Enhancing Treatment Effect Estimation: A Model Robust Approach Integrating Randomized Experiments and External Controls using the Double Penalty Integration Estimator

Yuwen Cheng[1] **Lili Wu**[2] **Shu Yang**[3]

[1]Statistics Dept., North Carolina State University
[2]Microsoft Research NYC
[3]Statistics Dept., North Carolina State University

Abstract

Randomized experiments (REs) are the cornerstone for treatment effect evaluation. However, due to practical considerations, REs may encounter difficulty recruiting sufficient patients. External controls (ECs) can supplement REs to boost estimation efficiency. Yet, there may be incomparability between ECs and concurrent controls (CCs), resulting in misleading treatment effect evaluation. We introduce a novel bias function to measure the difference in the outcome mean functions between ECs and CCs. We show that the ANCOVA model augmented by the bias function for ECs renders a consistent estimator of the average treatment effect, regardless of whether or not the ANCOVA model is correct. To accommodate possibly different structures of the ANCOVA model and the bias function, we propose a double penalty integration estimator (DPIE) with different penalization terms for the two functions. With an appropriate choice of penalty parameters, our DPIE ensures consistency, oracle property, and asymptotic normality even in the presence of model misspecification. DPIE is at least as efficient as the estimator derived from REs alone, validated through theoretical and experimental results.

1 INTRODUCTION

Randomized experiments (REs), which allow researchers to scientifically quantify the impact of an intervention on a particular outcome of interest, are widely employed in a variety of areas. To make informed decisions, technology businesses always conduct A/B testing to evaluate new technologies, using a randomized experiment to compare the performance of each new software implementation with the previous version. Meanwhile, in the medical domain, randomized clinical trials (RCTs) ensuring no systematic differences between treatment groups are the cornerstone of treatment effect evaluation. When analyzing data from REs, analysis of covariance (ANCOVA) is a popular method that can provide consistent results, even if the model is misspecified. REs often require the use of external data to analyze treatment effects better: for example, A/B testing is time-consuming and requires a reasonably high number of users; thus, it is crucial to do a preliminary offline evaluation of external data to implement new interventions more efficiently and eliminate ineffective ones in advance [Gilotte et al., 2018]; meanwhile, if data from earlier clinical stages (Phase I or II) indicate that the product under investigation has a favorable benefit-risk profile in a disease area with unmet healthcare needs, then it is possible to design an RCT with a larger treatment group and a relatively smaller concurrent control (CC) group. Because the small CC group cannot provide sufficient power to the trial, it is reasonable to augment RCTs with external controls (ECs) from earlier trials [Yuan et al., 2019]. In this paper, we propose a new method that combines the ECs with CCs to improve average treatment effect (ATE) estimation.

Since Pocock [1976], who first introduced historical controls to incorporate external data into analysis, numerous statistical methods have been developed. Specifying a set of covariates in advance and then calculating the propensity score for matching, stratification, or weighting [Greenland et al., 1999, Rubin and Thomas, 1996, Rubin, 2007, Hernán and Robins, 2016] is typical. However, these methods rely on the exchangeability assumption that there are no unmeasured confounders between ECs and CCs, which is unlikely in real-world applications. Additionally, Bayesian approaches [Spiegelhalter, 2004, Hobbs et al., 2013, Schmidli et al., 2014, Ibrahim and Chen, 2000, Hobbs et al., 2012, Neuenschwander et al., 2009] can handle datasets combining both ECs and CCs: appropriate priors can be selected for incorporating the ECs after evaluating the relationship between the ECs and the CCs. Nevertheless, these methods can result in type I error inflation [Viele et al., 2014]. Building upon

Accepted for the 39th Conference on Uncertainty in Artificial Intelligence (UAI 2023).

the work of Stuart and Rubin [2008] and Yang et al. [2022], who introduced a parametric bias function to adjust for the outcome heterogeneity between the control groups due to unmeasured confounders, Wu and Yang [2022] advanced the idea by using the sieve estimation approach [Chen, 2007] to estimate the unknown outcome model and bias function. The bias function in their approach measures the difference between experimental data and observational data. However, Wu and Yang [2022] did not fully leverage the advantage of REs [Wang et al., 2023] since these prior works were based on the assumption that the outcome model was correctly specified. In this paper, we extend the concept of the bias function to handle cases of possible model misspecification. Our bias function measures the difference between the EC outcome mean function and the working model in REs and guarantees consistency even when the outcome mean model is not correctly specified, providing a robust solution to the challenge of model misspecification. This is of great practical significance as simple models such as ANCOVA are commonly used despite the possibility of misspecification.

We adopt the nonparametric sieve estimation approach [Chen, 2007] to accurately estimate the bias function, therefore it is important to utilize feature selection techniques. These techniques tackle the high-dimensional aspect of the basis functions used in sieve estimation and address the possibility of the working outcome mean model containing irrelevant covariates. To resolve this, penalized terms can be added to each parameter in the objective function for optimization along with regularization parameters. However, using the same regularization terms for both the REs and ECs can cause issues due to their differing levels of sparsity and magnitude. Therefore, we implement different penalized terms for the unknown bias function and the working outcome mean model, considering their distinct levels of sparsity and magnitude.

Multiple methods have employed different penalties for different goals. However, they focused on decomposing one function into different parts and applying different penalties to those parts. Chernozhukov et al. [2017] proposed the Lava estimator by decomposing the signals into a dense part and a sparsity part, and then applying different penalties to each component; Bühlmann and Ćevid [2020] proposed a spectral deconfounding approach to estimate sparse parameters given hidden variables, and demonstrated the Lava method [Chernozhukov et al., 2017] as one of their special cases; Xing et al. [2021], further, focused on the estimation of multivariate regression with hidden variables, and demonstrated their method can be viewed as the multivariate generalization of the Lava approach [Chernozhukov et al., 2017]. In addition to decomposing the parameters into sparse and dense parts, Wang and Zhou [2019] decomposed the function into an easy-to-interpret part and an uninterpretable part, and then applied double penalties. Our approach makes use of double penalties to deal with the possible different structures of the working outcome mean function and bias function to consistently select useful terms and enhance the efficiency of the ATE estimator.

Different bias functions can utilize varying penalties, depending on their structure. Our study focuses on the use of the Smoothly Clipped Absolute Deviation (SCAD) penalty [Fan and Li, 2001, Fan and Peng, 2004] to illustrate the theorem in the context of variable selection. This is because SCAD offers both oracle properties and asymptotic normality by selecting the appropriate regularization parameters. Nevertheless, the existing results are limited to situations where the models are correctly specified. To overcome this limitation, we present a novel proof for the SCAD penalty that extends its desirable properties to scenarios involving potential misspecification of models.

Our main contributions can be summarized as follows:

a) We present a novel bias function to combine REs and ECs and use sieve estimation [Chen, 2007] to provide a flexible and computationally feasible way of estimating the unknown bias function. Our ATE estimator for REs is consistent regardless of the specification of the working outcome mean model.

b) We introduce the Double Penalty Integration Estimator (DPIE), which employs different penalized terms for the unknown bias function and the working outcome mean model to differentiate their different levels of sparsity and magnitude. We prove that DPIE guarantees consistency for the parameters that minimize the least squares loss and has the oracle property of only selecting non-zero parameters and exhibiting asymptotic normality under the SCAD penalties.

c) We demonstrate that combining different data sources results in a more efficient estimated ATE than using only REs, as long as the number of basis function terms in the bias function is fewer than that of the working outcome mean function. The oracle property of DPIE ensures the selection of relevant basis terms when the outcome mean function is more complex and less smooth than the bias function, leading to improved efficiency. On the other hand, using single penalties may result in a loss of the oracle property and failure to select useful basis terms, leading to decreased efficiency.

The rest of the paper is organized as follows. We introduce the basic idea in Section 2. Section 3 introduces the proposed DPIE estimator and derives the theorem. We conduct simulations for comparison in Section 4. Section 5 applies the proposed estimators to an observational study from the National Supported Work (NSW) and Current Population Survey (CPS). Finally, we conclude the paper with a discussion in Section 6.

2 PROBLEM SETUP

Denote $X \in \mathcal{X} \subset \mathbb{R}^d$ as the vector of pre-treatment covariates, $A \in \{0,1\}$ as the binary treatment, and $Y \in \mathbb{R}$ as the outcome of interest. Following the potential outcomes framework [Splawa-Neyman et al., 1990, Rubin, 1974], let $Y(a)$ be the potential outcome for the subject given the treatment $a, a = 0, 1$.

In real life, one can use previous trials or real-world data as ECs to supplement the REs. Assuming two data sources are accessible: the RE data source having n independent and identically distributed (i.i.d.) subjects $\{(X_i, A_i, Y_i) : i \in \mathcal{I}_{\mathrm{RE}}\}$ with n_1 concurrent treatments $\{(X_i, 1, Y_i) : i \in \mathcal{I}_{\mathrm{CT}}\}$ and n_0 concurrent controls $\{(X_i, 0, Y_i) : i \in \mathcal{I}_{\mathrm{CC}}\}$, and the EC data source with m i.i.d. subjects with $\{(X_i, 0, Y_i) : i \in \mathcal{I}_{\mathrm{EC}}\}$. Let $N = n + m$ be the total sample size. Define S as the indicator of the subject in the REs: $S_i = 1$ for $i \in \mathcal{I}_{\mathrm{RE}}$ and $S_i = 0$ for $i \in \mathcal{I}_{\mathrm{EC}}$. Then the ATE is $\tau = \mathbb{E}\{Y(1) - Y(0) \mid S = 1\}$. Further, let $e(X) = \mathbb{P}(A = 1 \mid X, S = 1)$ be the propensity score and also define the conditional outcome mean function as $\mu_{a,s}(X) = \mathbb{E}(Y \mid X, A = a, S = s)$ for $a = 0, 1$ and $s = 0, 1$.

One of the fundamental challenges to identifying the ATE is that $Y(1)$ and $Y(0)$ cannot be observed simultaneously. To overcome this issue, we make the following three common assumptions in the causal inference literature [Rubin, 1978]:

Assumption 1 $\{Y(0), Y(1)\} \perp\!\!\!\perp A \mid X, S = 1$ *almost surely, where* $\perp\!\!\!\perp$ *means "independent of".*

Assumption 2 $Y = Y(1)A + Y(0)(1 - A)$.

Assumption 3 *There exist constants* c_1 *and* c_2 *such that* $0 < c_1 \le e(X) \le c_2 < 1$ *almost surely.*

Assumption 1 states that the treatment assignment is unconfounded in the REs. Assumption 2 guarantees potential outcomes are unaffected by the received treatments. Under Assumptions 1 and 2, with $X = x$, we have $\tau(x) = \mathbb{E}\{Y(1) - Y(0) \mid X = x, S = 1\} = \mathbb{E}(Y \mid X = x, A = 1, S = 1) - \mathbb{E}(Y \mid X = x, A = 0, S = 1)$, and the conditional outcome mean function under the randomized experiments is $\mu_{a,1}(X) = \mathbb{E}\{Y(a) \mid X, S = 1\} = \mathbb{E}(Y \mid X, A = a, S = 1)$. Assumption 3 implies a sufficient overlap of the covariate distribution between the treatment groups, then averaging the treatment effect on the distribution of X is feasible, thus the ATE is $\tau = \mathbb{E}\{\tau(X)\} = \mathbb{E}\{\mathbb{E}(Y \mid X, A = 1, S = 1) - \mathbb{E}(Y \mid X, A = 0, S = 1)\} = \mathbb{E}\{\mu_{1,1}(X) - \mu_{0,1}(X)\}$.

To confirm the overlap of the covariate spaces of ECs and CCs, hence enabling the utilization of ECs to augment RCTs to boost estimation efficiency, we consider the following assumption:

Assumption 4 $\mathbb{P}(S = s|X) > 0$ *for all* s *almost surely.*

The Analysis of covariance (ANCOVA) model is a powerful tool for estimating the ATE in REs. The randomization design allows for the ATE estimator $\hat{\tau}$ to be consistent and asymptotically normal under arbitrary misspecification of the ANCOVA model [Wang et al., 2023]. Following the common practice, we use the ANCOVA model as the working model in REs. To enhance the model's generality, we incorporate a k_1-dimension basis function of X $p_\mu(X) = \{p_{\mu,1}(X), \ldots, p_{\mu,k_1}(X)\}^{\mathrm{T}}$ into the ANCOVA model as $\bar{\mu}_{A,1}(X;\beta) = \beta_{\mathrm{int}} + \beta_A A + \beta_X^{\mathrm{T}} p_\mu(X)$, where β is a $K_1 = (k_1 + 2)$-dimensional parameter $(\beta_{\mathrm{int}}, \beta_A, \beta_X^{\mathrm{T}})^{\mathrm{T}}$. Under the ANCOVA model, it is common to utilize ordinary least squares estimators for parameter estimation. Denote $\beta_* = (\beta_{\mathrm{int}*}, \beta_{A*}, \beta_{X*}^{\mathrm{T}})^{\mathrm{T}}$ as the minimizer of $\mathbb{E}[\{Y - \bar{\mu}_{A,1}(X;\beta)\}^2 \mid S = 1]$. Importantly, for $S = 1$, β_{A*} is the ATE τ regardless of the correctness of the working model.

To use ECs to supplement CCs, it is crucial to remove biases of EC data due to possible incomparability between ECs and CCs. We define the bias function as

$$b_0(X) = \mathbb{E}(Y \mid X, A = 0, S = 0) - \bar{\mu}_{0,1}(X;\beta).$$

If the working model $\bar{\mu}_{0,1}(X;\beta)$ is correctly specified, the bias function reduces to $\mathbb{E}(Y \mid X, A = 0, S = 0) - \mathbb{E}(Y \mid X, A = 0, S = 1)$, which measures the difference of the conditional mean of the control outcome given X between ECs and CCs. In this case, if X captures all confounders of S and Y, then $\mathbb{E}(Y \mid X, A = 0, S = 1) = \mathbb{E}(Y \mid X, A = 0, S = 0)$, and thus $b_0(X) \equiv 0$; otherwise, $b_0(X) \neq 0$. This special case is discussed in Wu and Yang [2022], but their analysis requires the outcome model to be correctly specified. In contrast, our setup does not necessitate a correctly specified outcome model.

Further, let the ANCOVA model augmented by the bias function $b_0(X)$ be $\bar{\mu}_{A,S}(X;\beta) = \beta_{\mathrm{int}} + \beta_A A + \beta_X^{\mathrm{T}} p_\mu(X) + (1 - S) b_0(X)$, then

$$\begin{aligned}
&\bar{\mu}_{0,0}(X;\beta) = \beta_{\mathrm{int}} + \beta_X^{\mathrm{T}} p_\mu(X) + b_0(X) \\
=&\bar{\mu}_{0,1}(X;\beta) + \mathbb{E}(Y \mid X, A = 0, S = 0) - \bar{\mu}_{0,1}(X;\beta) \\
=&\mathbb{E}(Y \mid X, A = 0, S = 0).
\end{aligned}$$

An important implication is that even if the outcome working model is misspecified, incorporating the bias function $b_0(X)$ ensures that $\bar{\mu}_{0,0}(X;\beta)$ recovers the true outcome mean under ECs. Denote $\beta_* = (\beta_{\mathrm{int}*}, \beta_{A*}, \beta_{X*}^{\mathrm{T}})^{\mathrm{T}}$ as the minimizer of $\mathbb{E}[\{Y - \bar{\mu}_{A,S}(X;\beta)\}^2]$. The following theorem demonstrates that β_{A*} still identifies the ATE τ in the combined RE and EC data.

Theorem 1 *(Identification) Under the augmented ANCOVA model, we have* $\beta_{A*} = \tau$.

Theorem 1 provides a vehicle to integrate REs and ECs for robust estimation of the ATE. Consistent estimation of τ still depends on an accurate approximation of unknown $b_0(X)$. Thus, we adopt the method of sieves [Chen, 2007]. Denote $p_b(X)$ as the K_2-dimension basis functions. Based on Theorem S1 in the Supplementary Material, there exists a K_2-vector δ_* such that the uniform convergence $p_b^{\mathrm{T}}(X)\delta_* \to b_0(X)$ and therefore the uniform convergence

$$\begin{aligned}&\beta_{\text{int}} + \beta_A A + \beta_X^{\mathrm{T}} p_\mu(X) + (1-S)p_b^{\mathrm{T}}(X)\delta_* \\ \to&\beta_{\text{int}} + \beta_A A + \beta_X^{\mathrm{T}} p_\mu(X) + (1-S)b_0(X)\end{aligned}$$

hold as $K_2 \to \infty$. Then our final working model becomes

$$\bar{\mu}_{A,S}(X;\beta,\delta) = \beta_{\text{int}}+\beta_A A+\beta_X^{\mathrm{T}} p_\mu(X)+(1-S)\delta^{\mathrm{T}} p_b(X).$$

We consider using the least squares loss function to obtain estimators for β_* and δ_*. To overcome the risk of overfitting in sieve estimation, where high-dimensional basis functions are used, it is necessary to add regularizers to the target function. Additionally, as the working models $\bar{\mu}_{A,1}(X;\beta)$ may contain irrelevant covariates, adding regularizers to the target functions is recommended to select proper covariates. In the subsequent section, we thoroughly explore the significance of the structural properties of regularizers to effectively accommodate the inherent characteristics of the problem at hand.

3 A DOUBLE PENALTY REGULARIZATION METHOD

3.1 MAIN IDEA

The common penalty regularization methods use the same regularization parameter for the two penalty functions on β_* and δ_*. However, it is important to note that β_* and δ_* may have distinct complexities, and it is beneficial to apply different penalties, instead of the same penalty, to both parameters.

To begin with, we denote $P_\lambda(\gamma) = \lambda P(\gamma)$ as the penalty function with regularization parameter λ for any parameter γ and any penalization $P(\cdot)$. There are various choices for the penalty function, like Lasso [Tibshirani, 1996], Smoothly Clipped Absolute Deviation (SCAD) [Fan and Li, 2001, Fan and Peng, 2004] penalties, or use black-box methods like the random forest. To overcome the limitations caused by adding the same penalty to all parameters, we set up different penalties for β and δ, separately. In other words, we provide a double penalty regularization method (DPIE) for the whole parameters set (β,δ) based on the integration data. Hence, the penalized least squares estimator using double penalties is

$$\begin{aligned}\left(\hat{\beta},\hat{\delta}\right) = \underset{\beta,\delta}{\operatorname{argmin}}\Bigg[&\sum_{i=1}^{N}\{Y_i - \bar{\mu}_{A_i,S_i}(X_i;\beta,\delta)\}^2 \\ &+ N\sum_{j=1}^{K_1} P_{\lambda_{1,j}}(|\beta|) + N\sum_{j=1}^{K_2} P_{\lambda_{2,j}}(|\delta|)\Bigg].\end{aligned}$$

There are multiple ways to search two regularization parameters λ_1 and λ_2. One simple way is to define a scaling tuning parameter sc as $sc = \lambda_2/\lambda_1$, then one can use cross-validation to choose sc given a particular search range and within each sc value, one can also use cross-validation to choose λ_2. Both cross-validation steps can use the software `cv.ncvreg` function in the R package `ncvreg`, which finds the tuning parameter based on the minimum cross-validated error.

On the other hand, if we only use the RE data, we have

$$\begin{aligned}\hat{\beta}_{\text{RE}} = \underset{\beta}{\operatorname{argmin}}\Bigg[&\sum_{i=1}^{N} S_i\{Y_i - \bar{\mu}_{A_i,S_i}(X_i;\beta,\delta)\}^2 \\ &+ n\sum_{j=1}^{K_1} P_{\lambda_{1,j}}(|\beta|)\Bigg].\end{aligned}$$

3.2 THEORETICAL ANALYSIS

The goal of this section is to derive the statistical properties of the DPIE. More interestingly, we aim to show the DPIE is at least as efficient as the ATE estimator based only on the REs.

For concreteness, we will focus on using the SCAD penalty to illustrate the theorem for variable selection. It is worth mentioning that our setup can also be developed for different penalty functions, but we will not be discussing those in this study. When the working function $\bar{\mu}_{A,S}(X;\beta,\delta)$ is correctly specified, the penalized maximum likelihood estimator with the SCAD penalty performs both oracle properties as well as asymptotic normality by selecting the appropriate regularization parameters λ [Fan and Li, 2001, Fan and Peng, 2004]. In the following, we will show the DPIE derived under the penalized least squares loss function with double SCAD penalties also has the oracle property and asymptotic normality under the working model $\bar{\mu}_{A,S}(X;\beta,\delta)$.

Following the framework in Fan and Li [2001], Fan and Peng [2004], we rewrite the working model as

$$\begin{aligned}&\bar{\mu}_{A,S}(X;\beta,\delta) \\ =&\beta_{\text{int}} + \beta_A A + \beta_X^{\mathrm{T}} p_\mu(X) + (1-S)\delta^{\mathrm{T}} p_b(X) \\ =&p^{\mathrm{T}}\theta,\end{aligned}$$

where $\theta^{\mathrm{T}} = (\beta^{\mathrm{T}},\delta^{\mathrm{T}})$ and $p = \{1, A, p_\mu^{\mathrm{T}}(X), (1-S)p_b^{\mathrm{T}}(X)\}^{\mathrm{T}}$ with dimension $K = K_1 + K_2$. In REs, we

have $p^{\mathrm{T}}\theta = p_\mu^{\mathrm{T}}\beta$, and in ECs, we have $p^{\mathrm{T}}\theta = p_\mu^{\mathrm{T}}\beta + p_b^{\mathrm{T}}\delta$. Denote g as the unknown true density function of (Y, p) and f as the density function such that minimizing the least squares loss is equivalent to maximizing the quasi-log-likelihood function; the Gaussian distribution is one such example. Denote $P_\lambda(\theta)$ as the SCAD penalty function:

$$P_\lambda^{'}(\theta) = \lambda\left\{\mathbf{1}(\theta \le \lambda) + \frac{(a\lambda - \theta)_+}{(a-1)\lambda}\mathbf{1}(\theta > \lambda)\right\}$$

for some $a > 2$ and $\theta > 0$, where $P_\lambda^{'}(\theta)$ is the first order derivative of $P_\lambda(\theta)$. Then one can rewrite the penalized least squares estimator as the penalized quasi-likelihood estimator

$$\begin{aligned}\left(\hat{\beta}, \hat{\delta}\right) &= \underset{\beta,\delta}{\operatorname{argmax}}\, Q(\beta, \delta) \\ &= \underset{\beta,\delta}{\operatorname{argmax}}\left[\sum_{i=1}^{N}\{\ln f(Y_i, p_i, \beta, \delta)\}\right. \\ &\quad \left. - N\sum_{j=1}^{K_1} P_{\lambda_{1,j}}(|\beta|) - N\sum_{j=1}^{K_2} P_{\lambda_{2,j}}(|\delta|)\right].\end{aligned}$$

Let

$$\alpha_N = \max_{1\le j_1\le K_1, 1\le j_2\le K_2}\left\{\begin{matrix} P_{\lambda_1}^{'}(|\beta_{*,j_1}|), P_{\lambda_2}^{'}(|\delta_{*,j_2}|), \\ \beta_{*,j_1}\neq 0, \delta_{*,j_2}\neq 0\end{matrix}\right\},$$

$$b_N = \max_{1\le j_1\le K_1, 1\le j_2\le K_2}\left\{\begin{matrix} P_{\lambda_1}^{''}(|\beta_{*,j_1}|), P_{\lambda_2}^{''}(|\delta_{*,j_2}|), \\ \beta_{*,j_1}\neq 0, \delta_{*,j_2}\neq 0\end{matrix}\right\},$$

where $P_\lambda^{''}(\theta)$ is the second-order derivative of $P_\lambda(\theta)$. We present the regularity conditions on the penalty functions given by Fan and Peng [2004]:

Assumption 5 *Let the values of $\beta_{*,1}, \ldots, \beta_{*,s_1}$ be nonzero and $\beta_{*,s_1+1}, \ldots, \beta_{*,K_1}$ be zero. Similarly, let the values of $\delta_{*,1}, \ldots, \delta_{*,s_2}$ be nonzero and $\delta_{*,s_2+1}, \ldots, \delta_{*,K_2}$ be zero. Then β_*, δ_* satisfy:*

$$\min_{1\le j\le s_1}|\beta_{*,j}|/\lambda_1 \to \infty,\ \min_{1\le j\le s_2}|\delta_{*,j}|/\lambda_2 \to \infty,$$
$$\max_{s_1+1\le j\le K_1}|\beta_{*,j}|/\lambda_1 \to 0,\ \max_{s_2+1\le j\le K_2}|\delta_{*,j}|/\lambda_2 \to 0,$$

as $N \to \infty$.

Fan and Peng [2004] showed under Assumption 5, the SCAD penalties have $\alpha_N = 0$ and $b_N = 0$ as N large enough, where the former ensures the existence of root-N/K-consistent penalized likelihood estimator, and the latter ensures the penalty function does not have much more influence on the penalized likelihood functions, making the penalty estimator have the same efficiency as the maximum likelihood estimator.

Denote $\|v\|_p$ as the $\mathcal{L}_p$-norm of a vector v. Based on these assumptions, we can provide the consistency and the asymptotic normality of the estimated parameters.

Theorem 2 *Suppose that the density function $f(Y, p, \beta, \delta)$ and the true density function $g(p, Y)$ satisfy Assumptions S1–S10 on the Supplementary Material, and the SCAD penalty functions $P_{\lambda_1}(\cdot), P_{\lambda_2}(\cdot)$ satisfy Assumption 5. If $K^4/N \to 0$ as $N \to \infty$, then there is a local maximizer $\left(\hat{\beta}, \hat{\delta}\right)$ of $Q(\beta, \delta)$ such that $\|\hat{\beta} - \beta_*\|_2 = O_p\left\{(K/N)^{1/2}\right\}, \|\hat{\delta} - \delta_*\|_2 = O_p\left\{(K/N)^{1/2}\right\}$.*

Denote $\beta_* = (\beta_{*1}^{\mathrm{T}}, \beta_{*2}^{\mathrm{T}})^{\mathrm{T}}$, where $\beta_{*1} \neq 0$ with s_1 dimensions and $\beta_{*2} = 0$. Similarly, denote $\delta_* = (\delta_{*1}^{\mathrm{T}}, \delta_{*2}^{\mathrm{T}})^{\mathrm{T}}$, $\theta_* = (\theta_{*1}^{\mathrm{T}}, \theta_{*2}^{\mathrm{T}})^{\mathrm{T}}$ where $\delta_{*1} \neq 0$ with s_2 dimensions and $\delta_{*2} = 0$, and $\theta_{*1} \neq 0$ with $s = s_1 + s_2$ dimensions and $\theta_{*2} = 0$. Then we have the following theorem:

Theorem 3 *Under Assumption 5 and Assumptions S1–S10 in the Supplementary Material, if $\lambda_1, \lambda_2 \to 0$, $\sqrt{N/K}\lambda_1 \to \infty, \sqrt{N/K}\lambda_2 \to \infty$ and $K^5/N \to 0$ as $N \to \infty$, then with probability tending to 1, $\hat{\beta}, \hat{\delta}$ in Theorem 2 must satisfy*

a) (Sparsity) $\hat{\beta}_2 = 0, \hat{\delta}_2 = 0$.

b) (Asymptotic normality)

$$\begin{aligned}&\sqrt{N}WA^{1/2}(\theta_{*1})\left(\hat{\theta}_1 - \theta_{*1}\right) \\ &\to \mathcal{N}\left(0, WA^{-1/2}(\theta_{*1})B(\theta_{*1})A^{-1/2}(\theta_{*1})W^{\mathrm{T}}\right)\end{aligned}$$

in distribution, where W is a $q \times s$ matrix such that $WW^{\mathrm{T}} \to G$, and G is a $q \times q$ nonnegative symmetric matrix. For simplicity, the specific forms of W, $A(\theta)$ and $B(\theta)$ are deferred to the Supplementary Materials.

Theorems 2-3 demonstrate that under proper selection of tuning parameters, the estimator $\hat{\theta}$ is consistent and asymptotic normal for θ_*.

However, using a single λ may fail to achieve such desirable results. To emphasize the importance of adding different penalties, we provide a simple analytical calculation: let $\beta_{1*} = O_p(N^{-1/2}), \beta_{2*} = O_p(N^{-1}), \delta_{1*} = O_p\left(N^{-1/10}\right), \delta_{2*} = O_p\left(N^{-1/3}\right)$. Assuming $\lambda_1 = N^\epsilon$ and $\lambda_2 = N^\gamma$ such that $\beta_{1*}, \beta_{2*}, \delta_{1*}$ and δ_{2*} all satisfy Assumption 5:

$$\frac{N^{-1/2}}{N^\epsilon} \to \infty,\ \frac{N^{-1}}{N^\epsilon} \to 0,$$
$$\frac{N^{-1/10}}{N^\gamma} \to \infty,\ \frac{N^{-1/3}}{N^\gamma} \to 0.$$

Hence, we have $-1 < \epsilon < -1/2$ and $-1/3 < \gamma < -1/10$, which means we cannot find a common λ for β_* and δ_*. If the magnitudes of parameters are much different, one penalty cannot satisfy the requirements for consistency and oracle properties. A toy numerical experiment in the Supplemental Materials demonstrates that utilizing different

penalties for β_* and δ_* yields superior performance compared to using a single penalty, particularly when β_* and δ_* have different magnitudes. We also illustrate this in the Simulation.

It should be noted that Theorem 2-3 can be expanded to incorporate different forms of f, not just the form equivalent to the least squares loss. This means we can extend to other losses, not solely the least squares loss. If f isn't the form equivalent to the least squares loss, under certain regularity conditions of f and g, the $\hat{\theta}$ that maximizes the penalized quasi-log-likelihood function $Q(\theta)$ will converge to $\tilde{\theta}_*$, where $\tilde{\theta}_*$ minimizes the Kullback-Leibler Information Criterion (KLIC) between f and g, $KLIC(g : f, \theta) = \mathbb{E}\left[\log\left\{g(Y,p)/f(Y,p,\theta)\right\}\right]$. Detailed discussions are included in the Supplementary Material.

3.3 COMPARISON BETWEEN THE DPIE AND THE RE-ONLY ESTIMATOR

We now show the advantage of the DPIE based on the combined data compared with the estimator based only on the REs. To facilitate comparison, we consider a simplifying assumption that the residual error ϵ in $Y = \bar{\mu}_{A,S}(X) + \epsilon$ is homogeneous with $\mathbb{V}(\epsilon) = \sigma^2$.

Theorem 4 *Under Assumption 5 and Assumptions S1–S10 in Supplementary Material, if $\lambda_1, \lambda_2 \to 0$, $\sqrt{N/K}\lambda_1 \to \infty$, $\sqrt{N/K}\lambda_2 \to \infty$ and $K^5/N \to 0$ as $N \to \infty$. For the least squares estimate $\hat{\beta}$ and $\hat{\delta}$,*

a) in combined data, $\hat{\tau} - \tau \to \mathcal{N}\left\{0, \mathbb{V}(\hat{\tau})\right\}$ in distribution;

b) in the RE data, $\hat{\tau}_{\mathrm{RE}} - \tau \to \mathcal{N}\left\{0, \mathbb{V}(\hat{\tau}_{\mathrm{RE}})\right\}$ in distribution; and

c) $\mathbb{V}(\hat{\tau}_{\mathrm{RE}}) \geq \mathbb{V}(\hat{\tau})$ and the inequality holds iff $p_\mu = Mp_b(X)$ for some matrix M.

Theorem 4 shows that the estimate of τ will remain accurate, regardless of the validity of the ANCOVA working model. When incorporating ECs, the estimator is no less efficient than the estimator obtained from the REs alone.

4 SIMULATION

In this section, first, we illustrate the importance of adding different penalties for different data sources, then we compare the proposed estimator of τ with existing competitors combining the REs and the ECs.

4.1 SIMULATION STUDY 1

We generate two data sources with sample sizes $n = m = 1000$. Covariates $X \in \mathbb{R}^{50}$ are generated by $X_d \sim \mathrm{Uniform}\left[1-\sqrt{3}, 1+\sqrt{3}\right], d = 1, \ldots, 50$, and outcome is generated by $Y = X^{\mathrm{T}}\beta_0 + (1-S)X^{\mathrm{T}}\delta_0 + \epsilon$, where $\epsilon \sim \mathcal{N}(0,1)$. We simulate $T = 100$ Monte Carlo times, and specify the true $\beta_0 = (1, \ldots, 50)^{\mathrm{T}}/50$.

To examine the instances where various penalties are required and validate the argument presented in Section 3, we specify three cases for δ_0:

a) $\|\delta_0\|_1 \geq \|\beta_0\|_1$ and half of parameters in δ_0 equal to zero: $c\|\beta_0\|_1 = \|\delta_0\|_1$ and $c = 1, 3, 5, 7, 9$.

b) $\|\delta_0\|_1 < \|\beta_0\|_1$ and half of parameters in δ_0 equal to zero: $c\|\beta_0\|_1 = \|\delta_0\|_1$ and $c = 0.1, 0.3, 0.5, 0.7, 0.9$.

c) Vary the sparsity level of δ_0 while ensuring that its magnitude satisfies $\|\delta_0\|_1 = \|\beta_0\|_1$: the number of variables in δ_0 equal to zero one by one.

In each case, we compare the results based on

a) Results based on the combined data and use the double SCAD penalty (denoted as "DPIE").

b) Results based on the combined data and use the single SCAD penalty (denoted as "SPIE").

c) Results based only on the RE data (denoted as "RE").

All results are based on re-fitting models with the parameters chosen in each method, and compare the results based on the mean squared error $MSE = \sqrt{d^{-1}\sum_{i=1}^{d}\left(\hat{\beta}_i - \beta_{0,i}\right)^2}$ and the percentage of incorrectly selecting more (denoted as "Over-select") and fewer parameters (denoted as "Under-select"). Figure 1 shows the MSE results and the percentage of Under-select and Over-select in case a). When the magnitudes of two parameters differ, using different penalties improves accuracy when compared to using the same penalties for all parameters. Moreover, the gained accuracy improves as the magnitude difference increases. The right panel of Figure 1 shows the percentage of incorrectly selecting more or fewer variables. When $\|\delta_0\|_1 > \|\beta_0\|_1$, using the same penalties makes it difficult to select β_0, resulting in a large MSE. These findings are consistent with the theoretical results in Section 3. Case b) for $\|\delta_0\|_1 < \|\beta_0\|_1$ shows a similar phenomenon, and thus the results are deferred to the Supplementary Material.

In contrast, in Case c), where we vary the sparsity levels of δ_0 while keeping the magnitude the same ($\delta_0 = \beta_0$), the DPIE and SPIE methods demonstrate similar performances. Refer to the figure in the Supplementary Material for a visual representation. This finding also aligns with the theoretical result in Section 3, where we only need to restrict the magnitude of different parameters to guarantee consistency and oracle properties.

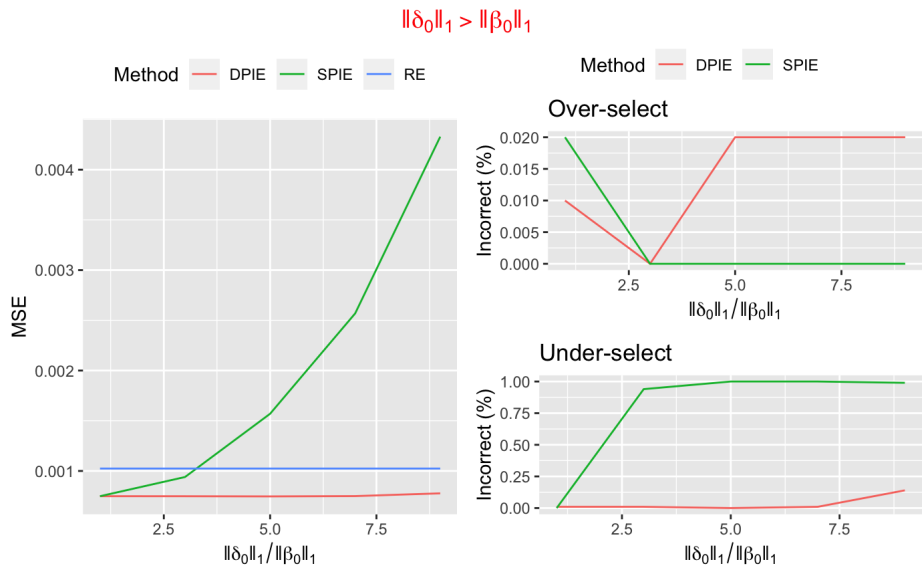

Figure 1: Simulation results based on 100 Monte Carlo times. The left panel shows the MSE versus the magnitude ratio between δ_0 and β_0. The right panel shows the percentage of wrongly choosing more and fewer parameters, separately.

4.2 SIMULATION STUDY 2

We now compare the proposed estimator of τ with existing competitors combining the REs and the ECs. We generate REs and ECs with sample sizes $n = m = 1000$. Covariates $X \in \mathbb{R}^2$ are generated by $X_d \sim \text{Uniform}\,[-1.5, 1.5]\,, d = 1, 2$. The treatments A in the REs are generated by $\text{Binomial}(1000, 0.5)$. We consider two settings for generating outcomes:

S1 $Y = -1.5X_1^2 - 1.5X_2 + 2A + (1-S)(10X_1^2 + 4X_2^3) + \epsilon$, where $\epsilon \sim \mathcal{N}(0, 1)$;

S2 $Y = -1.5X_1^2 - 1.5e^{X_2} + 2A + (1-S)(10X_1^2 + 4X_2^3) + \epsilon$, where $\epsilon \sim \mathcal{N}(0, 1)$.

In each case, we approximate the $\bar{\mu}_{1,1}(X;\beta)$, $\bar{\mu}_{0,1}(X;\beta)$ and $b_0(X)$ using the power series basis functions with the power up to three. We use the double SCAD penalty method to select important features of $\bar{\mu}_{1,1}(X;\beta)$, $\bar{\mu}_{0,1}(X;\beta)$ and $b_0(X)$, where, in Setting S1, the working models are correct, while in Setting S2, the working models are misspecified for $\bar{\mu}_{1,1}(X;\beta)$, $\bar{\mu}_{0,1}(X;\beta)$. After selecting parameters, we estimate the variance using the linear regression estimated variance of $\hat{\tau}$. We compare our method with the power prior Bayesian method [Lin et al., 2019] and the Matching procedure [Stuart and Rubin, 2008].

Table 1 shows the absolute bias of the estimated ATE $\hat{\tau}$, true variances v, estimated variances $\hat{v}$, MSE, and 95% Wald confidence intervals. In both settings, combining EC and RE data improves accuracy and efficiency. The Bayesian method uses the estimated probability of trial inclusion $\mathbb{P}(S = 1 \mid X)$ to adjust the EC, which borrows less information from ECs than correctly estimating the bias function,

Table 1: The absolute bias, estimated variance, the true variance, MSE and the 95% Wald confidence intervals of $\hat{\tau}$, compared with the Bayesian methods (denoted as $|\hat{\tau} - \tau|_B$ and v_B) and the Matching method (denoted as $|\hat{\tau} - \tau|_M$ and v_M) in two settings (denoted as S1, S2).

$\times 10^{-3}$	S1		S2	
	EC+RE	RE	EC+RE	RE
$\lvert\hat{\tau} - \tau\rvert$	2.43	4.05	4.42	4.67
$\lvert\hat{\tau} - \tau\rvert_B$	110.7		111.5	
$\lvert\hat{\tau} - \tau\rvert_M$	646		645	
v_B	326		328	
v_M	457		457	
v	3.33	3.94	3.33	3.97
$\hat{v}$	3.27	4.00	3.30	4.01
MSE	3.34	3.96	3.35	3.99
CI	94.6%	95.6%	95%	95.8%

resulting in worse results. The matching procedure in Stuart and Rubin [2008] measures the difference between the ECs and CCs in two stages: first they used ECs to match CCs, balancing covariates between the ECs and CCs in this process. They then determined the bias value, δ, between ECs and CCs using matched groups of CCs and ECs. This δ is constant for all X and may not be accurate. In contrast, our methods use a bias function, $b_0(X)$, which adapts to different X values and accounts for all differences, irrespective of whether they arise from covariates or the outcome. Consequently, our approach is more effective than the two-stage method proposed by Stuart and Rubin (2008). As shown in Table 1, Stuart and Rubin [2008] approach has a larger bias compared to our method and the Bayesian method due to

the less accurate bias term, δ.

5 REAL DATA ANALYSIS

We apply the proposed DPIE estimator as well as other methods in Section 4 to the data from the National Supported Work (NSW) study. This study aims at evaluating the effect of a job training program on future earnings, containing an experimental sample from a randomized evaluation of the NSW program, and a nonexperimental sample from the Current Population Survey (CPS) program. 15992 external control units are included in the original CPS dataset, whereas 260 random control units are included in the NSW dataset. We use the Matching procedure [Abadie and Imbens, 2006] to match each random control unit with 2 external control units without replacement, therefore, we use 520 external control units (ECs) and 260 random control units (CCs) as the control group, and 260 random treatment units as the treatment group.

This analysis includes the eight original covariates from the NSW and CPS datasets (age, education, Black, Hispanic, married, having no college degree (denoted as "nodeg"), real earnings in 1974 (denoted as "re74"), and real earnings in 1975 (denoted as "re75") as well as their 2-way interactions. The outcome of interest is the real earnings in 1978 (denoted as "re78"). For a better regression, we divide all the real earnings (re74, re75, re78) by 1000, scale all covariates between 0 and 1, and omit variables with the same value across observations. There are 86 covariates in total, with 43 covariates in the bias function and 43 covariates in the outcome mean function. Accordingly, we use the mean of real earnings in 1978 in the REs as the true value, 1.794.

Table 2 shows the estimated control mean $\hat{\tau}$ (reported as "Est") in the random group, i.e., $S = 1$, along with its standard error (reported as " se") and 95% Wald confidence intervals using the proposed DPIE estimator, the SPIE estimator and the SCAD estimator only based on the RE data. The number of variables selected in the outcome mean model (reported as "#var_$\bar{\mu}_{A,S}$")and in the bias function are also reported (reported as "#var_b_0"). Even the bias function is as complex as the outcome mean model, the DPIE improves efficiency by increasing the sample size, resulting in the standard error of DPIE being smaller than that of RE in this real data scenario. On the other hand, the SPIE estimator has a larger bias than the DPIE, because the magnitude of the outcome mean function is much smaller than that of the bias function, leading to a biased estimate compared with the DPIE, which is consistent with our simulation results shown in Figure 1. Based on the DPIE estimator, the estimated average treatment effect is 1.704.

Table 2: The first panel shows estimated $\hat{\tau}$ and corresponding standard error, bias, 95% Wald confidence interval and the number of selected variables in the outcome mean model. The second panel shows estimated variables in the outcome mean model $\mu_{0,1}(X)$ and the bias function $b_0(X)$ based on the DPIE estimator.

	Est	se	bias	#var_$\bar{\mu}_{A,S}$	#var_b_0
DPIE	1.857 (0.746 , 2.969)	0.567	0.063	4	4
SPIE	1.626 (0.582 , 2.671)	0.533	0.168	5	1
RE	1.698 (0.455, 2.941)	0.634	0.097	4	/

6 DISCUSSION

We introduce a bias function to measure the discrepancy between the ECs and the working model in REs and use sieve estimation and feature selection techniques to handle the high-dimensional nature of the basis functions and to prevent irrelevant covariates from being included in the outcome mean model. We propose a double penalty integration estimator (DPIE) that takes advantage of the different levels of smoothness of the outcome mean and bias functions. Our results demonstrate that the DPIE is consistent, has the oracle property, and is asymptotically normal when the penalty parameters are selected appropriately. Moreover, our estimator is robust to model misspecification and is at least as efficient as the REs alone.

We provide a general framework with a broad class of choices for combining multiple datasets and employing flexible penalized regression procedures. Combining several treatments for a more accurate estimation of the value functions in policy evaluation and individual treatment regimes is a direct extension of our method. In addition, our outcome Y can be extended to multiple types, including survival [Lee et al., 2022] and zero-inflation outcomes [Yu et al., 2021]. In lieu of better estimating the outcome mean function to enhance the ATE estimate, one may directly combine the bias function and the heterogeneous treatment effects (HTEs; Yang et al. [2023]), which are the causal effects of a treatment given the characteristics of the subjects, to obtain a more accurate estimate of the HTEs. Evaluating the HTEs is the primary question in many domains, including precision medicine and tailored policy recommendations [Colnet et al., 2020, Chu et al., 2023]. Finally, we exclusively consider the SCAD penalty in our theoretical study. The SCAD penalty addresses consistency, oracle property, and asymptotic normality of some local minimizer of the penalized loss. However, it doesn't ensure the uniqueness of the solution or provide methods for identifying the specific local minimizer with the desired properties among a large pool of potential local minimizes [Zhang, 2010]. This gap between theory and practice presents an interesting avenue for future research. To address this concern, we propose several potential approaches: Fan et al. [2014] introduced a general

procedure based on the LLA algorithm and derive a lower bound on the probability that a specific local solution exactly matches the oracle estimator, which could be applicable in real-world scenarios; Kim and Kwon [2012] provided conditions for determining the uniqueness of a local minimizer. Additionally, we recommend varying initial values in R's `ncvfit` function, and selecting the estimate that minimizes error. Alternatively, using unpenalized estimated covariates as initial values can be considered. Moreover, A general theoretical framework for multiple penalties, such as the adaptive Lasso [Zou, 2006] and minimax concave penalty [Zhang, 2010] of double penalty selection, would therefore be desirable.

References

Alberto Abadie and Guido W Imbens. Large sample properties of matching estimators for average treatment effects. *Econometrica*, 74(1):235–267, 2006.

Peter Bühlmann and Domagoj Ćevid. Deconfounding and causal regularisation for stability and external validity. *International Statistical Review*, 88:S114–S134, 2020.

Xiaohong Chen. Large sample sieve estimation of semi-nonparametric models. *Handbook of Econometrics*, 6: 5549–5632, 2007.

Victor Chernozhukov, Christian Hansen, and Yuan Liao. A lava attack on the recovery of sums of dense and sparse signals. *The Annals of Statistics*, 45(1):39–76, 2017.

Jianing Chu, Wenbin Lu, and Shu Yang. Targeted optimal treatment regime learning using summary statistics. *Biometrika*, 2023. doi: 10.1093/biomet/asad020.

Bénédicte Colnet, Imke Mayer, Guanhua Chen, Awa Dieng, Ruohong Li, Gaël Varoquaux, Jean-Philippe Vert, Julie Josse, and Shu Yang. Causal inference methods for combining randomized trials and observational studies: a review, 2020. URL `https://arxiv.org/abs/2011.08047`.

Jianqing Fan and Runze Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456): 1348–1360, 2001.

Jianqing Fan and Heng Peng. Nonconcave penalized likelihood with a diverging number of parameters. *The Annals of Statistics*, 32(3):928–961, 2004.

Jianqing Fan, Lingzhou Xue, and Hui Zou. Strong oracle optimality of folded concave penalized estimation. *The Annals of Statistics*, 42(3):819 – 849, 2014.

Alexandre Gilotte, Clément Calauzènes, Thomas Nedelec, Alexandre Abraham, and Simon Dollé. Offline a/b testing for recommender systems. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, pages 198–206, 2018.

Sander Greenland, Judea Pearl, and James M Robins. Causal diagrams for epidemiologic research. *Epidemiology*, 10 (1):37–48, 1999.

Miguel A Hernán and James M Robins. Using big data to emulate a target trial when a randomized trial is not available. *American Journal of Epidemiology*, 183(8): 758–764, 2016.

Brian P Hobbs, Daniel J Sargent, and Bradley P Carlin. Commensurate priors for incorporating historical information in clinical trials using general and generalized linear models. *Bayesian Analysis*, 7(3):639–674, 2012.

Brian P Hobbs, Bradley P Carlin, and Daniel J Sargent. Adaptive adjustment of the randomization ratio using historical control data. *Clinical Trials*, 10(3):430–440, 2013.

Joseph G Ibrahim and Ming-Hui Chen. Power prior distributions for regression models. *Statistical Science*, 15(1): 46–60, 2000.

Yongdai Kim and Sunghoon Kwon. Global optimality of nonconvex penalized estimators. *Biometrika*, 99(2):315–325, 2012.

Dasom Lee, Shu Yang, and Xiaofei Wang. Doubly robust estimators for generalizing treatment effects on survival outcomes from randomized controlled trials to a target population, 2022. URL `https://arxiv.org/abs/2201.06595`.

Junjing Lin, Margaret Gamalo-Siebers, and Ram Tiwari. Propensity-score-based priors for bayesian augmented control design. *Pharmaceutical Statistics*, 18(2):223–238, 2019.

Beat Neuenschwander, Michael Branson, and David J Spiegelhalter. A note on the power prior. *Statistics in Medicine*, 28(28):3562–3566, 2009.

Stuart J Pocock. The combination of randomized and historical controls in clinical trials. *Journal of Chronic Diseases*, 29(3):175–188, 1976.

Donald B Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688–701, 1974.

Donald B Rubin. Bayesian inference for causal effects: The role of randomization. *The Annals of Statistics*, 6(1): 34–58, 1978.

Donald B Rubin. The design versus the analysis of observational studies for causal effects: parallels with the design of randomized trials. *Statistics in Medicine*, 26(1):20–36, 2007.

Donald B Rubin and Neal Thomas. Matching using estimated propensity scores: relating theory to practice. *Biometrics*, 52(1):249–264, 1996.

Heinz Schmidli, Sandro Gsteiger, Satrajit Roychoudhury, Anthony O'Hagan, David Spiegelhalter, and Beat Neuenschwander. Robust meta-analytic-predictive priors in clinical trials with historical control information. *Biometrics*, 70(4):1023–1032, 2014.

David J Spiegelhalter. Incorporating bayesian ideas into health-care evaluation. *Statistical Science*, 19(1):156–174, 2004.

Jerzy Splawa-Neyman, D. M. Dabrowska, and T. P. Speed. On the application of probability theory to agricultural experiments. essay on principles. section 9. *Statistical Science*, 5(4):465–472, 1990.

Elizabeth A Stuart and Donald B Rubin. Matching with multiple control groups with adjustment for group differences. *Journal of Educational and Behavioral Statistics*, 33(3):279–306, 2008.

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

Kert Viele, Scott Berry, Beat Neuenschwander, Billy Amzal, Fang Chen, Nathan Enas, Brian Hobbs, Joseph G Ibrahim, Nelson Kinnersley, Stacy Lindborg, et al. Use of historical control data for assessing treatment effects in clinical trials. *Pharmaceutical Statistics*, 13(1):41–54, 2014.

Bingkai Wang, Ryoko Susukida, Ramin Mojtabai, Masoumeh Amin-Esmaeili, and Michael Rosenblum. Model-robust inference for clinical trials that improve precision by stratified randomization and covariate adjustment. *Journal of the American Statistical Association*, 118(542): 1152–1163, 2023.

Wenjia Wang and Yi-Hui Zhou. A double penalty model for interpretability, 2019. URL `https://arxiv.org/abs/1909.06263`.

Lili Wu and Shu Yang. Integrative r-learner of heterogeneous treatment effects combining experimental and observational studies. In *Proceedings of the First Conference on Causal Learning and Reasoning*, volume 177 of *Proceedings of Machine Learning Research*, pages 904–926, 2022.

Bing Xing, Ning Yang, and Xu Yaosheng. Adaptive estimation of multivariate regression with hidden variables. *The Annals of Statistics*, 50(2):640 – 672, 2021.

Shu Yang, Donglin Zeng, and Xiaofei Wang. Improved inference for heterogeneous treatment effects using real-world data subject to hidden confounding, 2022. URL `https://arxiv.org/abs/2007.12922`.

Shu Yang, Chenyin Gao, Donglin Zeng, and Xiaofei Wang. Elastic integrative analysis of randomized trial and real-world data for treatment heterogeneity estimation. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 2023. doi: 10.1093/jrsssb/qkad017.

Miao Yu, Wenbin Lu, Shu Yang, and Pulak Ghosh. Multiplicative structural nested mean model for zero-inflated outcomes. *Biometrika*, 103(1):1–17, 2021.

Jiacheng Yuan, Jeen Liu, Ray Zhu, Ying Lu, and Ulo Palm. Design of randomized controlled confirmatory trials using historical control data to augment sample size for concurrent controls. *Journal of Biopharmaceutical Statistics*, 29(3):558–573, 2019.

Cun-Hui Zhang. Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, 38(2): 894–942, 2010.

Hui Zou. The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476): 1418–1429, 2006.