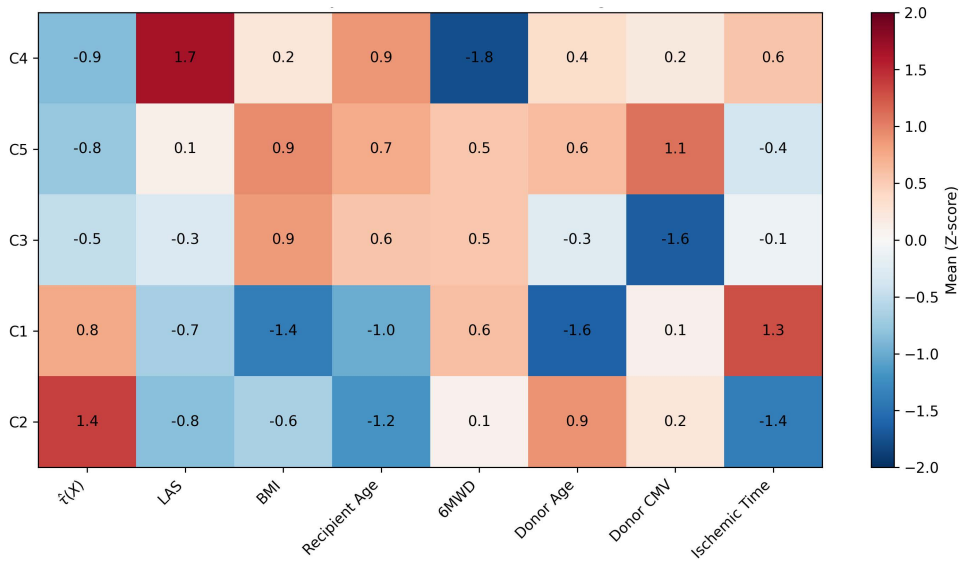

Figure S1: Heatmap of standardized cluster-level feature means derived from k-means clustering ($k = 5$ selected via average silhouette width). Clusters (rows) are ordered by increasing mean individualized treatment effect $\hat{\tau}(\boldsymbol{X})$ from bottom to top. Feature values (columns) are standardized (z-scores) to highlight relative differences: red indicates above-average and blue indicates below-average values. Consistent with GATE analyses, clusters with the highest projected benefit from BLT are characterized by younger recipient age, lower BMI, and lower Lung Allocation Scores (LAS).

Table S1: Baseline patient and donor characteristics used as confounders by transplant type.

Covariates	**BLT** (n = 9758)	**SLT** (n = 4548)
Recipient Age (years), Mean (SD)	53.4 (13.5)	63.5 (7.5)
Recipient BMI, Mean (SD)	24.8 (4.7)	26.3 (4.1)
Lung Allocation Score, Mean (SD)	40.8 (13.5)	40.3 (11.1)
Diabetes, n (%)	1999 (20.5%)	836 (18.4%)
TRR FeV1 %, Mean (SD)	37.2 (20.7)	43.6 (20.4)
TRF FEV1, Mean (SD)	78.2 (24.0)	65.4 (20.1)
Gender Female, n (%)	4081 (41.8%)	1610 (35.4%)
Gender Male, n (%)	5677 (58.2%)	2938 (64.6%)
Recipient Height (cm), Mean (SD)	170.0 (9.9)	170.6 (9.6)
Hispanic, n (%)	684 (7.0%)	305 (6.7%)
O_2 Requirement at Rest, Mean (SD)	5.6 (5.3)	4.7 (4.3)
O_2 Delivered, Mean (SD)	7.0 (6.6)	5.7 (5.2)
Ventilator Use, n (%)	465 (4.8%)	71 (1.6%)
6-Minute Walk Distance, Mean (SD)	773.8 (426.1)	794.6 (411.4)
Donor Age (years), Mean (SD)	34.7 (14.0)	34.9 (13.8)
Black Donor, n (%)	1798 (18.4%)	861 (18.9%)
Donor BMI, Mean (SD)	26.2 (5.6)	26.2 (5.3)
Donor Cigarette Use, n (%)	5499 (56.4%)	3164 (69.6%)
CMV Serology Positive, n (%)	6093 (62.4%)	2793 (61.4%)
Donor TBI (Cause of Death), n (%)	4236 (43.4%)	2006 (44.1%)
Donor Diabetes, n (%)	701 (7.2%)	315 (7.0%)
Distance to TX Center (NM), Mean (SD)	204.0 (246.3)	216.6 (253.3)
Donor Gender Female, n (%)	3857 (39.5%)	1683 (37.0%)
Donor Gender Male, n (%)	5901 (60.5%)	2865 (63.0%)
Donor Height (cm), Mean (SD)	171.7 (10.2)	172.3 (9.9)
Donor PCO_2, Mean (SD)	76.1 (40.1)	75.6 (34.0)
Allo Positive, n (%)	6926 (71.0%)	2949 (64.8%)
HgtR, Mean (SD)	1.0 (0.06)	1.0 (0.05)
Ischemic Time (hours), Mean (SD)	5.6 (1.76)	4.3 (1.52)
Sex Match, n (%)	6708 (68.7%)	3143 (69.1%)
Race Match, n (%)	5494 (56.3%)	2548 (56.0%)

Table S2: Cluster-level mean values of individualized treatment effects and key baseline features. Clusters with higher $\hat{\tau}(X)$ (C2, C1) exhibit lower LAS, lower BMI, and younger recipient age.

Cluster	Size	$\hat{\tau}(X)$	LAS	Isch(h)	6MWD(m)	BMI	Age(R)	Age(D)	CMV(%)
C2	3420	32.27	33.46	4.62	741.27	22.63	29.45	36.19	69.18
C1	2014	25.85	36.12	5.91	861.92	20.15	32.42	31.27	64.00
C3	3199	12.53	42.16	5.24	848.53	27.62	58.21	33.91	0.00
C5	4426	9.72	48.43	5.11	846.97	27.92	61.17	35.62	100.00
C4	1137	8.65	72.79	5.56	304.59	25.58	63.54	35.12	67.02

Note: LAS: Lung Allocation Score; Isch(h): ischemic time (hours); 6MWD(m): 6-min walk distance; BMI: recipient BMI; Age(R/D): recipient/donor age; CMV(%): donor CMV+.

Table S3: Estimation of GATE (BLT vs. SLT) across prespecified clinical subgroups.

Variable	**Subgroup**	**n**	**GATE**	**SE**	**95% CI**
Lung Allocation Score	<35	4,983	26.27	1.67	(22.99, 29.54)
	35–45	6,253	14.21	1.62	(11.04, 17.38)
	>45	2,960	12.32	1.34	(9.70, 14.95)
Ischemic Time	<4h	3,604	19.69	1.05	(17.63, 21.75)
	4–6h	6,880	19.07	0.80	(17.50, 20.65)
	>6h	3,712	14.56	3.22	(8.25, 20.86)
6-Min Walk Dist.	<250 m	1,862	15.73	1.87	(12.06, 19.40)
	250–350 m	666	17.73	2.90	(12.04, 23.42)
	>350 m	11,668	18.44	1.12	(16.23, 20.64)
Recipient BMI	<18.5	1,080	26.54	3.07	(20.53, 32.55)
	18.5–24.9	5,542	22.65	2.02	(18.68, 26.62)
	25–29.9	5,243	16.00	1.17	(13.70, 18.30)
	≥30	2,331	7.77	1.53	(4.77, 10.77)
Recipient Age	<50	3,067	24.96	1.64	(21.75, 28.17)
	50–64	6,859	17.62	1.29	(15.10, 20.14)
	65–69	3,032	13.96	2.74	(8.59, 19.32)
	70–74	1,090	13.41	3.50	(6.54, 20.28)
	≥75	148	12.58	6.52	(−0.19, 25.36)
Donor Age	Q1	3,568	17.27	2.23	(12.90, 21.65)
	Q2	3,827	20.21	1.55	(17.18, 23.24)
	Q3	3,335	16.01	1.93	(12.22, 19.80)
	Q4	3,466	18.43	2.00	(14.52, 22.34)
Donor CMV Status	CMV−	5,353	16.86	1.82	(13.29, 20.43)
	CMV+	8,843	18.77	1.09	(16.64, 20.91)

Table S4: Log-MSE (standard deviation) across simulation settings. (Numerical results for Figure 3).

Scenario I

		$n = 1000$		
Method	**p=20, σ=1**	**p=40, σ=1**	**p=20, σ=3**	**p=40, σ=3**
R–Lasso	0.252 (0.119)	0.292 (0.130)	0.534 (0.187)	0.824 (0.202)
Rev–Lasso	0.251 (0.119)	0.292 (0.130)	0.532 (0.189)	0.822 (0.202)
R–XGBoost	0.294 (0.150)	0.376 (0.145)	1.019 (0.169)	1.121 (0.168)
Rev–XGBoost	0.283 (0.144)	0.370 (0.133)	1.014 (0.166)	1.118 (0.160)
R–KRR	0.204 (0.121)	0.458 (0.118)	0.677 (0.187)	0.958 (0.176)
Rev–KRR	0.146 (0.123)	0.448 (0.121)	0.669 (0.192)	0.954 (0.171)
R–DNN	−0.682 (0.249)	−0.381 (0.207)	0.355 (0.237)	0.628 (0.183)
deepHTL	−0.711 (0.278)	−0.402 (0.246)	0.333 (0.246)	0.607 (0.185)
		$n = 2000$		
Method	**p=20, σ=1**	**p=40, σ=1**	**p=20, σ=3**	**p=40, σ=3**
R–Lasso	0.152 (0.074)	0.184 (0.077)	0.319 (0.118)	0.387 (0.132)
Rev–Lasso	0.152 (0.074)	0.183 (0.078)	0.319 (0.117)	0.387 (0.132)
R–XGBoost	−0.169 (0.131)	0.056 (0.124)	0.694 (0.155)	0.790 (0.151)
Rev–XGBoost	−0.188 (0.123)	−0.058 (0.125)	0.679 (0.150)	0.776 (0.147)
R–KRR	−0.382 (0.114)	0.183 (0.094)	0.356 (0.132)	0.668 (0.133)
Rev–KRR	−0.436 (0.103)	0.150 (0.095)	0.342 (0.138)	0.676 (0.137)
R–DNN	−0.847 (0.101)	−0.769 (0.096)	0.059 (0.143)	0.154 (0.128)
deepHTL	−0.919 (0.108)	−0.874 (0.111)	−0.101 (0.147)	0.113 (0.128)

Scenario II

		$n = 1000$		
Method	**p=20, σ=1**	**p=40, σ=1**	**p=20, σ=3**	**p=40, σ=3**
R–Lasso	0.410 (0.337)	0.423 (0.355)	0.459 (0.358)	0.455 (0.351)
Rev–Lasso	0.410 (0.337)	0.424 (0.352)	0.463 (0.358)	0.452 (0.352)
R–XGBoost	0.389 (0.375)	0.468 (0.365)	0.817 (0.368)	0.841 (0.353)
Rev–XGBoost	0.319 (0.369)	0.406 (0.372)	0.546 (0.377)	0.551 (0.368)
R–KRR	0.410 (0.356)	0.439 (0.360)	0.488 (0.371)	0.460 (0.357)
Rev–KRR	0.272 (0.372)	0.412 (0.357)	0.446 (0.358)	0.434 (0.349)
R–DNN	−0.110 (0.365)	0.353 (0.379)	0.382 (0.358)	0.401 (0.345)
deepHTL	−0.148 (0.481)	0.241 (0.394)	0.307 (0.357)	0.337 (0.345)
		$n = 2000$		
Method	**p=20, σ=1**	**p=40, σ=1**	**p=20, σ=3**	**p=40, σ=3**
R–Lasso	0.413 (0.337)	0.391 (0.327)	0.405 (0.325)	0.383 (0.308)
Rev–Lasso	0.413 (0.337)	0.392 (0.327)	0.406 (0.326)	0.382 (0.309)
R–XGBoost	0.193 (0.428)	0.299 (0.373)	0.623 (0.351)	0.550 (0.338)
Rev–XGBoost	0.156 (0.422)	0.285 (0.373)	0.460 (0.348)	0.409 (0.332)
R–KRR	0.081 (0.435)	0.403 (0.328)	0.427 (0.333)	0.405 (0.321)
Rev–KRR	−0.097 (0.467)	0.367 (0.331)	0.388 (0.326)	0.377 (0.310)
R–DNN	−0.269 (0.468)	−0.189 (0.309)	0.306 (0.334)	0.309 (0.308)
deepHTL	−0.292 (0.475)	−0.218 (0.320)	0.265 (0.333)	0.269 (0.309)

References

Chang, C., Jaki, T., Sadiq, M. S., Kuhlemeier, A. A., Feaster, D., Cole, N., Lamont, A., Oberski, D., Desai, Y., and Horn, M. L. V. (2021). A permutation test for assessing the presence of individual differences in treatment effects. *Statistics in Medicine*, 40(25):5690–5701.

Chen, T. and Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W. K., and Robins, J. M. (2018). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68.

Curth, A. and van der Schaar, M. (2021). Nonparametric estimation of heterogeneous treatment effects: From theory to learning algorithms. In *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 1810–1818.

Cybenko, G. (1989). Approximations by superpositions of sigmoidal functions. *Mathematics of Control, Signals, and Systems*, 2(4):303–314.

Dai, M. and Stern, H. S. (2022). A u-statistic-based test of treatment effect heterogeneity. *Journal of Nonparametric Statistics*, 34(1):141–163.

D'Amour, A., Ding, P., Feller, A., Lei, L., and Sekhon, J. (2021). Overlap in observational studies with high-dimensional covariates. *Journal of Machine Learning Research*, 22(209):1–54.

Davies, R. B. (1980). The distribution of a linear combination of χ^2 random variables. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 29(3):323–333.

Force, S. D., Kilgo, P., Neujahr, D. C., Pelaez, A., Pickens, A., Fernandez, F. G., Miller, D. L., and Lawrence, C. (2011). Bilateral lung transplantation offers better long-term survival, compared with single-lung transplantation, for younger patients with idiopathic pulmonary fibrosis. *Annals of Thoracic Surgery*, 91(1):244–249.

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

Hayes, D., Cherikh, W. S., Chambers, D. C., and Harhay, M. O. (2022). The impact of center volume on outcomes in lung transplantation. *The Annals of Thoracic Surgery*, 113(2):392–400.

Hornik, K., Stinchcombe, M., and White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural Networks*, 2(5):359–366.

Koch, B., Vock, D. M., and Wolfson, J. (2018). Covariate selection with group lasso and doubly robust estimation of causal effects. *Biometrics*, 74(1):8–17.

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(10):4156–4165.

Liu, D., Lin, X., and Ghosh, D. (2007). Semiparametric regression of multidimensional genetic pathway data: Least-squares kernel machines and linear mixed models. *Biometrics*, 63(4):1079–1088.

Lu, Z., Pu, H., Wang, F., Hu, Z., and Wang, L. (2017). The expressive power of neural networks: A view from the width. In *Advances in Neural Information Processing Systems (NIPS)*, pages 6231–6239.

Mi, X. et al. (2021). A deep learning semiparametric regression for adjusting complex confounding structures. *The Annals of Applied Statistics*, 15(3):1086–1100.

Mi, X., Zou, F., and Zhu, R. (2019). Bagging and deep learning in optimal individualized treatment rules. *Biometrics*, 75(2):674–684.

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

Oliveira, N. C. D., Osaki, S., Maloney, J., Cornwell, R. D., and Meyer, K. C. (2012). Lung transplant for interstitial lung disease: outcomes for single versus bilateral lung transplantation. *Interactive CardioVascular and Thoracic Surgery*, 14(3):263–267.

Piccart-Gebhart, M. J. et al. (2005). Trastuzumab after adjuvant chemotherapy in her2-positive breast cancer. *New England Journal of Medicine*.

Reck, M. et al. (2016). Pembrolizumab versus chemotherapy for pd-l1–positive non–small-cell lung cancer. *New England Journal of Medicine*.

RECOVERY Collaborative Group (2021). Dexamethasone in hospitalized patients with covid-19. *New England Journal of Medicine*.

Rigdon, J., Baiocchi, M., and Basu, S. (2018). Preventing false discovery of heterogeneous treatment effect subgroups in randomized trials. *Trials*, 19(1):382.

Robinson, P. M. (1988). Root-n-consistent semiparametric regression. *Econometrica*, 56(4):931–954.

Satterthwaite, F. E. (1946). An approximate distribution of estimates of variance components. *Biometrics Bulletin*, 2(6):110–114.

Schaffer, J. M., Singh, S. K., Joyce, D. L., Reitz, B. A., Zamanian, R. T., and Mallidi, H. R. (2015). Single- vs double-lung transplantation in patients with chronic obstructive pulmonary disease and idiopathic pulmonary fibrosis since the implementation of lung allocation based on medical need. *JAMA*, 313(9):936–948.

Subramanian, D. P., Cherikh, W. S., and Chambers, D. C. (2018). Bilateral versus single lung transplantation: are two lungs better than one? *Journal of Thoracic Disease*, 11(Suppl 14):S1792.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288.

Vovk, V. (2013). Kernel ridge regression. In *Empirical Inference: Festschrift in Honor of Vladimir N. Vapnik*, pages 105–116. Springer Berlin Heidelberg.

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

Young, R. P., Hopkins, R., and Eaton, T. E. (2007). Forced expiratory volume in one second: not just a lung function test but a marker of premature death from all causes. *European Respiratory Journal*, 30(4):616–622.

Zou, B., Mi, X., Wan, S., Wu, D., Xenakis, J. G., Hu, J., and Zou, F. (2025). A deep

neural network two-part model and feature importance test for semi-continuous data. *The Annals of Applied Statistics*, 19(2):1314–1331.