On Causal Inference in Real World Settings

HARVARD UNIVERSITY
Graduate School of Arts and Sciences

DISSERTATION ACCEPTANCE CERTIFICATE

The undersigned, appointed by the

Department of Biostatistics

have examined a dissertation entitled

"On Causal Inference in Real World Settings"

presented by Larry Han

candidate for the degree of Doctor of Philosophy and hereby certify that it is worthy of acceptance.

Signature Tianxi Cai
Tianxi Cai (May 1, 2023 20:57 EDT)

Typed name: Prof. Tianxi Cai

Signature lorenzo trippa
lorenzo trippa (May 2, 2023 17:53 EDT)

Typed name: Prof. Lorenzo Trippa

Signature Sebastian Schneeweiss
Sebastian Schneeweiss (May 5, 2023 08:53 EDT)

Typed name: Prof. Sebastien Schneeweiss

Signature Rui Duan
Rui Duan (May 5, 2023 09:05 EDT)

Typed name: Prof. Rui Duan

Date: May 1, 2023

On Causal Inference in Real World Settings

A DISSERTATION PRESENTED
BY
LARRY HAN
TO
THE DEPARTMENT OF BIOSTATISTICS

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY
IN THE SUBJECT OF
BIOSTATISTICS

HARVARD UNIVERSITY
CAMBRIDGE, MASSACHUSETTS
MAY 2023

Dissertation advisors: Professors Tianxi Cai and Lorenzo Trippa Larry Han

On Causal Inference in Real World Settings

Abstract

In the present dissertation, we consider three classical and yet modern topics in causal inference – surrogate markers, multi-source federated learning, and sensitivity analysis. In each case, present-day obstacles in real world settings make estimation and inference of causal estimands a challenging endeavor.

In Chapter 1, we tackle the problem of how to identify and validate surrogate markers using real-world data (RWD). There is a need to develop statistical methods to evaluate the proportion of treatment effect (PTE) explained by surrogates in RWD, which have become increasingly common. To address this knowledge gap, we propose inverse probability weighted (IPW) and doubly robust (DR) estimators of an optimal transformation of the surrogate and the corresponding PTE measure. We demonstrate that the proposed estimators are consistent and asymptotically normal, and the DR estimator is consistent when either the propensity score model or outcome regression model is correctly specified. In two RWD settings, we show that our method can identify and validate surrogate markers for inflammatory bowel disease (IBD).

Chapter 2 is focused on federated learning of causal effects in multi-source settings. We develop a Federated Adaptive Causal Estimation (FACE) framework to incorporate heterogeneous data from multiple sites to provide treatment effect estimation and inference for a flexibly specified target population of interest. To safely incorporate source sites and avoid negative transfer, we introduce an adaptive weighting procedure via a penalized regression, which achieves both consistency and optimal efficiency. Our strategy is communication-efficient and privacy-preserving, allowing participating sites to

only share summary statistics once with other sites. We conduct both theoretical and numerical evaluations of FACE, and apply it to conduct a comparative effectiveness study of BNT162b2 (Pfizer) and mRNA-1273 (Moderna) vaccines on COVID-19 outcomes in U.S. veterans using electronic health records from five VA regional sites.

In Chapter 3, we develop a novel framework to conduct sensitivity analysis at the design stage of complex clinical trials. Sensitivity analyses are useful to assess the dependence of important design operating characteristics with respect to various unknown parameters. Two crucial components of sensitivity analyses are (i) the choice of a set of plausible simulation scenarios and (ii) the list of operating characteristics of interest. We propose a robust approach to choose the set of scenarios for inclusion in design sensitivity analyses. We maximize a utility criterion that formalizes whether a specific set of sensitivity scenarios is adequate to summarize how the operating characteristics of the trial design vary across all plausible values of the unknown parameters. Then, we use optimization techniques to select the best set of simulation scenarios (according to the criteria specified by the investigator) to exemplify the operating characteristics of the trial design. We illustrate our proposal in three trial designs of increasing complexity.

Contents

List of Tables

List of Figures

The following authors contributed to chapters in this dissertation:

- Chapter 1: Larry Han, Xuan Wang, and Tianxi Cai
- Chapter 2: Larry Han, Jue Hou, Kelly Cho, Rui Duan, and Tianxi Cai
- Chapter 3: Larry Han, Andrea Arfe, and Lorenzo Trippa

This dissertation is dedicated to Yeye, Baba, Mama, and Jiejie. History humbles you. Without their love and support, none of this work would have been possible.

Acknowledgments

First and foremost, I thank God for His sovereignty in my life. Apart from Him, I can do nothing. I thank Him for guiding me in each step of my life, including bringing me to Boston, where I could experience His saving grace.

To my family, I love you. Without you, this dissertation would not have been possible. More importantly, you have helped shape me into the person I am today, through love and discipline. I thank my father, Bajin Han, for instilling in me the importance of passion and persistence. I thank my mother, Xiaomin Li, for showing me that it is possible to manifest gentle kindness while having an unparalleled work ethic. I thank my sister, Bo Han, for being a fantastic older sibling to me.

To Professor Tianxi Cai, I offer my deepest gratitude for your unwavering support, generosity, and patience throughout the course of my doctoral journey. I can think of no better mentor to have trained under; truly, your unwavering passion and insightful guidance have made all the difference in my development as a researcher and future mentor to others. You have taught me not only how to answer questions, but how to ask the right questions that can change the world for the better.

To Professor Lorenzo Trippa, I thank you for your dedication to helping me develop into a more precise thinker and writer.

To Professor Rui Duan, I am thankful for all of your advice and feedback, and thank you for setting an excellent example for all aspiring junior researchers. I had so much fun talking about grant ideas and research projects and look forward to many collaborations to come.

To Professor Sebastian Schneeweiss, I am appreciative of all your helpful suggestions on my research direction and projects. You helped me see the 'bigger picture', taught me about the regulatory landscape, and made sure that the problems I solved would have real world impact.

To Professor Jose Zubizarreta, I offer my sincere thanks for your mentorship and guidance, and especially for your words of encouragement in good times and challenging ones alike.

To my cohort-mates, especially Christina Howe, Gopal Kotecha, Patrick Emedom-Nnamdi, Jemar Bather, and Luli Zou, I thank you for your friendship. We did it!

To my friends in the Department of Biostatistics, Xihao Li, Xiao Wu, Yige Li, Alex Levis, Stephanie Wu, Jenny Lee, Wenying Deng, Isabella Nogues, Lara Maleyeff, Andy Shi, Linda Harrison, Alex Ocampo, Helian Feng, Jaffer Zaidi, Maya Mathur, Kaitlyn Cook, Jonno Larson, Dustin Rabideau, Harrison Reeder, Lucy Zhu Shen, Yi Zhang, Bijan Niknam, and Tian Gu, I thank you for all of the fun conversations, delicious meals, and lasting memories we made as we navigated the good times and difficult times together.

To my fellow Cai labmates, Jue Hou, Xuan Wang, Weijing Tang, Chuan Hong, Isabella Nogues, Molei Liu, Yuri Ahuja, Aaron Sonabend, Xiudi Li, David Cheng, and Jessica Gronsbell, I thank you for everything you have taught me – I cannot have imagined a better group of amazing researchers to learn from.

To my amazing teachers, Andrea Rotnitzky, Jamie Robins, Miguel Hernan, Jose Zubizarreta, Kosuke Imai, Rajarshi Mukherjee, Rui Wang, Marcello Pagano, Brent Coull, Bob Gray, Cyrus Mehta, J.P. Onnela, Sebastien Haneuse, L.J. Wei, David Wypij, Lorenzo Trippa, Curtis Huttenhower, Christoph Lange, and Heather Mattie, I thank you for your dedication to educating the next generation. I have learned so much from you!

To the graduate studies team, Paige Williams, Sebastien Haneuse, Rajarshi Mukherjee, and Erin Lake, I thank you for shepherding us through the Ph.D. program and constantly seeking to improve the department.

To Jelena Follweiler, Trevor Bierig, Shaina Andelman, and Elizabeth Solinga, I don't know how you do it, but you manage to keep the department functioning. I thank you for your exceptional professionalism and kindness.

To my church, Antioch Baptist Church, I am grateful beyond words and indebted to the sacrifices made by those who came before me, including Pastor Paul Kim and Chaplain Rebekah Kim JDSN, Pastor David Um and Angela Um SMN, Pastor Dan Cho, Pastor Sang Peter Chin, Pastor Thomas Chen, David Dominguez H, Tony Qian H, Ed Kao H, and my fellow "Young Adult Concord Cave of Adullam" brothers, Nathan Pak H, Brendan Liu H, Theo Huang H, Paul Kim H, Brian Louis, Sean Hwang, and Joe Kim.

0

Introduction

In the United States, the passage of the 21st Century Cures Act in 2016 mandated the use of data generated from the routine operation of healthcare to inform decisions made by regulators, providers, and patients [13]. Evidence generated from such real world data (RWD), termed real world evidence (RWE), has gained tremendous attention from the biomedical community [108,61]. To date, RWE has been used to support decision-making related to the primary approval of new medications through the formation of external controls; support supplemental indications and pediatric approvals; support

adaptive or accelerated approvals; and assess drug safety[9,95,107]. As evidence of the US Food and Drug Administration's (FDA) commitment to using RWE, the Sentinel Initiative was launched in 2008 as a national system of insurance claims databases used to assess the safety and performance of medical products. Despite the growing use of postmarket RWD to evaluate medical product safety, the use of RWD from electronic health records (EHRs), pragmatic randomized controlled trials (RCTs), and other healthcare database analyses to support effectiveness and approval decisions is still nascent. Indeed, while the FDA, European Medicines Agency, and other regulatory agencies have a strong familiarity and comfort in designing, implementing, and translating results from RCTs into policies and guidelines, this trust is understandably attenuated with RWD studies[38,40,105]. The present dissertation aims to narrow this trust gap by providing a suite of novel statistical methods and easily implementable tools, to accurately, robustly, and efficiently assess treatment effects using RWD, and more transparently communicate strategies for using RWD stemming from EHRs and RCTs. The methods proposed in this research are broadly applicable in different disease domains and expand the current capabilities and use scenarios of RWD.

In Chapter 1, we explore the question of how to identify and validate valid surrogate markers using real-world data (RWD). In comparative effectiveness research (CER), leveraging short-term surrogates to infer treatment effects on long-term outcomes can guide policymakers evaluating new treatments. Numerous statistical procedures for identifying surrogates have been proposed for RCTs, but no methods currently exist to evaluate the proportion of treatment effect (PTE) explained by surrogates in RWD, which have become increasingly common. To address this knowledge gap, we propose inverse probability weighted (IPW) and doubly robust (DR) estimators of an optimal transformation of the surrogate and the corresponding PTE measure. We demonstrate that the proposed estimators are consistent and asymptotically normal, and the DR estimator is consistent when either the propensity score model or outcome regression model is correctly specified. Our proposed estimators are evaluated through extensive simulation studies. In two RWD settings, we show that our method

can identify and validate surrogate markers for inflammatory bowel disease (IBD).

Chapter 2 is dedicated to federated learning of causal effects. Federated learning of causal estimands may greatly improve estimation efficiency by leveraging data from multiple study sites, but robustness to heterogeneity and model mis-specifications is vital for ensuring validity. We develop a Federated Adaptive Causal Estimation (FACE) framework to incorporate heterogeneous data from multiple sites to provide treatment effect estimation and inference for a flexibly specified target population of interest. FACE accounts for site-level heterogeneity in the distribution of covariates through density ratio weighting. To safely incorporate source sites and avoid negative transfer, we introduce an adaptive weighting procedure via a penalized regression, which achieves both consistency and optimal efficiency. Our strategy is communication-efficient and privacy-preserving, allowing participating sites to only share summary statistics once with other sites. We conduct both theoretical and numerical evaluations of FACE, and apply it to conduct a comparative effectiveness study of BNT162b2 (Pfizer) and mRNA-1273 (Moderna) vaccines on COVID-19 outcomes in U.S. veterans using EHRs from five VA regional sites. We show that compared to traditional methods, FACE meaningfully increases the precision of treatment effect estimates, with reductions in standard errors ranging from 26% to 67%.

In Chapter 3, we focus on how to conduct rigorous sensitivity analysis at the design stage of complex clinical trials. The use of simulation-based sensitivity analyses is fundamental to evaluating and comparing candidate designs for future clinical trials. In this context, sensitivity analyses are especially useful to assess the dependence of important design operating characteristics with respect to various unknown parameters. Typical examples of operating characteristics include the likelihood of detecting treatment effects and the average study duration, which depend on parameters that are unknown until after the onset of the clinical study, such as the distributions of the primary outcomes and patient profiles. Two crucial components of sensitivity analyses are (i) the choice of a set of plausible simulation scenarios and (ii) the list of operating characteristics of interest. We propose a new

approach to choosing the set of scenarios for inclusion in design sensitivity analyses. We maximize a utility criterion that formalizes whether a specific set of sensitivity scenarios is adequate to summarize how the operating characteristics of the trial design vary across all plausible values of the unknown parameters. Then, we use optimization techniques to select the best set of simulation scenarios to exemplify the operating characteristics of the trial design. We illustrate our proposal in three trial designs of increasing complexity.

1

Identifying Surrogate Markers in Comparative Effectiveness Research

1.1 Introduction

We are motivated by the need to develop statistical methods for evaluating surrogate markers in comparative effectiveness research (CER). A surrogate marker is an outcome measure that can be used

as a substitute for a primary outcome, such that changes caused by a therapy on a surrogate marker are expected to reflect changes in the primary outcome [36]. The use of valid surrogate markers to infer treatment effects on long term outcomes has the potential to reduce cost and expedite the approval of new therapies [22,128,16,24]. We aim to use real world data (RWD), such as electronic health records (EHRs), clinical registries, and cross-trial data, to identify and validate surrogates when comparing treatments.

In particular, we assess biologic therapies for inflammatory bowel disease (IBD) [103], where potential surrogate markers such as the likelihood of nonresponse score or the non-invasive partial Mayo score can be measured earlier and more cheaply than the primary outcome of clinical response [2,70]. While multiple treatments for IBD have shown efficacy in placebo-controlled trials [115], there is interest in identifying and validating surrogate markers of the treatment effect between biologic therapies. However, there have been few head-to-head comparisons of biologic therapies, and surrogate markers have not been validated for studying the treatment effect on different outcomes of interest. One of the first such head-to-head trials, a phase 3b trial of vedolizumab vs. adalimumab, showed that vedolizumab was superior with respect to achieving clinical remission [104]. However, randomized clinical trials (RCTs) of this type are typically prohibitively costly due to the large sample size that is needed to detect presumably small treatment effects [11]. The lack of direct comparisons for different therapies has resulted in the choice of treatment being influenced by factors often unrelated to treatment performance [2]. More generally, novel treatments frequently show great promise in placebo-controlled trials but head-to-head comparisons in RCTs are often missing. Pharmaceutical companies have historically avoided head-to-head trials for fear of losing market share from unfavorable results and have instead favored placebo-controlled trials for a greater chance of declaring superiority [34].

In addition to requiring long-term follow-up of patients to observe a sufficient number of events to detect treatment effects, RCTs are also often limited to narrowly defined patient populations with results that are not always generalizable to broader populations of interest. These shortcomings are es-

pecially pronounced in urgent health crises when potential treatments must be assessed rapidly. Coupled with the explosive growth in observational healthcare data[80,48,25], there is growing interest in using RWD generated from observational studies to evaluate surrogate markers[43].

In settings where it is costly or infeasible to observe outcomes directly, machine-learning model predictions can be used as surrogates[89]. For example, International Classification of Diseases (ICD) codes in EHR data have been used as surrogate outcomes to predict postoperative hospital mortality[129]. However, directly using these surrogate outcomes may be problematic due to bias or differences in variability compared to the true outcomes, and subsequent analyses based on these surrogate outcomes can result in poor post-prediction inference[122]. It is important to determine the strength of the surrogate outcome for the true outcome, both to inform decisions about whether to use these surrogates in future studies and because many statistical methods require the surrogate to be strong[88,91].

Since Prentice (1989)[90] originally proposed a definition and operational criteria for identifying valid surrogate markers, many statistical methods have been developed to make inference about the proportion of treatment effect (PTE) explained by a surrogate in RCT settings[39,73,126,88,91,124]. Freedman (1992)[39] proposed a parametric model-based estimate assuming two outcome regression models (including and excluding the surrogate), which rarely hold simultaneously[73]. Wang (2002)[126] proposed alternative measures of PTE that examined what the treatment effect would have been if the surrogate had the same distribution across treatment groups. Parast (2016)[88] proposed a fully nonparametric estimation procedure for the PTE defined in Wang (2002)[126]. More recently, Wang (2020)[124] proposed an alternative non-parametric PTE estimator by identifying a single optimal transformation of the surrogate to predict the outcome with weaker assumptions than those required by Parast (2016)[88]. However, all of these existing PTE estimators are derived for data from RCTs and are therefore not directly applicable to RWD where treatment assignment A may depend on confounding factors $\mathbf{X}$. Price (2018)[91] proposed finding optimal functions of a surrogate separately for each treatment group under the constraint that the Prentice criterion is satisfied[90]. However, their opti-

mal surrogate is guaranteed to attain perfect surrogacy even when the original surrogate is not a valid surrogate, and their focus is not on evaluating surrogacy [1,2,3].

In this paper, we aim to address the gap in identifying and validating surrogate markers using RWD by proposing novel inverse probability weighted (IPW) and doubly robust (DR) estimators for an optimal transformation function and corresponding PTE estimators. We propose flexible semi-nonparametric models for the relationship between outcome Y and surrogate S given baseline covariates $\mathbf{X}$ and a propensity score (PS) model for A given $\mathbf{X}$. We also propose perturbation resampling methods for variance and confidence interval estimation. We establish the asymptotic properties of the proposed estimators, including double robustness of the proposed estimator in that it is consistent when either the PS model or the outcome regression (OR) model is correctly specified. Our simulation studies demonstrate that the proposed estimators and inference procedures perform well in finite samples. We illustrate the utility of our method to identify and validate surrogate markers for IBD in two different types of RWD analyses.

1.2 Methods

1.2.1 Setting, Notation, and Identification

Let Y be the primary outcome and S be the surrogate marker, both of which may be discrete or continuous. Throughout, the notation takes S to be continuous, but all derivations and theoretical results remain valid if S is discrete by replacing density functions with probability mass functions. We denote $\{Y^{(a)}, S^{(a)}\}$ as the respective potential primary outcome and surrogate marker under treatment $A = a$, where $A = 1$ and $A = 0$ denote the treatment and the control group, respectively. With RWD, only $Y_i = A_i Y_i^{(1)} + (1 - A_i) Y_i^{(0)}$ and $S_i = A_i S_i^{(1)} + (1 - A_i) S_i^{(0)}$ can be observed for an individual i, and the treatment assignment A_i may depend on baseline confounding factors $\mathbf{X}_i$. For

identifiability, we require the standard causal assumptions [101,60]

$$\pi_a(\mathbf{x}) \equiv P(A = a \mid \mathbf{X} = \mathbf{x}) \in (0, 1) \tag{1.1}$$

$$\left(Y^{(1)}, Y^{(0)}, S^{(1)}, S^{(0)}\right) \perp A \mid \mathbf{X} \tag{1.2}$$

Assumption (1.1) states that within all covariate levels, patients may receive either treatment so that the PS is bounded away from 0 and 1. Assumption (1.2) implies that $\mathbf{X}$ includes all confounders that can affect the primary outcome and treatment simultaneously, or the surrogate and treatment simultaneously. In other words, we assume that observed covariates $\mathbf{X}$ contain all confounders of the effects of treatment A on surrogate S and primary outcome Y, such that treatment A is as good as randomized within levels of covariates $\mathbf{X}$. This "no unmeasured confounding" assumption has been made previously, for example, by Agniel (2020)[1]. We assume that the RWD for analysis consist of n independent and identically distributed random variables $\{\mathbf{D_i} = (Y_i, S_i, A_i, \mathbf{X}_i^{\intercal})^{\intercal}, i = 1, ..., n\}$.

1.2.2 Target Parameter and Leveraging Surrogates

The average treatment effect (ATE) on Y is defined as

$$\Delta = \mu_1 - \mu_0, \quad \text{where } \mu_a = E(Y^{(a)}) = \int E(Y \mid A = a, \mathbf{X}) d\mathbb{F}(\mathbf{X}),$$

and $\mathbb{F}(\mathbf{x}) = P(\mathbf{X} \leq \mathbf{x})$. Note that to obtain the mean potential outcome μ_a, we must integrate the conditional outcome model over the density of the covariates $\mathbf{X}$. Without loss of generality, we assume that the ATE is non-negative, i.e., $\Delta \geq 0$. To approximate Δ based on the treatment effect on S, we define a transformation function $g_{\text{opt}}(\cdot)$ such that the treatment effect on the transformed surrogate, $\Delta_{g_{\text{opt}}} = E[g_{opt}(S^{(1)}) - g_{opt}(S^{(0)})]$, can be used to predict the treatment effect on the primary outcome,

$\Delta = E[Y^{(1)} - Y^{(0)}]$. Rather than using S directly, the optimal transformation of the surrogate $g_{\text{opt}}(S)$ aims to recover the difference in the primary outcomes between patients in the two treatment groups. More formally, the optimality of g_{opt} is with respect to minimizing the mean squared error

$$\mathcal{L}(g_{opt}) = E\left[\left(Y^{(1)} - Y^{(0)}\right) - \left\{g_{\text{opt}}\left(S^{(1)}\right) - g_{\text{opt}}\left(S^{(0)}\right)\right\}\right]^2. \tag{1.3}$$

Alternative loss functions can and have been considered. For example, Price (2018)[91] proposed an alternative approach that identifies treatment-specific transformations of the surrogate. Following Wang (2020)[124], in Appendix 1, we show that under a working independence assumption that

$$(Y^{(1)}, S^{(1)}) \perp (Y^{(0)}, S^{(0)}), \tag{1.4}$$

g_{opt} takes the form

$$g_{opt}(s) = m(s) + \lambda \mathcal{P}_0(s),$$

where

$$m(s) = m_1(s)\mathcal{P}_1(s) + m_0(s)\mathcal{P}_0(s), \quad m_a(s) = E(Y^{(a)} \mid S^{(a)} = s),$$

$$\mathcal{P}_a(s) = f_a(s)(f_0(s) + f_1(s))^{-1}, \quad f_a(s) = dF_a(s)/ds,$$

$$\lambda = \frac{\int \{m_0(s) - m_1(s)\}\mathcal{P}_1(s)dF_0(s)}{\int \mathcal{P}_0(s)dF_0(s)} = \frac{\mu_0 - \int m(s)dF_0(s)}{\int \mathcal{P}_0(s)dF_0(s)},$$

and $F_a(s) = P(S^{(a)} \leq s)$.

Note that the optimal transformation g_{opt} is invariant to treatment A and covariates $\mathbf{X}$, whose role is made clear in the estimation of the conditional mean functions $m_a(s)$ and densities $f_a(s)$ in Section 3. The optimal transformation g_{opt} can be interpreted as the conditional mean function, $m(s)$ that is shifted by a scaled posterior probability function of $A = 0 \mid S$, where λ determines the degree of the shift. For example, when there is no treatment effect on the conditional expectation of $Y \mid S$

such that $m_0(s) = m_1(s)$, then $\lambda = 0$ and g_{opt} reduces to $m(s)$. Practically, the shape of g_{opt} can provide useful information about the effect of S on Y. In particular, existing methods[88,126] require that the relationship between Y and S be monotone, which can often be violated. Our proposed estimation strategy evaluates the PTE explained for $g_{\text{opt}}(S)$ rather than S, highlighting the robustness of our method.

Remark 1. *Since only one of* $(Y^{(1)}, S^{(1)})$ *and* $(Y^{(0)}, S^{(0)})$ *can be observed for an individual and the correlation structure is not identifiable, we minimize the MSE-type loss (1.3) under the working independence assumption (1.4). A similar assumption was considered by Robins and Richardson (2010)[97], who showed that it would hold under a minimal sufficient causal model in the sufficient causal framework. Although restrictive and unlikely to hold in practice, it is used to help derive* g_{opt} *and is not needed to interpret the proposed PTE measure or for valid inference. In our simulation studies, we show that* g_{opt} *and PTE estimates are robust against violations of (1.4).*

By employing the transformation g_{opt} such that the treatment effect on Y is optimally approximated by the treatment effect on $g_{\text{opt}}(S)$, we can naturally quantify the PTE as the ratio of treatment effect on $g_{\text{opt}}(S)$ vs Y, i.e.,

$$\text{PTE}_{g_{\text{opt}}} \equiv \Delta_{g_{\text{opt}}}/\Delta.$$

Given the form of the PTE, as in Wang (2020)[124], it holds that $\text{PTE}_{g_{\text{opt}}} \in [0, 1]$ provided that

(A1) $\mathbb{S}_1(u) \geq \mathbb{S}_0(u)$ for all u,

(A2) $\mathbb{M}_1(u) \geq \mathbb{M}_0(u)$ for all u in the common support of $g_{opt}(S^{(1)})$ and $g_{opt}(S^{(0)})$,

where $\mathbb{S}_a(u) = P\{g_{opt}(S^{(a)}) \geq u\}$ and $\mathbb{M}_a(u) = E(Y^{(a)} \mid g_{opt}(S^{(a)}) = u)$, for $a = 0, 1$ (Appendix 2). Assumptions (A1) and (A2) are weaker than those required in previous literature, which has required monotonicity, $m_1(s) > m_0(s)$ for all s and the same surrogate support for $S^{(1)}$ and $S^{(0)}$[88] to

ensure that the PTE is between 0 and 1.

Remark 2. *In Appendix 2, we show that* $\Delta \geq \Delta_{g_{\text{opt}}} \geq 0$*, so that* $\Delta = 0$ *implies that* $\Delta_{g_{\text{opt}}} = 0$ *as well. Thus, our proposed PTE measure avoids the surrogate paradox, since it is never the case that the treatment effect on the surrogate is positive* $(\Delta_{g_{\text{opt}}} > 0)$ *but the treatment effect on the primary outcome is negative* $(\Delta < 0)$*, regardless of the correlation between the surrogate and primary outcome*[120]*. It is worth noting that Price (2018)*[91] *avoid the surrogate paradox without making these assumptions by defining an optimal transformation of the surrogate under the Prentice definition constraint*[90]*. However, Agniel (2020)*[1] *recently demonstrated that the resolution of Price (2018)*[91] *can be too strong, i.e., the surrogate paradox is resolved for all surrogates, even those that are completely unrelated to the outcome, and the power to detect treatment effects can actually increase as the surrogate weakens.*

1.3 Estimation Procedures

Given the above framework, our goal in this section is to construct inverse probability weighted (IPW) and doubly robust (DR) estimators for g_{opt} and $\text{PTE}_{g_{\text{opt}}}$ using RWD. Estimation of $g_{opt}(s)$ and $\text{PTE}_{g_{\text{opt}}}$ using RWD is more challenging than using RCT data because we cannot directly estimate $m(s)$, λ, and $\mathcal{P}_a(s)$ due to confounding. We propose an IPW estimator and a DR estimator for g_{opt} and $\text{PTE}_{g_{\text{opt}}}$ accounting for the effects of $\mathbf{X}$ on A, Y and S. We first present the simpler IPW estimator and then the DR estimator. For both estimators, we fit a parametric model for the PS model $\pi_1(\mathbf{X})$, denoted by $\pi_1(\mathbf{X};\alpha)$, where α is a finite dimensional parameter that can be estimated as the standard maximum likelihood estimator, $\widehat{\alpha}$. A simple example is a logistic regression model $\pi_1(\mathbf{X};\alpha) = G\{\alpha^{\intercal}\Phi(\mathbf{X})\}$, where $G(x) = e^x/(1+e^x)$ and $\Phi(\mathbf{X})$ is a vector of basis functions of $\mathbf{X}$ to account for potential non-linear effects.

1.3.1 IPW Estimation

To construct an IPW estimator for g_{opt}, we first obtain IPW kernel smoothed estimators for $m_a(s)$ and $f_a(s)$ respectively as

$$\widehat{m}_a(s) = \frac{\sum_{i=1}^n K_h(S_i - s) Y_i \widehat{\omega}_{ai}}{\sum_{i=1}^n K_h(S_i - s)\widehat{\omega}_{ai}} \quad \text{and} \quad \widehat{f}_a(s) = \frac{\sum_{i=1}^n K_h(S_i - s)\widehat{\omega}_{ai}}{\sum_{i=1}^n \widehat{\omega}_{ai}},$$

where $\widehat{\omega}_{ai} = I(A_i = a)/\pi_a(\mathbf{X}_i, \widehat{\alpha})$, $K_h(\cdot) = h^{-1}K(\cdot/h)$, $K(\cdot)$ is a symmetric density function and bandwidth $h = O(n^{-\nu})$ with $\nu \in (1/4, 1/2)$. Throughout, when S is discrete, kernel estimators $K_h(S_i - s)$ can be replaced with the indicator $I(S_i = s)$. Then $m(\cdot)$, $\mathcal{P}_a(\cdot)$ and λ may be estimated as

$$\widehat{m}(s) = \sum_{a=0}^{1} \widehat{m}_a(s)\widehat{\mathcal{P}}_a(s), \quad \widehat{\mathcal{P}}_a(s) = \frac{\widehat{f}_a(s)}{\widehat{f}_1(s) + \widehat{f}_0(s)}, \quad \hat{\lambda} = \frac{\int (\widehat{m}_0(s) - \widehat{m}_1(s))\widehat{\mathcal{P}}_1(s)\widehat{f}_0(s)ds}{\int \widehat{\mathcal{P}}_0(s)\widehat{f}_0(s)ds},$$

respectively. Subsequently, we construct plug-in estimators for $g_{\text{opt}}(s)$, $\Delta_{g_{\text{opt}}}$, Δ and $\text{PTE}_{g_{\text{opt}}}$ as

$$\hat{g}_{\text{opt}}(s) = \widehat{m}(s) + \hat{\lambda}\widehat{\mathcal{P}}_0(s),$$

$$\widehat{\Delta}_{\hat{g}_{\text{opt}}} = \widehat{\mu}_{1,\hat{g}_{\text{opt}}} - \widehat{\mu}_{0,\hat{g}_{\text{opt}}},$$

$$\widehat{\Delta} = \widehat{\mu}_1 - \widehat{\mu}_0,$$

and

$$\widehat{\text{PTE}}_{\hat{g}_{\text{opt}}} = \frac{\widehat{\Delta}_{\hat{g}_{\text{opt}}}}{\widehat{\Delta}},$$

where $\widehat{\mu}_{a,g_{\text{opt}}} = \frac{\sum_{i=1}^n g_{\text{opt}}(S_i)\widehat{\omega}_{ai}}{\sum_{i=1}^n \widehat{\omega}_{ai}}$ and $\widehat{\mu}_a = \frac{\sum_{i=1}^n Y_i\widehat{\omega}_{ai}}{\sum_{i=1}^n \widehat{\omega}_{ai}}$.

We show in Appendix 3 of the Supplementary Materials that when $\pi_1(\mathbf{x}; \alpha)$ is correctly specified, $\widehat{\text{PTE}}_{\hat{g}_{\text{opt}}}$ is consistent for $\text{PTE}_{g_{\text{opt}}}$. We also show that $\sqrt{n}(\widehat{\text{PTE}}_{\hat{g}_{\text{opt}}} - \text{PTE}_{g_{\text{opt}}})$ converges in distribution to a normal distribution with mean 0 and variance σ^2, where the form of σ^2 is derived in Appendix 3.

1.3.2 Doubly Robust Estimation

When the PS model is misspecified, the IPW estimator is likely to be biased. To achieve improved robustness and efficiency gains, we propose DR estimators for g_{opt} and $\text{PTE}_{g_{\text{opt}}}$. Following Robins (1994)[98], for any counterfactual random variable $U^{(a)}$, a DR estimator for its mean $E(U^{(a)})$ can be constructed as $n^{-1}\sum_{i=1}^{n}\left\{\widehat{\omega}_{ai}U_i-(\widehat{\omega}_{ai}-1)\widehat{\varphi}_a(\mathbf{X}_i)\right\}$, where $\widehat{\varphi}_a(\mathbf{X}_i)$ is an estimator for $E(U_i^{(a)}\mid\mathbf{X}_i)$ derived under a specified model. This estimator is DR in the sense that it is consistent for $E(U^{(a)})$ when either the PS model for $\pi_a(\mathbf{X})$ or the outcome regression (OR) model for $E(U_i^{(a)}\mid\mathbf{X}_i)$ is correctly specified. Deriving a DR estimator for $\text{PTE}_{g_{\text{opt}}}$ is more challenging since $g_{\text{opt}}(S)$ involves conditional mean functions of $Y^{(a)}\mid S^{(a)}$ and density functions of $S^{(a)}$ for $a=0,1$.

To construct a DR estimator for $g_{\text{opt}}(s)=m(s)+\lambda\mathcal{P}_0(s)$, we propose the following DR estimators for $m_a(s)$ and $f_a(s)$ respectively,

$$\widehat{m}_{a,\text{DR}}(s)=\frac{\widehat{\mathcal{M}}_{a,\text{DR}}(s)}{\widehat{f}_{a,\text{DR}}(s)}, \tag{1.5a}$$

$$\widehat{\mathcal{M}}_{a,\text{DR}}(s)=n^{-1}\sum_{i=1}^{n}\left\{K_h(S_i-s)Y_i\widehat{\omega}_{ai}-(\widehat{\omega}_{ai}-1)\widehat{\psi}^{\dagger}_{a,\text{m}}(s;\mathbf{X}_i)\widehat{\psi}^{\dagger}_{a,\text{f}}(s;\mathbf{X}_i)\right\}, \tag{1.5b}$$

$$\widehat{f}_{a,\text{DR}}(s)=n^{-1}\sum_{i=1}^{n}\left\{K_h(S_i-s)\widehat{\omega}_{ai}-(\widehat{\omega}_{ai}-1)\widehat{\psi}_{a,\text{f}}(s;\mathbf{X}_i)\right\}, \tag{1.5c}$$

where $\widehat{\psi}_{a,\text{m}}(\mathbf{x})$ and $\widehat{\psi}_{a,\text{f}}(s;\mathbf{x})$ are the respective estimators for

$$\psi_{a,\text{m}}(s;\mathbf{x})=E(Y_i^{(a)}\mid S_i^{(a)}=s,\mathbf{X}_i=\mathbf{x})=E(Y_i\mid A_i=a,S_i=s,\mathbf{X}_i=\mathbf{x}) \text{ and}$$
$$\psi_{a,\text{f}}(s;\mathbf{x})=\frac{\partial P(S_i^{(a)}\le s\mid\mathbf{X}_i=\mathbf{x})}{\partial s}.$$

In Appendix 4 of the Supplementary Materials, we show that $\widehat{m}_{a,\text{DR}}(s)$ and $\widehat{f}_{a,\text{DR}}(s)$ are consistent for $m_a(s)$ and $f_a(s)$ if either $\sup_{\mathbf{x}}|\widehat{\pi}_a(\mathbf{x})-\pi_a(\mathbf{x})|\to 0$ in probability or $\sup_{\mathbf{x},s}\{|\widehat{\psi}_{a,\text{m}}(s;\mathbf{x})-\psi_{a,\text{m}}(s;\mathbf{x})|+$

$|\widehat{\psi}_{a,\mathrm{f}}(s;\mathbf{x}) - \psi_{a,\mathrm{f}}(s;\mathbf{x})|\} \to 0$ in probability. In other words, double robustness of the optimal transformation g_{opt} can be achieved by either correctly specifying the PS model or correctly specifying the conditional outcome model and the conditional surrogate model.

Since the true form of the models $S \mid A, \mathbf{X}$ and $Y \mid A, S, \mathbf{X}$ are not known in RWD settings, simple parametric models are likely to produce biased estimates, with the degree of bias depending on the extent to which the models are misspecified. Alternatively, nonparametric estimators can be used. However, due to the curse of dimensionality, the convergence rates may be slow with less smoothness and more covariates. Therefore, to balance model flexibility and model interpretability, and to minimize assumptions on the dependency structure between S and Y, we construct flexible estimators $\widehat{\psi}_{a,\mathrm{m}}(s;\mathbf{x})$ and $\widehat{\psi}_{a,\mathrm{f}}(s;\mathbf{x})$ through semi-non-parametric models for $Y^{(a)} \mid S^{(a)}, \mathbf{X}$ and $S^{(a)} \mid \mathbf{X}$. We are able to handle multiple confounders by implementing a two-step estimator that first reduces potentially high-dimensional $\mathbf{X}$ into $\mathbf{X}^{\mathsf{T}}\hat{\gamma}_a$, where $\hat{\gamma}_a$ are the estimated covariate effects, through a generalized regression model (GRM) and then estimates the conditional density of $S \mid \mathbf{X}^{\mathsf{T}}\hat{\gamma}_a$ using the method of Hall (2004)[44]. Specifically, we fit a GRM for $S_i \mid A_i = a, \mathbf{X}_i$:

$$S_i = \mathcal{D}_a \odot \mathcal{H}_a(\mathbf{X}_i^{\mathsf{T}}\gamma_a, \varepsilon_{ia}) \quad \text{with} \quad P(\varepsilon_{ai} \le e \mid \mathbf{X}_i) = \mathcal{F}_a(e), \tag{1.6}$$

where the composite function $\mathcal{D}_a \odot \mathcal{H}_a$ assumes that $\mathcal{D}_a(\cdot)$ is an increasing function and $\mathcal{H}_a(\cdot,\cdot)$ is a strictly increasing function of each of its arguments, and the unknown covariate effects $\gamma_a = (\gamma_{a1}, \ldots, \gamma_{ap})^{\mathsf{T}}$ are constrained to the unit sphere $\Omega : \{\gamma : \|\gamma\|_2 = 1\}$ for identifiability[45]. With the given γ_a under GRM and the no-unmeasured-confounders assumption, $\psi_{a,\mathrm{f}}(s;\mathbf{x})$ can be estimated non-parametrically via kernel smoothing. To estimate γ_a, we propose an adapted maximum rank correlation (MRC) estimator $\widehat{\gamma}_a = \mathrm{argmax}_{\gamma\in\Omega} \left\{ \sum_{i\neq j, A_i = A_j = a} I(\mathbf{X}_i^{\mathsf{T}}\gamma > \mathbf{X}_j^{\mathsf{T}}\gamma) I(S_i > S_j) \right\}$, which Sherman (1993)[109] showed to be consistent and asymptotically normal for γ_a. Subsequently, we es-

timate $\psi_{a,\mathsf{f}}(s,\mathbf{x})$ as

$$\widehat{\psi}_{a,\mathsf{f}}(s;\mathbf{x}) = \frac{\sum_{i=1}^{n} K_\zeta(\widehat{\gamma}_a\mathbf{X}_i - \widehat{\gamma}_a^{\mathsf{T}}\mathbf{x})K_h(S_i - s)}{\sum_{i=1}^{n} K_\zeta(\widehat{\gamma}_a\mathbf{X}_i - \widehat{\gamma}_a^{\mathsf{T}}\mathbf{x})}. \tag{1.7}$$

To estimate $\psi_{a,\mathsf{m}}(s;\mathbf{x})$, we fit a varying-coefficient generalized linear model (VGLM):

$$E(Y_i \mid A_i = a, S_i = s, \mathbf{X}_i) = M\{\beta_a(S_i)^{\mathsf{T}}\vec{\mathbf{X}}_i\}, \tag{1.8}$$

where $M(\cdot)$ is a known smooth link function, $\vec{\mathbf{x}} = (1, \mathbf{x}^{\mathsf{T}})^{\mathsf{T}}$ for any vector $\mathbf{x}$ and $\beta_a(s)$ is an unknown $p+1$ dimensional unspecified smooth coefficient functions[47]. We may estimate $\beta_a(s)$ as $\widehat{\beta}_a(s)$, the solution to the estimating equation $\widehat{\mathbf{U}}_a(\beta;s) = n^{-1}\sum_{i=1}^{n} I(A_i = a)K_h(S_i - s)\vec{\mathbf{X}}_i\left\{Y_i - M(\beta^{\mathsf{T}}\vec{\mathbf{X}}_i)\right\} = 0$. Then we estimate $\psi_{a,\mathsf{m}}(s;\mathbf{x})$ as

$$\widehat{\psi}_{a,\mathsf{m}}(s,\mathbf{x}) = M\{\widehat{\beta}_a(s)^{\mathsf{T}}\vec{\mathbf{x}}\}. \tag{1.9}$$

These estimators (1.7) and (1.9) can then be plugged into (1.5b) and (1.5c) to construct $\widehat{f}_{a,\mathsf{DR}}(s)$ and $\widehat{m}_{a,\mathsf{DR}}(s)$ as in (1.5a).

Based on $\widehat{m}_{a,\mathsf{DR}}(s)$ and $\widehat{f}_{a,\mathsf{DR}}(s)$, we obtain a DR estimator for $g_{opt}(s)$ as: $\widehat{g}_{\mathsf{DR}}(s) = \widehat{m}_{\mathsf{DR}}(s) + \widehat{\lambda}_{\mathsf{DR}}\widehat{\mathcal{P}}_{0,\mathsf{DR}}(s)$, where $\widehat{m}_{\mathsf{DR}}(s) = \sum_{a=0}^{1}\widehat{m}_{a,\mathsf{DR}}(s)\widehat{\mathcal{P}}_{a,\mathsf{DR}}(s)$,
$\widehat{\lambda}_{\mathsf{DR}} = \frac{\int\left\{\widehat{m}_{0,\mathsf{DR}}(s) - \widehat{m}_{1,\mathsf{DR}}(s)\right\}\widehat{\mathcal{P}}_{1,\mathsf{DR}}(s)\widehat{f}_{0,\mathsf{DR}}(s)ds}{\int\widehat{\mathcal{P}}_{0,\mathsf{DR}}(s)\widehat{f}_{0,\mathsf{DR}}(s)ds}$, and $\widehat{\mathcal{P}}_{a,\mathsf{DR}}(s) = \frac{\widehat{f}_{a,\mathsf{DR}}(s)}{\widehat{f}_{0,\mathsf{DR}}(s)+\widehat{f}_{1,\mathsf{DR}}(s)}$, for $a = 0, 1$. We can now construct a DR plug-in estimator for $\Delta_{g_{\mathsf{opt}}}$ as $\widehat{\Delta}_{\widehat{g},\mathsf{DR}} = \widehat{\mu}_{1,\widehat{g},\mathsf{DR}} - \widehat{\mu}_{0,\widehat{g},\mathsf{DR}}$, where

$$\widehat{\mu}_{a,\widehat{g},\mathsf{DR}} = n^{-1}\sum_{i=1}^{n}\left\{g_{\mathsf{opt}}(S_i)\widehat{\omega}_{ai} - (\widehat{\omega}_{ai} - 1)\hat{\zeta}_{a,g_{\mathsf{opt}}}(\mathbf{X}_i)\right\},$$

and $\hat{\zeta}_{a,g_{\mathsf{opt}}}(\mathbf{x}) = \int g_{\mathsf{opt}}(s)\widehat{\psi}_{a,\mathsf{f}}(s,\mathbf{x})ds$ is an estimator for $\zeta_{a,g_{\mathsf{opt}}}(\mathbf{x}) = E\{g_{\mathsf{opt}}(S_i^{(a)}) \mid \mathbf{X}_i = \mathbf{x}\}$ derived under the GRM. The plug-in estimator for $\Delta_{g_{\mathsf{opt}}}$ is DR in the sense that it is consistent when either the PS model or conditional surrogate model is correctly specified.

Similarly, we obtain $\widehat{\Delta}_{\mathsf{DR}} = \widehat{\mu}_{1,\mathsf{DR}} - \widehat{\mu}_{0,\mathsf{DR}}$ to estimate Δ, where

$$\widehat{\mu}_{a,\mathsf{DR}} = n^{-1} \sum_{i=1}^{n} \left\{ Y_i \widehat{\omega}_{ai} - (\widehat{\omega}_{ai} - 1) \hat{\zeta}_a(\mathbf{X}_i) \right\},$$

where $\hat{\zeta}_a(\mathbf{x}) = \int \widehat{\psi}_{a,\mathsf{m}}(s;\mathbf{x}) \widehat{\psi}_{a,\mathsf{f}}(s;\mathbf{x}) ds$ is an estimator for $\zeta_a(\mathbf{x}) = E(Y_i^{(a)} \mid \mathbf{X}_i = \mathbf{x})$. The plug-in estimator for Δ is DR in the sense that it is consistent when either the PS model or the conditional outcome model is correctly specified.

Finally, we estimate $\mathrm{PTE}_{g_{\mathsf{opt}}}$ as $\widehat{\mathrm{PTE}}_{\widehat{g},\mathsf{DR}} = \widehat{\Delta}_{\widehat{g},\mathsf{DR}} / \widehat{\Delta}_{\mathsf{DR}}$. Following similar arguments as given in Appendix 4, it is not difficult to show that $\widehat{\Delta}_{\widehat{g},\mathsf{DR}}$, $\widehat{\Delta}_{\mathsf{DR}}$ and $\widehat{\mathrm{PTE}}_{\widehat{g},\mathsf{DR}}$ are DR estimators for $\Delta_{g_{\mathsf{opt}}}$, Δ, and $\mathrm{PTE}_{g_{\mathsf{opt}}}$, respectively. The steps needed to construct the DR estimator are summarized in Algorithm 1.

Algorithm 1: Estimation procedure for constructing DR estimators.

1. Estimate the conditional density, $S \mid A, \mathbf{X}$, by fitting a GRM as in (1.6), obtain the MRC estimator, and calculate $\widehat{\psi}_{a,\mathsf{f}}(s;\mathbf{x})$ via kernel smoothing as in (1.7).

2. Estimate the conditional mean, $E(Y \mid A, S, \mathbf{X})$, by fitting a VCGLM as in (1.8) and solve the corresponding estimating equation for $\widehat{\psi}_{a,\mathsf{m}}(s,\mathbf{x})$ as in (1.9).

3. Calculate the plug-in estimates $\widehat{g}_{\mathsf{DR}}(s)$, $\hat{\Delta}_{\widehat{g},\mathsf{DR}}$, $\hat{\Delta}_{\mathsf{DR}}$, and $\widehat{\mathrm{PTE}}_{\widehat{g},\mathsf{DR}}$.

1.4 Perturbation Resampling

In practice, we can estimate σ^2 empirically by estimating the influence functions or via perturbation resampling similar to those employed in Wang (2020)[124]. Given the complexity of the doubly

robust estimator, we propose to estimate the variability and construct confidence intervals of our proposed estimators using a perturbation-resampling approach[63,117]. For resampling, we generate $\{\mathbf{V}^{[b]} = (V_1^{[b]}, ..., V_n^{[b]})^\intercal, b = 1, ..., B\}$, which are $n \times B$ independent and identically distributed non-negative random variables from a known distribution with unit mean and unit variance, such as the unit exponential distribution. For each set of $\mathbf{V} = (V_1, ..., V_n)^\intercal$, we let $\bar{\mathcal{V}}_i = V_i/(n^{-1}\sum_{i=1}^n V_i)$ and perturb each observation i by $\bar{\mathcal{V}}$. In Appendix 5, we provide the detailed perturbation resampling procedure for both the IPW and DR estimators. Operationally, we generate a large number, say $B = 500$, realizations for $\mathbf{V}$ and then obtain B realizations of the perturbed statistics of interest. Standard error estimates and confidence intervals can then be constructed based on empirical quantiles of these realizations.

1.5 Simulation Studies

We conduct simulation studies to evaluate the finite sample performance of our proposed estimators compared to several existing methods. Namely, we compare our IPW and DR estimators to the following estimators:

1. the PTE estimator of Freedman (1992)[39], denoted $\widehat{\text{PTE}}_{\text{F,naive}}$;
2. the PTE estimator given in Parast (2016)[88], denoted $\widehat{\text{PTE}}_{\text{P}}$; and
3. the PTE estimator of Wang (2020)[124], denoted $\widehat{\text{PTE}}_{\text{W}}$.

Note that all of the above estimators assume that treatment is randomly assigned and do not take into account baseline covariates $\mathbf{X}$ in their models. We let $n = 400$ and $n = 1000$ and choose $K(\cdot)$ as a Gaussian kernel. We set the bandwidth $h = h_{opt}n^{-c_0}$, $c_0 = 0.11$ where $h_{opt} = 1.06n^{-1/5}$ to satisfy the undersmoothing assumption[106]. We compute the true population parameters via Monte Carlo, under the counterfactual models used to generate the data, with $N = 100,000$ averaged over 100

replications. All results are summarized based on 500 simulated datasets for each configuration, and $B = 500$ resampling replications were used for variance and interval estimation.

We consider two general settings. In Setting I, the surrogate is moderately strong and both assumptions (A1) and (A2) hold so that the surrogate paradox is avoided. In Setting II, the surrogate is weak and the working independence assumption (1.4) is violated. The relationship between S and $E(Y \mid S = s)$ are visualized in the Supplementary Materials.

Specifically, in Setting I, we generate a 3-dimensional baseline covariate vector $\mathbf{X}_i = (X_{i1}, X_{i2}, X_{i3})^\intercal$ as $X_{i1} \sim \mathcal{N}(0, 0.04)$, $X_{i2} \sim \text{Gamma}(2, 2)$ and $X_{i3} \sim \text{Uniform}(-1, 1)$, and

$$S_i^{(0)} = \gamma_{0[1]}^\intercal \vec{\mathbf{X}}_i + \varepsilon_i, \quad S_i^{(1)} = \gamma_{1[1]}^\intercal \vec{\mathbf{X}}_i + \varepsilon_i,$$

$$Y_i^{(0)} = 0.5S_i^{(0)} + \beta_{0[1]}^\intercal \vec{\mathbf{X}}_i + X_{1i}X_{2i} + X_{2i}X_{3i} + e_i, \quad Y_i^{(1)} = 0.3S_i^{(1)} + \beta_{1[1]}^\intercal \vec{\mathbf{X}}_i + X_{1i}X_{2i} + X_{2i}X_{3i} + e_i,$$

where $\varepsilon_i \sim \mathcal{N}(0, 1)$, $e_i \sim \mathcal{N}(0, 0.04)$, $\gamma_{0[1]} = (0, 0.5, 1, -0.5)^\intercal$, $\gamma_{1[1]} = (0, 1, 0.5, 2)^\intercal$, $\beta_{0[1]} = (0, 0.2, -0.3, -0.5)^\intercal$, and $\beta_{1[1]} = (0, 1, -0.5, 0.2)^\intercal$.

We generate $A_i \mid \mathbf{X}_i$ from the PS model

$$P(A_i = 1 \mid \mathbf{X}_i) = \text{expit}\{-0.8X_{i1} + 0.7X_{i2} - \log(X_{i3}) + 0.6X_{i1}X_{i3}\},$$

so that 58% receive treatment $A = 1$. Under this setting, $\Delta_{g_{\text{opt}}} = 0.29$ and $\Delta = 0.54$ so that the true potential outcomes PTE is 0.539, i.e., S is a moderately strong surrogate for Y. We consider scenarios in which we correctly specify both the PS and OR models, misspecify the PS model by omitting the interaction term X_1X_3, misspecify the OR model by omitting the variable X_2 and all interaction terms including X_2, and misspecify both models.

In Setting II, we consider a relatively weak surrogate and generate data such that the effect of S on Y is also non-linear. We generate baseline covariates $\mathbf{X}_i = (X_{i1}, X_{i2}, X_{i3})^\intercal$ from $X_{i1} \sim \mathcal{N}(0, 1)$,

$X_{i2} \sim \text{Gamma}(2,2)$, and $X_{i3} \sim \text{Uniform}(0,5)$. Given $\mathbf{X}_i$, we generate

$$S_i^{(0)} = \gamma_{0[2]}^\intercal \vec{\mathbf{X}}_i + \varepsilon, \quad S_i^{(1)} = \gamma_{1[2]}^\intercal \vec{\mathbf{X}}_i + \varepsilon_i,$$

$$Y_i^{(0)} = 100 + \beta_{0[2]}(S_i^{(0)})^\intercal \mathbf{X}_i + e_i, \quad Y_i^{(1)} = 50 + \beta_{1[2]}(S_i^{(1)})^\intercal \mathbf{X}_i + e_i,$$

where $\varepsilon_i \sim \mathcal{N}(0,4)$ and $e_i \sim \mathcal{N}(0,1)$ and we let $\gamma_{0[2]} = (100,1,5,0)^\intercal$, $\gamma_{1[2]} = (100,2,4,0)^\intercal$, $\beta_{0[2]}(s) = (s, -2\log(s), 25)^\intercal$, and $\beta_{1[2]}(s) = (s, -3\log(s), -14)^\intercal$.

We generate $A_i \mid \mathbf{X}_i$ from the same PS model as in setting 1, but because of the different covariate distributions, here 50% receive treatment $A = 1$. Under this data generating mechanism, $\Delta_{g_{\text{opt}}} = 5.7$ and $\Delta = 26.7$, resulting in $\text{PTE}_{g_{\text{opt}}} = 0.214$. We consider scenarios in which we correctly specify both the PS and OR models, misspecify the PS model by omitting the $\log(X_3)$ term, misspecify the OR model by omitting X_2, and misspecify both models.

We first summarize the results for Setting I. In Figure 1.1, we plot the empirical biases, the empirical standard error (ESE) compared to the average of the estimated standard error (ASE), and empirical coverage probabilities of the 95% pointwise confidence intervals (CIs) for $g_{opt}(\cdot)$ based on the DR estimator $\widehat{g}_{\text{DR}}(\cdot)$ estimated with sample size $n = 1000$ under four specification scenarios. When at least one of the two models is correctly specified, the point estimates for $g_{opt}(\cdot)$ present negligible bias, the ASEs are close to the ESEs, probabilities of the 95% CIs are close to their nominal level. When both models are misspecified, bias is observed in the tails, the ASE somewhat underestimates the ESE, and the coverage probabilities of the 95% confidence intervals are somewhat below the nominal level. Results for $n = 400$ bear similar patterns and can be found in the Supplementary Materials.

In Table 1, we summarize results for PTE estimation obtained via the proposed method and other existing methods. When at least one of the PS and OR models are correctly specified, our proposed DR estimator displays negligible bias and nominal coverage. The IPW estimator $\widehat{\text{PTE}}_{\widehat{g}}$ has substantial bias when the PS model is misspecified and the DR estimator also presents bias when both models

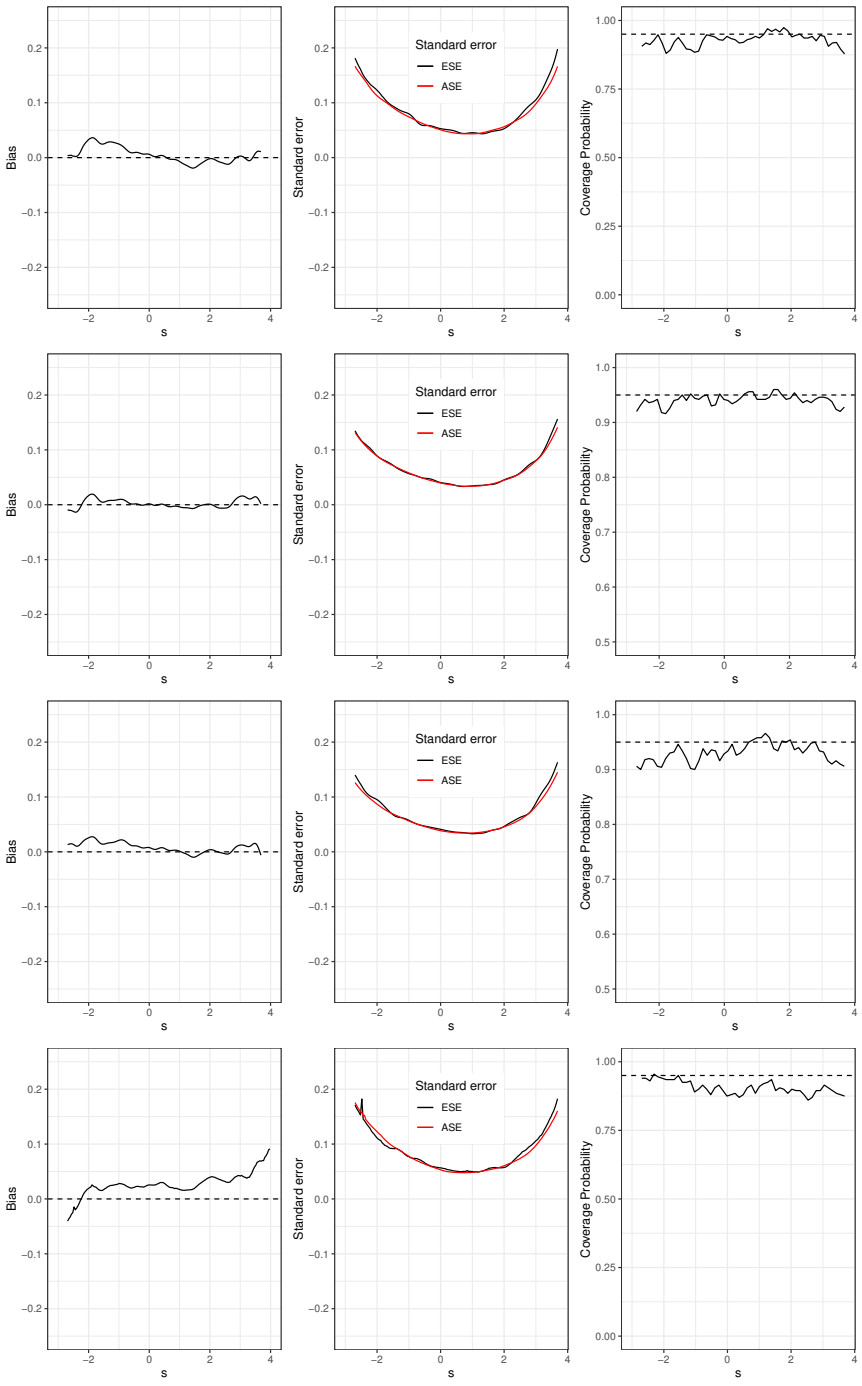


Figure 1.1: Empirical bias, empirical standard error (ESE) versus the average of the estimated standard error (ASE), and coverage probabilities of the 95% confidence intervals for $\widehat{g}_{\text{opt}}(s)$ when $n = 1000$ and (Row 1) both models are correctly specified, (Row 2) PS model is misspecified, (Row 3) OR model is misspecified, (Row 4) both models are misspecified.

are incorrect, as expected. In addition, the IPW estimator is less efficient compared to the DR estimator when the PS model is correctly specified. The estimator from Wang (2020)[124], $\widehat{\text{PTE}}_{\text{W}}$ shows considerable bias and below nominal coverage, and $\widehat{\text{PTE}}_{\text{F,naive}}$ and $\widehat{\text{PTE}}_{\text{P}}$ show substantial bias.

Table 1.1: Bias for PTE estimators whose target parameter is $\text{PTE}_{g_{\text{opt}}} = 0.539$, Empirical Standard Error (ESE), Average of the Estimated Standard Errors (ASE) and Empirical Coverage Probabilities of the 95% CIs of Estimators under Different Model Scenarios for Setting I.

Size	Estimator	Scenario	Bias	ESE	ASE	Coverage
n = 400	$\widehat{\text{PTE}}_{\text{F,naive}}$	No $\mathbf{X}$	-0.159	0.114	-	-
	$\widehat{\text{PTE}}_{\text{P}}$	No $\mathbf{X}$	-0.223	0.101	-	-
	$\widehat{\text{PTE}}_{\text{W}}$	No $\mathbf{X}$	-0.105	0.110	0.110	0.729
	$\widehat{\text{PTE}}_{\widehat{g}}$	PS Correct	0.006	0.087	0.088	0.930
		PS Misspecified	-0.059	0.107	0.109	0.916
	$\widehat{\text{PTE}}_{\widehat{g},\text{DR}}$	Both Correct	0.003	0.079	0.079	0.940
		PS Misspecified	-0.005	0.074	0.079	0.954
		OR Misspecified	-0.007	0.084	0.082	0.940
		Both Misspecified	-0.091	0.119	0.115	0.779
n = 1000	$\widehat{\text{PTE}}_{\text{F,naive}}$	No $\mathbf{X}$	-0.151	0.064	-	-
	$\widehat{\text{PTE}}_{\text{P}}$	No $\mathbf{X}$	-0.204	0.061	-	-
	$\widehat{\text{PTE}}_{\text{W}}$	No $\mathbf{X}$	-0.103	0.181	0.174	0.803
	$\widehat{\text{PTE}}_{\widehat{g}}$	PS Correct	-0.006	0.052	0.053	0.950
		PS Misspecified	-0.087	0.069	0.072	0.768
	$\widehat{\text{PTE}}_{\widehat{g},\text{DR}}$	Both Correct	-0.006	0.051	0.050	0.944
		PS Misspecified	-0.003	0.050	0.050	0.956
		OR Misspecified	0.001	0.050	0.052	0.948
		Both Misspecified	-0.107	0.074	0.072	0.635

Table 2 shows that under Setting II, $\widehat{\text{PTE}}_{\widehat{g}}$ is consistent when the PS model is correctly specified and $\widehat{\text{PTE}}_{\widehat{g},\text{DR}}$ is consistent when either the PS model or OR model is correctly specified. However, other literature estimators, $\widehat{\text{PTE}}_{\text{F,naive}}$, and $\widehat{\text{PTE}}_{\text{P}}$ all estimate the true PTE as being close to 0. This is likely due to the non-monotone relationship between Y and S in this setting, and the fact that these estimators use S directly rather than $g_{\text{opt}}(S)$ in estimating the treatment effect.

Table 1.2: Bias for PTE estimators whose target parameter is $\text{PTE}_{g_{\text{opt}}} = 0.214$, Empirical Standard Error (ESE), Average of the Estimated Standard Errors (ASE), and Empirical Coverage Probabilities of the 95% CIs of Estimators under Different Model Scenarios for Setting II.

Size	Estimator	Scenario	Bias	ESE	ASE	Coverage
n = 400	$\widehat{\text{PTE}}_{\text{F,naive}}$	No $\mathbf{X}$	-0.166	0.031	-	-
	$\widehat{\text{PTE}}_{\text{P}}$	No $\mathbf{X}$	-0.179	0.044	-	-
	$\widehat{\text{PTE}}_{\text{W}}$	No $\mathbf{X}$	-0.015	0.042	0.041	0.980
	$\widehat{\text{PTE}}_{\widehat{g}}$	PS Correct	0.006	0.050	0.048	0.936
		PS Misspecified	0.024	0.053	0.052	0.894
	$\widehat{\text{PTE}}_{\widehat{g},\text{DR}}$	Both Correct	0.002	0.048	0.049	0.946
		PS Misspecified	0.000	0.047	0.046	0.952
		OR Misspecified	0.005	0.048	0.046	0.940
		Both Misspecified	-0.017	0.058	0.060	0.845
n = 1000	$\widehat{\text{PTE}}_{\text{F,naive}}$	No $\mathbf{X}$	-0.210	0.005	-	-
	$\widehat{\text{PTE}}_{\text{P}}$	No $\mathbf{X}$	-0.230	0.104	-	-
	$\widehat{\text{PTE}}_{\text{W}}$	No $\mathbf{X}$	-0.011	0.032	0.030	0.970
	$\widehat{\text{PTE}}_{\widehat{g}}$	PS Correct	-0.004	0.030	0.030	0.948
		PS Misspecified	0.020	0.030	0.031	0.914
	$\widehat{\text{PTE}}_{\widehat{g},\text{DR}}$	Both Correct	0.005	0.028	0.029	0.942
		PS Misspecified	0.004	0.029	0.029	0.946
		OR Misspecified	0.007	0.029	0.028	0.940
		Both Misspecified	-0.024	0.051	0.045	0.853

1.6 Real World Data Applications

In this section, we use our proposed estimators to examine the comparative effectiveness of biologic therapies, and we validate two surrogate markers in two different RWD settings of interest. In the first application, we validate an algorithm-derived score for the likelihood of nonresponse based on narrative text data in an EHR setting. In the second application, we validate a non-invasive partial Mayo score in a cross-trial comparison. In both data applications, the number of covariates are limited such that there may be unmeasured residual confounding for the relationship between treatment A and the surrogate S and primary outcome Y. However, we proceed with the analyses to illustrate our

method and to show that using the PTE estimator of Wang (2020)[124] that does not adjust for baseline confounders can result in qualitatively different conclusions on the strength of surrogates. In both data applications, race/ethnicity was defined as a self-disclosed, mutually exclusive categorical variable of non-Hispanic white, non-Hispanic black, non-Hispanic Asian, Hispanic, or other. However, due to the predominance of non-Hispanic white ($\approx$ 90%), we used a binary race/ethnicity variable of non-Hispanic white or not. Race/ethnicity was used as a confounder since the propensity of receiving treatment can be influenced by race/ethnicity. For example, it has been documented that racial bias in pain perception can result in differential pain treatment recommendations[53].

1.6.1 Application I: EHR Data

Our data consisted of 1451 IBD patient records from Mass General Brigham for patients who initiated adalimumab (1060) or infliximab (391) between December 1998 and June 2010. We excluded 211 patients who did not have surrogate outcomes S or who were missing data on 6 commonly used baseline confounders: age, sex, race/ethnicity, prior hospitalizations, prior anti-TNF status, and Charlson comorobidity score[3]. Our complete case analysis consisted of 1240 patients (971 on adalimumab and 269 on infliximab). It appears that the treatment groups are somewhat imbalanced (Table 3), with patients receiving adalimumab tending to be slightly older (mean age: 35.8 vs 34.3), less likely to be male (0.42 vs 0.48), more likely to be white (0.93 vs 0.83), have a lower Charlson score 2.49 vs 2.87), more likely to have had a prior hospitalization (0.49 vs 0.44), and prior anti-TNF (0.29 vs 0.12).

We applied our proposed estimators to examine the surrogacy of a likelihood of nonresponse score at 6 months for the number of narrative mentions of abdominal pain at 1 year, which is an important outcome of interest. The likelihood of nonresponse score is an algorithm-derived score that weights textual information such as the number of narrative mentions for diarrhea and fatigue[2]. The likelihood of nonresponse score has been shown to differentiate symptomatic nonresponse from response and was correlated with higher IBD-surgery and hospitalization rates[2].

Table 1.3: Baseline characteristics (mean (SD)) of 1240 IBD patients from MGB who initiaited adalimumab or infliximab.

	Infliximab $n_0 = 971$	Adalimumab $n_1 = 269$
Age	34.3 (16.6)	35.8 (14.0)
Sex		
Male	0.48 (0.50)	0.42 (0.50)
Race/Ethnicity		
White	0.83 (0.49)	0.93 (0.29)
Charlson Score	2.87 (3.22)	2.49 (2.75)
Prior Hospitalization	0.44 (0.50)	0.49 (0.50)
Prior anti-TNF Status	0.12 (0.32)	0.29 (0.45)

The estimated $g_{opt}(\cdot)$ along with point-wise CIs based on the IPW (red) and DR (black) estimators are similar. The estimated transformation function appears to be non-linear, with a positive trend between s and $\widehat{g}_{opt}(s)$, as shown in Figure 1.2. The DR estimate for the treatment effect is $\widehat{\Delta} = 39.9$ (i.e., patients initiating adalimumab had a greater expected number of narrative mentions of abdominal pain at 1 year) and the corresponding treatment effect on the transformed surrogate $\Delta_{g_{\text{opt}}}$ is estimated as $\widehat{\Delta}_{\widehat{g},\text{DR}} = 28.5$, resulting in a PTE estimate of 0.72 with a 95% CI of $(0.52, 0.92)$, suggesting that the likelihood of nonresponse score at 6 months is a strong surrogate for abdominal pain at 1 year. Our finding that patients initiating adalimumab had a greater number of narrative mentions of abdominal pain at 1 year compared to patients initiating infliximab is consistent with previous literature [2]. The results for the IPW PTE estimate are similar with a PTE estimate of 0.72 and a slightly wider 95% CI of $(0.51, 0.93)$. However, Wang (2020) assumes that treatment is randomized and thus under-estimates both the treatment effect on the optimal transformation $\Delta_{g_{\text{opt}}}$ and the treatment effect Δ, but more severely under-estimates $\Delta_{g_{\text{opt}}}$, resulting in a deflated estimate of the PTE = $\Delta_{g_{\text{opt}}}/\Delta$. Indeed, the Wang (2020) method estimates Δ to be 17.5 and $\Delta_{g_{\text{opt}}}$ to be 7.0, giving a biased PTE estimate of 0.41 with a 95% CI of $(0.14, 0.67)$, suggesting a weak-to-moderate surrogate rather than a strong surrogate.

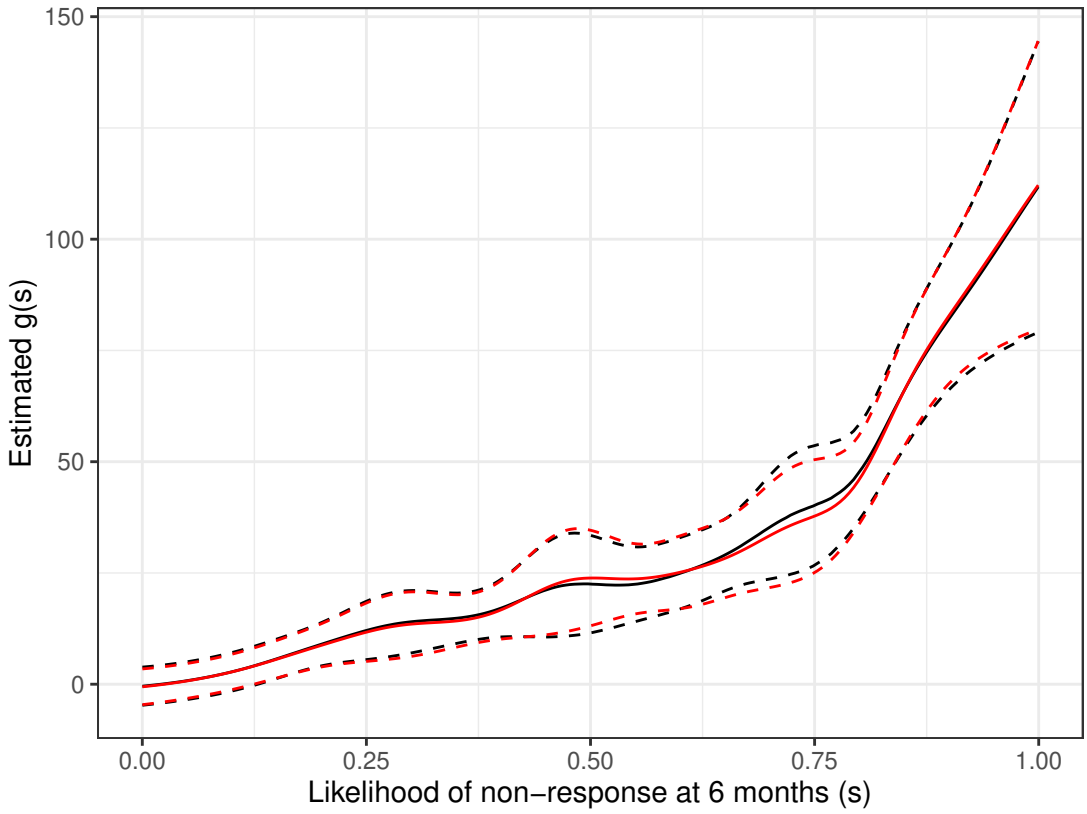


Figure 1.2: Estimated $g_{\text{opt}}(s)$ based on IPW (red) and DR (black) estimators and pointwise 95% confidence intervals for the likelihood of nonresponse score at 6 months (surrogate) in an EHR comparison of adalimumab and infliximab for 1240 IBD patients

1.6.2 Application II: Cross-trial Treatments

Previous research has shown that a cheap and non-invasive partial Mayo score may be a good surrogate for the expensive and invasive full Mayo score[70,23,2]. The partial Mayo score is a composite score that can be measured within weeks after receiving therapy and ranges in value from 0 to 9. It is based on a patient's self-assessed stool frequency (0-3), rectal bleeding (0-3), and a physician's global assessment (0-3). The full Mayo score ranges from 0 to 12 and requires an invasive endoscopy score evaluating mucosal appearance and is typically collected much later in the trial.

To illustrate the utility of our methods, we examine the surrogacy of the partial Mayo score at week 6 on the primary outcome of the full Mayo score at week 54 among patients with moderate-to-severe ulcerative colitis (UC). We compare head-to-head trials of two biologic therapies, infliximab and golimumab[118]. Treatment randomization is broken by combining data from *treatment arms only* in two separate trials with similar inclusion criteria, one comparing infliximab against a placebo (NCT00036439) and another comparing golimumab against a placebo (NCT00488631). To adjust for confounding bias, we consider baseline covariates $\mathbf{X}$ including patient age, sex, race/ethnicity, and a health status score ranging from 0 to 100. Data were obtained from the Yale University Open Data Access (YODA) database[99]. It appears that the treatment groups are somewhat imbalanced (Table 4), with patients receiving infliximab tending to be slightly older (mean age: 41.4 vs 39.9 years), more likely to be male (0.61 vs 0.57), white (0.95 vs 0.88), and have a higher health status score (57.8 vs 55.1), indicating more severe disease.

Table 1.4: Baseline characteristics (mean (SD)) of 381 moderate-to-severe UC patients who initiaited infliximab or golimumab.

	Golimumab $n_0 = 216$	Infliximab $n_1 = 165$
Age	39.9 (13.2)	41.4 (13.7)
Sex		
Male	0.57 (0.50)	0.61 (0.49)
Race/Ethnicity		
White	0.88 (0.32)	0.95 (0.22)
Health Score	55.1 (20.3)	57.8 (20.9)

The ranges of S and Y in the two treatment groups are $\{1, 2, \ldots, 9\}$ and $\{1, 2, \ldots, 12\}$ respectively, and the distributions are provided in Figure 1.3.

The analysis focused on the 381 patients who had complete information on the partial Mayo score at week 6, the full Mayo score at week 54, and baseline covariates, with 216 patients in the golimumab group and 165 in the infliximab group. We applied the proposed methods to examine $g_{opt}(\cdot)$ of the

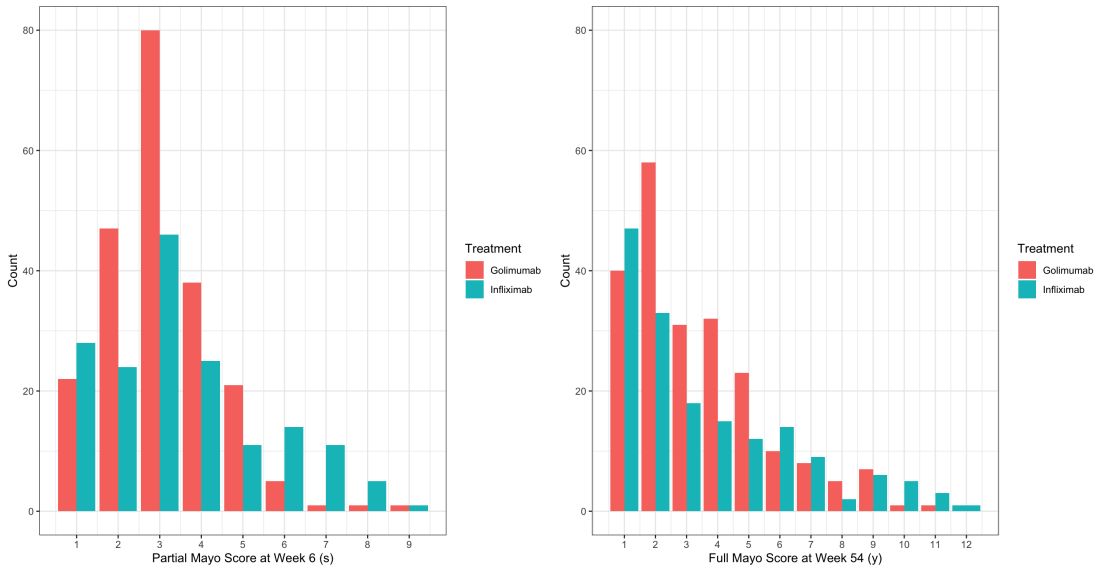


Figure 1.3: (Left) Histogram of the partial Mayo score at week 6 (surrogate) in the two treatment groups; (Right) Histogram of the full Mayo score at week 54 (primary outcome) in the two treatment groups

surrogate for predicting the treatment response as quantified by the full Mayo score. The estimated $g_{opt}(\cdot)$ along with point-wise CIs based on the IPW (red) and DR (black) estimators are similar. The estimated transformation function appears to be slightly non-linear, with a clear positive trend between s and $\widehat{g}_{opt}(s)$, as shown in Figure 1.4.

The DR estimator for the treatment effect is estimated as $\widehat{\Delta} = 1.19$ in favor of golimumab and the corresponding treatment effect on the optimally transformed outcome $\Delta_{g_{\text{opt}}}$ is estimated to be $\widehat{\Delta}_{\widehat{g},\text{DR}} = 1.09$. This results in a DR PTE estimate of 0.89 with a 95% CI of $(0.45, 1.33)$, suggesting that the partial Mayo score at week 6 is a strong surrogate for the full Mayo score at week 54. The results for the IPW PTE estimate are similar at 0.90 with a slightly wider 95% CI of $(0.43, 1.37)$. However, Wang (2020) again under-estimates both the treatment effect on the optimal transformation $\Delta_{g_{\text{opt}}}$ and the treatment effect Δ, but more severely under-estimates $\Delta_{g_{\text{opt}}}$, resulting in a deflated estimate of the PTE = $\Delta_{g_{\text{opt}}}/\Delta$. Wang (2020) estimates Δ to be 0.24 and $\Delta_{g_{\text{opt}}}$ to be 0.09, resulting in a biased PTE estimate of 0.39 with a 95% CI of $(-0.04, 0.83)$, suggesting a weak-to-moderate surrogate rather

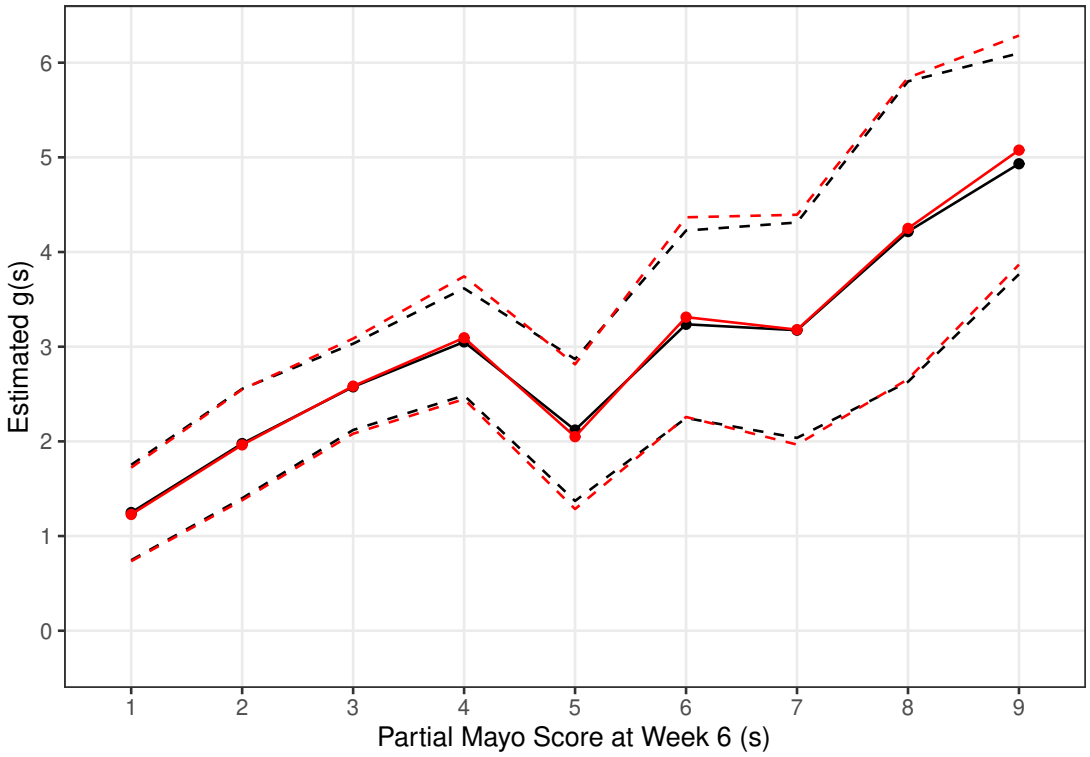


Figure 1.4: Estimated $g_{opt}(s)$ based on IPW (red) and DR (black) estimators and pointwise 95% confidence intervals for the partial Mayo score at week 6 (surrogate) in a cross-trial comparison of infliximab and golimumab for 361 UC patients

than a strong surrogate.

1.7 Discussion

Despite the growth of RWD, robust and flexible statistical methods to identify and validate surrogate markers in such data settings are lacking. There is great interest in leveraging RWD, including EHRs, registry data, and cross-trial data, to inform the design of shorter and cheaper clinical trials through the use of valid surrogate markers. Motivated by the need for statistical methods in CER for surrogate marker evaluation, we propose novel IPW and DR estimators for the optimal transformation function and the corresponding PTE explained by a surrogate in RWD settings. In two separate ap-

plications, we validate two surrogate markers for outcomes of interest in IBD. Using EHR data from Mass General Brigham, we validate an algorithm-derived likelihood of nonresponse score at 6 months as a surrogate for the number of narrative mentions of abdominal pain at 1 year. In a second example, we validate a partial Mayo score at week 6, which does not require an invasive endoscopy procedure, as a strong surrogate for the full Mayo score at week 54 in a cross-trial study, supplementing evidence from previous placebo-controlled trials[70,23,2]. These findings may be particularly useful in informing future cross-trial designs for biological therapies.

One important question is how one should use the estimated PTE to identify or validate a proposed surrogate marker. In practice, the primary use of the PTE estimate is to determine whether a proposed surrogate is of 'high quality', in the sense that it can explain a large proportion of the treatment effect on the primary outcome of interest. While there are no official criteria on what constitutes a high-quality surrogate, previous work has proposed calling a surrogate "high quality" if the lower bound of the 95% CI of the PTE estimate is above 0.50 (Lin, 1997). If a proposed surrogate meets this criterion or some similar threshold, then future studies can use the surrogate to make inferences on the treatment effect on the primary outcome. For example, when the primary outcome is not available or may be costly to obtain, then the treatment effect on the transformed surrogate Δ_g may be used instead to test for the treatment effect on the primary outcome Δ. Finally, a validated surrogate can be useful in the design of future studies, where sample size and power calculations can be based on the treatment effect on the transformed surrogate Δ_g.

Our approach has some limitations. First, our proposed plug-in estimators for PTE use the same data to estimate both g_{opt} and PTE given g, which may result in overfitting bias. However, in simulation studies, the bias appears small compared to the standard error, even with modest sample sizes. For small sample sizes, cross-validation can be used, in which separate data is used to estimate g_{opt} and PTE given g. Second, we fit our PS models using logistic regression with specified basis functions, but alternative approaches like gradient boosting, super learner, and other machine learning classifiers can

be used. Third, our approach is tailored for a single surrogate. Future extensions of our method can be developed to incorporate multiple surrogates, for example, through a surrogate index approach[6,8]. Another interesting question is the relative efficiency gain of the DR estimator compared to the IPW estimator when the level of covariate overlap varies. In observational studies, the unconfoundedness assumption is more plausible when one adjusts for a richer set of covariates[102], the intuition being that including these covariates decreases the likelihood of unmeasured confounding. However, this can present an issue for covariate overlap, because if these covariates can nearly perfectly predict treatment assignment, then propensity scores will not be bounded away from zero and one[33]. Papers in the semiparametric estimation literature have shown that the convergence rate of estimators depends on the level of covariate overlap[66,54], and it would be interesting to examine this effect on efficiency in our setting.

Acknowledgements

This study, carried out under the Yale University Open Data Access Project (YODA) Project 2019-4092, used data obtained from YODA, which has an agreement with Janssen Research & Development, L.L.C. The interpretation and reporting of research using this data is solely the responsibility of the authors and does not necessarily represent the official views of the YODA Project or Janssen Research & Development, L.L.C. Larry Han was supported by the Clinical Orthopedic and Musculoskeletal Education and Training (COMET) Program, NIAMS grant T32 AR055885.

Data Availability Statement

The data that support the findings of this study are not publicly available but may be requested through an application to YODA.

Supplementary Materials

The supplementary materials contain six appendices. Appendix A.1 provides a derivation of the optimal transformation function. Appendix A.2 provides a derivation of the bounded PTE measure and avoidance of the surrogate paradox when assumptions (A1) and (A2) are satisfied. Appendix A.3 provides a proof for consistency and asymptotic normality of $\widehat{\text{PTE}}_{\widehat{g}}$. Appendix A.4 proves that our proposed DR estimators are consistent when either the PS model or the OR models are correctly specified. Appendix A.5 provides details on perturbation resampling. Appendix A.6 provides additional figures.

2

Federated Adaptive Causal Estimation (FACE) of Target Treatment Effects

2.1 Introduction

Multi-center, federated causal inference is of great interest, particularly when studying novel treatments, rare diseases, or in times of urgent health crises. For example, the COVID-19 pandemic has

highlighted the need for novel approaches to efficiently and safely evaluate the effectiveness of novel therapies and vaccines, leveraging data from multiple healthcare systems to ensure the generalizability of findings. Over the past few years, many research networks and data consortia have been built to facilitate multi-site studies and have been actively contributing to COVID-19 studies, including the Observational Health Data Sciences and Informatics (OHDSI) consortium [56] and the Consortium for Clinical Characterization of COVID-19 by EHR [15].

Analyzing data collected from multiple healthcare systems, however, is highly challenging for several reasons. Various sources of heterogeneity exist in terms of (i) differences in the underlying population of each dataset and (ii) policy level variations of treatment assignment. Since treatment effects may differ across different patient populations, it would be of interest to infer the average treatment effect (ATE) for specific target populations. However, the presence of heterogeneity and potential model mis-specification poses great difficulty to ensure valid estimates for the target average treatment effect (TATE). Furthermore, patient-level data typically cannot be shared across healthcare centers, which brings additional practical challenges. To overcome these challenges, we propose a Federated Adaptive Causal Estimation (FACE) framework that aims to incorporate heterogeneous data from multiple sites to make inference about the TATE, while accounting for heterogeneity and data-sharing constraints.

Most existing literature on federated learning has focused on regression and classification models [19,71,18,68,72,125,30]. Limited federated learning methods currently exist to make causal inference with multiple heterogeneous studies. Recently, [130] proposed federated inverse probability weighted (IPW) estimation of the ATE specifically for an entire study population. Although [130] provided multiple methods for point estimation and variance estimation, the choice of the proper method depends on prior knowledge about model homogeneity and specification, which are difficult to verify in practice. No empirical study in [130] was provided to test the robustness of the approach to the covariate shift assumption. In addition, their methods cannot be used to estimate the ATE of a target population that

differs from the full study population.[121] proposed a Bayesian approach that models potential outcomes as random functions distributed by Gaussian processes. Their focus is also on the population ATE rather than any particular target population, and their approach requires specifying parameters and hyperparameters of Gaussian processes and modeling between-site covariate correlations through kernel functions, which can be numerically intensive. Compared to these approaches, our approach estimates the TATE in a particular target population and accounts for the heterogeneity across populations without requiring prior information on the source data distribution or the validity of model specifications. Our approach further safeguards against incorporating source datasets that may introduce bias to the estimation of the TATE, known as negative transfer[87,127].

Another related strand of literature concerns the generalizability and transportability of randomized clinical trials to EHR studies. For example, Stuart et al.[113],[112],[111] assessed the generalizability of results from randomized trials to target populations of interest.[26,29], and[65] all focused on extending inferences about treatments from a randomized trial to a new target population by using different weighting schemes. For a comprehensive review of statistical methods for generalizability and transportability, see Degtiar & Rose[27]. However, to date, no literature in generalizability and transportability has sought to leverage observational data from a potentially large number of source sites in a data-adaptive manner to obtain unbiased, efficient, and robust estimation of target treatment effects.

The major contributions of FACE can be summarized as follows. First, FACE allows for flexibility in the specification of the target population. For example, the target population in a research network can be defined as the underlying population of a given healthcare center, or multiple healthcare centers that share certain properties (e.g., geographic location), or the overall population combining all sites. This flexibility provides stakeholders and policymakers at different levels with information on their respective target populations. Second, using a semiparametric density ratio weighting approach, FACE allows the distribution of covariates to be heterogeneous across sites. Third, FACE protects against negative transfer through an adaptive integration strategy which anchors on the target data

and computes data-adaptive weights for source sites. In doing so, FACE can achieve optimal efficiency while maintaining consistency, and it is robust to the distribution of data and potential model mis-specifications in the source sites. Moreover, FACE is a communication-efficient federated algorithm that allows each participating site to keep their data stored locally and only share summary statistics once with other sites.

The remainder of the paper is organized as follows. In Section 2.2, we introduce the problem setting, notation, and assumptions required for identification of the TATE. In Section 2.3, we describe the proposed FACE framework for estimating the TATE. We introduce the in-site estimators based on the target population and source populations separately in Sections 2.3.1 and 2.3.2 and present the adaptive and distributed integration in Section 2.3.3. In Section 2.4, we provide the theoretical guarantees of FACE, including double robustness, asymptotic normality, and relative efficiency. In Section 2.5, we conduct extensive simulations for various numbers of sites, data generating mechanisms, and show robustness to mis-specification of different models. In Section 2.6, we apply FACE to conduct a comparative effectiveness study of COVID-19 vaccines using the EHRs from five federated Veterans Affairs (VA) sites. We conclude in Section 2.7 with key takeaways and directions for future research.

2.2 Setting and Notation

For the i-th observation, we denote the outcome as $Y_i \in \mathbb{R}$, the p-dimensional baseline covariate vector as $\mathbf{X}_i = (X_{i1}, ..., X_{ip})^\top \in \mathcal{X} \subset \mathbb{R}^p$, and the indicator for binary treatment as $A_i \in \{0, 1\}$. There are $J \geq 1$ target sites and another $K \geq 0$ source sites. Let $\mathcal{T} \subseteq [J + K]$ indicate sites that are in the target population and $\mathcal{S} \subset [J + K]$ indicate sites that are in the source population, where $[K] = \{1, ..., K\}$ for any integer K. Under the federated learning setting, a total of N observations are stored at $J + K$ study sites, where the k-th site has sample size n_k, and $N = \sum_{k=1}^{J+K} n_k$. Let R_i be a

site indicator such that $R_i = k$ indicates the i-th patient in the k-th site. Indexing the site by a single integer R_i, we assume that each observation may only belong to one site. We summarize the observed data at each site k as $\mathscr{D}_k = \{(Y_i, \mathbf{X}_i^\top, A_i, R_i)^\top, R_i = k\}$, and consider a federated data setting where each site has access to its own patient-level data but can share only summary statistics with other sites. We denote the index set for each site as $\mathcal{I}_k = \{i : R_i = k\}$. The data included in the target sites are denoted by $\mathscr{D}_{\mathcal{T}}$. For simplicity of notation, we use $(Y, \mathbf{X}, A, R)$ without subscripts to state general assumptions and conclusions.

Under the potential outcomes framework[81,100], we denote $Y^{(a)}$ as the potential outcome of patients under treatment $A = a$, $a = 0, 1$. Our goal is to estimate the TATE for a specified target population $\mathcal{T}$,

$$\Delta_{\mathcal{T}} = \mu_{\mathcal{T}}^{(1)} - \mu_{\mathcal{T}}^{(0)}, \quad \mu_{\mathcal{T}}^{(a)} = \mathbb{E}(Y^{(a)} \mid R \in \mathcal{T}), \tag{2.1}$$

where the expectation is taken over the distribution in the target population. The target population can be specified at multiple levels (e.g., single site, multiple sites, all sites) corresponding to different targets of real world interest. This distinction between target and source sites also distinguishes our setting from that of[130], in which the target population always contains all participating sites.

To identify the TATE, we make the following standard assumptions[60,49] throughout the paper:

Assumption 1 *For a positive constant $\varepsilon > 0$, $a \in \{0, 1\}$, and $\mathbf{x} \in \mathcal{X}$,*

(a) Consistency: $Y = Y^{(A)}$.

(b) Overlapping of treatment arms: $\mathbb{P}(A = a \mid \mathbf{X} = \mathbf{x}, R = k) \in (\varepsilon, 1 - \varepsilon)$, $k \in [J + K]$.

(c) Overlapping of site populations: $\mathbb{P}(R = k \mid \mathbf{X} = \mathbf{x}) > \varepsilon$, $k \in [J + K]$.

(d) Ignorability: $\left(Y^{(1)}, Y^{(0)}\right) \perp\!\!\!\perp (A, R) \mid \mathbf{X}$ *for* $R \in \{\mathcal{T}, \mathcal{S}^*\}$ *for some* $\mathcal{S}^* \subseteq \mathcal{S}$.

Remark 1 *Assumption 1(d) implies that the underlying true treatment response pattern is shared across target sites and an unspecified subset of source sites $\mathcal{S}^* \subseteq \mathcal{S}$ so that the treatment effect estimates from $\mathcal{T}$ and $\mathcal{S}^*$ can be safely combined to estimate the TATE. Our adaptive selection and aggregation step in FACE, as detailed in Section 2.3.3, is designed to incorporate these source sites $\mathcal{S}^*$ for precision gain while preventing negative transfer from other source sites $\mathcal{S} \setminus \mathcal{S}^*$. Assumption 1(d) may be violated, for example, when the target and source populations differ along unobserved features, such that controlling for observed confounders is insufficient.* [83] *consider such a setting. They assume that the distribution of potential outcomes across target and source populations are the same conditioning on observed confounders* $\mathbf{X}$ *and unmeasured effect modifiers* $\mathbf{U}$ *and derive bounds for the TATE by assuming a sensitivity model that directly implies a bound on the unobserved distribution shift ratio. Since violations of the transportability assumption are in general untestable, many works have also proposed sensitivity analysis for how much violation of the assumption can result in transportability bias* [4,82].

We denote the specified models for the site-specific propensity score (PS) and outcome regression (OR) as:

$$\text{PS}: \quad \mathbb{P}(A = a \mid R = k, \mathbf{X}) = \pi_k(a, \mathbf{X}; \alpha_k), \tag{2.2}$$

$$\text{OR}: \quad \mathbb{E}(Y \mid R = k, A = a, \mathbf{X}) = m(a, \mathbf{X}; \beta_{a,k}). \tag{2.3}$$

For the target sites, we require $E(Y^{(a)} \mid R = k, \mathbf{X})$ to be shared but do not require α_k to be the same across $\mathcal{T}$. Under possible model mis-specifications, we allow either (i) the outcome models in (2.3) to be correctly specified with $\beta_{a,k} = \beta_a$, or (ii) the PS models in (2.2) to be correctly specified, for $k \in \mathcal{T}$.

Since the distribution of the covariates $\mathbf{X}$ can be heterogeneous across sites, we characterize the difference in covariate distributions between a target site $k_t \in \mathcal{T}$ and a source site $k_s \in \mathcal{S}$ through a density ratio

$$\omega_{k_t,k_s}(\mathbf{x}) = \frac{f(\mathbf{X} \mid R = k_t)}{f(\mathbf{X} \mid R = k_s)} = \frac{\mathbb{P}(R = k_t \mid \mathbf{X} = \mathbf{x})\mathbb{P}(R = k_s)}{\mathbb{P}(R = k_s \mid \mathbf{X} = \mathbf{x})\mathbb{P}(R = k_t)}.$$

We choose flexible semiparametric models for the density ratio

$$\omega_{k_t,k_s}(\mathbf{X};\gamma_{k_t,k_s}) = \exp\{\gamma_{k_t,k_s}^{\mathsf{T}}\psi(\mathbf{X})\}, \tag{2.4}$$

where $\psi : \mathbb{R}^p \mapsto \mathbb{R}^q$ is a vector-valued basis function with an intercept term. One may specify a range of basis functions to capture potential non-linearity in the density ratio model to improve the robustness of the estimation for $\omega_{k_t,k_s}(\mathbf{x})$.

Remark 2 *The exponential tilt density ratio model* (2.4) *is widely used to account for heterogeneity between two distributions*[92,93,32]*. By including higher-order terms of* $\mathbf{x}$ *in* $\psi(\mathbf{x})$*, higher-order differences such as variance and skewness can be captured. We propose in Section 2.3 a communication-efficient approach to estimate* γ_{k_t,k_s} *in covariate distributions between a target site and source site without sharing individual-level data. In the simulation study and real-data example, we have selected the exponential tilt model with* $\psi(\mathbf{x}) = \mathbf{x}$*, which recovers the whole class of natural exponential family distributions, including the normal distribution with mean shift, Bernoulli distribution for binary covariates, etc.*

2.3 Method

In this section, we detail the FACE method. We start with an overview of its main workflow, where a schematic illustration can be found in Figure S1 of the Supplementary Materials. In step 1, each target site calculates summary statistics of its covariate distribution, $\bar{\psi}_k = n_k^{-1}\sum_{i\in\mathcal{I}_k}\psi(\mathbf{X}_i)$ for $k \in \mathcal{T}$, a key quantity for estimating the density ratio model to balance covariate distributions, and broadcasts them to all source sites, along with its OR parameters $\{\hat{\beta}_{a,k}, a = 0, 1\}$. Each target site also constructs a doubly robust estimator[7] for its site-specific ATE, obtains additional summary statistics needed for the adaptive aggregation, and shares them with the leading analysis center (AC) (see Section 2.3.1). In Step 2, each source site uses the summary statistics of the target site ($\bar{\psi}_k$ from $k \in \mathcal{T}$) to fit its density

ratio model and construct an augmentation term $\widehat{\delta}_{\mathcal{T},k}$ for $k \in \mathcal{S}$ for the TATE. Each source site shares the augmentation term, together with additional summary statistics needed for the aggregation, to the AC (see Section 2.3.2). In Step 3, the AC performs the aggregation with estimators and parameters from Steps 1 and 2 to obtain the final FACE estimator, $\widehat{\Delta}_{\mathcal{T},\text{FACE}}$ (see Section 2.3.3). Overall, each site is only required to share information one time with other sites.

We detail each step of FACE in Sections 2.3.1-2.3.3 with generic models. Each site will need to fit both the OR models and the PS model using its own local data. Standard regression models such as logistic regression and generalized linear models can be used. Non-linear basis functions can be included to incorporate non-linear effects. For $k \in [J+K]$, we denote the estimated PS as $\pi_k(a, \mathbf{X}; \hat{\alpha}_k)$ and the predicted outcome for treatment a as $m(a, \mathbf{X}; \hat{\beta}_a)$, where $\hat{\alpha}_k$ and $\hat{\beta}_a$ can be achieved via classical estimation methods such as maximum likelihood estimation or estimating equations. An example with logistic regression models is given in Section 2.3.5.

2.3.1 Step 1: Estimation Using Target Data

The initial doubly robust TATE estimator is obtained from the site-specific ATE of the target sites. Within target sites, we compute the doubly robust TATE[7], $\widehat{\Delta}_{\mathcal{T},k} = \hat{M}_k + \widehat{\delta}_{\mathcal{T},k}$, where

$$\hat{M}_k = n_k^{-1} \sum_{i \in \mathcal{I}_k} \left\{ m(1, \mathbf{X}_i; \hat{\beta}_{1,k}) - m(0, \mathbf{X}_i; \hat{\beta}_{0,k}) \right\} \text{ for } k \in \mathcal{T}$$

is the OR model based estimate of the TATE, and

$$\widehat{\delta}_{\mathcal{T},k} = n_k^{-1} \sum_{i \in \mathcal{I}_k} \frac{(-1)^{1-A_i}}{\pi_k(A_i, \mathbf{X}; \hat{\alpha}_k)} \{ Y_i - m(A_i, \mathbf{X}_i; \hat{\beta}_{A_i,k}) \} \text{ for } k \in \mathcal{T}, \tag{2.5}$$

is the augmentation term that guards against mis-specification of the OR model. In addition, we calculate summary statistics for the $k \in \mathcal{T}$ target site covariate distribution, $\bar{\psi}_k = n_k^{-1} \sum_{i \in \mathcal{I}_k} \psi(\mathbf{X}_i)$.

The AC can construct the initial TATE estimate,

$$\widehat{\Delta}_{\mathcal{T},\mathcal{T}} = N_{\mathcal{T}}^{-1} \sum_{k \in \mathcal{T}} n_k \widehat{\Delta}_{\mathcal{T},k},$$

with summary data from target sites, $\{\widehat{\Delta}_{\mathcal{T},k}, n_k : k \in \mathcal{T}\}$. The consistency of $\widehat{\Delta}_{\mathcal{T},\mathcal{T}}$ is ensured when either the PS or OR is consistently estimated for each $k \in \mathcal{T}$.

Remark 3 *Here, we estimate β_a in each target site $k \in \mathcal{T}$ as $\hat{\beta}_{a,k}$. Alternatively, one could estimate β_a jointly at the cost of one additional round of communication between target sites. A jointly estimated β_a could benefit from efficiency gain under certain model specification conditions. Previous literature have developed distributed methods for aggregating estimates of β_a [19,57,31]. In practice, one should balance the advantage of potential efficiency gain with the cost of an additional cross-site communication.*

To facilitate optimal aggregation, we also share the estimators for the variance-covariance of scaled estimators, $\sqrt{n_k}(\hat{M}_k, \widehat{\delta}_{\mathcal{T},k}, \bar{\psi}_k, \hat{\beta}_{1,k}, \hat{\beta}_{0,k})$, which we denote as $\hat{\Sigma}_k$ for the target sites $k \in \mathcal{T}$. Variance estimation $\hat{\Sigma}_k$ for $k \in \mathcal{T}$ can be conducted through classical influence functions or bootstrapping within site. The exact role of the matrix in the aggregation will be unveiled after introducing the optimal combination weights in (2.8), which is the centerpiece of the adaptive aggregation step.

2.3.2 Step 2: Estimation Using Source Data

To safely use source data to assist in estimating $\Delta_{\mathcal{T}}$, we further account for the covariate shifts between the source sites and the target sites by tilting the source sites to the target population through the density ratios $\omega_{k_t,k_s}(\mathbf{X}; \gamma_{k_t,k_s})$. If individual-level data could be shared, estimating $\hat{\gamma}_{k_t,k_s}$ could be achieved by constructing a pseudo-likelihood function as in [92]. However, such an estimator cannot be directly obtained in a federated data setting. Instead, we propose a simple estimating equation approach that can be calculated in each source site $k_s \in \mathcal{S}$ using its data, along with summary statistics $\bar{\psi}_{k_t}$ obtained

from the target sites $k_t \in \mathcal{T}$. Specifically, we estimate γ_{k_t,k_s} as

$$\hat{\gamma}_{k_t,k_s} : \text{ solution to } n_{k_s}^{-1} \sum_{i \in \mathcal{I}_{k_s}} \omega_{k_t,k_s}\left(\psi(\mathbf{X}_i); \gamma_{k_t,k_s}\right) \psi(\mathbf{X}_i) = \bar{\psi}_{k_t}. \tag{2.6}$$

Remark 4 *Our approach is related to recent work that adjusts for observed differences in covariate distributions between a target population and the population that actually receives treatments*[50,114]*.*[50] *construct minimax linear weights that achieve approximate sample balance as in (6) uniformly over an absolutely convex class $\mathcal{M}$. They show that when $\mathcal{M}$ is selected appropriately, the solution to (6) converges in empirical mean square to the functional's Riesz representer, i.e., the unique square-integrable function that satisfies the corresponding population balance condition for all square-integrable functions*[51]*. Relatedly,*[114] *propose regularized calibrated estimators in the high-dimensional setting under minimal sparsity assumptions.*

For each source site, we construct a site augmentation term similar to the augmentation term in (2.5) for the target sites but with an additional density ratio weight

$$\widehat{\delta}_{\mathcal{T},k_s} = n_{k_s}^{-1} \sum_{i \in \mathcal{I}_{k_s}} \omega_{k_t,k_s}(\mathbf{X}_i; \hat{\gamma}_{k_t,k_s}) \frac{(-1)^{1-A_i}}{\pi_k(A_i, \mathbf{X}_i; \hat{\alpha}_k)} \{Y_i - m(A_i, \mathbf{X}_i; \hat{\beta}_{A_i,k_t})\} \text{ for } k_s \in \mathcal{S}.$$

We use the OR estimates from target sites $\hat{\beta}_{A_i,k_t}$ to ensure robustness when the OR is mis-specified. See Remark 5 for details.

Then, the site-specific augmentation terms $\widehat{\delta}_{\mathcal{T},k_s}$ are shared back to the AC, together with (i) $\hat{\sigma}^2_{k_s}$, an estimate for the scaled conditional variance $n_{k_s}\text{Var}\left(\widehat{\delta}_{\mathcal{T},k_s} \mid \mathscr{D}_{\mathcal{T}}\right)$, and (ii) $\hat{\mathbf{d}}_{k_t,k_s}$, an estimate for the partial derivatives of $\widehat{\delta}_{\mathcal{T},k_s}$ with respect to $\bar{\psi}_{k_t}, \hat{\beta}_{1,k_t}$, and $\hat{\beta}_{0,k_t}$. The role of $\hat{\mathbf{d}}_{k_t,k_s}$ in the aggregation will be explained in (2.8). Both $\hat{\sigma}^2_{k_s}$ and $\hat{\mathbf{d}}_{k_t,k_s}$ can be constructed from classical influence functions. Alternatively, $\hat{\sigma}^2_{k_s}$ can be estimated by bootstrapping within site and $\hat{\mathbf{d}}_{k_t,k_s}$ can be estimated by numerical derivatives.

Remark 5 *Combining the source site augmentation term $\widehat{\delta}_{\mathcal{T},k_s}$ with the initial TATE OR estimator from the target sites $\hat{M}_{\mathcal{T}}$, we obtain the $k_s \in \mathcal{S}$ source site estimators $\widehat{\Delta}_{\mathcal{T},k_s} = \hat{M}_{\mathcal{T}} + \widehat{\delta}_{\mathcal{T},k_s}$ as*

$$\widehat{\Delta}_{\mathcal{T},k_s} = N_{\mathcal{T}}^{-1} \sum_{k_t \in \mathcal{T}} n_{k_t} \Bigg(n_{k_t}^{-1} \sum_{i \in \mathcal{I}_{k_t}} \{ m(1, \mathbf{X}_i; \hat{\beta}_{1,k_t}) - m(0, \mathbf{X}_i; \hat{\beta}_{0,k_t}) \} + n_{k_s}^{-1} \sum_{i \in \mathcal{I}_{k_s}} \omega_{k_t,k_s}(\mathbf{X}_i; \hat{\gamma}_{k_t,k_s}) \frac{(-1)^{1-A_i}}{\pi_{k_s}(A_i, \mathbf{X}_i; \hat{\alpha}_{k_s})} \{ Y_i - m(A_i, \mathbf{X}_i; \hat{\beta}_{A_i,k_t}) \} \Bigg).$$

When the underlying OR model in the $k_s \in \mathcal{S}$ source site is the same as in the target population, the estimator $\widehat{\Delta}_{\mathcal{T},k_s}$ is doubly robust in the following sense: either (i) the OR model is consistent for all $k \in \{\mathcal{T}, k_s\}$, or (ii) the PS and density ratio models are consistent for the source site. Shifts in covariate distributions may induce heterogeneity in OR estimates across sites under mis-specified OR models, even if the conditional distribution $Y \mid A, \mathbf{X}$ is shared. To achieve robustness toward mis-specified OR, it is important to use the same $\hat{\beta}_{a,k_t}$ for $\hat{M}_{\mathcal{T}}$ and $\widehat{\delta}_{\mathcal{T},k_s}$ so that we may rely on the correct PS and density ratio models for consistency according to the alternative representation

$$\begin{aligned} N_{\mathcal{T}}^{-1} \sum_{k_t \in \mathcal{T}} n_{k_t} \Bigg\{ & n_{k_s}^{-1} \sum_{i \in \mathcal{I}_{k_s}} \omega_{k_t,k_s}(\mathbf{X}_i; \hat{\gamma}_{k_t,k_s}) \frac{(-1)^{1-A_i}}{\pi_{k_s}(A_i, \mathbf{X}_i; \hat{\alpha}_{k_s})} Y_i \\ & + n_{k_t}^{-1} \sum_{i \in \mathcal{I}_{k_t}} m(1, \mathbf{X}_i; \hat{\beta}_{1,k_t}) - n_{k_s}^{-1} \sum_{i \in \mathcal{I}_{k_s}} \omega_{k_t,k_s}(\mathbf{X}_i; \hat{\gamma}_{k_t,k_s}) \frac{A_i}{\pi_{k_s}(1, \mathbf{X}_i; \hat{\alpha}_{k_s})} m(1, \mathbf{X}_i; \hat{\beta}_{1,k_t}) \\ & - n_{k_t}^{-1} \sum_{i \in \mathcal{I}_{k_t}} m(0, \mathbf{X}_i; \hat{\beta}_{0,k_t}) + n_{k_s}^{-1} \sum_{i \in \mathcal{I}_{k_s}} \omega_{k_t,k_s}(\mathbf{X}_i; \hat{\gamma}_{k_t,k_s}) \frac{1 - A_i}{\pi_{k_s}(0, \mathbf{X}_i; \hat{\alpha}_{k_s})} m(0, \mathbf{X}_i; \hat{\beta}_{0,k_t}) \Bigg\}. \end{aligned}$$

To protect against negative transfer from source sites with biased TATE estimators, we combine information from each source site with the target sites through our adaptive aggregation step in Section 2.3.3.

2.3.3 Step 3: Adaptive Aggregation

In the final step, we obtain our FACE estimator by adaptively aggregating the initial TATE estimator $\widehat{\Delta}_{\mathcal{T},\mathcal{T}}$ and the source site estimators $\widehat{\Delta}_{\mathcal{T},k_s}$. Denote $\widehat{\delta}_{\mathcal{T},\mathcal{T}} = N_{\mathcal{T}}^{-1}\sum_{k\in\mathcal{T}} n_k \widehat{\delta}_{\mathcal{T},k}$. The AC can estimate $\Delta_{\mathcal{T}}$ by taking a linear combination of the initial TATE estimator $\widehat{\Delta}_{\mathcal{T},\mathcal{T}}$ and the source site estimators $\widehat{\Delta}_{\mathcal{T},k_s}$, where the weights are estimated to make an optimal bias-variance tradeoff. Indeed, the proposed FACE estimator is an "anchor and augmentation" estimator"

$$\hat{\Delta}_{\mathcal{T},\mathsf{FACE}} = \widehat{\Delta}_{\mathcal{T},\mathcal{T}} + \sum_{k_s\in\mathcal{S}} \eta_{k_s}\{\widehat{\Delta}_{\mathcal{T},k_s} - \widehat{\Delta}_{\mathcal{T},\mathcal{T}}\} = \widehat{\Delta}_{\mathcal{T},\mathcal{T}} + \sum_{k_s\in\mathcal{S}} \eta_{k_s}\{\widehat{\delta}_{\mathcal{T},k_s} - \widehat{\delta}_{\mathcal{T},\mathcal{T}}\},$$

which anchors on the initial TATE estimator $\widehat{\Delta}_{\mathcal{T},\mathcal{T}}$ and is augmented with source site estimators $\widehat{\Delta}_{\mathcal{T},k_s}$, with the weights $\{\eta_{k_s}, k_s \in \mathcal{S}\}$ to be estimated in a data-adaptive fashion to filter out potentially biased source site estimators. The second expression of $\hat{\Delta}_{\mathcal{T},\mathsf{FACE}}$ in (2.3.3) shows how the parameters from Steps 1 and 2 are used in the construction of the FACE estimator.

Moreover, the aggregation of the remaining unbiased source site augmentation terms should also minimize the estimation variance. Under the federated learning setting, the key to evaluate the variance of (2.3.3) is to decompose it into contributions from separate sites so that they can be estimated within each site. For any subset of $\mathcal{S}$, $\mathcal{S}' \subseteq \mathcal{S}$, we consider the following decomposition

$$\begin{aligned}
&\mathrm{Var}\left\{\widehat{\Delta}_{\mathcal{T},\mathcal{T}} + \sum_{k_s\in\mathcal{S}'} \eta_{k_s}(\widehat{\Delta}_{\mathcal{T},k_s} - \widehat{\Delta}_{\mathcal{T},\mathcal{T}})\right\} \\
&\approx \sum_{k_s\in\mathcal{S}'} \eta_{k_s}^2 \mathrm{Var}\left(\widehat{\delta}_{\mathcal{T},k_s} \mid \mathscr{D}_{\mathcal{T}}\right) \\
&+ \sum_{k_t\in\mathcal{T}} \mathrm{Var}\left\{\left(\frac{n_{k_t}}{N_{\mathcal{T}}}, \frac{n_{k_t} - n_{k_t}\sum_{k_s\in\mathcal{S}'}\eta_{k_s}}{N_{\mathcal{T}}}, \sum_{k_s\in\mathcal{S}'}\eta_{k_s}\mathbf{d}_{k_t,k_s}^{\mathsf{T}}\right)(\hat{M}_{\mathcal{T}}, \widehat{\delta}_{k_t}, \bar{\psi}_{k_t}^{\mathsf{T}}, \hat{\beta}_{1,k_t}^{\mathsf{T}}, \hat{\beta}_{0,k_t}^{\mathsf{T}})^{\mathsf{T}}\right\}, \quad (2.7)
\end{aligned}$$

where $\mathbf{d}_{k_t,k_s}$ is the limit for $\hat{\mathbf{d}}_{k_t,k_s}$, which is the partial derivative of $\widehat{\delta}_{\mathcal{T},k_s}$ with respect to the broadcast

estimators $\bar{\psi}_{k_t}, \hat{\beta}_{1,k_t}$ and $\hat{\beta}_{0,k_t}$. We decouple the dependence of the source site augmentation terms $\widehat{\delta}_{\mathcal{T},k_s}$ on the target sites by subtracting the first order approximation of the dependence $\left(\bar{\psi}_{k_t}^{\mathsf{T}}, \hat{\beta}_{1,k_t}^{\mathsf{T}}, \hat{\beta}_{0,k_t}^{\mathsf{T}}\right) \mathbf{d}_{k_t,k_s}$. The resulting $\widehat{\delta}_{\mathcal{T},k_s} - \mathbf{d}_{k_t,k_s}^{\mathsf{T}} \bar{\psi}_{k_t}$ is asymptotically independent of the target sites.

Since including information from source sites $\mathcal{S} \setminus \mathcal{S}^*$ may lead to biases, we adopt an adaptive combination strategy similar to the one given in [20] for combining data from a randomized trial and an observation study. Here, we overcome the additional challenge of data sharing constraints, and we propose the following adaptive L_1 penalized optimal aggregation

$$\hat{\eta} = \underset{\eta \in \mathbb{R}^K}{\arg\min}\, N \left[\sum_{k_s \in \mathcal{S}} \eta_{k_s}^2 \frac{\hat{\sigma}_{k_s}^2}{n_{k_s}} + \sum_{k_t \in \mathcal{T}} \hat{\mathbf{h}}_{k_t}(\eta)^{\mathsf{T}} \frac{\hat{\Sigma}_{k_t}}{n_{k_t}} \hat{\mathbf{h}}_{k_t}(\eta) \right] + \lambda \sum_{k_s \in \mathcal{S}} |\eta_{k_s}| \left(\widehat{\delta}_{\mathcal{T},k_s} - \widehat{\delta}_{\mathcal{T},\mathcal{T}} \right)^2, \tag{2.8}$$

where

$$\hat{\mathbf{h}}_{k_t}(\eta) = \left(\frac{n_{k_t}}{N_{\mathcal{T}}}, \frac{n_{k_t} - n_{k_t} \sum_{k_s \in \mathcal{S}^*} \eta_{k_s}}{N_{\mathcal{T}}}, \sum_{k_s \in \mathcal{S}} \eta_{k_s} \hat{\mathbf{d}}_{k_t,k_s}^{\mathsf{T}} \right)^{\mathsf{T}},$$

with $\hat{\Sigma}_{k_t}$ estimated from Step 1 and $\hat{\sigma}_{k_s}^2$ and $\hat{\mathbf{d}}_{k_t,k_s}$ estimated from Step 2. The multiplicative N factor is required to stabilize the loss. Choosing $\lambda \asymp N^{\nu}$ with $\nu \in (0, 1/2)$, we achieve the following oracle property for selection and aggregation: (i) biased source site augmentation terms have zero weights with high probability; (ii) regularization on the weights for unbiased source site augmentation terms is asymptotically negligible ($\ll N^{-1/2}$). Analogous to the phenomenon in meta-analysis, the estimation uncertainty of $\hat{\eta}$ has no asymptotic effect on the aggregated estimator.

Using the variance estimator (stabilized by "N" factor likewise)

$$\hat{\mathcal{V}} = N \left\{ \sum_{k_s \in \mathcal{S}} \hat{\eta}_{k_s} \frac{\hat{\sigma}_{k_s}^2}{n_{k_s}} + \sum_{k_t \in \mathcal{T}} \hat{\mathbf{h}}_{k_t}(\hat{\eta})^{\mathsf{T}} \frac{\hat{\Sigma}_{k_t}}{n_{k_t}} \hat{\mathbf{h}}_{k_t}(\hat{\eta}) \right\} \tag{2.9}$$

and the $1-\alpha/2$ quantile for the standard normal distribution $\mathcal{Z}_{\alpha/2}$, we construct the $(1-\alpha) \times 100\%$

confidence interval

$$\hat{\mathcal{C}}_\alpha = \left[\hat{\Delta}_{\mathcal{T},\mathsf{FACE}} - \sqrt{\hat{\mathcal{V}}/N}\mathcal{Z}_{\alpha/2}, \hat{\Delta}_{\mathcal{T},\mathsf{FACE}} + \sqrt{\hat{\mathcal{V}}/N}\mathcal{Z}_{\alpha/2}\right]. \tag{2.10}$$

The full FACE workflow is summarized in Algorithm 1.

Algorithm 1: FACE under generic model specifications

Data: J target sites $k_t \in \mathcal{T}$, K source sites $k_s \in \mathcal{S}$, and a Leading AC

for *Target* $k_t \in \mathcal{T}$ **do**

  Estimate $\alpha_{k_t}, \beta_{a,k_t}$ to calculate the initial TATE $\widehat{\Delta}_{\mathcal{T},k_t}$ its augmentation $\widehat{\delta}_{\mathcal{T},k_t}$ and the variance estimator $\hat{\Sigma}_{k_t}$ and transfer to the leading AC. Calculate $\bar{\psi}_{k_t}$ and broadcast to source sites along with $\hat{\beta}_{a,k_t}$.

end

for *Source sites* $k_s \in \mathcal{S}$ **do**

  Estimate γ_{k_t,k_s} and α_{k_s} to calculate the site-specific augmentation $\widehat{\delta}_{\mathcal{T},k_s}$ and transfer to the leading AC. Calculate $\hat{\sigma}^2_{k_s}$, $\hat{\mathbf{d}}_{k_t,k_s}$ and transfer to the leading AC.

end

for *Leading AC* **do**

  Estimate η by solving the penalized regression in (2.8). Construct the final global estimator as $\hat{\Delta}_{\mathcal{T},\mathsf{FACE}}$ by (2.3.3). Calculate the global estimator variance by (2.9) and construct 95% CI.

end

Result: Global TATE estimate, $\hat{\Delta}_{\mathcal{T},\mathsf{FACE}}$ and 95% CI

Remark 6 *Our aggregation procedure is communication-efficient and privacy-protected, whereas aggregation procedures given in the current literature such as those in*[20] *require sharing individual-level influence functions. Equation* (2.8) *is constructed using summary statistics, which provides a federated learning solution when individual-level data sharing is forbidden.*

2.3.4 Tuning Parameters

To choose an optimal tuning parameter λ, we propose a sample splitting approach that does not require sharing individual-level data. In each site, the data is first split into training and validation datasets, keeping the same proportion within each site. In the training datasets, Algorithm 1 is implemented to obtain the summary statistics ($\hat{\Sigma}_{k_t}$, $\hat{\mathbf{d}}_{k_s}$, $\hat{\sigma}^2_{k_s}$, $\widehat{\delta}_{\mathcal{T},k_s}$, and $\widehat{\delta}_{\mathcal{T},\mathcal{T}}$) needed for Equation (2.8). The AC selects a grid of λ values and calculates $\hat{\eta}(\lambda)$ by solving the penalized regression in (2.8). The upper and lower bounds on the grid of λ values can be left unrestricted; in practice, we have found that searching between 0.01 to 100 to be sufficiently large to provide good finite sample performance. In parallel, the validation datasets are used to obtain summary statistics denoted by ($\tilde{\Sigma}_{k_t}$, $\tilde{\mathbf{d}}_{k_t k_s}$, $\tilde{\sigma}^2_{k_s}$, $\tilde{\delta}_{\mathcal{T},k_s}$ and $\tilde{\delta}_{\mathcal{T},\mathcal{T}}$). These summary statistics are calculated using the validation datasets and plugging in the parameters estimated from the corresponding training datasets. The AC sets the value of the optimal tuning parameter, λ_{opt}, to be the value corresponding to the $\hat{\eta}$ that minimizes $Q(\hat{\eta})$ in the validation datasets, defined as

$$Q(\hat{\eta}) = N^V \left[\sum_{k_s \in \mathcal{S}} \hat{\eta}^2_{k_s} \frac{\tilde{\sigma}^2_{k_s}}{n^V_{k_s}} + \sum_{k_t \in \mathcal{T}} \tilde{\mathbf{h}}_{k_t}(\hat{\eta})^\intercal \frac{\tilde{\Sigma}_{k_t}}{n^V_{k_t}} \tilde{\mathbf{h}}_{k_t}(\hat{\eta}) \right],$$

where N^V, $n^V_{k_s}$, and $n^V_{k_t}$ are the sample sizes for validation data from all sites, source site k_s, and target site k_t, respectively.

2.3.5 FACE Under Logistic Regression Models

As an example, we illustrate FACE under logistic regression models with Y being binary, $J + K = 5$ total sites and $\mathcal{T} = \{1\}$ as the target site. For notational ease, let $\mathbf{X}$ be the vector of covariates with an intercept term. We fit logistic regression models with link $g(x) = 1/(1 + e^{-x})$ and loss $\ell(y, x) = \log(1 + e^x) - yx$ for all PS and OR models. For simplicity, we let $\psi(\mathbf{X}) = \mathbf{X}$.

In Step 1, we calculate the mean covariate vector in the target site $k_t = 1$ as $\bar{\psi}_{\mathcal{T}} = \frac{1}{n_1}\sum_{i\in\mathcal{I}_1}\mathbf{X}_i$ and transfer it to sites 2 through 5. Then, we estimate the models for $k_t = 1$

$$\hat{\alpha}_1 = \underset{\alpha\in\mathbb{R}^{p+1}}{\arg\min}\,\frac{1}{n_1}\sum_{i\in\mathcal{I}_1}\ell(A_i,\alpha^\top\mathbf{X}_i),\ \hat{\beta}_{a,1} = \underset{\beta\in\mathbb{R}^{p+1}}{\arg\min}\,\frac{1}{n_1}\sum_{i\in\mathcal{I}_1}I(A_i=a)\ell(Y_i,\alpha^\top\mathbf{X}_i).$$

Using the estimated models, we obtain the initial estimator and its augmentation term

$$\begin{aligned}\hat{M}_{\mathcal{T}} &= \frac{1}{n_1}\sum_{i\in\mathcal{I}_1}\left\{g\left(\hat{\beta}_{1,1}^\top\mathbf{X}_i\right) - g\left(\hat{\beta}_{0,1}^\top\mathbf{X}_i\right)\right\},\\ \widehat{\delta}_{\mathcal{T},\mathcal{T}} &= \frac{1}{n_1}\sum_{i\in\mathcal{I}_1}\left[\frac{A_i}{g\left(\hat{\alpha}_1^\top\mathbf{X}_i\right)}\left\{Y_i - g\left(\hat{\beta}_{1,1}^\top\mathbf{X}_i\right)\right\} - \frac{1-A_i}{g\left(-\hat{\alpha}_1^\top\mathbf{X}_i\right)}\left\{Y_i - g\left(\hat{\beta}_{0,1}^\top\mathbf{X}_i\right)\right\}\right]\end{aligned}$$

and $\widehat{\Delta}_{\mathcal{T},\mathcal{T}} = \hat{M}_{\mathcal{T}} + \widehat{\delta}_{\mathcal{T},\mathcal{T}}$. The variance covariance matrix estimator $\hat{\Sigma}_1$ can be calculated as $\hat{\Sigma}_1 = n_1^{-1}\sum_{i\in\mathcal{I}_1}\hat{\mathbf{U}}_i\hat{\mathbf{U}}_i^\top$ through the estimated influence functions, where $\hat{\mathbf{U}}_i = (\hat{\zeta}_i, \hat{\xi}_i, \psi(\mathbf{X}_i)^\top, \hat{\upsilon}_{1,i}, \hat{\upsilon}_{0,i})^\top$, and the exact form of $\hat{\xi}_{i,1}, \hat{\zeta}_i$ and $\hat{\upsilon}_{a,i}$ are given in the Supplement B.3.4.

In Step 2, we estimate the models for $k_s = 2,\ldots,5$

$$\hat{\alpha}_{k_s} = \underset{\alpha\in\mathbb{R}^{p+1}}{\arg\min}\,n_{k_t}^{-1}\sum_{i\in\mathcal{I}_{k_s}}\ell(A_i,\alpha^\top\mathbf{X}_i),\ \hat{\gamma}_{k_s} = \underset{\gamma\in\mathbb{R}^{p+1}}{\arg\min}\,n_{k_t}^{-1}\sum_{i\in\mathcal{I}_{k_s}}\exp(\gamma^\top\mathbf{X}_i) - \gamma^\top\bar{\psi}_{\mathcal{T}}.$$

Using the estimated models, we obtain the site-specific augmentations

$$\widehat{\delta}_{\mathcal{T},k_s} = n_{k_s}^{-1}\sum_{i\in\mathcal{I}_{k_s}}e^{\hat{\gamma}_{k_s}^\top\mathbf{X}_i}\left[\frac{A_i}{g\left(\hat{\alpha}_{k_s}^\top\mathbf{X}_i\right)}\left\{Y_i - g\left(\hat{\beta}_{1,1}^\top\mathbf{X}_i\right)\right\} - \frac{1-A_i}{g\left(-\hat{\alpha}_{k_s}^\top\mathbf{X}_i\right)}\left\{Y_i - g\left(\hat{\beta}_{0,1}^\top\mathbf{X}_i\right)\right\}\right].$$

along with the partial derivative of $\widehat{\delta}_{\mathcal{T},k_s}$ with respect to $\bar{\psi}_{\mathcal{T}}$, $\hat{\mathbf{d}}_{k_s} = (\hat{\mathbf{d}}_{k_s,\psi}^{\mathsf{T}}, \hat{\mathbf{d}}_{k_s,\beta_1}^{\mathsf{T}}, \hat{\mathbf{d}}_{k_s,\beta_0}^{\mathsf{T}})^{\mathsf{T}}$, as

$$\hat{\mathbf{d}}_{k_s,\psi} = -\left\{ n_{k_t}^{-1} \sum_{i\in\mathcal{I}_{k_s}} e^{\hat{\gamma}_{k_s}^{\mathsf{T}}\mathbf{X}_i} \mathbf{X}_i\mathbf{X}_i^{\mathsf{T}} \right\}^{-1} n_{k_t}^{-1} \sum_{i\in\mathcal{I}_{k_s}} e^{\hat{\gamma}_{k_s}^{\mathsf{T}}\mathbf{X}_i} \frac{(-1)^{1-A_i}}{g\left(\hat{\alpha}_{k_s}^{\mathsf{T}}\mathbf{X}_i\right)} \left\{ Y_i - g\left(\hat{\beta}_{A_i,k_s}^{\mathsf{T}}\mathbf{X}_i\right) \right\} \mathbf{X}_i,$$

$$\hat{\mathbf{d}}_{k_s,\beta_a} = (-1)^a n_{k_t}^{-1} \sum_{i\in\mathcal{I}_{k_s}} e^{\hat{\gamma}_{k_s}^{\mathsf{T}}\mathbf{X}_i} \frac{\mathrm{I}(A_i = a)}{g\left\{(-1)^{1-a}\hat{\alpha}_{k_s}^{\mathsf{T}}\mathbf{X}_i\right\}} g'\left(\hat{\beta}_{A_i,k_s}^{\mathsf{T}}\mathbf{X}_i\right)\mathbf{X}_i.$$

The variance estimator $\hat{\sigma}_{k_s}^2$ can be calculated as $\hat{\sigma}_{k_s}^2 = n_{k_t}^{-1}\sum_{i\in\mathcal{I}_{k_s}} \hat{\xi}_{i,k_s}^2$ through the estimated influence function, where the form of $\hat{\xi}_{i,k_s}$ is given in the Supplement B.3.4.

In Step 3, we use $\hat{\Sigma}_1$, $\hat{\mathbf{d}}_{k_s}$, $\hat{\sigma}_{k_s}^2$, $\widehat{\delta}_{\mathcal{T},k_s}$ and $\widehat{\delta}_{\mathcal{T},\mathcal{T}}$ to solve the adaptive selection and aggregation (2.8), which leads to $\hat{\Delta}_{\mathcal{T},\mathsf{FACE}}$ and the confidence interval $\hat{\mathcal{C}}_\alpha$.

2.4 Theoretical Guarantees

In this section, we provide the theoretical results for the FACE estimator. We start with a high-level theory for a generic choice of models in Section 2.4.1. Then, we discuss the efficiency gain from leveraging source sites in Section 2.4.2. We give in Section B.2 a detailed set of conditions corresponding to the realization of Section 2.3.5. In our asymptotic theory, N is allowed to grow but the distribution for $(Y, \mathbf{X}^{\mathsf{T}}, A, R)^{\mathsf{T}}$ and $J + K$ are fixed.

2.4.1 Theory for General FACE

To compress notation, we combine the broadcast parameters and their asymptotic limits as

$$\hat{\theta}_{k_t} = \left(\bar{\psi}_{k_t}^{\mathsf{T}}, \hat{\beta}_{1,k_t}^{\mathsf{T}}, \hat{\beta}_{0,k_t}^{\mathsf{T}}\right)^{\mathsf{T}}, \ \bar{\theta}_{k_t} = \left(\mathbb{E}\{\psi(\mathbf{X})^{\mathsf{T}} \mid R = k_t\}, \bar{\beta}_{1,k_t}^{\mathsf{T}}, \bar{\beta}_{0,k_t}^{\mathsf{T}}\right)^{\mathsf{T}}. \tag{2.11}$$

Regularity conditions are detailed in the following

Assumption 2 *For absolute constants $M, \varepsilon > 0$,*

(a) (Regularity of estimators) The estimators $\hat{M}_{\mathcal{T}}$, $\widehat{\delta}_{\mathcal{T},k_t}$, $\hat{\beta}_{a,k_t}$ and $\widehat{\delta}_{\mathcal{T},k_s}$ admit the following asymptotically linear representations

$$\begin{aligned}
\sqrt{N_{\mathcal{T}}}(\hat{M}_{\mathcal{T}} - \bar{M}_{\mathcal{T},\mathcal{T}}) &= \frac{1}{\sqrt{N_{\mathcal{T}}}} \sum_{k_t \in \mathcal{T}} \sum_{i \in \mathcal{I}_{k_t}} \zeta_i + o_p(1), \\
\sqrt{N_{\mathcal{T}}}(\widehat{\delta}_{\mathcal{T},\mathcal{T}} - \bar{\delta}_{\mathcal{T},\mathcal{T}}) &= \frac{1}{\sqrt{N_{\mathcal{T}}}} \sum_{k_t \in \mathcal{T}} \sum_{i \in \mathcal{I}_{k_t}} \xi_{i,\mathcal{T}} + o_p(1), \\
\sqrt{n_{k_s}}(\widehat{\delta}_{\mathcal{T},k_s} - \bar{\delta}_{\mathcal{T},k_s}) &= \frac{1}{\sqrt{n_{k_s}}} \sum_{i \in \mathcal{I}_{k_s}} \xi_{i,k_s} + \sqrt{n_{k_s}} \sum_{k_t \in \mathcal{T}} \bar{\mathbf{d}}_{k_t,k_s}^{\mathsf{T}} \left(\hat{\theta}_{k_t} - \bar{\theta}_{k_t} \right) + o_p(1), \\
\sqrt{n_{k_t}} \left(\hat{\beta}_{a,k_t} - \bar{\beta}_{a,k_t} \right) &= \frac{1}{\sqrt{n_{k_t}}} \sum_{i \in \mathcal{I}_{k_t}} v_{i,a} + o_p(1).
\end{aligned}$$

with bounded asymptotic limits $\bar{M}_{\mathcal{T},\mathcal{T}}$, $\bar{\delta}_{\mathcal{T},\mathcal{T}}$, $\bar{\delta}_{\mathcal{T},k_s}$, $\bar{\mathbf{d}}_{k_t,k_s}$ and iid mean zero random variables ζ_i, $\xi_{i,\mathcal{T}}$, ξ_{i,k_s}.

(b) (Compact support) The covariates $\mathbf{X}$ and their functions $\psi(\mathbf{X})$ in the density ratio are in compact sets $\mathbf{X} \in [-M, M]^p$ and $\psi(\mathbf{X}) \in [-M, M]^q$ almost surely.

(c) (Stable variance) The variance of ξ_{i,k_s} is in the set $[\varepsilon, M]$. The variance-covariance matrix

$$\Sigma_{k_t} = \mathrm{Var}\left\{ \left(\zeta_i, \xi_{i,\mathcal{T}}, \psi(\mathbf{X}_i)^{\mathsf{T}}, v_{i,1}^{\mathsf{T}}, v_{i,0}^{\mathsf{T}} \right)^{\mathsf{T}} \mid R = k_t \right\}$$

has eigenvalues all in $[\varepsilon, M]$ for some positive constant ε and M.

(d) (Regularity of auxiliary estimators) The estimators $\hat{\Sigma}_{k_t}$, $\hat{\sigma}_{k_s}^2$, $\hat{\mathbf{d}}_{k_s}$ are $\sqrt{N}$-consistent

$$\sum_{k_t \in \mathcal{T}} \left\| \hat{\Sigma}_{k_t} - \Sigma_{k_t} \right\| + \sum_{k_s \in \mathcal{S}} \left\{ \left| \hat{\sigma}_{k_s}^2 - \mathrm{Var}(\xi_{i,k_s} \mid R_i = k_s) \right| + \left\| \hat{\mathbf{d}}_{k_s} - \bar{\mathbf{d}}_{k_s} \right\| \right\} = O_p\left(N^{-1/2} \right).$$

(e) (Root-N rate consistency) For each target site $k_t \in \mathcal{T}$, at least one of the two models is correctly specified:

-i the PS model is consistently estimated:

$$\sup_{a=0,1} \sup_{\|\mathbf{x}\|_\infty \leq M} \sum_{k_t \in \mathcal{T}} |\mathbb{P}(A = a \mid \mathbf{X} = \mathbf{x}, R = k_t) - \pi_k(a, \mathbf{x}; \hat{\alpha}_{k_t})| = O_p\left(N^{-1/2}\right).$$

-ii the OR model is consistently estimated:

$$\sup_{a=0,1} \sup_{\|\mathbf{x}\|_\infty \leq M} \sum_{k_t \in \mathcal{T}} \left|\mathbb{E}(Y \mid A = a, \mathbf{X} = \mathbf{x}, R = k_t) - m_{k_t}(a, \mathbf{x}; \hat{\beta}_{a,k_t})\right| = O_p\left(N^{-1/2}\right).$$

Assumptions 2(a) and 2(e) are the typical regularity conditions under classical parametric models. They can be verified in two steps: 1) asymptotic normality of model estimators [119] and 2) local expansion of the estimators. Assumption 2(c) regulates the scale of variability of the data, which leads to a stable variance for $\hat{\Delta}_{\mathcal{T},\mathsf{FACE}}$. Assumption 2(e) ensures identification of the true TATE by anchoring on $\widehat{\Delta}_{\mathcal{T},\mathcal{T}}$[7]. Note that in the setting of multiple target sites, Assumption 2(e) allows for each target site to have different correct model specifications for either the OR model or the PS model.

We now state the theory for the general FACE estimation.

Theorem 1 *Under Assumptions 1 and 2, the FACE estimator is consistent and asymptotically normal with consistent variance estimation $\hat{\mathcal{V}}$,*

$$\sqrt{N/\hat{\mathcal{V}}}\left(\hat{\Delta}_{\mathcal{T},\mathsf{FACE}} - \Delta_{\mathcal{T}}\right) \rightsquigarrow \mathcal{N}(0,1).$$

We use $\rightsquigarrow$ for convergence in distribution.

Theorem 1 implies that (2.10) provides asymptotically honest coverage.

Corollary 1 *Under Assumptions 1 and 2, the coverage rate of the confidence interval* (2.10) *approaches the nominal level asymptotically*

$$\lim_{N\to\infty} \mathbb{P}\left(\Delta_{\mathcal{T}} \in \hat{\mathcal{C}}_\alpha\right) = 1-\alpha$$

A key step in the proof of Theorem 1 is the analysis of the L_1 penalized adaptive selection and aggregation (2.8). We are able to establish the oracle property[35], i.e., the data-driven selection and aggregation through (2.8) is asymptotically equivalent to the process with a priori selection and optimal aggregation. The problem is different from the typical penalized regression, so we develop a new proof strategy. We first analyze the optimal combination with oracle selection, in which the biased augmentations are excluded. For unbiased augmentations, $\widehat{\Delta}_{\mathcal{T},k_s} - \widehat{\Delta}_{\mathcal{T},\mathcal{T}} = O_p\left(N^{-1/2}\right)$, so the penalty term is asymptotically negligible, i.e., $\lambda(\widehat{\Delta}_{\mathcal{T},k_s} - \widehat{\Delta}_{\mathcal{T},\mathcal{T}})^2 = o_p\left(N^{-1/2}\right)$ when λ is chosen such that $\lambda \asymp N^{\nu}$ with $\nu \in (0, 1/2)$. Thus, the estimated combination converges to the asymptotic limit at the regular $N^{-1/2}$ rate. Finally, we show that the estimated combination with oracle selection also solves the original problem with high probability.

Remark 7 *For consistency of $\hat{\Delta}_{\mathcal{T},\mathsf{FACE}}$, we require that the PS or OR model is correct for the target sites but allow the models for the source sites and density ratio to be mis-specified. To meaningfully leverage information from source sites for the TATE, we would expect that many $k_s \in \mathcal{S}$ among the source sites (i) satisfy the ignorability condition 1(d) and (ii) either the OR model $m(a)$ is correct, or both the PS π_{k_s} and the density ratio ω_{k_t,k_s} models are correct. For source sites satisfying the conditions above, their site-specific augmentations are unbiased and thus contribute to the efficiency improvement of $\hat{\Delta}_{\mathcal{T},\mathsf{FACE}}$.*

2.4.2 Relative Efficiency

Notice that we recover the initial TATE estimator $\widehat{\Delta}_{\mathcal{T},\mathcal{T}}$ from (2.3.3) if $\hat{\eta} = 0$. Since we are minimizing the post-aggregation variance, the optimal solution must be no worse than any alternative

solutions. If there exists informative source sites in $\mathcal{S}'$, as defined in Assumption 3, improvement in the efficiency of FACE compared to the target only estimator is guaranteed.

Assumption 3 *For a nonempty set $\mathcal{S}' \subseteq \mathcal{S}$, one of the following holds*

(a) (i) Correct OR: the OR model is consistently estimated:

$$\sup_{a=0,1} \sup_{\|\mathbf{x}\|_\infty \le M} \sum_{k_t \in \mathcal{T}} \left| \mathbb{E}(Y \mid A = a, \mathbf{X} = \mathbf{x}, R = k_t) - m_{k_t}(a, \mathbf{x}; \hat{\beta}_{a,k_t}) \right| = O_p\left(N^{-1/2}\right);$$

(ii) Consistent weighting: the PS and density ratio models are consistently estimated:

$$\begin{aligned} \sup_{a=0,1} \sup_{\|\mathbf{x}\|_\infty \le M} & \sum_{k_s \in \mathcal{S}'} \left| \mathbb{P}(A = a \mid \mathbf{X} = \mathbf{x}, R = k_s) - \pi_{k_s}(a, \mathbf{x}; \hat{\alpha}_{k_s}) \right| \\ & + \sum_{k_t \in \mathcal{T}} \sum_{k_s \in \mathcal{S}'} \left| \frac{\mathbb{P}(R = k_t \mid \mathbf{X} = \mathbf{x}) \mathbb{P}(R = k_s)}{\mathbb{P}(R = k_s \mid \mathbf{X} = \mathbf{x}) \mathbb{P}(R = k_t)} - \omega_{k_t,k_s}(\mathbf{x}; \hat{\gamma}_{k_t,k_s}) \right| = O_p\left(N^{-1/2}\right). \end{aligned}$$

(b) Informative source: Let $\vartheta = (\psi(\mathbf{X})^\intercal, v_1^\intercal, v_0^\intercal)^\intercal$ be the combined influence function for broadcast estimators. For all $k_s \in \mathcal{S}'$

$$\left| Cov\left(\frac{\zeta + \xi_{\mathcal{T}}}{\mathbb{P}(R \in \mathcal{T})}, -\frac{\xi_{\mathcal{T}}}{\mathbb{P}(R \in \mathcal{T})} + \sum_{k_t \in \mathcal{T}} \frac{\mathrm{I}(R = k_t)}{\mathbb{P}(R = k_t)} \left(\psi(\mathbf{X})^\intercal, v_1^\intercal, v_0^\intercal\right) \bar{\mathbf{d}}_{k_t,k_s} \mid R \in \mathcal{T} \right) \right| \ge \varepsilon.$$

The two model consistency conditions in Assumption 3(a) ensure the consistency of the doubly robust estimator $\widehat{\Delta}_{\mathcal{T},k_s}$. Assumption 3(b) characterizes the informativeness of a source site k_s such that the updated direction $\left(\widehat{\delta}_{\mathcal{T},k_s} - \widehat{\delta}_{\mathcal{T},\mathcal{T}}\right)$ is correlated with the initial $\widehat{\Delta}_{\mathcal{T},\mathcal{T}}$. The covariance in the condition is likely to be negative with the opposite sign of $\xi_{\mathcal{T}}$.

Proposition 1 *Under the conditions of Theorem 1, the asymptotic variance of $\hat{\Delta}_{\mathcal{T},\mathsf{FACE}}$ is no larger than that of $\widehat{\Delta}_{\mathcal{T},\mathcal{T}}$. Moreover, if Assumption 3 holds, the asymptotic variance of $\hat{\Delta}_{\mathcal{T},\mathsf{FACE}}$ is strictly smaller than that of $\widehat{\Delta}_{\mathcal{T},\mathcal{T}}$.*

Proposition 1 offers a guarantee on the relative efficiency in general settings. As the exact efficiency gain may take different forms under general settings, we showcase the efficiency gain with a clear interpretation under a simple ideal setting. When models are correctly specified, we have an explicit form for the oracle optimal combination $\bar{\eta}$ and the improvement in estimation efficiency for the TATE.

Assumption 4 *The PS, OR, and density ratio models are consistently estimated at $\sqrt{N}$ rate:*

$$\sup_{a=0,1}\sup_{\|\mathbf{x}\|_\infty\le M}\sum_{k=1}^{K}\left|\mathbb{P}(A=a\mid\mathbf{X}=\mathbf{x},R=k)-\pi_k(a,\mathbf{x};\hat{\alpha}_k)\right| + \sum_{k_t\in\mathcal{T}}\left|\mathbb{E}(Y\mid A=a,\mathbf{X}=\mathbf{x},R=k_t)-m_{k_t}(a,\mathbf{x};\hat{\beta}_{a,k_t})\right| + \sum_{k_t\in\mathcal{T}}\sum_{k_s\in\mathcal{S}}\left|\frac{\mathbb{P}(R=k_t\mid\mathbf{X}=\mathbf{x})\mathbb{P}(R=k_s)}{\mathbb{P}(R=k_s\mid\mathbf{X}=\mathbf{x})\mathbb{P}(R=k_t)}-\omega_{k_t,k_s}(\mathbf{x};\hat{\gamma}_{k_t,k_s})\right| = O_p\left(N^{-1/2}\right).$$

Proposition 2 *Suppose $\mathcal{T}=\{1\}$ and $\mathcal{S}=\{2\}$. Denote*

$$\begin{aligned}
\mathcal{V}_m^2 &= \mathrm{Var}\left\{(-1)^{1-A}m(A,\mathbf{X};\bar{\beta}_a)-\Delta_{\mathcal{T}}\mid R=1\right\},\\
\mathcal{V}_{\mathcal{T}}^2 &= \mathrm{Var}\left[\frac{(-1)^{1-A}}{\pi(A,\mathbf{X};\bar{\alpha}_1)}\left\{Y-m(A,\mathbf{X};\bar{\beta}_a)\right\}\mid R=1\right],\\
\mathcal{V}_{\mathcal{S}}^2 &= \mathrm{Var}\left[\omega_{1,2}(\mathbf{X};\bar{\gamma}_{1,2})\frac{(-1)^{1-A}}{\pi(A,\mathbf{X};\bar{\alpha}_2)}\left\{Y-m(A,\mathbf{X};\bar{\beta}_a)\right\}\mid R=2\right].
\end{aligned} \tag{2.12}$$

Under Assumptions 1-4, the optimal combination asymptotically approaches

$$\bar{\eta}=\frac{n_{\mathcal{S}}\mathcal{V}_{\mathcal{T}}^2}{n_{\mathcal{S}}\mathcal{V}_{\mathcal{T}}^2+n_{\mathcal{T}}\mathcal{V}_{\mathcal{S}}^2}.$$

The efficiency of FACE relative to the initial TATE estimator is

$$1+\frac{\mathcal{V}_{\mathcal{T}}^4}{\mathcal{V}_m^2\mathcal{V}_{\mathcal{T}}^2+n_{\mathcal{T}}\left(\mathcal{V}_m^2+\mathcal{V}_{\mathcal{T}}^2\right)\mathcal{V}_{\mathcal{S}}^2/n_{\mathcal{S}}}.$$

Resulting from independence under the ideal setting, the weights $\{1-\bar{\eta}, \bar{\eta}\}$ coincide with the inverse variance weights for $\{\widehat{\delta}_{\mathcal{T},1}, \widehat{\delta}_{\mathcal{T},2}\}$. According to Proposition 2, the relative efficiency of FACE is monotone increasing in $n_{\mathcal{S}}/\mathcal{V}_{\mathcal{S}}^2$. When $n_{\mathcal{S}}$ increases, the relative efficiency approaches $1+\mathcal{V}_{\mathcal{T}}^2/\mathcal{V}_m^2$. In that case, the asymptotic variance of FACE approaches $\mathcal{V}_m^2/\mathbb{P}(R\in\mathcal{T})$, the asymptotic variance of $\hat{M}_{\mathcal{T}}$. Under the ideal setting, the two components in the initial TATE estimator, outcome regression $\hat{M}_{\mathcal{T}}$ and augmentation $\widehat{\delta}_{\mathcal{T},\mathcal{T}}$, are independent. The FACE estimator includes the source site data to improve the augmentation component, leading to a reduction in its asymptotic variance.

2.5 Simulation Studies

We study the finite sample performance of the FACE estimator and make comparisons with an estimator that leverages target data only and a sample-size adjusted estimator that does not adaptively weight different sites. In the simulation studies, we take the target population to be a single site. We examine the empirical bias, empirical standard error (ESE), average of the estimated standard error (ASE), and coverage probability (CP) of the 95% CI over 1,000 simulations. We vary the number of source sites $K\in\{5,10,50\}$, the true OR, PS, and density ratio models, and the number of source sites with correctly specified models.

2.5.1 Data Generation

To allow for heterogeneity in the covariate distribution between sites, the covariates in each site $\mathbf{X}_{kp}$ are generated from a skewed normal distribution, $\mathbf{X}_{kp}\sim\mathcal{SN}(\mathbf{x};\kappa_{kp},\varphi_{kp}^2,\nu_{kp})$, where $k=1,...,J+K$ indexes the sites and $p=1,...,10$ indexes the ten covariates, κ_{kp} is the location parameter, φ_{kp} is the scale parameter, and ν_{kp} is the skewness parameter. For all sites, we let $\kappa_{k\cdot}\in(0.10,0.15)$ and $\varphi_{k\cdot}=(1,...,1)$. For the target site, we set $\nu_{k\cdot}=0$. For the source sites, we let $\nu_{k\cdot}\in\{-0.25,0.25\}$. Under these settings, the exponential tilt model provides a good approximation quality for projecting

the source site covariate distribution to the target site. We fix the sample size in the target site and source sites to be $n_{k_t} = n_{k_s} = 200$.

The true potential outcomes are generated as

$$Y_k(a) = [(\mathbf{X}_k - \mu_1)^\intercal, (\mathbf{X}_k^{\circ 2})^\intercal](\beta_{1a}^\intercal, \beta_{2a}^\intercal)^\intercal + 3I(a = 1) + \varepsilon_k, \quad \varepsilon_k \sim \mathcal{N}(0, 1), \quad a = 0, 1,$$

where $\mathbf{X}_k^{\circ 2}$ denotes $\mathbf{X}_k$ squared element-wise, $\beta_{11} = (0.4, .., 1.2)$, and $\beta_{10} = (0.4, .., 1.2)$ with equally-spaced increments for a length 10.

The true PS model is generated as

$$A_k \mid \mathbf{X} = \mathbf{x} \sim \text{Bernoulli}(\pi_k), \quad \pi_k = \text{expit}(\mathbf{X}_k \alpha_{1k} + \mathbf{X}_k^{\circ 2} \alpha_{2k}),$$

where for the target site, $\alpha_{11} = (0.4, ..., -0.4)$, with equally-spaced decrements for a length 10 and $\alpha_{21} = 0$. For the source sites, $\alpha_{1k} = (0.5, ..., -0.5)$, with equally-spaced decrements for a length 10 and $\alpha_{2k} = 0$. For all sites, we fit linear regression models for the OR and logistic regression models for the PS, where we only include the linear terms of the covariates $\mathbf{X}_k$.

2.5.2 Simulation Settings

Since the specified OR and PS models do not include the quadratic terms, we consider a correct OR by setting $\beta_{21} = \beta_{20} = 0$; a correct PS by setting $\alpha_{2k} = 0$; a mis-specified OR by setting $\beta_{21} = (0.2, .., 0.4)$ and a mis-specified PS by setting $\alpha_{2k} = (0.12, ..., -0.12)$.

We consider the following settings. In Setting 1, we examine the scenario where both the OR and PS models are correctly specified for all sites. In Setting 2, we mis-specify the OR while keeping the PS correctly specified for all sites. In Setting 3, we mis-specify the PS and correctly specify the OR for all sites. In Setting 4, the OR and PS models are mis-specified for half of the source sites. To examine

the effect of increasing the number of mis-specified source sites, in Setting 5, the OR and PS models are mis-specified in all of the source sites.

In each setting, we choose the tuning parameter λ by the distributed cross validation procedure described in Section 2.3.4 from $\{0, 10^{-4}, 10^{-3}, 10^{-2}, 0.1, 0.25, 0.5, 1, 2, 5, 10\}$, where we split the simulated datasets in each site into two equally sized training and validation datasets.

2.5.3 Simulation Results

In Table 2.1, we summarize the bias, ESE, ASE, and CP of the 95% CI of a target-only estimator (Target), a sample-size weighted estimator (SS), and FACE over 1, 000 simulations across Settings 1-5. The results show that FACE performs well in all settings, with minimal bias, substantially reduced variance compared to the Target estimator, and nominal coverage. The SS estimator performs well in Settings 1-3 where each source site estimator is consistent, but performs poorly in Settings 4-5 when some or all of the source sites are biased for the TATE. On the other hand, FACE is able to data-adaptively drop source sites that display large bias. Even in Setting 5, when the OR and PS models are mis-specified in all of the source sites, FACE displays only minimal bias even when $K = 50$ and close to nominal coverage. Given that the sample size in each site is $n_{k_t} = n_{k_s} = 200$, $K = 50$ is a relatively large number of sites. Our theory requires K to be fixed, so bias can be introduced when K is large since the difference between the estimated and optimal weights grows with K. However, such bias reduces if we increase the sample size, which has been validated in an additional simulation with sample size increasing to 400.

Further, as displayed in Figure 2.1, FACE shows decreasing variance as the number of source sites K increases, showing the potential benefit of leveraging additional source sites. The precision gain holds across different model mis-specification scenarios (Settings 1-4).

Table 2.1: Bias, Empirical Standard Error (ESE), Average of the Estimated Standard Error (ASE), and Coverage Probability (CP) of the 95% CI of estimators over $1,000$ simulations in four model specification settings.

	Number of Source Sites											
	$K = 5$				$K = 10$				$K = 50$			
	Bias	ESE	ASE	CP	Bias	ESE	ASE	CP	Bias	ESE	ASE	CP
Setting 1												
Target	-0.01	0.79	0.79	0.95	0.00	0.78	0.79	0.96	-0.02	0.77	0.79	0.95
SS	0.05	0.54	0.55	0.95	0.01	0.40	0.40	0.95	0.01	0.29	0.29	0.95
FACE	0.01	0.56	0.54	0.95	0.05	0.50	0.48	0.96	0.05	0.45	0.44	0.96
Setting 2												
Target	-0.02	0.79	0.80	0.96	0.02	0.82	0.81	0.95	0.00	0.80	0.81	0.96
SS	-0.05	0.55	0.56	0.95	0.01	0.40	0.40	0.95	0.01	0.29	0.30	0.95
FACE	0.01	0.58	0.58	0.96	0.06	0.51	0.49	0.96	0.06	0.46	0.44	0.95
Setting 3												
Target	-0.04	0.78	0.78	0.94	-0.03	0.78	0.79	0.95	-0.03	0.80	0.79	0.95
SS	-0.08	0.58	0.58	0.95	-0.02	0.42	0.42	0.96	-0.02	0.31	0.31	0.94
FACE	0.00	0.56	0.56	0.95	0.01	0.50	0.50	0.96	0.02	0.46	0.44	0.95
Setting 4												
Target	-0.04	0.79	0.81	0.95	0.00	0.81	0.81	0.96	0.01	0.81	0.81	0.96
SS	0.76	0.22	0.22	0.15	0.85	0.15	0.14	0.07	0.87	0.11	0.11	0.00
FACE	0.01	0.60	0.59	0.96	0.05	0.52	0.51	0.96	0.06	0.48	0.45	0.96
Setting 5												
Target	-0.03	0.79	0.80	0.95	0.01	0.80	0.80	0.95	-0.01	0.81	0.81	0.96
SS	0.82	0.37	0.36	0.18	0.94	0.24	0.24	0.05	0.98	0.18	0.19	0.01
FACE	0.05	0.72	0.73	0.94	0.06	0.65	0.65	0.92	0.09	0.59	0.57	0.91

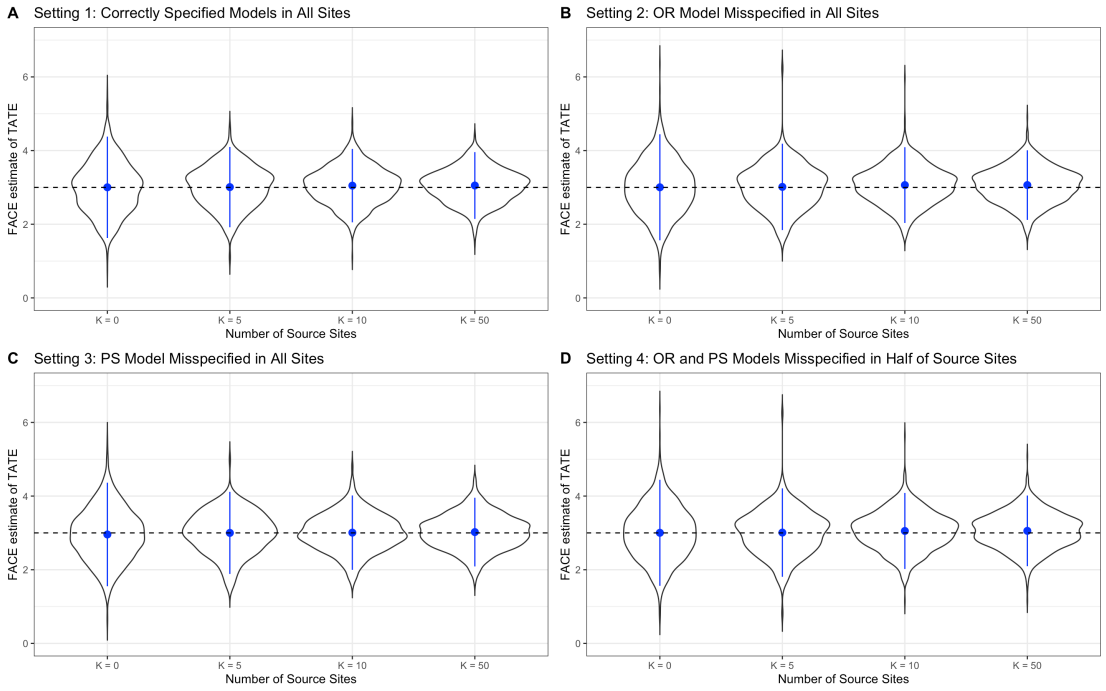


Figure 2.1: Simulated FACE estimates of the TATE across 1,000 simulations in Settings 1-4 with $K = 0, 5, 10, 50$. $K = 0$ corresponds to the Target only estimator. Blue dots (lines) are means (95% CIs). The dotted black line is the true TATE of 3.

2.6 Comparative Effectiveness of COVID-19 Vaccines

To illustrate FACE, we study the comparative effectiveness of BNT162b2 (Pfizer) versus mRNA-1273 (Moderna) for the prevention of COVID-19 outcomes in five VA sites. It is of interest to understand the real world effectiveness of these vaccines, but head-to-head comparisons have been rare. A recent emulated target trial using the EHRs of US veterans showed that the 24-week risk of COVID-19 outcomes was low for patients who received either vaccine, but lower for veterans assigned to Moderna compared to Pfizer[28]. Utilizing FACE, we examine the TATE in a federated data setting where the target population of interest is one of five sites (North Atlantic, Southwest, Midwest, Continental, or Pacific) in the VA healthcare system. Our problem is more challenging than that of[28] or[74] due to the federated data setting and the different target populations of interest that we are able to study.

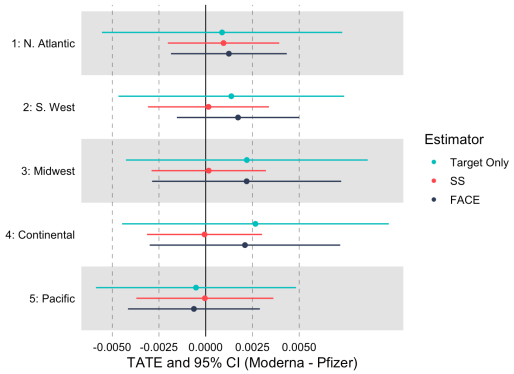
Inclusion criteria included veteran status, at least 18 years of age by January 1, 2021, no previously documented COVID-19 infection, no previous COVID-19 vaccination, and documented two-dose COVID-19 vaccination with either Pfizer or Moderna between January 1 and March 24, 2021. For each eligible veteran, follow-up began on the day that the second dose of vaccine was received (baseline) and ended on the day of death, 120 or 180 days after baseline, or the end of the study time period (September 24, 2021). The outcomes of interest were documented SARS-CoV-2 infection either 120 or 180 days after baseline and death with COVID-19 infection either 120 or 180 days after baseline.

Among the 608,359 eligible veterans, 293,137 (48.2%) received Pfizer and 315,222 (51.8%) received Moderna. Baseline characteristics among the two groups were similar within site. Across sites, there was heterogeneity in race (a larger proportion of Asians in the Pacific), and ethnicity (a larger Hispanic population in the Southwest and Pacific). Baseline characteristics in each of the five sites is summarized in Supplementary Tables 1 and 2. All models were adjusted for age, sex, race, ethnicity, residence, and important comorbidities: chronic lung disease (including asthma, bronchitis, and chronic obstructive pulmonary disease), cardiovascular disease (including acute myocardial in-

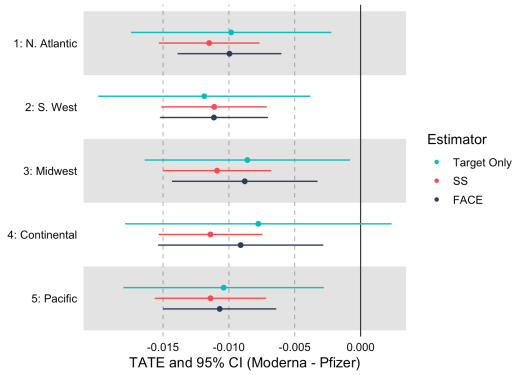
farction, cardiomyopathy, coronary heart disease, heart failure, and peripheral vascular disease), hypertension, type 2 diabetes, chronic kidney disease, autoimmune diseases (including HIV infection, rheumatoid arthritis, etc.), and obesity (defined as body mass index of 30 or greater).

The raw event rates for documented COVID-19 infection within 180 days of receiving the second dose for Pfizer (Moderna) in the five sites were 2.81% (1.93%) in the North Atlantic, 3.58% (3.23%) in the Southwest, 2.25% (2.08%) in the Midwest, 2.97% (2.36%) in the Continental, and 2.80% (1.43%) in the Pacific. The raw event rates for death with COVID-19 infection within 180 days of receiving the second dose for Pfizer (Moderna) were 0.37% (0.06%) in the North Atlantic, 0.36% (0.23%) in the Southwest, 0.18% (0.21%) in the Midwest, 0.21% (0.26%) in the Continental, and 0.11% (0.09%) in the Pacific.

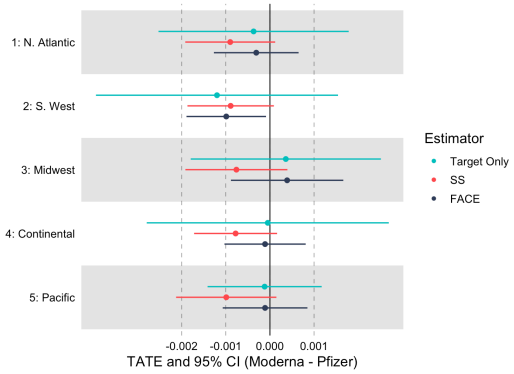
Figure 2.2 shows the TATE estimates for the four outcomes of interest: (a) 120-day COVID-19 infection, (b) 180-day COVID-19 infection, (c) 120-day death with COVID-19 infection, and (d) 180-day death with COVID-19 infection. For each outcome, the target population is taken to be one of the five sites. Three estimators are compared along with their 95% confidence interval: (i) a doubly robust estimator that only uses target site data (Target Only), (ii) a sample-size weighted estimator that leverages each site where η_k is taken to be n_k/N(SS), $k = 1, \ldots, 5$, and (iii) the FACE estimator. Our results indicate that the FACE estimator tracks the Target Only estimator more closely compared to the SS estimator. Compared to the Target Only estimator, the FACE estimator has substantially tighter confidence intervals, resulting in qualitatively different conclusions in certain cases, e.g., 180-day COVID-19 infection in the Continental site, 120-day death with COVID-19 infection in the Southwest site, and 180-day death with COVID-19 infection in the Midwest, North Atlantic, and Southwest sites. Using FACE, our results show that veterans who received Moderna had an approximately 1% lower rate of 180-day COVID-19 infection compared to Pfizer, and this difference appeared consistent across sites.



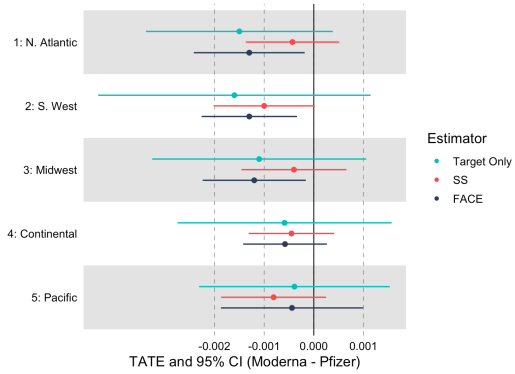
(a) TATE for COVID-19 infection (120 days)



(b) TATE for COVID-19 infection (180 days)



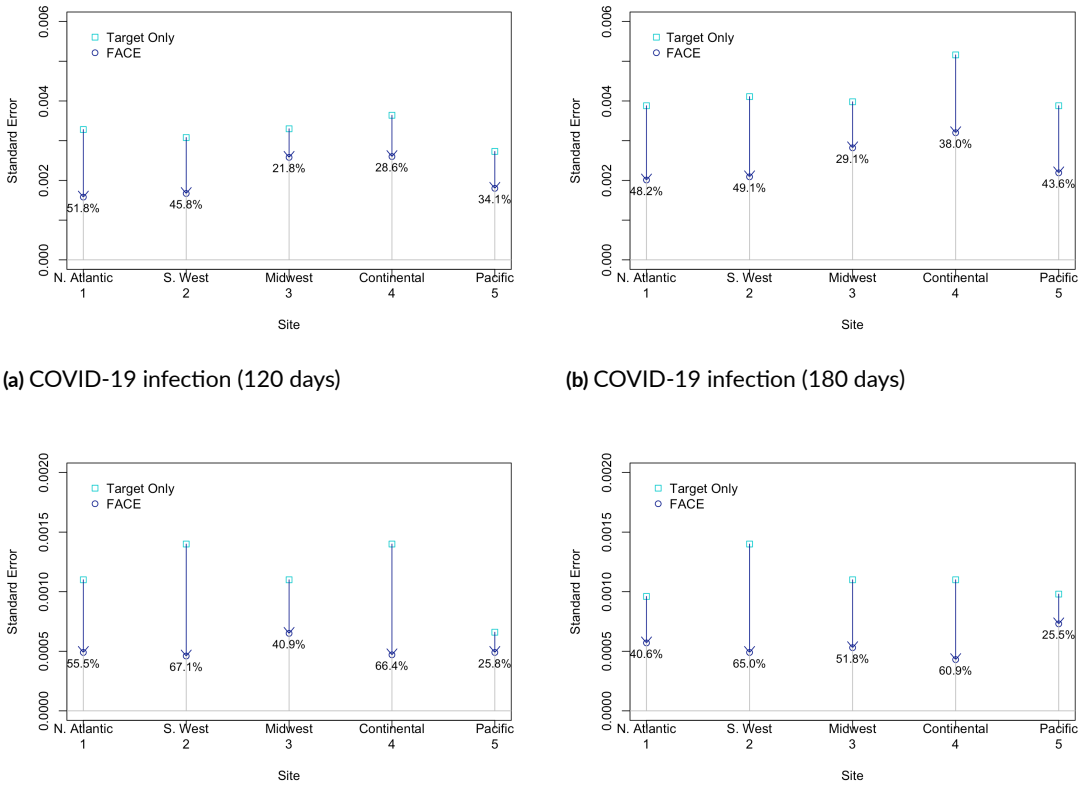
(c) TATE for COVID-19 death (120 days)



(d) TATE for COVID-19 death (180 days)

Figure 2.2: TATE estimates for the comparative effectiveness of Moderna vs. Pfizer vaccines for four outcomes.

Figure 2.3 visualizes the efficiency gain in using FACE compared to the Target Only estimator. For each of the four outcomes of interest, FACE meaningfully reduces the standard error of the TATE estimate for each target site, with the percentage reduction ranging from 25.5% to 67.1%.



(a) COVID-19 infection (120 days)

(b) COVID-19 infection (180 days)

(c) COVID-19 death (120 days)

(d) COVID-19 death (180 days)

Figure 2.3: Gain in efficiency for TATE estimate using FACE vs Target Only estimator. For each site, the percent reduction in SE is calculated for each of the four outcomes.

2.7 Discussion

In this paper, we have developed FACE to leverage heterogeneous data from multiple study sites to more precisely estimate treatment effects for a target population of interest. FACE accounts for het-

erogeneity in the distribution of covariates through a density ratio weighting approach and protects against distributional heterogeneity and model mis-specification of the source sites through an adaptive integration strategy. It improves upon the precision of the target-population only estimator by leveraging source population information without inducing bias. FACE is privacy-preserving and communication-efficient, requiring only one round communication of aggregated summary statistics between sites. In addition to providing theoretical double robustness and efficiency guarantees, FACE does not rely on prior knowledge of model stability or correct model specification, which is a substantial improvement on current federated methods for causal inference [130]. We also obtained promising results from a real world analysis of COVID-19 outcomes for veterans assigned to either Pfizer or Moderna vaccines among five federated VA sites.

FACE can easily be generalized to the setting where some sites have RCT data. In such a setting, one could define the target population as the set of trial participants. When the RCT data is treated as the anchoring site, the target site PS model is known, so the target site estimator for the TATE is consistent, and the global adaptive estimator is likely to be more reliable. Our FACE framework can thus be viewed as a contribution to recent work on using observational studies to reduce the variance associated with treatment effect estimates from experimental studies [5]. For greater generalizability, participants for whom there is only observational data can be taken to be the target population. FACE can also be adapted to target different causal parameters of interest, such as the average treatment effect of the treated (ATT).

Future work may consider focusing on developing methods for estimands defined by subpopulations of interest. For example, the conditional average treatment effect (CATE) is an important estimand of real world interest, particularly for understanding benefits and dangers of treatments for underrepresented groups and fairness research.

Supplementary Materials

The Supplementary Materials are divided into four sections. In Section B.1, we illustrate the workflow of FACE to construct a global estimator in a federated data setting. In Section B.2, we provide a mild set of sufficient conditions for the necessary regularity conditions to hold in the special case with logistic regression models for the nuisance functions. In Section B.3, we provide proofs for the theoretical results in Section 4 of the main paper. In Section B.4, we provide supplementary tables corresponding to the real data analysis.

3

Robust and Optimal Sensitivity Analyses (ROSA) of Clinical Trial Designs

3.1 Introduction

Clinical trial designs are becoming increasingly complex to meet the multifaceted needs and goals of precision medicine. Examples of complex designs include adaptive seamless phase i/ii designs for eval-

uating, early in the treatment development process, the dosing, safety, and activity of new drugs[52]. Also, adaptive randomized trials with frequent interim looks at the data can evaluate one or more therapies simultaneously while attempting to minimize trial duration and resources[116,12]. Additional examples of complex designs have been implemented in biomarker-stratified trials to evaluate the efficacy of a therapy and possible variations of treatment effects across patient subgroups[78].

When planning a new trial, it is necessary to predict and evaluate several operating characteristics. Relevant operating characteristics can include the likelihood of selecting an effective dose with low toxicity in a phase i/ii study, the probability of detecting treatment effects in a randomized study, the expected trial duration, costs, and other metrics to evaluate designs that often enroll patients from different subgroups. Multiple operating characteristics typically need to be examined jointly in order to evaluate the relevant trade-offs achieved by candidate designs, such as balancing the accuracy in estimating treatment effects and the expected study duration.

The obvious challenge for evaluating a candidate design is that the vector of operating characteristics of the study design is not known and it is difficult to estimate before the onset of the trial. Indeed, the operating characteristics are usually a function of a vector of unknown parameters that identify the distribution of all relevant variables to be captured during the trial. For example, unknown parameters can include the enrollment and drop-out rates, the magnitude of treatment effects, and the prevalence of predictive biomarkers in the trial population. Uncertainty on these parameters makes it non-trivial to evaluate whether a candidate design is appropriate for implementing the new study.

Sensitivity analyses are commonly used to account for uncertainty on unknown parameters and operating characteristics when evaluating a candidate design. They typically proceed in three steps. First, a set of plausible scenarios, i.e., specific values of the vector of unknown parameters, is selected. Next, the corresponding operating characteristics are computed using trial simulations or analytic results. Finally, based on the computed operating characteristics and their variations across the set of scenarios, the investigators evaluate whether the candidate design is appropriate to achieve the aims of

the study. Throughout the manuscript, we use the terms *sensitivity analysis* or *simulation report* to indicate a set of scenarios and the associated operating characteristics which are computed to illustrate how the operating characteristics vary across plausible values of unknown parameters.

Producing a simulation report to effectively evaluate a study design has been recommended as one of the key supporting documents for interacting with the FDA[76,37]. However, it can be difficult to select the set of unknown parameters, especially if the dimension of the vector of unknown parameters is moderate to high (say $\geq$ 5). For the investigators, it might be unclear if the selected scenarios are adequate to illustrate the variations of the operating characteristics across potential values of the unknown parameters. Similarly, for regulators, there may be skepticism as to whether the selected scenarios are chosen to highlight positive aspects of the trial design without pointing at its limitations and negative aspects[96]. Another subtle challenge is the choice of the number of scenarios. Indeed, a large number of scenarios (say 100) may simplify the task of representing how the operating characteristics vary across potential values of the unknown parameters, but a simulation report that contains too many scenarios makes it difficult to interpret and communicate the included results.

We propose a method to choose an optimal set of scenarios for a simulation report that will provide relevant operating characteristics. This decision is based on a utility criterion, which formalizes the ability of any set of scenarios to represent the map between the unknown parameters and the operating characteristics. In some cases, we will consider a restriction of the unknown parameter space to focus only on regions of higher uncertainty or plausible values of the unknown parameters. The utility criterion assigns high (low) utility to a set of scenarios if the table of potential unknown parameters and operating characteristics is an accurate (inaccurate) summary of how the design's operating characteristics vary across the considered parameter space. We call the set of scenarios that maximizes the utility criterion the Representative and Optimal Sensitivity Analysis (ROSA) scenarios. To select the ROSA scenarios, we introduce a computational procedure that leverages (i) flexible regression methods like neural networks (NNs)[41] and (ii) optimization algorithms like simulated annealing[10]. Our

approach is applicable to any trial design, regardless of the number of unknown parameters and the number of operating characteristics.

To illustrate the method, we conduct sensitivity analyses for three trial designs. The first is a two-arm randomized design that aims to test and estimate the effects of an experimental treatment compared to the standard of care (SOC). The second is a multi-stage randomized trial that leverages an auxiliary/intermediate/surrogate outcome S measured shortly after randomization for interim decisions and a primary outcome Y with a longer ascertainment time[84]. The third is a biomarker-adaptive enrichment design similar to the design of the TAPPAS trial[78], a randomized phase iii trial comparing TRC105 and pazopanib versus pazopanib alone in patients with advanced angiosarcoma[62,64]. In the first design, we consider a single unknown parameter and a single operating characteristic, whereas the latter two designs consider multiple unknown parameters and multiple operating characteristics.

3.2 Selecting Sensitivity Scenarios

3.2.1 Notation and problem set-up

We introduce our procedure to select K sensitivity scenarios $\theta_1, ..., \theta_K \in \Theta$, where Θ is the set of potential values of the unknown parameters θ. We assume that Θ is a bounded subset of $\mathbb{R}^d$ and use the notation $||\cdot||_2$ to indicate the Euclidean norm on $\mathbb{R}^d$. We will restrict Θ to a subset Θ' when there is sufficient prior information from completed studies or clinical experience. We identify ROSA scenarios $\theta_1^*, ..., \theta_K^*$ as the scenarios that maximize a utility criterion $\mathcal{U}$

$$\theta_1^*, ..., \theta_K^* = \text{argmax}_{\theta_1,...,\theta_K} \mathcal{U}(\theta_1, ..., \theta_K), \tag{3.1}$$

where

$$\mathcal{U}(\theta_1, ..., \theta_K) = -\max_{\theta' \in \Theta} \left\{ \min_{k=1,...,K} D[\mathbf{f}(\theta'), \mathbf{f}(\theta_k)] \right\}. \tag{3.2}$$

We can symmetrically define the corresponding loss function $\mathcal{L} = -\mathcal{U}$ by inverting the sign in equation (3.2). Here, $D[\mathbf{f}(\theta'), \mathbf{f}(\theta_k)]$ is a metric between the operating characteristics $\mathbf{f}(\theta') = (f_1(\theta'), \ldots f_R(\theta'))$ and $\mathbf{f}(\theta_k) = (f_1(\theta_k,) \ldots, f_R(\theta_k))$. We will consider metrics of the form

$$D[\mathbf{f}(\theta'), \mathbf{f}(\theta_k)] = \sum_{r=1}^{R} w_r ||f_r(\theta') - f_r(\theta_k)||_2,$$

where $w_1, \ldots, w_R$ are non-negative weights that sum to one. The weights can be user-specified to calibrate the relative importance of different operating characteristics. Setting the weights to $1/R$ results in equal weighting for each operating characteristic.

We can now provide an explicit interpretation of the utility function $\mathcal{U}$ in equation (3.2). Consider a set of scenarios $\{\theta_1, ..., \theta_K\}$ – the order of the entries is not relevant – and an arbitrary scenario θ' in Θ. For $1 \leq k \leq K$, the metric $D[\mathbf{f}(\theta'), \mathbf{f}(\theta_k)]$ is a summary of the differences between the operating characteristics at θ' and the same operating characteristics when we consider the k-th scenario θ_k. Therefore, $\min_{k=1,...,K} D[\mathbf{f}(\theta'), \mathbf{f}(\theta_k)]$ can be viewed as an approximation error between $\mathbf{f}(\theta')$ and a similar vector of operating characteristics selected among our K options $\mathbf{f}(\theta_1), ..., \mathbf{f}(\theta_K)$. Expression (3.2) identifies through the maximization operator the worst-case (with highest approximation error) that we can obtain by varying θ' in Θ. We maximize the utility function $\mathcal{U}$ and use $\theta_1^*, ..., \theta_K^*$ to indicate the ROSA scenarios. Alternative utility criteria and loss functions are described later in the manuscript.

3.2.2 An example with a geometric interpretation

To provide a geometric interpretation of the utility criterion $\mathcal{U}$, we illustrate how one set of K scenarios can be preferable to a different set of K scenarios (Figure 3.1). Specifically, suppose we aim to design a single-arm trial with an interim analysis that allows for early-stopping for futility. The goal of the trial is to compare the response rate of an experimental drug θ_1 with that of the SOC θ_0 at the end of the study.

However, because study patients only receive the experimental drug, the response rate under the SOC θ_0 is estimated ($\widehat{\theta}_0$) before the onset of the study, for example using data from a previous trial. At the interim analysis, the trial may stop for futility if the preliminary evidence of positive treatment effects $\Delta_{interim}$ is insufficient to continue the study. During the final analysis, the null hypothesis $H_0 : \theta_1 \leq \widehat{\theta}_0$ (the experimental therapy is not superior to the historical control) is tested against the alternative hypothesis $H_1 : \theta_1 > \widehat{\theta}_0$ (the experimental therapy is superior to the historical control). In this design, $\theta = (\theta_0, \theta_1)$ are the unknown parameters, and $\Theta = [0,1]^2$. Suppose that there are two operating characteristics of interest: (i)f_1, the probability of a true positive result when the experimental drug has beneficial effects compared to the SOC (f_1 is equal to zero if the treatment effects are null or negative) and (ii)f_2, the expected sample size.

The left panel of Figure 3.1 is a representation of Θ. We are interested in the two operating characteristics of the single-arm design. Two sets of $K = 6$ scenarios are proposed. The first set of scenarios $\{\theta_1^1, ..., \theta_6^1\}$ (blue points) is chosen by varying both unknown parameters at the same time, while the second set $\{\theta_1^2, ..., \theta_6^2\}$ (red points) is chosen by varying only θ_0 while fixing the value of θ_1. The two sets of scenarios, the corresponding operating characteristics, and associated loss $\mathcal{L} = -\mathcal{U}$ are represented in the right panel of Figure 3.1. The first set of scenarios (blue points) is preferred over the second set (red points) because it is more representative of the variation of the operating characteristics over Θ. Geometrically, the loss $\mathcal{L}(\theta_1^1, ..., \theta_6^1)$ associated with the blue points is identical to the minimum radius of the circles with centers $\mathbf{f}(\theta_1^1), ..., \mathbf{f}(\theta_6^1)$ (see Figure 3.1) necessary to cover the operating characteristics surface $\mathbf{f}(\Theta)$.

3.2.3 Estimating the operating characteristics

We describe an algorithm to numerically approximate the operating characteristics $\mathbf{f}(\theta)$ for every $\theta \in \Theta$. This is necessary to solve the optimization problem in equation (3.2). Indeed, in most cases the function $\mathbf{f}(\theta)$ cannot be computed in closed form.

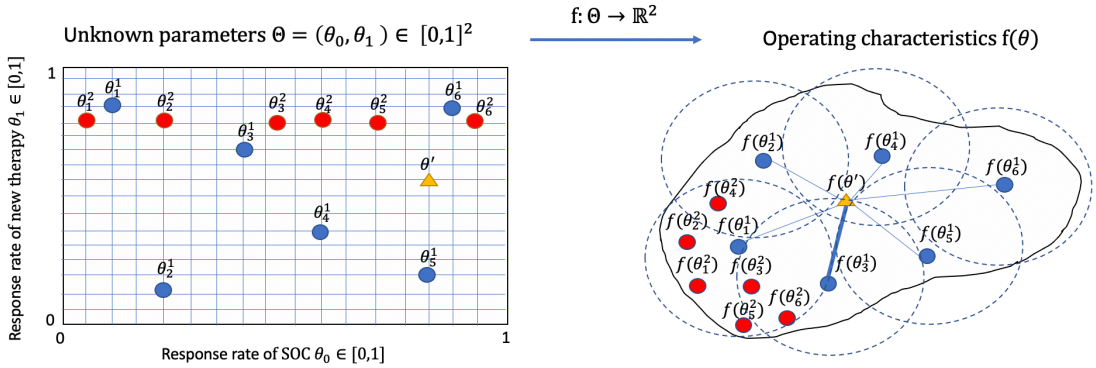


Figure 3.1: Geometric representation of an arbitrary scenario θ' and two proposed sets of scenarios. (Left) Parameter space $\Theta = [0,1]^2$ with arbitrary scenario θ' (orange triangle) and two sets of proposed scenarios $\{\theta_1^1, \ldots, \theta_6^1\}$ (blue points) and $\{\theta_1^2, \ldots, \theta_6^2\}$ (red points). (Right) Operating characteristic surface $\mathbf{f}(\Theta)$ with the corresponding operating characteristics for θ' and the two proposed sets of scenarios. The radius of the dotted circles (with blue points as centers) is the value of the loss $\mathcal{L}$ associated with the blue points. ROSA scenarios minimize the loss $\mathcal{L}$, which in turn is equal to the radius of the dotted circles that cover the operating characteristic surface $\mathbf{f}(\Theta)$.

We briefly outline our four-step procedure. In the first step, we choose a large number J (say $J = 1000$) of training scenarios $\theta_1', \ldots, \theta_J'$. In the second step, we use Monte Carlo simulations to obtain estimates $\bar{\mathbf{f}}(\theta_1'), \ldots, \bar{\mathbf{f}}(\theta_J')$ of $\mathbf{f}(\theta_1'), \ldots, \mathbf{f}(\theta_J')$. In the third step, we train a flexible regression model – we use NNs in our implementation – based on the data points $(\theta_1', \bar{\mathbf{f}}(\theta_1')), \ldots, (\theta_J', \bar{\mathbf{f}}(\theta_J'))$. The output of this step is a regression function $\hat{\mathbf{f}}(\theta)$ that is easy to compute at any $\theta \in \Theta$ and that approximates $\mathbf{f}(\theta)$. In the fourth step, we validate the regression model based on J' (say $J' = 200$) independent simulations $(\theta_1'', \bar{\mathbf{f}}(\theta_1'')), \ldots, (\theta_{J'}'', \bar{\mathbf{f}}(\theta_{J'}''))$. Steps 1-3 of this procedure are summarized in Algorithm 1. Step 4 is described in Algorithm 2.

In more detail, in step 1, to select the training scenarios $\theta_1', \ldots, \theta_J'$, we randomly select J scenarios in Θ using Latin hypercube sampling (LHS)[77]. LHS generates J scenarios by first partitioning the d unknown parameter dimensions into J non-overlapping intervals and selecting one value from each interval at random. The J values obtained for the first unknown parameter θ_1 are randomly paired with the J values obtained for the second θ_2, and so on, for all d unknown parameters to form J d-tuples, which constitute the training scenarios $\theta_1', \ldots, \theta_J'$. In practice, users may ascribe greater weight to dif-

ferent partitions of the grid, and extensions of standard LHS can accommodate these implementation choices [17].

In step 2, we estimate the operating characteristics of the trial design. For simplicity, we consider operating characteristics defined as expected values (e.g., bias, power, duration of the trial, etc.), but the algorithm can be easily modified to consider other operating characteristics. Specifically, we assume that $\mathbf{f}(\theta) = \mathbb{E}_\theta[\phi(Z, \theta)]$ for some function ϕ, where the random vector Z represents the data generated during the trial – including the collection of treatment assignment indicators and realized patient outcomes – under scenario θ. For example, ϕ can be the indicator that captures if a null hypothesis of interest has been correctly rejected at the end of the study, or the duration of the simulated trial. In practice, to estimate $\mathbf{f}(\theta)$, we proceed as follows. First, for each of the training scenarios θ^t_j, $1 \leq j \leq J$, we simulate M (say $M = 200$) clinical trials following the trial design. We then use the M scenario-specific simulated trials to compute the estimate

$$\bar{\mathbf{f}}(\theta^t_j) = M^{-1} \sum_{m=1}^{M} \phi(Z_{j,m}, \theta^t_j), \quad 1 \leq j \leq J,$$

where $Z_{j,m}$ is the m^{th} trial dataset simulated under the j^{th} training scenario θ^t_j.

In step 3, we have only two inputs, the scenarios θ^t_j and the estimates $\bar{\mathbf{f}}(\theta^t_j)$, $1 \leq j \leq J$, to fit a function $\hat{\mathbf{f}}(\theta)$. For example, one could use NNs, splines [14], or Gaussian processes [94]. We use NN regression functions in our applications because these are easy to compute using widely available software and have been demonstrated to have good empirical performance [69,55,41].

In step 4 (Algorithm 2), we investigate the differences between $\hat{f}$ and f. Specifically, we first select at random J' validation scenarios $\theta^v_1, \ldots, \theta^v_{J'}$ independently with respect to previous computations (step 1-3) and simulate M' trials (say $M' = 500$) for each $\theta^v_{j'}$, $1 \leq j' \leq J'$. Based on the results of the simu-

Algorithm 2: Obtaining a function $\hat{f}$ that approximates the operating characteristic function $\mathbf{f}$

Input: Trial design, Parameter space Θ, J, M

Step 1: Select J scenarios $\theta'_1, ..., \theta'_J \in \Theta$

Step 2: **for** $j = 1$ *to* J **do**

  Simulate M trials

  Obtain approximate operating characteristics $\bar{f}(\theta'_j) = M^{-1} \sum_{m=1}^{M} \phi(Z_{j,m}, \theta'_j)$, where ϕ is a function of $Z_{j,m}$, the m^{th} trial dataset simulated under the j^{th} scenario, and the corresponding parameter θ_j

end

Step 3: Obtain an approximation of the operating characteristics $\hat{f}$ by training a regression algorithm, for example a NN model, and using the data points $(\theta'_j, \bar{f}(\theta'_j)), 1 \leq j \leq J$

Output: Function $\hat{\mathbf{f}}(\theta)$

lated trials, for each j', we then compute Monte Carlo estimates $\bar{\mathbf{f}}(\theta^v_{j'}) = M'^{-1} \sum_{m'=1}^{M'} \phi(Z_{j',m'}, \theta^v_{j'})$ of the operating characteristics $\mathbf{f}(\theta^v_{j'})$. For several important operating characteristics (e.g., average sample size, expected duration, power, type 1 error), the estimator $\bar{\mathbf{f}}(\theta^v_{j'}) = M'^{-1} \sum_{m'=1}^{M'} \phi(Z_{j',m'}, \theta^v_{j'})$ is unbiased. Finally, we compare the estimates $\bar{f}(\theta^v_{j'})$ and the independent estimates $\hat{f}(\theta^v_{j'})$. We use summary statistics and graphs to evaluate the differences $\hat{f}(\theta^v_{j'}) - \bar{f}(\theta^v_{j'})$. If the approximation $\hat{f}(\theta^v_{j'})$ is not adequate, we can use a different regression methodology, increase the number (M, M') of trials, or increase the number J of training scenarios in Algorithm 1.

3.2.4 Approximating the loss function

After computing $\hat{f}$ (Algorithm 1) and validating its accuracy (Algorithm 2), we use it to approximate the loss function $\mathcal{L}(\theta_1, ..., \theta_K)$. To proceed, we choose a diffuse and finite subset of the parameter

Algorithm 3: Validating the approximation of the operating characteristics $\hat{f}$

Input: Approximation of the operating characteristics $\hat{f}$, Trial design, J'
Randomly select J' scenarios $\theta^v_1, \ldots, \theta^v_{J'} \in \Theta$ independently from previous computations (Algorithm 1)
for $j' = 1$ *to* J' **do**
    Simulate M' trials $Z_{j',m'}$
    Compute $\bar{f}(\theta^v_{j'}) = M'^{-1} \sum_{m'=1}^{M'} \phi(Z_{j',m'}, \theta^v_{j'})$
    Compute $\hat{f}(\theta^v_{j'})$
end
Output: Set of differences $\hat{f}(\theta^v_{j'}) - \bar{f}(\theta^v_{j'})$ and scatterplots to jointly visualize the operating characteristic estimates $\bar{f}(\theta^v_{j'})$ and the independent estimates $\hat{f}(\theta^v_{j'})$, $1 \leq j' \leq J'$. Compute summaries of the differences (e.g., median, range, or other descriptive statistics).
Interpretation: Differences between $\bar{f}(\theta^v_{j'})$ and the independent estimates $\hat{f}(\theta^v_{j'})$, $1 \leq j' \leq J'$, consistently close to zero provide evidence that $\hat{f}$ is an accurate approximation of f

space $\Theta^F \subset \Theta$. For example Θ^F can include 100,000 random points from a distribution with support Θ. When Θ^F contains a large number of random points that are distributed over Θ, under minimal assumptions (e.g., compact Θ and operating characteristics with bounded range),

$$\begin{aligned}\mathcal{L}(\theta_1, \ldots, \theta_K) &= \max_{\theta \in \Theta} \left\{ \min_{k=1,\ldots,K} D[f(\theta), f(\theta_k)] \right\} \\ &\approx \max_{\theta' \in \Theta^F} \left\{ \min_{k=1,\ldots,K} D[\hat{f}(\theta'), \hat{f}(\theta_k)] \right\} = \hat{\mathcal{L}}(\theta_1, \ldots, \theta_K).\end{aligned}$$

To summarize, we can approximate the loss function $\mathcal{L}(\theta_1, \ldots, \theta_K)$ over the entire parameter space Θ by $\hat{\mathcal{L}}(\theta_1, \ldots, \theta_K)$ using a diffuse and finite subset Θ^F.

3.2.5 Optimization by simulated annealing

We now aim to approximately minimize the loss function $\hat{\mathcal{L}}$. To illustrate the need for approximate solutions, consider the setting of a single unknown parameter ($d = 1$), a finite Θ, and an easy-to-compute loss function $\mathcal{L}$. Even in this simple setting, identifying $\theta_1^*, \ldots, \theta_K^* \in \Theta$ can be challenging. For example, to select $K = 10$ representative scenarios $\theta_1^*, \ldots, \theta_K^*$ from 1000 points $\{\theta_j; 1 \leq j \leq 1000\} = \Theta$, the loss function $\hat{\mathcal{L}}$ would need to be calculated for 2.63×10^{23} different possible sets $\{\theta_1, \ldots, \theta_K\}$. In what follows, we describe the use of simulated annealing (Algorithm 3), a simple strategy to reduce the outlined computational burden, regardless if Θ is finite or not[67,10,110].

The simulated annealing algorithm proceeds as follows. First, initial scenarios $\theta_1^1, \ldots, \theta_K^1$ are proposed, for example by sampling $\theta_1^1, \ldots, \theta_K^1$ from a probability distribution with support Θ. Then, iteratively for $1 \leq i \leq I$, the current scenarios $\theta_1^i, \ldots, \theta_K^i$ are perturbed by adding to them Gaussian noise variables $z_1^i, \ldots, z_K^i$, thus obtaining new proposed scenarios $\theta_1', \ldots, \theta_K'$ (this step is represented by the "Perturb" operator in Algorithm 3). At each iteration, the proposed scenarios $\theta_1', \ldots, \theta_K'$ can either be accepted (i.e., $[\theta_1^{i+1}, \ldots, \theta_K^{i+1}] \leftarrow [\theta_1', \ldots, \theta_K']$) or rejected (i.e., $[\theta_1^{i+1}, \ldots, \theta_K^{i+1}] \leftarrow [\theta_1^i, \ldots, \theta_K^i]$). The acceptance or rejection of the proposed scenarios is stochastic, with probability ρ_i (defined below), which is a function of $\hat{\mathcal{L}}(\theta_1', \ldots, \theta_K')$ and $\hat{\mathcal{L}}(\theta_1^i, \ldots, \theta_K^i)$.

The acceptance probability ρ_i is equal to 1 when $\hat{\mathcal{L}}(\theta_1', \ldots, \theta_K') < \hat{\mathcal{L}}(\theta_1^i, \ldots, \theta_K^i)$. That is, if the proposed scenarios decrease the current loss value, then the proposed scenarios are accepted. If instead $\hat{\mathcal{L}}(\theta_1', \ldots, \theta_K') \geq \hat{\mathcal{L}}(\theta_1^i, \ldots, \theta_K^i)$, then ρ_i is

$$\rho_i = \exp\left(\frac{\hat{\mathcal{L}}(\theta_1^i, \ldots, \theta_K^i) - \hat{\mathcal{L}}(\theta_1', \ldots, \theta_K')}{T_i}\right),$$

where T_i, $0 \leq i \leq I$, is a decreasing sequence of positive real numbers often called the "cooling schedule" of the algorithm. A common cooling schedule is $T_i = T_0 \cdot r^{i-1}$, where T_0 is a constant and

$r \in (0,1)$ is a multiplicative contraction, but other forms are possible [110]. In our applications, we use a piecewise-constant cooling schedule [58].

After simulating the outlined Markov Chain for a fixed number I of iterations, the final set of scenarios $\{\theta_1^{I+1}, ..., \theta_K^{I+1}\}$ approximately minimizes the loss function $\hat{\mathcal{L}}$ [10]. In our ROSA implementation, we use multiple independent replicates of Algorithm 3, with different initial scenarios $\theta_1^1, ..., \theta_K^1$, to investigate convergence of the Markov chain. Intuitively, if the independent chains converge, then the corresponding loss values of the approximate optima $\hat{\mathcal{L}}(\theta_1^{I+1}, ..., \theta_K^{I+1})$ should be nearly identical.

Algorithm 4: Pseudocode for simulated annealing to obtain ROSA scenarios

Initialize the values of $\theta_1^1, ..., \theta_K^1$, e.g., by sampling from a distribution over Θ
Best proposal $\leftarrow \theta_1^1, ..., \theta_K^1$
for $i = 1$ *to* I **do**
    New proposal $\theta_1', ..., \theta_K' \leftarrow \text{Perturb}(\theta_1^i, ..., \theta_K^i)$
    if $\hat{\mathcal{L}}(\theta_1', ..., \theta_K') \leq \hat{\mathcal{L}}(\theta_1^i, ..., \theta_K^i)$ **then**
        Define $\theta_j^{i+1} = \theta_j'$ for every $j = 1, \ldots, K$;
    else
        Compute the acceptance probability
        $\rho_i = \exp\left([\hat{\mathcal{L}}(\theta_1^i, ..., \theta_K^i) - \hat{\mathcal{L}}(\theta_1', ..., \theta_K')]/T_i\right)$
        Sample $U_i \sim \text{Uniform}(0,1)$
        If $U_i \leq \rho_i$, define $\theta_j^{i+1} = \theta_j'$ for every $j = 1, \ldots, K$;
        Otherwise $\theta_j^{i+1} = \theta_j^i$ for every $j = 1, \ldots, K$.
end
Output: $\theta_1^{I+1}, ..., \theta_K^{I+1}$, $\hat{\mathcal{L}}(\theta_1^{I+1}, ..., \theta_K^{I+1})$

3.3 Applications: Sensitivity Analyses of Three Trial Designs

We illustrate the ROSA approach by performing sensitivity analyses for three designs of different complexity levels. In each example, we describe the design of the trial, the unknown parameters,

and the operating characteristics of interest. By illustrating the ROSA methodology in three trial designs, we show its flexibility with potential applications to evaluate nearly any clinical trial design. Indeed, ROSA only requires the possibility of simulating the trials under potential unknown parameters $\theta \in \Theta = \mathbb{R}^d$ and the definition of the operating characteristics of interest.

3.3.1 Application 1: Two-arm RCT

In the first example, we will only consider a single unknown parameter (i.e., $\theta \in \mathbb{R}$) and a single operating characteristic $f(\theta)$ that can be computed analytically. In this case, the optimal set of scenarios $\{\theta_1^*, ..., \theta_K^*\}$ can be computed exactly, without resorting to approximation methods. This simple and stylized setting is useful to highlight the similarity of the approximations and selected scenarios computed by ROSA with their exact counterparts.

Trial design

We consider the design of a two-arm randomized trial (1:1 randomization ratio) with a sample of $n = 30$ patients. For each $i = 1, \ldots, n$, we let $A_i = 0$ or 1 if the i-th study patient is assigned to the control or experimental arm. The outcomes of the n study patients are $Y_1, \ldots, Y_n$, which we assume to be independent and normally distributed. If $A_i = a$ then Y_i has mean $\mu_a = 100 + 15a$ and standard deviation σ equal to 30. In the analysis of the study, a z-statistic will be used to test the null hypothesis $H_0 : \mu_1 - \mu_0 \leq 0$ against the alternative $H_1 : \mu_1 - \mu_0 > 0$ at 5% significance level.

Aim of the sensitivity analysis

The goal of the sensitivity analysis is to assesses the variation of the probability of rejecting H_0, a function $f(\theta)$ of the unknown treatment effect $\theta = \mu_1 - \mu_0 \in \Theta = \mathbb{R}$. For example, if we knew that $\theta = 13.5$, then $f(\theta) = 0.80$, but in general θ is an unknown value. Suppose we aim to identify $K = 3$

scenarios $\theta_1^*, \theta_2^*, \theta_3^*$ that maximize the utility $\mathcal{U}$, i.e.,

$$\theta_1^*, \theta_2^*, \theta_3^* = \text{argmax}_{\theta_1,\theta_2,\theta_3 \in \Theta} \mathcal{U}(\theta_1, \theta_2, \theta_3), \tag{3.3}$$

where $\mathcal{U}(\theta_1, \theta_2, \theta_3) = -\max_{\theta' \in \Theta} \min_{k=1,2,3} |f(\theta') - f(\theta_k)|$.

In this trial, we have a single unknown parameter ($\Theta = \mathbb{R}$), and the operating characteristic of interest is monotone, continuous, invertible, and ranges from 0 to 1. Therefore, it is straightforward to see that the optimal scenarios $\theta_1^*, \theta_2^*, \theta_3^*$ correspond to the operating characteristic values that evenly divide the interval $(0, 1)$. To be precise, $\{f(\theta_1^*), f(\theta_2^*), f(\theta_3^*)\} = \{1/6, 3/6, 5/6\}$; these are the three values of a regular grid on the interval $(0, 1)$. Figure 3.2A illustrates the optimal set of scenarios when $K = \{3, 5, 10\}$. Since $f(\theta)$ can be calculated exactly, the optimal scenarios $\theta_1^*, \theta_2^*, \theta_3^*$ can be obtained by computing the inverse function f^{-1} at the values $1/6$, $3/6$, and $5/6$. Specifically,

$$\{\theta_1^*, \theta_2^*, \theta_3^*\} = \left\{ \frac{\sigma(z_{f(\theta_1^*)} + z_{1-\alpha/2})}{\sqrt{n}}, \frac{\sigma(z_{f(\theta_2^*)} + z_{1-\alpha/2})}{\sqrt{n}}, \frac{\sigma(z_{f(\theta_3^*)} + z_{1-\alpha/2})}{\sqrt{n}} \right\},$$

where $z_{1-\alpha/2}$ is the $1-\alpha/2$ quantile of the standard normal distribution. The corresponding optimal scenarios are illustrated as red asterisks in Figure 3.2B.

Implementing and benchmarking ROSA

The exact computation of the optimal set of scenarios provides a solid benchmark for an initial evaluation of ROSA (Algorthms 1-3). We can compare the exact solution with the results from ROSA, which has the advantage of being applicable to other designs and operating characteristics that are not available in closed form.

We implement our ROSA approach to identify $K = 3$ scenarios. We randomly select $J = 1000$ scenarios $\theta_1', \ldots, \theta_{1000}'$ with independent samples from the Uniform$(-5, 25)$ distribution. Note that

$f(-5) \approx 0$ and $f(25) \approx 1$. For each θ_j', $1 \leq j \leq 1000$, we simulate $M = 200$ trials to compute the estimate $\bar{f}(\theta_j') = 200^{-1} \sum_{m=1}^{200} \phi(Z_{j,m}, \theta_j')$, where $\phi(Z_{j,m}, \theta_j') \in \{0, 1\}$ either accepts or rejects the $H_0 : \theta_j' \leq 0$ for trial m and scenario j. Then, we compute a smooth function $\hat{f}(\theta)$ using the independent estimates $\bar{f}(\theta_j)$ and a NN with 3 hidden layers (8, 64, and 64 neurons respectively) and ReLU activation functions. Finally, to select three sensitivity scenarios, we use a simulated annealing algorithm based on an initial parameterization $T_1 = 1000$, temperature reduction factor $r = 0.8$, and final parameterization $T_{\min} = 0.1$ (c.f. Algorithm 3). We repeat these three steps (selection of scenarios, use of the NN, and optimization with simulated annealing) 20 times, each time initializing $\theta_1, \theta_2, \theta_3$ with independent random draws from the Uniform$(-5, 25)$ distribution. The results of the exact approach (red asterisks) compared with ROSA (blue points) are shown in Figure 3.2B. The scenarios $\theta_1^*, \theta_2^*, \theta_3^*$ selected by simulated annealing (blue dots) are close to the exact solution (red asterisks).

Choice of number K of scenarios

In practice, the decision regarding the number K of scenarios to report is left to the analyst. This choice can be supported by a graph like Figure 3.2C, which allows the investigator to determine the minimum number K of scenarios needed to guarantee a loss $\mathcal{L}(\theta_1^*, \ldots, \theta_K^*)$ no larger than a targeted threshold. For example, to guarantee a loss no larger than 0.050 in this example, we need to select at least 10 scenarios for the simulation report.

We ran ROSA with $K = 2, 3, 4, 5, 6, 7, 8, 9, 10, 20$, or 30, and compared the loss $\mathcal{L}$ in the resulting set of scenarios with that of the exact solution. The difference in the loss $\mathcal{L}$ of the exact and approximate optima was less than 1% across all K values that we considered (Figure 3.2C). Table 3.1 indicates that the computation time of the simulated annealing algorithm scales well as K increases and that, as expected, the loss $\mathcal{L}$ decreases as K increases. All analyses were run on a Windows laptop with an Intel(R) Core(TM) i7-7700HQ 2.80 GHz processor, 16GB RAM, and 6MB of cache memory.

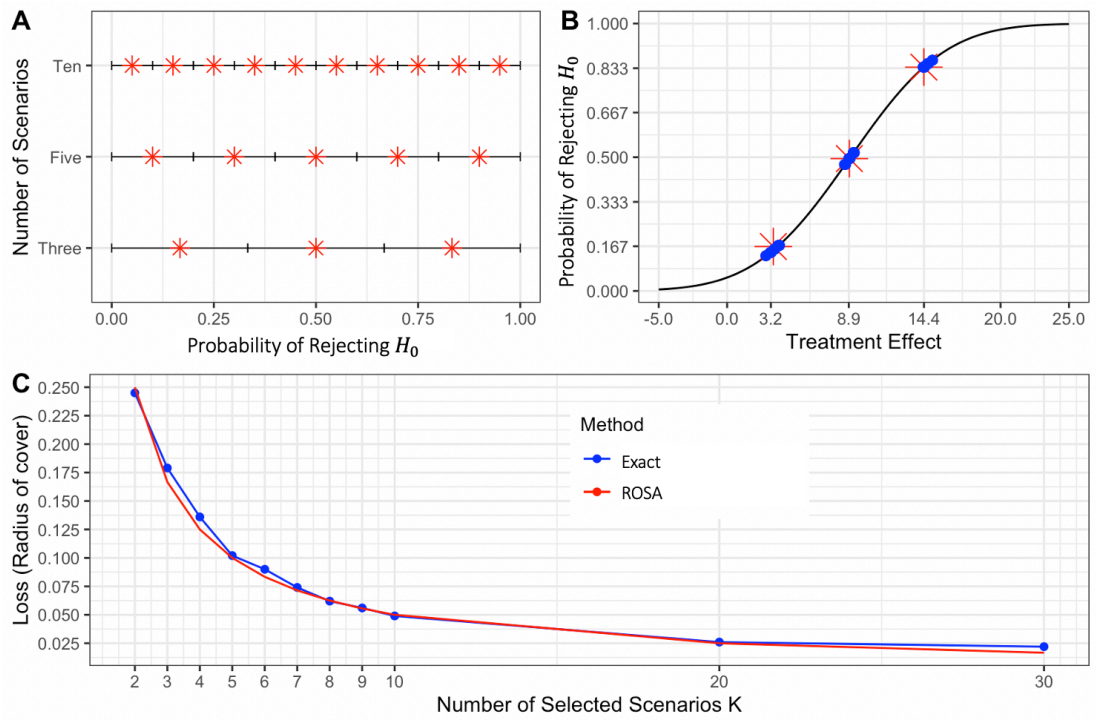


Figure 3.2: Sensitivity analysis of a RCT (operating characteristic: probability of rejecting H_0). **Panel A:** Exact solutions when $K = \{3, 5, 10\}$. **Panel B:** Comparison of $K = 3$ scenarios selected through exact calculation (red asterisks) and by 20 ROSA implementations with different initial proposals (blue points). **Panel C:** Graphical tool to choose the number K of sensitivity scenarios.

Number K of Scenarios	Time (seconds)	ROSA Loss $\mathcal{L}$	Min. Loss $\mathcal{L}$	Rel. Diff.
5	8.8	0.101	0.100	1.0%
6	8.8	0.084	0.083	0.7%
7	9.1	0.072	0.071	0.8%
8	9.2	0.062	0.0625	0.7%
9	9.1	0.056	0.056	0.6%
10	9.1	0.050	0.050	0.2%
20	10.1	0.025	0.025	0.5%
30	10.2	0.017	0.0167	0.8%

Table 3.1: ROSA computation time, ROSA loss $\mathcal{L}$, minimum (exact) loss $\mathcal{L}$, and relative difference in loss of ROSA scenarios compared to the exact solutions.

3.3.2 Application 2: Interim decisions based on auxiliary outcomes

In the second example, we consider sensitivity analyses with multiple unknown parameters and two operating characteristics. We illustrate the use of our computational procedures, including the operating characteristics approximation procedure (Algorithm 1), the validation procedure (Algorithm 2), and the simulated annealing optimization procedure (Algorithm 3). We investigate whether it is appropriate to fix the value of some of the unknown parameters across all sensitivity scenarios. Identical values for a subset of the unknown parameters can simplify the interpretation of the sensitivity analysis but can also introduce severe limitations in faithfully representing how the operating characteristics vary across plausible values of the unknown parameters.

Trial design

We consider a two-arm, two-stage randomized trial with a binary primary outcome Y and a binary auxiliary outcome S[84]. The primary outcome Y is available T_Y months after randomization, while the auxiliary outcome S is available after $T_S < T_Y$ months. For example, in glioblastoma trials, 12-month progression-free survival (PFS) and 24-month overall survival (OS) have been used as auxiliary and primary outcomes, respectively[46]. The approach that we illustrate is applicable for any value of T_Y

and $T_S < T_Y$.

We let N_a be the planned number of patients for arms $a = 0, 1$ (i.e., control and experimental arms) and indicate with p_a the response probability $P(Y = 1 \mid A = a)$. Similarly, let n_a be the planned number of patients assigned to arm a before the interim analysis, and q_a indicate the response probability $P(S = 1 \mid A = a)$. The difference $\Delta = p_1 - p_0$ is the treatment effect on Y. The primary aim of the trial is to test $H_0 : \Delta \leq 0$ versus $H_1 : \Delta > 0$, at level α. The final analysis of the study involves only the primary outcome Y, and the trial will use a standard Z-test, $Z_Y = \frac{\hat{p}_1 - \hat{p}_0}{\sqrt{\bar{p}(1-\bar{p})(N_1^{-1} + N_0^{-1})}}$, where $\hat{p}_a$ is the estimate of p_a and $\bar{p}$ is a weighted average of $\hat{p}_1$ and $\hat{p}_0$.

An interim analysis is conducted after the auxiliary outcomes S become available for n_a patients for arms $a = 0$ and 1 (i.e., T_S months after the enrollment of n_a patients on arms $a = 0$ and 1), with early-stopping for futility or continuation based on a summary of the auxiliary outcomes S. In several clinical settings, the treatment effect on S tends to be more pronounced than the treatment effect on Y. The interim analysis is based on the summary $Z_S = \frac{\hat{q}_1 - \hat{q}_0}{\sqrt{\bar{q}(1-\bar{q})(n_1^{-1} + n_0^{-1})}}$, where $\hat{q}_a$ is the estimate of q_a and $\bar{q}$ is a weighted average of $\hat{q}_1$ and $\hat{q}_0$. We replicate the design of[84], which calculates at the interim analysis the conditional power (CP) using the auxiliary outcome S to determine whether to stop the trial for futility or not. Specifically, the CP is calculated based on Z_S and the information fraction $t_S = \frac{N_1^{-1} + N_0^{-1}}{n_1^{-1} + n_0^{-1}}$ as

$$CP(t_S) = 1 - \Phi\left(\frac{z_{1-\alpha} - Z_S t_S^{1/2}}{\sqrt{1 - t_S}}\right),$$

where $z_{1-\alpha}$ is the $1 - \alpha$ quantile of the standard normal distribution and $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution. Here, we set the cut-off point to be 0.5 so that the trial continues when $CP(t_s) \geq 0.5$.

Aim of the sensitivity analysis

The complexity of the simulation report increases with K (the number of scenarios), d (the number of entries of the unknown parameters θ), and R (the number of operating characteristics $\mathbf{f}(\theta)$). Here the full set of unknown parameters $\Theta \subset \mathbb{R}^7$ include the enrollment rate $e \in (0, \infty)$, the response rates $p_a \in (0,1)$ for Y in $A = a$, the response rates $q_a \in (0,1)$ for S in $A = a$, and the correlation between Y and S in $A = a, \rho_a \in (-1,1)$.

Controlling the complexity of the simulation report is important to ensure high interpretability of the report, which will be discussed by several stakeholders. There are a few potential strategies to reduce the complexity of the simulation report. First, it is often possible to consider only a subset of the parameter space $\Theta' \subset \Theta$ based on prior knowledge of plausible values of the unknown parameters. For example, previous clinical studies can indicate a plausible range for the enrollment rate e, the response rates p_0 under the SOC, and other parameters that are expected to have minimal variations across trials. In addition, we can also consider fixing multiple entries of the K vectors $\theta_1, \ldots, \theta_K$ to some reference values. In this case the space from which we select scenarios $\theta_1, \ldots, \theta_K$ is further reduced to $\Theta'_{re} \subset \Theta'$. For example, if the operating characteristics have low sensitivity with respect to the correlation parameters ρ_a or the enrollment rate e of the study, then we can fix these unknown parameters to common values (i.e., estimates) across all K scenarios.

ROSA allows us to evaluate whether it is appropriate to assign the same value to one or more unknown parameters (e.g., ρ_0 and ρ_1) across all K scenarios. In other words, we evaluate a simulation report with all scenarios in a restricted subset $\Theta'_{re} \subset \Theta'$. A simulation report with scenarios in Θ'_{re} can potentially be easier to interpret compared to a report in which all d entries of θ vary across scenarios by reducing the number of dimensions d of the unknown parameters and pointing to the most relevant unknown parameters when discussing the variations of the operating characteristics across Θ'. We can select scenarios from the restriction $\Theta'_{re} \subset \Theta'$ only if the capability of the simulation

report of representing the operating characteristics variations across Θ' is preserved. Our case study investigates this aspect. The operating characteristics of interest in our case study are the probability of rejecting the null hypothesis of no treatment effect on Y at the end of the study and the average sample size.

Implementing and benchmarking ROSA

Using our ROSA procedure, we randomly select $J = 1000$ training scenarios using LHS and conduct $M = 500$ Monte Carlo simulations for each of the J training scenarios to obtain estimates of the operating characteristics across Θ'. Here Θ' is a product space with the enrollment rate $e \in (0.2, 1)$, the response rates $p_a \in (0.2, 0.4)$ for Y in $A = a$, the response rates $q_a \in (0.2, 0.4)$ for S in $A = a$, and the correlation between Y and S in $A = a, \rho_a \in (0, 0.6)$. For Θ'_{re}, we fix the enrollment rate $e = 0.5$ and the response rates $p_0 = q_0 = 0.3$ in the control groups.

We use a NN to obtain an interpolation of the operating characteristics. As described in Algorithm 4, to evaluate if the estimates of the operating characteristics are accurate, we compare them to independent Monte Carlo estimates of size M = 100,000 on a set of $J' = 200$ uniformly-distributed validation points spanning the plausible parameter space Θ'. The coefficient of determination R^2 in this comparison is 0.96, suggesting that the NN accurately estimates the operating characteristics.

We compare two simulation reports, and our goal is to provide stakeholders the simplified version if it accurately describes the operating characteristics. The first one includes scenarios from $\Theta' \subset \mathbb{R}^7$ restricted by prior knowledge from completed studies and clinical experience and the second includes scenarios from $\Theta'_{re} \subset \Theta'$ further restricted by fixing the value of some entries of θ as described above. We use simulated annealing to identify two sets of scenarios in Θ'_{re} and Θ', respectively. In both cases we minimize the same loss function $\mathcal{L}$ defined over K-tuples of Θ' points. We also calculate the loss $\mathcal{L}$ associated with these two optimal sets of scenarios from Θ' and Θ'_{re}. In Figure 3.3, we illustrate the difference in loss $\mathcal{L}$ between these two optimal sets; as expected, the loss $\mathcal{L}$ decreases as K increases.

We observe in Figure 3.3 that for any value of K, the loss $\mathcal{L}$ associated with the optimal set of scenarios restricted to Θ'_{re} is larger compared to the optimal scenarios in Θ'. However, the difference is modest, and the gain in interpretability of a sensitivity analysis report with fewer unknown parameters may be worth the slightly larger loss. For example, if an investigator requires the loss to be under a threshold of $\mathcal{L} = 0.2$, then it is sufficient to consider $K = 10$ scenarios, regardless of whether we consider scenarios selected from Θ' or Θ'_{re}.

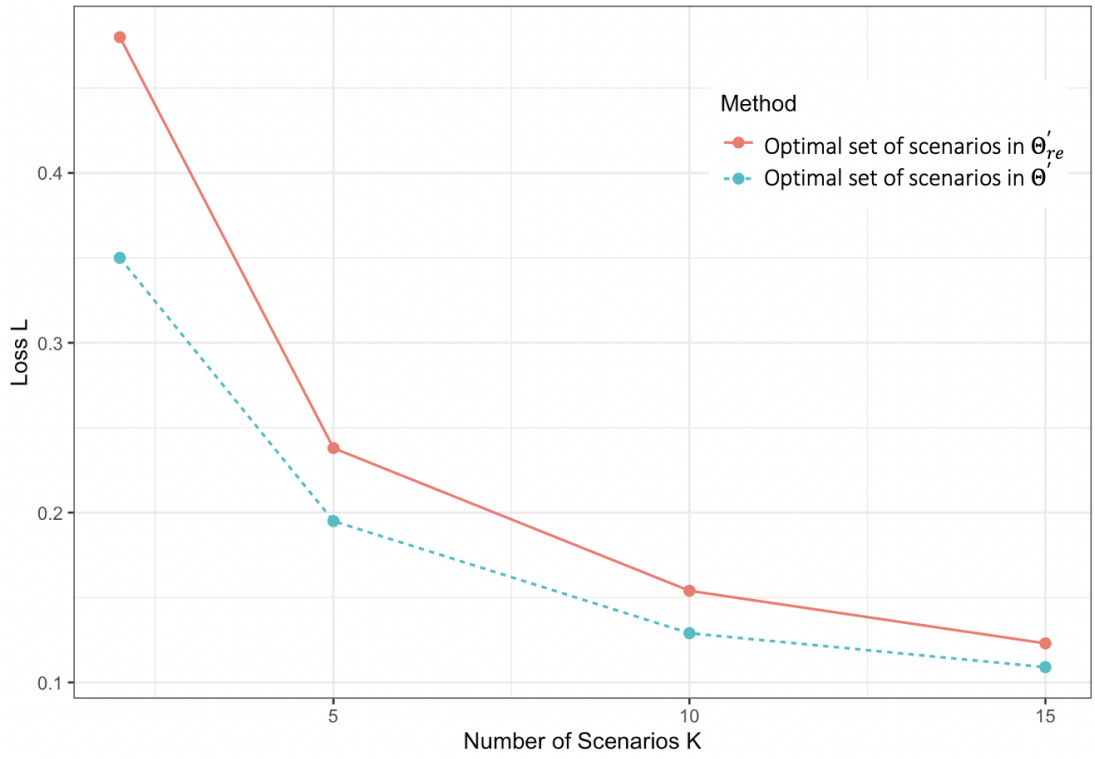


Figure 3.3: Clinical trial design with an interim analysis and an auxiliary endpoint. A graphical representation to choose the number of sensitivity scenarios $K \in \{2, 5, 10, 15\}$. We compare optimal sets of scenarios selected from $\Theta' \subset \mathbb{R}^7$ and from the lower-dimensional restriction $\Theta'_{re} \subset \Theta'$.

3.3.3 Application 3: Biomarker-driven adaptive enrichment

In the third example, we discuss sensitivity analyses dedicated to an adaptive trial with sub-populations defined by biomarkers, considering multiple unknown parameters and multiple operating characteristics of interest. As a motivating example, in several oncology trials, a major decision is whether to restrict patient enrollment to a targeted subgroup of patients (e.g., biomarker-positive subgroup) or to enroll a broader patient population. Enrolling only a biomarker-positive subgroup may deny a substantial number of patients access to an effective therapy, whereas enrolling a larger population may compromise the power to detect positive treatment effects. Several trial designs discussed in the literature attempt to address the outlined problem through interim looks at the data.

Trial design

We consider an adaptive two-stage enrichment trial design with one-to-one randomization[62,64,78]. The design is applicable in the setting where a biomarker-positive subgroup of patients is hypothesized to benefit more from the experimental treatment than the rest of the study population. The design includes a single interim analysis, and it uses progression-free survival (PFS) for interim decision-making, while overall survival (OS) is the endpoint for the final analysis, which occurs when a pre-specified number of events is reached. The interim analysis uses the estimated PFS hazard ratio (HR) to capture potential early signals of treatment effects. In the implementation of[62], which we replicate, the HR is estimated for both the overall population ($\hat{\theta}_{HR}$) and the biomarker-positive subgroup ($\hat{\theta}^{+}_{HR}$). An interim decision determines which group is enrolled and tested during the second stage of the trial:

A – `Promising results in the biomarker-positive population.` If the HR estimate $\hat{\theta}^{+}_{HR} < 0.6$ but $\hat{\theta}_{HR} \geq 0.8$, then the trial will continue enrolling only biomarker-positive patients and the final analysis will test H_0^+. Here H_0^+ is the null hypothesis of no differences in OS between treatment and control groups in the biomarker-positive population. The null hypothesis is rejected if $\omega_1 \Phi^{-1}(1-$

$p_1^+) + \omega_2 \Phi^{-1}(1 - p_2^+) < 1.96$, where p_1^+ (p_2^+) is a log-rank p-value computed using only OS data from patients randomized during the first (second) stage of the trial. The weights (ω_1, ω_2) and the standard normal cumulative distribution function Φ are used to summarize evidence of treatment effects from the two stages of the trial. We refer to[62] for details on the choice of (ω_1, ω_2) and other aspects of the final analysis.

B – `Promising results in the overall population only.` If $\hat{\theta}_{HR}^+ \geq 0.6$ but $\hat{\theta}_{HR} < 0.8$, then the trial will continue enrolling all patients and the final analysis will only test H_0^O, the null hypothesis of no differences in OS in the overall population. In this case the null hypothesis is tested using stage-specific OS log-rank p-values (p_1^O, p_2^O) and combining evidence from the two stages of the trial.

C – `Unpromising results.` If $\hat{\theta}_{HR}^+ \geq 0.6$ and $\hat{\theta}_{HR} \geq 0.8$, then the trial stops early for futility.

D – `Promising early results for both populations.` Lastly, if the estimated HR in the biomarker-positive subgroup $\hat{\theta}_{HR}^+ < 0.6$ and the overall population $\hat{\theta}_{HR} < 0.8$, then the trial will continue enrolling all patients and testing efficacy both in the overall population and in the biomarker-positive subgroup.

The potential conclusion at the final analysis are (i) to recommend the new treatment for biomarker-positive patients, (ii) recommend the new treatment for both biomarker-positive and biomarker-negative patients, or (iii) not recommend the experimental treatment for future patients.

Aims of the sensitivity analysis

We focus on the following three operating characteristics: (i) f_1, the probability of enrolling only biomarker-positive patients in the second stage, (ii) f_2, the probability of enrolling both biomarker-positive and biomarker-negative patients in the second stage, and (iii) f_3, the probability of no evidence of positive treatment effects, which is equal to the probability of not rejecting the null hypotheses.

We choose plausible intervals for the unknown parameters based on prior literature. Specifically, the recruitment rate $\theta_1 \in (0.5, 1)$ per week, the prevalence of the biomarker-positive subgroup $\theta_2 \in$

$(0.15, 0.25)$, the PFS HR comparing the treatment and control groups in the biomarker-positive subgroup $\theta_3 \in (0.5, 1.2)$, the PFS HR comparing treatment and control in the biomarker-negative subgroup $\theta_4 \in (0.6, 1.2)$, the OS HR comparing treatment and control in the biomarker-positive subgroup $\theta_5 \in (0.7, 1.2)$, the OS HR comparing treatment and control groups in the biomarker-negative subgroup $\theta_6 \in (0.8, 1.2)$, the correlation between OS and PFS in the biomarker-positive subgroup $\theta_7 \in (0.3, 0.6)$, and the correlation between OS and PFS in the biomarker-negative subgroup $\theta_8 \in (0.2, 0.7)$. Marginal exponential distributions using a mixture representation were used for simulating correlated OS and PFS times[79]. More flexible models such as the Weibull distribution can be considered.

Implementing and benchmarking ROSA

For the outlined two-stage trial with biomarker populations, our ROSA pipeline can be used to compute multiple simulation reports, varying both the list of operating characteristics f and the definition of Θ'. For example, one can fix the OS HRs in the biomarker-positive and negative populations to focus on the design sensitivity to other parameters, such as the PFS HRs. Similarly, the set of unknown parameters Θ' can be restricted to θ values with positive effects only for the biomarker-positive population. Importantly, one set of training simulations can be re-utilized to compute multiple sensitivity tables where the definitions of f and Θ' vary.

We examine the difference in the marginal losses

$$\mathcal{L}_r(\theta_1, ..., \theta_K) = \max_{\theta \in \Theta} \left\{ \min_{k=1,...,K} ||f_r(\theta) - f_r(\theta_k)||_2 \right\}, \quad 1 \leq r \leq R, \tag{3.4}$$

when the set of scenarios are chosen by optimizing different loss functions. For example, let $\mathcal{S}_r$ be the set of scenarios that minimize the marginal loss $\mathcal{L}_r$ in (3.4). Similarly, let $\mathcal{S}$ be the set of scenarios that minimize the joint loss $\mathcal{L} = -\mathcal{U}$ in (3.2). Then it is intuitive that $\mathcal{L}_r(\mathcal{S}_r) \leq \mathcal{L}_r(\mathcal{S}), 1 \leq r \leq R$.

In different words, the marginal losses $\mathcal{L}_r$ tend to be smaller when the set of scenarios is chosen to minimize $\mathcal{L}_r$ compared to a set of scenarios that minimizes $\mathcal{L}$ with the aim of representing multiple operating characteristics. If the discrepancy $\mathcal{L}_r(\mathcal{S}_r) - \mathcal{L}_r(\mathcal{S}), 1 \leq r \leq R$, is relatively small for all R total operating characteristics, then this indicates that it is reasonable to select a single set of scenarios $\mathcal{S}$ to illustrate how the R operating characteristics vary jointly across Θ. We describe the difference between the marginal losses $\mathcal{L}_r$, $r = 1, 2, 3$, when scenarios $\theta_1, \ldots, \theta_K$ in Θ' are chosen by optimizing $\mathcal{L}_r$ in (3.4) – optimum: $\mathcal{S}_r = \theta^*_{1,r}, \ldots, \theta^*_{K,r}$ – or by optimizing $\mathcal{L}$ as in (3.2) – optimum: $\mathcal{S} = \theta^*_1, \ldots, \theta^*_K$. Recall that $\mathcal{S}$ is computed with the goal of illustrating how multiple operating characteristics vary across Θ' while $\mathcal{S}_r$ optimizes the representation of a single operating characteristic f_r. The weights in (3.2) are $w_1 = w_2 = w_3 = 1/3$. In Figure 3.4 panel 1, we plot $\mathcal{L}_1(\mathcal{S}_1)$ in red and $\mathcal{L}_1(\mathcal{S})$ in blue. Similarly, in panel 2 we compare $\mathcal{L}_2(\mathcal{S}_2)$ and $\mathcal{L}_2(\mathcal{S})$, and in panel 3 we compare $\mathcal{L}_3(\mathcal{S}_3)$ and $\mathcal{L}_3(\mathcal{S})$. Our results indicate that for all three operating characteristics, $\mathcal{L}_r(\mathcal{S}) > \mathcal{L}_r(\mathcal{S}_r)$, $r = 1, 2, 3$; as expected, there is an increase of the marginal losses $\mathcal{L}_r$ when the set of scenarios is selected to illustrate jointly the variations of multiple operating characteristics across Θ'. However, this difference is small ($< 10\%$) for all $K \in \{2, 5, 10, 15\}$. Furthermore, for each $K \in \{2, 5, 10, 15\}$, the relative difference is similar across the three operating characteristics f_1, f_2, f_3 (Figure 3.4). This result supports the use of identical weights and of a single sensitivity table, with the same set of scenarios $\mathcal{S}$ to illustrate jointly all three operating characteristics.

3.4 Discussion

The evaluation of complex designs such as dose-finding studies[59], factorial trials[42], and response-adaptive trials[86] focuses on multiple operating characteristics, such as the level of toxicities, the probability of selecting the correct treatment arm, or frequentist operating characteristics, including power and false positive probabilities. During the design stage of a complex clinical trial, simulation reports

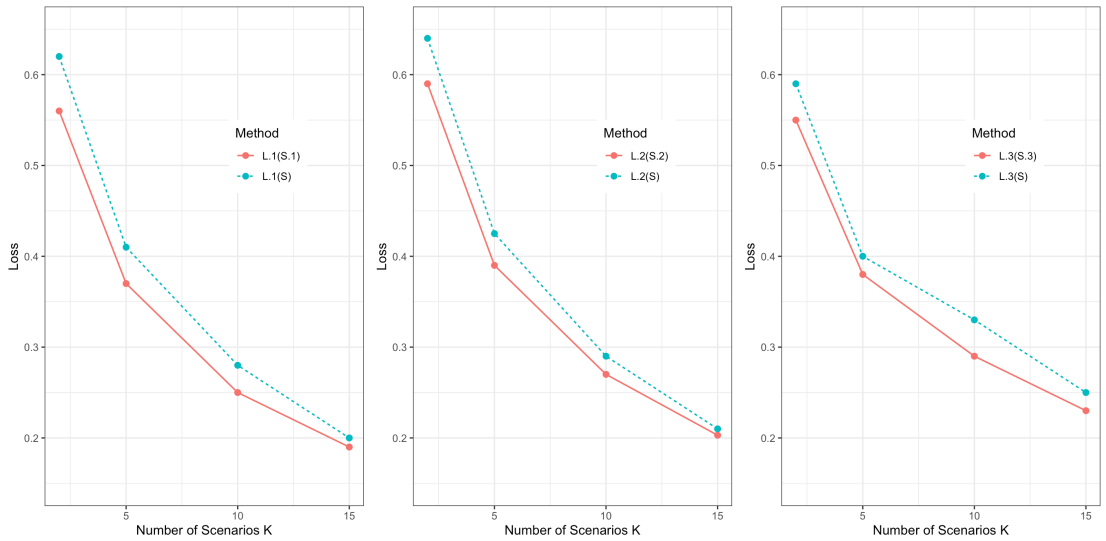


Figure 3.4: Marginal losses $\mathcal{L}_r$, $r = 1, 2, 3$ of different sets of scenarios $\mathcal{S}_r$ (red) and $\mathcal{S}$ (blue).

are typically produced to discuss sample size, interim analyses, and other major decisions with various stakeholders. The simulation report consists of one or a few tables dedicated to showcasing how major operating characteristics $\mathbf{f}(\theta)$ vary across potential values of unknown parameters in Θ. In most cases, the analyst focuses on subsets of plausible parameters $\Theta' \subset \Theta$, for example, values concordant with previous studies, or subsets of potential θ values of particular interest because of positive and clinically relevant treatment effects.

Simulations are fundamental in the design of complex trials since operating characteristics can rarely be obtained analytically and are crucial in the assessment of study designs for regulators, pharmaceutical companies and other stakeholders[37]. However, a limited number of scenarios or poorly chosen scenarios could be inadequate to highlight variations of the operating characteristics across plausible unknown parameters and can result in sub-optimal decisions. We propose ROSA as a useful tool that can support investigators at this design stage when selecting which and how many scenarios to include in these simulation reports.

We focus on choosing an informative number K of scenarios $\theta_1, \ldots, \theta_K$ among the plausible un-

known parameters to summarize the variations of key operating characteristics. Our approach minimizes an explicit loss function and uses established techniques for functional approximation (NNs) and numerical optimization (simulated annealing). We showcase our approach in three trials. Importantly, our approach is general and can be applied to nearly any clinical trial design. It only requires simulations to mimic the clinical trial under hypothetical scenarios.

Although our approach is general, we focused on loss functions $\mathcal{L}$ of a specific form (3.2). It is possible to consider different loss functions. For example, one could consider the loss function $\tilde{\mathcal{L}}(\theta_1, ..., \theta_K) = E_{\theta' \sim g(\cdot)} \left\{ \min_{k=1,...,K} D[\mathbf{f}(\theta'), \mathbf{f}(\theta_k)] \right\}$, where $g(\cdot)$ is a probability distribution on Θ (e.g., a posterior distribution obtained from previous data). The distribution g could be used to incorporate prior information about the unknown parameters in the selection of sensitivity scenarios. Moreover, the metric $D : \Theta^2 \to \mathbb{R}$ can be extended to capture both differences between operating characteristics at plausible values $\theta, \theta' \in \Theta$ and other aspects, such as the difference between expected values of the outcomes Y at θ and θ'.

One major challenge in the presentation of simulation reports is the need for simplicity and interpretability of the results. To this end, we considered fixing one or more unknown parameters to identical values across the K scenarios, which may be reasonable when there is a priori knowledge of certain unknown parameters. There are other ways to simplify a simulation report, such as removing operating characteristics that do not vary across plausible unknown parameters, or reporting only the range of the operating characteristics across Θ instead of presenting the operating characteristics for each representative scenario. Further, instead of maximizing the utility for a given number of scenarios and operating characteristics, future work can consider a penalty for using too many scenarios K.

Variations of the ROSA approach may also consider optimization algorithms other than simulated annealing and regression methods alternative to NN for approximating the operating characteristics across Θ.

Acknowledgements

The authors thank Cyrus Mehta and Christina Howe for helpful conversations and feedback that greatly enhanced the paper. Larry Han was supported by the Clinical Orthopedic and Musculoskeletal Education and Training (COMET) Program, NIAMS grant T32 AR055885. Lorenzo Trippa was supported by NIH grant R01LM013352.

Supplementary Material

The supplementary material includes a table of notation used in the paper.

4

Conclusion

In the present dissertation, we have considered how to make causal inferences in the presence of real world constraints, such as confounding bias, heterogeneity in data distributions, treatment guidelines, and coding practices, as well as privacy constraints that preclude the sharing of patient-level data. In this section, we will summarize the key contributions of this dissertation and detail some of the ongoing research that extends the current work.

4.1 Surrogate Markers and Semi-Supervised Learning

Surrogate markers are outcome measures that can be used as a substitute for a primary outcome. Changes caused by a therapy on a surrogate marker are expected to reflect changes in the primary outcome. The use of valid surrogate markers to infer treatment effects on long-term outcomes has the potential to reduce cost, expedite the approval of new therapies, and potentially reduce the invasiveness of procedures for patients. While there is rich literature on quantifying the effectiveness of a single surrogate marker in an RCT, there is a need for methods development to properly leverage RWD and to consider multiple surrogate markers when they are available.

In Chapter 1, we develop a method to use RWD (e.g., EHR and cross-trial data) to identify and validate surrogates in comparative effectiveness studies. We propose inverse probability weighted and doubly robust estimators for an optimal transformation function of the surrogate and the proportion of treatment effect explained (PTE) measure. We show that our proposed PTE measure avoids the surrogate paradox since it is never the case that the treatment effect on the surrogate is positive but the treatment effect on the primary outcome is negative, regardless of the correlation between the surrogate and primary outcome. We establish the consistency, asymptotic normality, and robustness of the proposed estimators. In two different data applications, we validate two surrogate markers for outcomes of interest in inflammatory bowel disease. These findings may be particularly useful in informing future cross-trial designs for biologic therapies. When there is a single surrogate, one can make inference about the optimal transformation function and the PTE measure nonparametrically, but this is not possible with multiple surrogates due to the curse of dimensionality. In a related research project, we develop a robust calibrated model fusion approach to allow for the incorporation of multiple surrogate markers. Our approach identifies an optimal combination of multiple surrogates without strictly relying on parametric assumptions while borrowing modeling strategies to avoid fully nonparametric estimation. In an analysis using data from the Diabetes Prevention Program study, we

show that it is beneficial to use three surrogates jointly to infer the treatment effect on the primary outcome, compared to any single surrogate individually. In another piece of work, we examine surrogacy from a principal stratification framework. Existing approaches cannot identify the ATE in all principal strata, defined on the joint potential values of the intermediate variables. For example, under non-compliance, existing methods for instrumental variable estimation and principal scores can identify the ATE in at most three of the four principal strata. We propose a new principal resampling technique for unbiased estimation of the ATE and proportions of each stratum without requiring the deterministic monotonicity assumption.

Estimating the time-specific risk of disease onset is difficult due to the presence of censoring and because survival outcomes are imperfectly measured in patient EHRs. Convenient surrogates of disease onset based on ICD-10 codes often exhibit temporal biases of the true event time and can result in power loss and invalid inference on treatment effects. In extensions to Chapter 1, we are developing semi-supervised estimation procedure for the ATE at a time point t, where there exists a small subset of patients for whom gold-standard outcome labels are available and a large subset of patients for whom silver-standard EHR-derived surrogates are available. We develop doubly robust survival curve estimators that are consistent if either the outcome model is correctly specified or the outcome labeling model is correctly specified. Further, we aim to propose a more efficient estimator that can leverage the rich information from EHR surrogates to maximize imputation precision in the unlabeled set. This method can be used to examine, for example, the ability of the influenza vaccine to decrease the time to heart failure in a cardiovascular patient population.

4.2 Federated and Transfer Learning

The growth of large research networks has facilitated multi-center collaborative research, which is particularly important when studying novel treatments, rare diseases, or in times of urgent health

crises. Integrative analysis of data from multiple sources is an important strategy for making more precise, timely, and generalizable decisions. Multi-source analyses can overcome potential biases from a single healthcare system and improve power due to the increased sample size. However, integrative analysis is highly challenging due to heterogeneity in covariate distributions, treatment guidelines, and underlying models across sources, as well as data privacy since individual patient data (IPD) typically cannot be shared.

I have been fortunate to collaborate with investigators in several distributed research networks, including the Consortium for Clinical Characterization of COVID-19 by EHR (4CE) and the U.S. Department of Veterans Affairs. Working closely with clinicians and policymakers at these centers motivated me to study the problem of how to estimate treatment effects in federated data settings. The development of the Federated Adaptive Causal Estimation (FACE) framework allows investigators to incorporate heterogeneous data from multiple sites to provide treatment effect estimation and inference for a flexibly specified target population of interest. To safely incorporate source sites and avoid negative transfer, we introduce an adaptive weighting procedure via a penalized regression of the influence functions, which achieves both consistency and optimal efficiency. FACE is communication-efficient and privacy-preserving, allowing participating sites to share summary statistics only once with other sites. In a comparative effectiveness study of vaccines on COVID-19 outcomes in U.S. veterans using EHRs from five VA sites, we show that FACE substantially increases the precision of treatment effect estimates, with reductions in standard errors ranging from 26% to 67%. In ongoing research, we extend the FACE framework to propose a multiply robust estimator of the target average treatment effect (TATE) to allow researchers at different sites to propose multiple candidate outcome and treatment models due to different local conditions such as treatment guidelines, hospital resources, or patient populations. In a separate follow-up paper, we shift our focus to the small sample problem and focus on how to better estimate treatment effects for underrepresented populations. We develop a federated transfer learning approach to leverage information from different populations and differ-

ent sites. Theoretical and simulation studies show that our proposed method is superior to using data from the underrepresented population alone, robust to model misspecifications, and efficient under relatively mild assumptions. In another extension, we focus on the setting where it is not clear a priori which subgroups are of primary interest; rather, the goal is to identify subgroups who would benefit most from a new treatment, which is important in resource-constrained settings (e.g. COVID-19 vaccines early in the pandemic). To this end, we develop a doubly robust federated causal tree approach to estimate heterogeneous treatment effects. Finally, in a related paper, we develop a causal framework for hospital quality measurement to properly adjust for differences in patient case mixes, identify relevant peer hospitals, and use only summary-level data when IPD cannot be shared. Our strategy allows us to make hospital comparisons on treatment-specific outcomes and permits flexibility in the specification of the target population.

The typical multi-source transfer learning problem leverages data from multiple source sites to help make predictions or causal inference for a target population of interest. Due to the challenge of obtaining accurate labels, there are often very few outcome labels. In ongoing research, we focus on the even more challenging setting where we do not observe any data on the target population, yet we aim to leverage multiple source populations to learn about the target population. This type of problem is encountered frequently in the real world. For example, in genomics, data is often separated into batches but heterogeneity across batches can lead to undesirable variation in the data. The goal in this setting is to identify batches that show low levels of concordance with the majority of the batches and adjust for such differences in downstream analyses. We aim to (i) provide the identification condition for extracting information from many source populations to make inference for an unseen and possibly heterogeneous target population, (ii) develop a general sampling algorithm for overcoming the post-selection problem with current methods, and (iii) implement the method for high-dimensional and low-dimensional prediction and causal inference with multiple data sources. We also aim to continue work on federated survival analysis with high dimensional and heterogeneous data, as well as to

develop federated and transfer learning approaches for conformal inference under covariate shift.

4.3 Sensitivity Analysis

To compare candidate designs for future clinical trials, simulation-based sensitivity analyses are often conducted. In this context, sensitivity analyses are used to assess the dependence of important design operating characteristics (OCs) with respect to various unknown parameters (UPs). In Chapter 3, we proposed a new Representative and Optimal Sensitivity Analysis (ROSA) approach to choose the set of scenarios (and its size) for inclusion in design sensitivity analyses. Our approach balances the need for simplicity and interpretability of OCs computed across several scenarios with the need to faithfully summarize how the OCs vary. We are developing the ROSA strategy into a practical tool for investigators, who might otherwise be unclear if scenarios which are selected ad hoc are adequate to illustrate the variations of the OCs across potential values of the UPs. Similarly, for regulators, this tool can be used to resolve any skepticism as to whether scenarios are chosen to highlight positive aspects of the trial design without pointing at its limitations. We are also interested in extending this framework to observational studies. In observational studies, multiple sources of bias must be accounted for (e.g., selection bias, measurement error, unmeasured confounding, etc.). Existing research has primarily considered single sources of bias, although recent work has shown that it is possible to bound the total composite bias due to multiple sources. We plan to develop more informative sensitivity measures, which can be especially useful when bias bounds are wide.

A

Proofs and Supplemental Materials for Identifying Surrogate Markers in Comparative Effectiveness Research

Overview of Supplementary Materials

The supplementary materials contain six appendices. Appendix A.1 provides a derivation of the optimal transformation function. Appendix A.2 provides a derivation of the bounded PTE measure and avoidance of the surrogate paradox when assumptions (A1) and (A2) are satisfied. Appendix A.3 provides a proof for consistency and asymptotic normality of $\widehat{\text{PTE}}_{\widehat{g}}$. Appendix A.4 proves that our proposed DR estimators are consistent when either the PS model or the OR models are correctly specified. Appendix A.5 provides details on perturbation resampling. Appendix A.6 provides additional figures.

A.1 Derivation of Optimal Transformation

Without loss of generality, assume that S is continuous with conditional densities $\dot{F}_a(s)$, given binary treatment $A = a, a = 0, 1$, with respect to the Lebesgue measure. The derivation is similar when S is discrete. Let $\widetilde{g}_{\text{opt}}(S) = g_{\text{opt}}(S) - m(S)$, where

$$m(s) = m_1(s)\mathcal{P}_1(s) + m_0(s)\mathcal{P}_0(s),$$

where $m_a(s) = E(Y^{(a)}|S^{(a)} = s)$ and $\mathcal{P}_a(s) = f_a(s)(f_0(s)+f_1(s))^{-1}, f_a(s) = dF_a(s)/ds$. Since $g_{\text{opt}}(S) = m(S) + \widetilde{g}_{\text{opt}}(S)$,

$$E\left[\left\{Y - g_{\text{opt}}(S)\right\}^2\right] = E\left[\left\{Y - m(S) - \widetilde{g}_{\text{opt}}(S)\right\}^2\right] = E\left[(Y - m(S))^2\right] - 2E[(Y - m(S))]\widetilde{g}_{\text{opt}}(S) + E[\widetilde{g}^2(S)].$$

Thus the problem is equivalent to finding a function $\widetilde{g}_{\text{opt}}(\cdot)$ that solves the constrained optimization:

$$\min_{\tilde{g}} E\left[\tilde{g}^2(S)\right] \quad \text{given} \quad E[\tilde{g}(S)|A = 0] = c$$

where $c := E[Y - m(S)|A = 0]$.

Our optimization problem is thus

$$\min_{\tilde{g}} \int \tilde{g}^2(s)\left\{f_0(s) + f_1(s)\right\} ds \quad \text{given} \quad \int \tilde{g}(s) dF_0(s) = c$$

which is equivalent to

$$\min_{\tilde{g}} \mathcal{L}(\tilde{g}) \quad \text{given} \quad \mathbb{G}(\tilde{g}) = c$$

where we used the functional notation

$$\mathcal{L}(\tilde{g}) = \int \tilde{g}^2(s)\left\{f_0(s) + f_1(s)\right\} ds, \quad \text{and} \quad \mathbb{G}(\tilde{g}) = \int \tilde{g}(s) dF_0(s).$$

Taking the Frechet derivatives of the functionals, we have that for all measurable h such that $\int h^2(s)\{f_0(s) + f_1(s)\}ds < \infty$,

$$\frac{d}{d\tilde{g}}[\mathcal{L}(\tilde{g}) - \lambda\mathbb{G}(\tilde{g})](h) = \int \widetilde{g}_{\text{opt}}(s)h(s)\left\{f_0(s) + f_1(s)\right\} ds - \lambda \int h(s) dF_0(s) = 0.$$

Setting $h = \delta(s)$, this implies that $\widetilde{g}_{\text{opt}}(s) = \lambda \mathcal{P}_0(s)$ for all s. Hence, by the constraint, we have

$$\lambda = \frac{c}{\int \mathcal{P}_0(s) dF_0(s)} = \frac{\mu_0 - \int m(s) dF_0(s)}{\int \mathcal{P}_0(s) dF_0(s)}.$$

We can simplify the expression for λ by first noting:

$$\begin{aligned}
\mu_0 = E(Y|A=0) &= E(Y^{(0)}) \\
&= \int y f(y|A=0) dy \\
&= \int \int y f(y|A=0, s) f(s|A=0) dy ds \\
&= \int m_0(s) f(s|A=0) ds \\
&= \int m_0(s) dF_0(s),
\end{aligned}$$

where the second-to-last equality follows because $m_0(s) = E(Y|S=s, A=0)$, and the last equality follows because $f(s|A=0) = dF_0(s)/ds$. Thus,

$$\begin{aligned}
\mu_0 - \int m(s) dF_0(s) &= \int m_0(s) dF_0(s) - \int [m_0(s)\mathcal{P}_0(s) + m_1(s)\mathcal{P}_1(s)] dF_0(s) \\
&= \int [m_0(s)(1 - \mathcal{P}_0(s)) - m_1(s)\mathcal{P}_1(s)] dF_0(s) \\
&= \int [m_0(s) - m_1(s)] \mathcal{P}_1(s) dF_0(s).
\end{aligned}$$

Hence,

$$\lambda = \frac{\int [m_0(s) - m_1(s)] \mathcal{P}_1(s) dF_0(s)}{\int \mathcal{P}_0(s) dF_0(s)}.$$

Finally, the optimal function $g_{\text{opt}}(\cdot)$ can be expressed as

$$g_{\text{opt}}(s) = m(s) + \lambda \mathcal{P}_0(s).$$

A.2 Bounded PTE and Avoidance of Surrogate Paradox

To ensure that $PTE \in [0,1]$ and to avoid the surrogate paradox situation, we show that the only assumptions required are that

(A1) $\mathbb{S}_1(u) \geq \mathbb{S}_0(u)$ for all u,

(A2) $\mathbb{M}_1(u) \geq \mathbb{M}_0(u)$ for all u in the common support of $g_{opt}(S^{(1)})$ and $g_{opt}(S^{(0)})$,

where $\mathbb{S}_a(u) = P\{g_{opt}(S^{(a)}) \geq u\}$ and $\mathbb{M}_a(u) = E(Y^{(a)} \mid g_{opt}(S^{(a)}) = u)$, for $a = 0, 1$.

By definition,

$$\Delta = E(Y^{(1)}) - E(Y^{(0)}) = \int \mathbb{M}_1(u) d\mathbb{F}_1(u) - \int \mathbb{M}_0(u) d\mathbb{F}_0(u).$$

Recall that $g_{opt}(s) = m(s) + \lambda \mathcal{P}_0(s)$, and

$$\Delta_{g_{opt}(S)} = E(g_{opt}(S^{(1)}) - g_{opt}(S^{(0)})) = \int g_{opt}(s)\{\dot{F}_1(s) - \dot{F}_0(s)\} ds,$$

where $m(s) = m_1(s)\mathcal{P}_1(s) + m_0(s)\mathcal{P}_0(s)$, $\lambda = \frac{\mu_0 - \int m(s) dF_0(s)}{\int \mathcal{P}_0(s) dF_0(s)}$, and $\mathcal{P}_a(s) = \frac{\dot{F}_a(s)}{\dot{F}_0(s) + \dot{F}_1(s)}$.

We can thus rewrite $\Delta_{g_{opt}(S)}$ as,

$$
\begin{aligned}
\Delta_{g_{opt}(S)} &= \int g_{opt}(s)\left\{\dot{F}_1(s) - \dot{F}_0(s)\right\} ds \\
&= \int m_1(s)\mathcal{P}_1(s)\left\{\dot{F}_1(s) - \dot{F}_0(s)\right\} ds + \int m_0(s)\mathcal{P}_0(s)\left\{\dot{F}_1(s) - \dot{F}_0(s)\right\} ds \\
&+ \frac{\int \mathcal{P}_0(s)\left\{\dot{F}_1(s) - \dot{F}_0(s)\right\} ds}{\int \mathcal{P}_0(s) dF_0(s)} \int \left\{-m_1(s) + m_0(s)\right\} \mathcal{P}_1(s) dF_0(s) \\
&= \int m_1(s)\left[\dot{F}_1(s) - \mathcal{P}_0(s)\dot{F}_1(s) - \mathcal{P}_1(s)\dot{F}_0(s)\frac{\int \mathcal{P}_0(s)\dot{F}_1(s)ds}{\int \mathcal{P}_0(s)dF_0(s)}\right] ds \\
&- \int m_0(s)\left[\dot{F}_0(s) - \mathcal{P}_0(s)\dot{F}_1(s) - \mathcal{P}_1(s)\dot{F}_0(s)\frac{\int \mathcal{P}_0(s)\dot{F}_1(s)ds}{\int \mathcal{P}_0(s)dF_0(s)}\right] ds \\
&:= \int m_1(s)\{dF_1(s) - d\mathcal{F}_{new}(s)\} - \int m_0(s)\{dF_0(s) - d\mathcal{F}_{new}(s)\}
\end{aligned}
$$

where $\mathcal{F}_{new}(s) = \frac{\int_{-\infty}^{s} \mathcal{P}_0(v)dF_1(v)}{\int \mathcal{P}_0(v)dF_0(v)}$. Note that $\mathcal{F}_{new}(\cdot)$ is a subdistribution since $\mathcal{F}_{new}(\infty) = \frac{\int_{-\infty}^{\infty} \mathcal{P}_0(v)dF_1(v)}{\int \mathcal{P}_0(v)dF_0(v)} \leq \frac{\int_{-\infty}^{\infty} \mathcal{P}_0(v)dF_0(v)}{\int \mathcal{P}_0(v)dF_0(v)} = 1$.

We thus have that

$$
\Delta_{g_{opt}} = \int \mathbb{M}_1(u)\{d\mathbb{F}_1(u) - d\mathbb{F}_{new}(u)\} - \int \mathbb{M}_0(u)\{d\mathbb{F}_0(u) - d\mathbb{F}_{new}(u)\}
$$

Examining the difference between Δ and $\Delta_{g_{opt}}$, we see that

$$
\Delta - \Delta_{g_{opt}} = \int \{\mathbb{M}_1(u) - \mathbb{M}_0(u)\}\dot{\mathbb{F}}_{new}(u)du,
$$

where $\mathbb{F}_a(u) = 1 - \mathbb{S}_a(u)$, $\dot{\mathbb{F}}_{new}(u) = \frac{\int_{-\infty}^{u} \mathbb{P}_0(v)d\mathbb{F}_1(v)}{\int \mathbb{P}_0(v)d\mathbb{F}_0(v)}$, $\mathbb{P}_a(v) = \frac{\dot{\mathbb{F}}_a(v)}{\dot{\mathbb{F}}_0(v)+\dot{\mathbb{F}}_1(v)}$, and $\dot{\mathbb{F}}_a(u) = \frac{d\mathbb{F}_a(u)}{du}$.

From an integration by parts, we see that

$$\Delta_{g_{opt}(S)} = \int u d\mathbb{F}_1(u) - \int u d\mathbb{F}_0(u) = \int \{\mathbb{S}_1(u) - \mathbb{S}_0(u)\} du.$$

By assumption (A1), we conclude that $\Delta_{g_{opt}(S)} \geq 0$. Further, since $\dot{\mathbb{F}}_{new}(u) \geq 0$, then the difference $\Delta - \Delta_{g_{opt}(S)} \geq 0$ under assumption (A2). It follows that

$$PTE = \frac{\Delta_{g_{opt}(S)}}{\Delta} \in [0, 1]$$

under assumptions (A1) and (A2).

Furthermore, we have guaranteed that $\Delta_{g_{opt}(S)} = 0$ when $\Delta = 0$, i.e. if there is no treatment effect on the primary outcome, we will not observe a treatment effect on the optimal transformation of the surrogate.

A.3 Consistency and Asymptotic Normality of $\widehat{\text{PTE}}_{\widehat{g}}$

We assume that all components of $(Y, S, A, \mathbf{X})$ are sub-gaussian, the true conditional mean function $\psi^{\dagger}_{a,\mathsf{m}}(s; \mathbf{x}) = E(Y^{(a)} \mid S^{(a)} = s, \mathbf{X} = \mathbf{x})$ and the true conditional density of $S^{(a)} \mid \mathbf{X} = \mathbf{x}$, $\psi^{\dagger}_{a,\mathsf{f}}(s; \mathbf{x})$, are continuously differentiable. We also assume that $S^{(a)}$ has a compact support and that $h = O(n^{-\nu})$ with $\nu \in (1/4, 1/2)$. In this section, we show that when the propensity score model is correctly specified, the proposed IPW kernel smoothed estimators $\widehat{g}(s)$ and $\widehat{\text{PTE}}_{\widehat{g}}$ are consistent for $g_{\mathsf{opt}}(s)$ and $\text{PTE}_{g_{\mathsf{opt}}}$, respectively. We will also show that $\sqrt{n}(\widehat{\text{PTE}}_{\widehat{g}} - \text{PTE}_{g_{\mathsf{opt}}})$ converges in distribution to a normal distribution with mean zero and variance σ^2, which we will derive.

To this end, we first show that $\widehat{m}_a(s)$ and $\widehat{f}_a(s)$ are consistent for $m_a(s)$ and $f_a(s)$, respectively. Without loss of generality, we prove the consistency of $\widehat{m}_a(s) \equiv \widehat{m}_a(s; \widehat{\alpha})$ for $m_a(s) = E(Y^{(a)} \mid S^{(a)} =$

$s) = E\{\psi^{\dagger}_{a,\mathsf{m}}(s;\mathbf{X})\}$, where $\widehat{m}_a(s;\alpha) = \frac{n^{-1}\sum_{i=1}^{n} K_h(S_i - s) Y_i I(A_i = a)/\pi_a(\mathbf{X}_i;\alpha)}{n^{-1}\sum_{i=1}^{n} K_h(S_i - s) I(A_i = a)/\pi_a(\mathbf{X}_i;\alpha)}$.

First, under the correct specification of the PS model, $\widehat{\alpha} \to \alpha_0$ in probability, where α_0 is the true parameter value. Hence

$$\max_i |\widehat{\omega}_{ai} - \omega_{ai}| \leq \sup_{\mathbf{x}} |\pi_a(\mathbf{x};\widehat{\alpha})^{-1} - \pi_a(\mathbf{x};\alpha_0)^{-1}| \to 0$$

in probability, where $\omega_{ai} = I(A_i = a)/\pi_a(\mathbf{X}_i;\alpha_0)$. It then follows from standard theory for non-parametric kernel estimators[75,85] and Taylor series expansions that

$$\begin{aligned}\sup_s |\widehat{m}_a(s;\widehat{\alpha}) - m_a(s)| &\leq \sup_s |\widehat{m}_a(s;\alpha_0) - m_a(s)| + \sup_{s,\alpha:\|\alpha-\alpha_0\|\leq c} \|\widehat{\mathbf{m}}'_a(s;\alpha)\|_2 \|\widehat{\alpha} - \alpha_0\|_2 \\ &= O_p\{(nh)^{-\frac{1}{2}}\sqrt{\log n} + h^2 + n^{-\frac{1}{2}}\} = o_p(1),\end{aligned}$$

where $\widehat{\mathbf{m}}'_a(s;\alpha) = \partial \widehat{m}_a(s;\alpha)/\partial\alpha$ and c is any small constant. Similarly, we have

$$\sup_s |\widehat{f}_a(s) - f_a(s)| = O_p\{(nh)^{-\frac{1}{2}}\sqrt{\log n} + h^2 + n^{-\frac{1}{2}}\} = o_p(1).$$

When $h = O(n^{-\nu})$ with $\nu \in (1/4, 1/2)$, it is not difficult to show that $\widehat{\lambda} - \lambda = O_p(n^{-\frac{1}{2}} + h^2) = O_p(n^{-\frac{1}{2}})$. It follows that

$$\sup_s |\widehat{g}(s) - g(s)| = O_p\{(nh)^{-\frac{1}{2}}\sqrt{\log n} + h^2 + n^{-\frac{1}{2}}\} = o_p(1).$$

Similarly, we may show that

$$|\widehat{\Delta}_{\widehat{g}} - \Delta_{g_{\mathsf{opt}}}| = \widehat{\Delta}_{\widehat{g}} - \widehat{\Delta}_{g_{\mathsf{opt}}} + \widehat{\Delta}_{g_{\mathsf{opt}}} - \Delta_{g_{\mathsf{opt}}} = \int \{\widehat{g}(s) - g_{\mathsf{opt}}(s)\} d\widehat{D}(s) + O_p(n^{-\frac{1}{2}}),$$

where $\widehat{D}(s) = n^{-1}\sum_{i=1}^{n}(\widehat{\omega}_{1i} - \widehat{\omega}_{0i}) I(S_i \leq s)$. It follows from the uniform convergence of $\widehat{g}(s) \to$

$g_{\text{opt}}(s)$ and $\widehat{D}(s) \to D(s) = P(S^{(1)} \le s) - P(S^{(0)} \le s)$ that $\widehat{\Delta}_{\widehat{g}} - \Delta_{g_{\text{opt}}} \to 0$ in probability. This, together with the consistency of $\widehat{\Delta}$ for Δ, implies the consistency of $\widehat{\text{PTE}}_{\widehat{g}}$ for $\text{PTE}_{g_{\text{opt}}}$.

We next establish the asymptotic normality of $\sqrt{n}(\widehat{\text{PTE}}_{\widehat{g}} - \text{PTE}_{g_{\text{opt}}})$. First, note that

$$\widehat{f}_a(s) - f_a(s) = n^{-1}\sum_{i=1}^{n}\left[\omega_{ai}\left\{K_h\left(S_i - s\right) - f_a(s)\right\} + \mathbf{f}'_a(s)\boldsymbol{\mathcal{U}}_{\alpha,i}\right] + o_p((nh)^{-1/2}),$$
$$\widehat{m}_a(s) - m_a(s) = n^{-1}\sum_{i=1}^{n}\left[\omega_{ai}K_h\left(S_i - s\right)\mathcal{U}_{m_a,i}(s) + \mathbf{m}'_a(s)\boldsymbol{\mathcal{U}}_{\alpha,i}\right] + o_p((nh)^{-1/2}),$$

where $\mathbf{m}'_a(s;\alpha) = \partial m_a(s;\alpha)/\partial\alpha, \mathbf{f}'_a(s;\alpha) = \partial f_a(s;\alpha)/\partial\alpha, \mathcal{U}_{m_a,i}(s) = f_a(s)^{-1}\{Y_i^{(a)} - m_a(s)\}$,

$$m_a(s;\alpha) = \frac{E\left\{\psi^{\dagger}_{a,\mathsf{m}}(s;\mathbf{X}_i)\psi^{\dagger}_{a,\mathsf{f}}(s;\mathbf{X}_i)\frac{\pi(\mathbf{X}_i;\alpha_0)}{\pi(\mathbf{X}_i;\alpha)}\right\}}{f_a(s;\alpha)}, \quad f_a(s;\alpha) = E\left\{\psi^{\dagger}_{a,\mathsf{f}}(s;\mathbf{X}_i)\frac{\pi(\mathbf{X}_i;\alpha_0)}{\pi(\mathbf{X}_i;\alpha)}\right\},$$

and $\widehat{\alpha} - \alpha_0 = n^{-1}\sum_{i=1}^{n}\boldsymbol{\mathcal{U}}_{\alpha,i} + o_p(n^{-\frac{1}{2}})$ following standard likelihood theory. It follows that

$$\begin{aligned}
\widehat{\mathcal{P}}_0(s) - \mathcal{P}_0(s) &= \frac{\hat{f}_0(s)f_1(s) - \hat{f}_1(s)f_0(s)}{\{f_1(s) + f_0(s)\}^2} + o_p((nh)^{-1/2}) \\
&= \frac{\mathcal{P}_1(s)\mathcal{P}_0(s)(\hat{f}_0(s) - f_0(s))}{f_0(s)} - \frac{\mathcal{P}_1(s)\mathcal{P}_0(s)(\hat{f}_1(s) - f_1(s))}{f_1(s)} + o_p((nh)^{-1/2}) \\
&= \mathcal{P}_1(s)\mathcal{P}_0(s)n^{-1}\sum_{i=1}^{n}\left[\omega_{0i}\left\{K_h(S_i - s)f_0(s)^{-1} - 1\right\} - \omega_{1i}\left\{K_h(S_i - s)f_1(s)^{-1} - 1\right\}\right. \\
&\left. + \boldsymbol{\mathcal{U}}^{\mathsf{T}}_{\alpha,i}\left\{\frac{\mathbf{f}'_0(s)}{f_0(s)} - \frac{\mathbf{f}'_1(s)}{f_1(s)}\right\}\right] + o_p((nh)^{-1/2}) \\
&= n^{-1}\sum_{i=1}^{n}\left[K_h(S_i - s)\mathcal{G}_{\mathcal{P}_0,i}(s) + (\omega_{1i} - \omega_{0i})\mathcal{P}_1(s)\mathcal{P}_0(s) + \boldsymbol{\mathcal{U}}^{\mathsf{T}}_{\alpha,i}\boldsymbol{\mathcal{B}}_{\mathcal{P}_0}(s)\right] + o_p((nh)^{-1/2}) \\
&= n^{-1}\sum_{i=1}^{n}\mathcal{U}_{\mathcal{P}_0,i} + o_p((nh)^{-1/2}),
\end{aligned}$$

where $\mathcal{U}_{\mathcal{P}_0,i}(s) = K_h(S_i - s)\mathcal{G}_{\mathcal{P}_0,i}(s) + (\omega_{1i} - \omega_{0i})\mathcal{P}_1(s)\mathcal{P}_0(s) + \boldsymbol{\mathcal{U}}^{\mathsf{T}}_{\alpha,i}\boldsymbol{\mathcal{B}}_{\mathcal{P}_0}(s), \mathcal{G}_{\mathcal{P}_0,i}(s) = \{\omega_{0i}/f_0(s) -$

$\omega_{1i}/f_1(s)\}\mathcal{P}_1(s)\mathcal{P}_0(s)$, $\boldsymbol{\mathcal{B}}_{\mathcal{P}_0}(s) = \{\mathbf{f}_0'(s)/f_0(s) - \mathbf{f}_1'(s)/f_1(s)\}\mathcal{P}_1(s)\mathcal{P}_0(s)$.

Similarly, we have $\widehat{\mathcal{P}}_1(s) - \mathcal{P}_1(s) = -(\widehat{\mathcal{P}}_0(s) - \mathcal{P}_0(s)) = -n^{-1}\sum_{i=1}^n \mathcal{U}_{\mathcal{P}_0,i}(s) + o_p(n^{-\frac{1}{2}})$.

Now

$$\begin{aligned}\widehat{m}(s) - m(s) &= \{\widehat{m}_1(s) - m_1(s)\}\mathcal{P}_1(s) + \widehat{m}_1(s)\{\widehat{\mathcal{P}}_1(s) - \mathcal{P}_1(s)\} \\ &\quad + \{\widehat{m}_0(s) - m_0(s)\}\mathcal{P}_0(s) + \widehat{m}_0(s)\{\widehat{\mathcal{P}}_0(s) - \mathcal{P}_0(s)\} \\ &= n^{-1}\sum_{i=1}^n \mathcal{U}_{m,i}(s) + o_p((nh)^{-1/2}),\end{aligned}$$

where

$$\begin{aligned}\mathcal{U}_{m,i}(s) &= \sum_{a=0}^1 \left\{\omega_{ai}K_h(S_i - s)\mathcal{U}_{m_a,i}(s) + \mathbf{m}_a'(s)\boldsymbol{\mathcal{U}}_{\alpha,i}\right\}\mathcal{P}_a(s) + \{m_0(s) - m_1(s)\}\mathcal{U}_{\mathcal{P}_0,i}(s) \\ &= K_h(S_i - s)\mathcal{G}_{m,i}(s) + \boldsymbol{\mathcal{U}}_{\alpha,i}^{\mathsf{T}}\boldsymbol{\mathcal{B}}_m(s) + (\omega_{1i} - \omega_{0i})\mathcal{A}_m(s),\end{aligned}$$

where

$$\mathcal{G}_{m,i}(s) = \sum_{a=0}^1 \omega_{ai}\mathcal{U}_{ma,i}(s)\mathcal{P}_a(s) + \mathcal{G}_{\mathcal{P}_0,i}(s)\{m_0(s) - m_1(s)\},$$

$$\boldsymbol{\mathcal{B}}_m(s) = \sum_{a=0}^1 \mathbf{m}_a'(s)\mathcal{P}_a(s) + \boldsymbol{\mathcal{B}}_{\mathcal{P}_0}(s)\{m_0(s) - m_1(s)\},$$

and

$$\mathcal{A}_m(s) = \mathcal{P}_1(s)\mathcal{P}_0(s)\{m_0(s) - m_1(s)\}.$$

Together with arguments given in Appendix B of Parast et al. [88], the fact that $h = o_p(n^{-1/4})$, and a Taylor series expansion for approximating $\int K_h(S_i - s)H(s)ds$ for any given smooth function H,

where we denote $\mu_{H_0} := \int H(s)dF_0(s)$, we have the following expansion:

$$\begin{aligned}
\int \widehat{m}(s)\widehat{f}_0(s)ds - \mu_{m0} &= \int \{\widehat{m}(s) - m(s)\}f_0(s)ds \\
&+ \int m(s)\left\{\widehat{f}_0(s) - f_0(s)\right\}ds + o_p(n^{-\frac{1}{2}}) \\
&= n^{-1}\sum_{i=1}^{n}\Bigg[\int K_h(S_i - s)\left\{\mathcal{G}_{m,i}(s)f_0(s) + \omega_{0i}m(s)\right\}ds \\
&+ (\omega_{1i} - \omega_{0i})\int \mathcal{A}_m(s)f_0(s)ds \\
&- \omega_{0i}\int m(s)f_0(s)ds + \boldsymbol{\mathcal{U}}_{\alpha,i}^{\mathsf{T}}\int\left\{m(s)\mathbf{f}_0'(s) + \boldsymbol{\mathcal{B}}_m(s)f_0(s)\right\}ds\Bigg] + o_p(n^{-1/2}) \\
&= n^{-1}\sum_{i=1}^{n}\Bigg[\mathcal{G}_{m,i}(S_i)f_0(S_i) + \omega_{0i}m(S_i) \\
&+ (\omega_{1i} - \omega_{0i})\int \mathcal{A}_m(s)f_0(s)ds - \omega_{0i}\int m(s)f_0(s)ds \\
&+ \boldsymbol{\mathcal{U}}_{\alpha,i}^{\mathsf{T}}\int\left\{m(s)\mathbf{f}_0'(s) + \boldsymbol{\mathcal{B}}_m(s)f_0(s)\right\}ds\Bigg] + o_p(n^{-1/2}).
\end{aligned}$$

Similarly, we have

$$\begin{aligned}
\int \widehat{P}_0(s)\widehat{f}_0(s)ds - \mu_{\mathcal{P}_0 0} &= \int\left\{\widehat{P}_0(s) - \mathcal{P}_0(s)\right\}f_0(s)ds \\
&+ \int \mathcal{P}_0(s)\left\{\widehat{f}_0(s) - f_0(s)\right\}ds + o_p(n^{-1/2}) \\
&= n^{-1}\sum_{i=1}^{n}\Bigg[\int K_h(S_i - s)\{\mathcal{G}_{\mathcal{P}_0,i}(s)f_0(s) + \omega_{0i}\mathcal{P}_0(s)\}ds \\
&+ (\omega_{1i} - \omega_{0i})\int \mathcal{P}_1(s)\mathcal{P}_0(s)f_0(s)ds - \omega_{0i}\int \mathcal{P}_0(s)f_0(s)ds \\
&+ \boldsymbol{\mathcal{U}}_{\alpha,i}^{\mathsf{T}}\int\left\{\mathcal{P}_0(s)\mathbf{f}_0'(s) + \boldsymbol{\mathcal{B}}_{\mathcal{P}_0}(s)f_0(s)\right\}ds\Bigg] + o_p(n^{-1/2}) \\
&= n^{-1}\sum_{i=1}^{n}\Bigg[\mathcal{G}_{\mathcal{P}_0,i}(S_i)f_0(S_i) + \omega_{0i}\mathcal{P}_0(S_i) + (\omega_{1i} - \omega_{0i})\int \mathcal{P}_1(s)\mathcal{P}_0(s)f_0(s)ds \\
&- \omega_{0i}\int \mathcal{P}_0(s)f_0(s)ds + \boldsymbol{\mathcal{U}}_{\alpha,i}^{\mathsf{T}}\int\left\{\mathcal{P}_0(s)\mathbf{f}_0'(s) + \boldsymbol{\mathcal{B}}_{\mathcal{P}_0}(s)f_0(s)\right\}ds\Bigg] + o_p(n^{-1/2}).
\end{aligned}$$

Since $\lambda = \frac{\mu_0 - \mu_{m0}}{\mu_{\mathcal{P}_0 0}}$ and $\widehat{\lambda} = \frac{\widehat{\mu}_0 - \widehat{\mu}_{m0}}{\widehat{\mu}_{\mathcal{P}_0 0}}$, it follows from above that

$$\widehat{\lambda} - \lambda = n^{-1} \sum_{i=1}^{n} \mathcal{U}_{\lambda,i} + o_p(n^{-\frac{1}{2}}),$$

where

$$\begin{aligned}
\widehat{\lambda} - \lambda &= \mu_{\mathcal{P}_0 0}^{-1} \left(\widehat{\mu}_0 - \mu_0\right) - \mu_{\mathcal{P}_0 0}^{-1} \lambda \left\{ \int \widehat{\mathcal{P}}_0(s) d\widehat{F}_0(s) - \mu_{\mathcal{P}_0 0} \right\} \\
&\quad - \mu_{\mathcal{P}_0 0}^{-1} \left\{ \int \widehat{m}(s) d\widehat{F}_0(s) - \mu_{m0} \right\} + o_p(n^{-1/2}) \\
&= \mu_{\mathcal{P}_0 0}^{-1} n^{-1} \sum_{i=1}^{n} w_{0i} \left(Y_i - \mu_0\right) \\
&\quad - \mu_{\mathcal{P}_0 0}^{-1} \lambda n^{-1} \sum_{i=1}^{n} \left\{ \mathcal{G}_{\mathcal{P}_0,i}(S_i) f_0(S_i) + \omega_{0i} \mathcal{P}_0(S_i) \right. \\
&\quad \left. + (\omega_{1i} - \omega_{0i}) \int \mathcal{P}_1(s) \mathcal{P}_0(s) f_0(s) ds - \omega_{0i} \int \mathcal{P}_0(s) f_0(s) ds \right\} \\
&\quad + \mu_{\mathcal{P}_0 0}^{-1} \lambda n^{-1} \sum_{i=1}^{n} \boldsymbol{\mathcal{U}}_{\alpha,i}^{\mathsf{T}} \int \left\{ \mathcal{P}_0(s) \mathbf{f}_0'(s) + \boldsymbol{\mathcal{B}}_{\mathcal{P}_0}(s) f_0(s) \right\} ds \\
&\quad - \mu_{\mathcal{P}_0 0}^{-1} n^{-1} \sum_{i=1}^{n} \left\{ \mathcal{G}_{m,i}(S_i) f_0(S_i) + \omega_{0i} m(S_i) \right. \\
&\quad \left. + (\omega_{1i} - \omega_{0i}) \int \mathcal{A}_m(s) f_0(s) ds - \omega_{0i} \int m(s) f_0(s) ds \right\} \\
&\quad + \mu_{\mathcal{P}_0 0}^{-1} n^{-1} \sum_{i=1}^{n} \boldsymbol{\mathcal{U}}_{\alpha,i}^{\mathsf{T}} \int \left\{ m(s) \mathbf{f}_0'(s) ds + \boldsymbol{\mathcal{B}}_m(s) f_0(s) \right\} ds + o_p(n^{-1/2}) \\
&= n^{-1} \sum_{i=1}^{n} \mathcal{U}_\lambda(\mathbf{D}_i) + o_p(n^{-1/2}).
\end{aligned}$$

Gathering the above expansions, we may obtain the form of $\widehat{g}(s) - g_{\text{opt}}(s)$ as

$$\begin{aligned}
\widehat{g}(s) - g_{\text{opt}}(s) &= \widehat{m}(s) - m(s) + (\widehat{\lambda} - \lambda)\mathcal{P}_0(s) + \lambda\left\{\widehat{\mathcal{P}}_0(s) - \mathcal{P}_0(s)\right\} + o_p((nh)^{-1/2}) \\
&= n^{-1}\sum_{i=1}^{n}\Big[K_h(S_i - s)\mathcal{G}_{m,i}(s) + \boldsymbol{\mathcal{U}}_{\alpha,i}^{\mathsf{T}}\boldsymbol{\mathcal{B}}_m(s) + (\omega_{1i} - \omega_{0i})\mathcal{A}_m(s) + \mathcal{P}_0(s)\mathcal{U}_\lambda(\mathbf{D}_i) \\
&\quad + \lambda\{K_h(S_i - s)\mathcal{G}_{\mathcal{P}_0,i}(s) + (\omega_{1i} - \omega_{0i})\mathcal{P}_1(s)\mathcal{P}_0(s) + \boldsymbol{\mathcal{U}}_{\alpha,i}^{\mathsf{T}}\boldsymbol{\mathcal{B}}_{\mathcal{P}_0}(s)\}\Big] + o_p((nh)^{-1/2}) \\
&= n^{-1}\sum_{i=1}^{n}\mathcal{U}_G(s;\mathbf{D}_i) + o_p((nh)^{-1/2}),
\end{aligned}$$

where

$$\begin{aligned}
\mathcal{U}_G(s;\mathbf{D}_i) &= K_h(S_i - s)\{\mathcal{G}_{m,i}(s) + \lambda\mathcal{G}_{\mathcal{P}_0,i}(s)\} + \boldsymbol{\mathcal{U}}_{\alpha,i}^{\mathsf{T}}\{\boldsymbol{\mathcal{B}}_m(s) + \lambda\boldsymbol{\mathcal{B}}_{\mathcal{P}_0}(s)\} \\
&\quad + (\omega_{1i} - \omega_{0i})\{\mathcal{A}_m(s) + \lambda\mathcal{P}_1(s)\mathcal{P}_0(s)\} + \mathcal{P}_0(s)\mathcal{U}_\lambda(\mathbf{D}_i).
\end{aligned}$$

To derive the asymptotic distribution for $\widehat{\text{PTE}}$, observe that

$$\begin{aligned}
&\widehat{\text{PTE}} - \text{PTE} \\
&= \int\left\{\widehat{g}(s) - g_{\text{opt}}(s)\right\} d\left\{\widehat{F}_1(s) - \widehat{F}_0(s)\right\} + \int g_{\text{opt}}(s)d\left\{\widehat{F}_1(s) - \widehat{F}_0(s)\right\} - \text{PTE} \\
&= \frac{1}{\Delta}n^{-1}\sum_{i=1}^{n}\Big[(\mathcal{G}_{m,i}(S_i) + \lambda\mathcal{G}_{\mathcal{P}_0,i}(S_i))(f_1(S_i) - f_0(S_i)) + (\omega_{1i} - \omega_{0i})g_{\text{opt}}(S_i) \\
&\quad + \int g_{\text{opt}}(s)\{\mathbf{f}_1'(s) - \mathbf{f}_0'(s)\}\boldsymbol{\mathcal{U}}_{\alpha,i}ds - \text{PTE} + \int\Big\{\boldsymbol{\mathcal{U}}_{\alpha,i}^{\mathsf{T}}\{\boldsymbol{\mathcal{B}}_m(s) + \lambda\boldsymbol{\mathcal{B}}_{\mathcal{P}_0}(s)\} \\
&\quad + (\omega_{1i} - \omega_{0i})\{\mathcal{A}_m(s) + \lambda\mathcal{P}_1(s)\mathcal{P}_0(s)\} + \mathcal{P}_0(s)\mathcal{U}_\lambda(\mathbf{D}_i)\Big\}(f_1(s) - f_0(s))ds\Big] \\
&\quad - \frac{\text{PTE}}{\Delta}n^{-1}\sum_{i=1}^{n}[\omega_{1i}(Y_i - \mu_1) - \omega_{0i}(Y_i - \mu_0)] + o_p(n^{-1/2}) \\
&= n^{-1}\sum_{i=1}^{n}\mathcal{U}_{\text{PTE}}(\mathbf{D}_i) + o_p(n^{-1/2}),
\end{aligned}$$

where

$$
\begin{aligned}
&\mathcal{U}_{\mathrm{PTE}}(\mathbf{D}_i) \\
&\quad = \frac{1}{\Delta}\Bigg[(\mathcal{G}_{m,i}(S_i) + \lambda \mathcal{G}_{\mathcal{P}_0,i}(S_i))(f_1(S_i) - f_0(S_i)) + (\omega_{1i} - \omega_{0i}) g_{\mathrm{opt}}(S_i) \\
&\qquad + \int g_{\mathrm{opt}}(s)\{\mathbf{f}_1'(s) - \mathbf{f}_0'(s)\}\boldsymbol{\mathcal{U}}_{\alpha,i} ds - \mathrm{PTE} + \int \Big\{\boldsymbol{\mathcal{U}}_{\alpha,i}^{\mathsf{T}}\{\boldsymbol{\mathcal{B}}_m(s) + \lambda \boldsymbol{\mathcal{B}}_{\mathcal{P}_0}(s)\} \\
&\qquad + (\omega_{1i} - \omega_{0i})\{\mathcal{A}_m(s) + \lambda \mathcal{P}_1(s)\mathcal{P}_0(s)\} + \mathcal{P}_0(s)\mathcal{U}_\lambda(\mathbf{D}_i)\Big\}(f_1(s) - f_0(s)) ds\Bigg] \\
&\qquad - \frac{\mathrm{PTE}}{\Delta}[\omega_{1i}(Y_i - \mu_1) - \omega_{0i}(Y_i - \mu_0)].
\end{aligned}
$$

Therefore, by the central limit theorem, $\sqrt{n}\left(\widehat{\mathrm{PTE}} - \mathrm{PTE}\right)$ converges in distribution to a normal with mean zero and variance $\sigma^2 = E\left\{\mathcal{U}_{\mathrm{PTE}}\left(\mathbf{D}_i\right)^2\right\}$.

A.4 Double Robustness

In this section, we prove that our proposed DR estimators are consistent when either the PS model or the OR models are correctly specified. Recall that we proposed the DR estimators for $m_a(s)$ and $f_a(s)$ as

$$
\begin{aligned}
\widehat{m}_{a,\mathsf{DR}}(s) &= \frac{\widehat{\mathcal{M}}_{a,\mathsf{DR}}(s)}{\widehat{f}_{a,\mathsf{DR}}(s)}, \\
\widehat{\mathcal{M}}_{a,\mathsf{DR}}(s) &= n^{-1}\sum_{i=1}^{n}\left\{K_h(S_i - s) Y_i \widehat{\omega}_{ai} - (\widehat{\omega}_{ai} - 1)\widehat{\psi}^{\dagger}_{a,\mathsf{m}}(s;\mathbf{X}_i)\widehat{\psi}^{\dagger}_{a,\mathsf{f}}(s;\mathbf{X}_i)\right\}, \\
\widehat{f}_{a,\mathsf{DR}}(s) &= n^{-1}\sum_{i=1}^{n}\left\{K_h(S_i - s)\widehat{\omega}_{ai} - (\widehat{\omega}_{ai} - 1)\widehat{\psi}_{a,\mathsf{f}}(s;\mathbf{X}_i)\right\},
\end{aligned}
$$

where $h = O(n^{-\nu})$ with $\nu \in (1/4, 1/2)$, $\widehat{\psi}_{a,\mathsf{m}}(s;\mathbf{x})$ and $\widehat{\psi}_{a,\mathsf{f}}(s;\mathbf{x})$ are estimators for the conditional mean $\psi^{\dagger}_{a,\mathsf{m}}(s;\mathbf{x})$ and the conditional density $\psi^{\dagger}_{a,\mathsf{f}}(s;\mathbf{x})$, respectively.

We now show that the estimators are consistent if either $\sup_{\mathbf{x}} |\pi_a(\mathbf{x};\widehat{\alpha}) - \pi_a(\mathbf{x})| \to 0$ in probability

or $\sup_{\mathbf{x},s}\{|\widehat{\psi}_{a,\mathsf{m}}(s;\mathbf{x}) - \psi^{\dagger}_{a,\mathsf{m}}(s;\mathbf{x})| + |\widehat{\psi}_{a,\mathsf{f}}(s;\mathbf{x}) - \psi^{\dagger}_{a,\mathsf{f}}(s;\mathbf{x})|\} \to 0$ in probability. Let $\bar{\alpha}$, $\bar{\psi}_{a,\mathsf{m}}(s;\mathbf{x})$, $\bar{\psi}_{a,\mathsf{f}}(s;\mathbf{x})$ denote the respective limits of $\widehat{\alpha}$, $\widehat{\psi}_{a,\mathsf{m}}(s;\mathbf{x})$ and $\widehat{\psi}_{a,\mathsf{f}}(s;\mathbf{x})$ under possible mis-specification of their respective models, $\bar{\pi}_a(\mathbf{x}) = \pi_a(\mathbf{x};\bar{\alpha})$, and $\bar{\omega}_{ai} = I(A_i = a)/\bar{\pi}_a(\mathbf{X}_i)$. Regardless of the adequacy of the models, by the central limit theorem and convergence of kernel smoothed estimators[85], we have that

$$\widehat{\alpha} - \bar{\alpha} = O_p(n^{-\frac{1}{2}})$$

and

$$\sup_{s,\mathbf{x}} |\widehat{\psi}_{a,\mathsf{f}}(s;\mathbf{x}) - \bar{\psi}_{a,\mathsf{f}}(s;\mathbf{x})| + \sup_{s,\mathbf{x}} |\widehat{\psi}_{a,\mathsf{m}}(s;\mathbf{x}) - \bar{\psi}_{a,\mathsf{m}}(s;\mathbf{x})| = o_p(1).$$

When the PS model is correctly specified, $\sup_s |\widehat{m}_a(s) - m_a(s)| + \sup_s |\widehat{f}_a(s) - f_a(s)| \to 0$ in probability as shown in Appendix 3. In addition, the augmentation terms

$$n^{-1}\sum_{i=1}^{n}(\widehat{\omega}_{ai} - 1)\widehat{\psi}_{a,\mathsf{m}}(s;\mathbf{X}_i) = n^{-1}\sum_{i=1}^{n}(\omega_{ai} - 1)\bar{\psi}_{a,\mathsf{m}}(s;\mathbf{X}_i) + O_p(\|\widehat{\alpha} - \alpha_0\|_2)$$

and

$$n^{-1}\sum_{i=1}^{n}(\widehat{\omega}_{ai} - 1)\widehat{\psi}_{a,\mathsf{f}}(s;\mathbf{X}_i) = n^{-1}\sum_{i=1}^{n}(\omega_{ai} - 1)\bar{\psi}_{a,\mathsf{f}}(s;\mathbf{X}_i) + O_p(\|\widehat{\alpha} - \alpha_0\|_2)$$

also converge to o in probability, regardless of the adequacy of the OR models. Therefore, under the correct specification of the PS model,

$$\sup_s |\widehat{m}_{a,\mathsf{DR}}(s) - m_a(s)| + \sup_s |\widehat{f}_{a,\mathsf{DR}}(s) - f_a(s)| \to 0$$

in probability.

We next establish the consistency of the DR estimators when the PS model may be mis-specified

but the OR models are correctly specified. First consider $\widehat{f}_{a,\mathsf{DR}}(s)$, which can be written as

$$\begin{aligned}\widehat{f}_{a,\mathsf{DR}}(s) &= n^{-1}\sum_{i=1}^{n}\left[\{K_h(S_i-s)-\psi^{\dagger}_{a,\mathsf{f}}(s;\mathbf{X}_i)\}\widehat{\omega}_{ai}-(\widehat{\omega}_{ai}-1)\{\widehat{\psi}_{a,\mathsf{f}}(s;\mathbf{X}_i)-\psi^{\dagger}_{a,\mathsf{f}}(s;\mathbf{X}_i)\}+\psi^{\dagger}_{a,\mathsf{f}}(s;\mathbf{X}_i)\right]\\ &= n^{-1}\sum_{i=1}^{n}\left[\{K_h(S_i-s)-\psi^{\dagger}_{a,\mathsf{f}}(s;\mathbf{X}_i)\}\widehat{\omega}_{ai}-(\widehat{\omega}_{ai}-1)\{\widehat{\psi}_{a,\mathsf{f}}(s;\mathbf{X}_i)-\psi^{\dagger}_{a,\mathsf{f}}(s;\mathbf{X}_i)\}\right]\\ &+f_a(s)+O_p(n^{-\frac{1}{2}})\\ &= n^{-1}\sum_{i=1}^{n}\varepsilon_{a,\mathsf{DR}}(s;\mathbf{D}_i)+f_a(s)-n^{-1}\sum_{i=1}^{n}(\bar{\omega}_{ai}-1)\{\widehat{\psi}_{a,\mathsf{f}}(s;\mathbf{X}_i)-\psi^{\dagger}_{a,\mathsf{f}}(s;\mathbf{X}_i)\}+O_p(n^{-\frac{1}{2}}),\end{aligned}$$

where $\varepsilon_{a,\mathsf{f}}(s;\mathbf{D}_i)=\{K_h(S_i-s)-\psi^{\dagger}_{a,\mathsf{f}}(s;\mathbf{X}_i)\}\bar{\omega}_{ai}$.

It follows from uniform convergence of kernel smoothed estimators[85] that

$$\sup_s|n^{-1}\sum_{i=1}^{n}\varepsilon_{a,\mathsf{DR}}(s;\mathbf{D}_i)-E\{\varepsilon_{a,\mathsf{DR}}(s;\mathbf{D}_i)\}|=o_p(1)$$

and

$$E\{\varepsilon_{a,\mathsf{DR}}(s;\mathbf{D}_i)\}=E\left[\bar{\omega}_{ai}\left\{\int K_h(S-s)\psi^{\dagger}_{a,\mathsf{f}}(S;\mathbf{X}_i)dS-\psi^{\dagger}_{a,\mathsf{f}}(s;\mathbf{X}_i)\right\}\right]=O(h^2).$$

This together with $\sup_{s,\mathbf{x}}|\widehat{\psi}_{a,\mathsf{f}}(s;\mathbf{x})-\psi^{\dagger}_{a,\mathsf{f}}(s;\mathbf{x})|=o_p(1)$ implies that $\sup_s|\widehat{f}_{a,\mathsf{DR}}(s)-f_a(s)|=o_p(1)$.

We have a similar consistency result for $\widehat{\mathcal{M}}_{a,\mathsf{DR}}(s)$, where

$$\begin{aligned}\widehat{\mathcal{M}}_{a,\mathsf{DR}}(s) &= n^{-1}\sum_{i=1}^{n}\left\{K_h(S_i-s)Y_i\bar{\omega}_{ai}-(\bar{\omega}_{ai}-1)\psi^{\dagger}_{a,\mathsf{m}}(s;\mathbf{X}_i)\psi^{\dagger}_{a,\mathsf{f}}(s;\mathbf{X}_i)\right\}+o_p(1)\\ &= n^{-1}\sum_{i=1}^{n}\left[\varepsilon_{a,\mathsf{m}}(s;\mathbf{D}_i)+\left\{K_h(S_i-s)\psi^{\dagger}_{a,\mathsf{m}}(s;\mathbf{X}_i)-\psi^{\dagger}_{a,\mathsf{m}}(s;\mathbf{X}_i)\psi^{\dagger}_{a,\mathsf{f}}(s;\mathbf{X}_i)\right\}\bar{\omega}_{ai}\right]\\ &+m_a(s)+o_p(1),\end{aligned}$$

and $\varepsilon_{a,\mathsf{m}}(s;\mathbf{D}_i)=K_h(S_i-s)\{Y_i-\psi^{\dagger}_{a,\mathsf{m}}(s;\mathbf{X}_i)\}\bar{\omega}_{ai}$.

Following the convergence of $\widehat{f}_{a,\mathsf{DR}}(s) \to f_a(s)$, $\widehat{\mathcal{M}}_{a,\mathsf{DR}}(s) \to m_a(s)f_a(s)$, $\widehat{\psi}_{a,\mathsf{m}}(s;\mathbf{x}) \to \psi^{\dagger}_{a,\mathsf{m}}(s;\mathbf{x})$ and $\widehat{\omega}_{ai} \to \bar{\omega}_{ai}$, we arrive at the consistency of $\widehat{m}_{a,\mathsf{DR}}(s)$ to $m_a(s)$ when the PS model may be misspecified but the OR models are correctly specified.

Thus, we get the double robustness properties for $\widehat{f}_{a,\mathsf{DR}}(s)$ and $\widehat{m}_{a,\mathsf{DR}}(s)$.

Since all remaining estimators relevant to $\widehat{g}_{\mathsf{DR}}(s)$ are plug-in estimators that are derived based on $\widehat{m}_{a,\mathsf{DR}}(s)$ and $\widehat{f}_{a,\mathsf{DR}}(s)$, we can conclude the double robustness of $\widehat{g}_{\mathsf{DR}}(s)$ for $g_{\mathsf{opt}}(s)$.

Finally, the PTE will be DR by standard arguments for the conditional mean estimators[98], where we construct a plug-in estimator for $\Delta_{g_{\mathsf{opt}}}$ as $\widehat{\Delta}_{\widehat{g},\mathsf{DR}} = \widehat{\mu}_{1,\widehat{g},\mathsf{DR}} - \widehat{\mu}_{0,\widehat{g},\mathsf{DR}}$, where

$$\widehat{\mu}_{a,g,DR} = n_a^{-1} \sum_{i:A_i=a} \left\{ \frac{g(S_i)}{\hat{\pi}_a(\mathbf{X}_i)} - \frac{I(A_i=a) - \hat{\pi}_a(\mathbf{X}_i)}{\hat{\pi}_a(\mathbf{X}_i)} \hat{\zeta}_{a,g}(\mathbf{X}_i) \right\},$$

where $\hat{\zeta}_{a,g}(\mathbf{x})$ is an estimator for $\zeta_{a,g}(\mathbf{x}) = E(g(S_i^{(a)}) \mid \mathbf{X}_i = \mathbf{x}) = E(g(S_i) \mid A_i = a, \mathbf{X}_i = \mathbf{x})$, and $n_a = \sum_{i=1}^n I(A_i = a), a = 0, 1$. Similarly, we define $\widehat{\Delta}_{\mathsf{DR}} = \widehat{\mu}_{1,\mathsf{DR}} - \widehat{\mu}_{0,\mathsf{DR}}$, where

$$\widehat{\mu}_{a,\mathsf{DR}} = n_a^{-1} \sum_{i:A_i=a} \left\{ \frac{Y_i}{\widehat{\pi}_a(\mathbf{X}_i)} - \frac{I(A_i=a) - \widehat{\pi}_a(\mathbf{X}_i)}{\widehat{\pi}_a(\mathbf{X}_i)} \hat{\zeta}_a(\mathbf{X}_i) \right\},$$

where $\hat{\zeta}_a(\mathbf{x})$ is an estimator for $\zeta_a(\mathbf{x}) = E(Y_i^{(a)} \mid \mathbf{X}_i = \mathbf{x}) = E(Y_i \mid A_i = a, \mathbf{X}_i = \mathbf{x})$.

A.5 Perturbation Resampling

In this section, we provide the detailed inference procedure for both the IPW and DR estimators based on perturbation resampling. Recall that we generate $\{\mathbf{V}^{[b]} = (V_1^{[b]}, ..., V_n^{[b]})^{\intercal}, b = 1, ..., B\}$, which are $n \times B$ independent and identically distributed non-negative random variables from a known distribution with unit mean and unit variance, such as the unit exponential distribution. For the IPW

estimators, for each set of $\mathbf{V} = (V_1, ..., V_n)^\intercal$, we let $\bar{\mathcal{V}}_i = V_i/(n^{-1}\sum_{i=1}^n V_i)$,

$$\widehat{m}_a^*(s) = \frac{\sum_{i=1}^n K_h(S_i - s) Y_i \bar{\mathcal{V}}_i \widehat{\omega}_{ai}^*}{\sum_{i=1}^n K_h(S_i - s)\bar{\mathcal{V}}_i \widehat{\omega}_{ai}^*}, \quad \widehat{f}_a^*(s) = \frac{\sum_{i=1}^n K_h(S_i - s)\bar{\mathcal{V}}_i \widehat{\omega}_{ai}^*}{\sum_{i=1}^n \bar{\mathcal{V}}_i \widehat{\omega}_{ai}^*}, \quad \widehat{\omega}_{ai}^* = \frac{I(A_i = a)}{\pi(\mathbf{X}_i, \widehat{\alpha}^*)},$$

where $\widehat{\alpha}^*$ is obtained by fitting a weighted logistic regression $A_i \sim G\{\alpha^\intercal \Phi(\mathbf{X}_i)\}$ with weights $\{\bar{\mathcal{V}}_i\}$. The perturbed counterparts of $\widehat{m}(\cdot)$, $\hat{\mathcal{P}}_a(\cdot)$ and $\widehat{\lambda}$ are obtained as

$$\widehat{m}^*(s) = \sum_{a=0}^{1} \widehat{m}_a^*(s)\widehat{\mathcal{P}}_a^*(s), \quad \widehat{\mathcal{P}}_a^*(s) = \frac{\widehat{f}_a^*(s)}{\widehat{f}_1^*(s) + \widehat{f}_0^*(s)}, \quad \widehat{\lambda}^* = \frac{\int \{\widehat{m}_0^*(s) - \widehat{m}_1^*(s)\}\widehat{\mathcal{P}}_1^*(s)\widehat{f}_0^*(s)ds}{\int \widehat{\mathcal{P}}_0^*(s)\widehat{f}_0^*(s)ds},$$

respectively. Subsequently, we construct the perturbed counterparts of $\widehat{g}(s)$, $\widehat{\Delta}_{g_{\text{opt}}}$, $\widehat{\Delta}$ and $\widehat{\text{PTE}}$ as $\widehat{g}^*(s) = \widehat{m}^*(s) + \hat{\lambda}^*\widehat{\mathcal{P}}_0^*(s)$ $\widehat{\Delta}_{\widehat{g}^*}^* = \widehat{\mu}_{1,\widehat{g}^*}^* - \widehat{\mu}_{0,\widehat{g}^*}^*$, $\widehat{\Delta}^* = \widehat{\mu}_1^* - \widehat{\mu}_0^*$, and $\widehat{\text{PTE}}_{\widehat{g}^*}^* = \frac{\widehat{\Delta}_{\widehat{g}^*}^*}{\widehat{\Delta}^*}$, where $\widehat{\mu}_{a,g}^* = \frac{\sum_{i=1}^n g(S_i)\bar{\mathcal{V}}_i\widehat{\omega}_{ai}^*}{\sum_{i=1}^n \bar{\mathcal{V}}_i\widehat{\omega}_{ai}^*}$ and $\widehat{\mu}_a^* = \frac{\sum_{i=1}^n Y_i\bar{\mathcal{V}}_i\widehat{\omega}_{ai}^*}{\sum_{i=1}^n \bar{\mathcal{V}}_i\widehat{\omega}_{ai}^*}$.

For the DR estimators, for each set of $\mathbf{V}$, we let

$$\widehat{m}_{a,\text{DR}}^*(s) = \frac{\widehat{\mathcal{M}}_{a,\text{DR}}^*(s)}{\widehat{f}_{a,\text{DR}}^*(s)}, \quad \widehat{\mathcal{M}}_{a,\text{DR}}^*(s) = n^{-1}\sum_{i=1}^n \bar{\mathcal{V}}_i \left\{K_h(S_i - s)Y_i\widehat{\omega}_{ai}^* - (\widehat{\omega}_{ai}^* - 1)\widehat{\psi}_{a,\text{m}}^*(s;\mathbf{X}_i)\widehat{\psi}_{a,\text{f}}^*(s;\mathbf{X}_i)\right\},$$

$$\widehat{f}_{a,\text{DR}}^*(s) = n^{-1}\sum_{i=1}^n \bar{\mathcal{V}}_i \left\{K_h(S_i - s)\widehat{\omega}_{ai}^* - (\widehat{\omega}_{ai}^* - 1)\hat{\psi}_{a,f}^*(\mathbf{X}_i)\right\},$$

where

$$\widehat{\psi}_{a,\text{f}}^*(s,\mathbf{x}) = \frac{\sum_{i=1}^n \bar{\mathcal{V}}_i K_\varsigma(\mathbf{X}_i^\intercal\widehat{\gamma}_a^* - \mathbf{x}^\intercal\widehat{\gamma}_a^*)K_h(S_i - s)}{\sum_{i=1}^n \bar{\mathcal{V}}_i K_\varsigma(\widehat{\gamma}_a^*\mathbf{X}_i - \mathbf{x}^\intercal\widehat{\gamma}_a^*)}, \quad \widehat{\psi}_{a,\text{m}}^*(s,\mathbf{x}) = M\{\widehat{\beta}_a^*(s)^\intercal\vec{\mathbf{x}}\},$$

$\widehat{\gamma}_a^* = \text{argmax}_{\gamma\in\Omega}\{\sum_{i\neq j, A_i=A_j=a}\bar{\mathcal{V}}_i\bar{\mathcal{V}}_j I(\mathbf{X}_i^\intercal\gamma > \mathbf{X}_j^\intercal\gamma)I(S_i > S_j)\}$ and $\widehat{\beta}_a^*(s)$ is the solution to

$$\widehat{\mathbf{U}}_a^*(\beta; s) \equiv n^{-1}\sum_{i=1}^n \bar{\mathcal{V}}_i I(A_i = a)K_h(S_i - s)\vec{\mathbf{X}}_i \left\{Y_i - M(\beta^\intercal\vec{\mathbf{X}}_i)\right\} = 0.$$

We construct the perturbed counterparts of $\widehat{g}_{\mathsf{DR}}(s)$, $\hat{\Delta}_{\mathsf{DR}}$, $\hat{\Delta}_{\widehat{g},\mathsf{DR}}$, and $\widehat{\mathrm{PTE}}_{\widehat{g},\mathsf{DR}}$ respectively as:

$$\widehat{g}^*_{\mathsf{DR}}(s) = \widehat{m}^*_{\mathsf{DR}}(s) + \widehat{\lambda}^*_{\mathsf{DR}}\widehat{\mathcal{P}}^*_{0,\mathsf{DR}}(s), \quad \widehat{\Delta}^*_{\mathsf{DR}} = \widehat{\mu}^*_{1,\mathsf{DR}} - \widehat{\mu}^*_{0,\mathsf{DR}}, \quad \widehat{\Delta}^*_{\widehat{g}^*_{\mathsf{DR}},\mathsf{DR}} = \widehat{\mu}^*_{1,\widehat{g}^*_{\mathsf{DR}},\mathsf{DR}} - \widehat{\mu}^*_{0,\widehat{g}^*_{\mathsf{DR}},\mathsf{DR}},$$

and $\widehat{\mathrm{PTE}}^*_{\widehat{g}^*_{\mathsf{DR}},\mathsf{DR}} = \widehat{\Delta}^*_{\widehat{g}^*_{\mathsf{DR}},\mathsf{DR}}/\widehat{\Delta}^*_{\mathsf{DR}}$, where $\widehat{m}^*_{\mathsf{DR}}(s) = \sum_{a=0}^{1}\widehat{m}^*_{a,\mathsf{DR}}(s)\widehat{\mathcal{P}}^*_{a,\mathsf{DR}}(s)$,

$$\widehat{\lambda}^*_{\mathsf{DR}} = \frac{\int\left\{\widehat{m}^*_{0,\mathsf{DR}}(s) - \widehat{m}^*_{1,\mathsf{DR}}(s)\right\}\widehat{\mathcal{P}}^*_{1,\mathsf{DR}}(s)\widehat{f}^*_{0,\mathsf{DR}}(s)ds}{\int\widehat{\mathcal{P}}^*_{0,\mathsf{DR}}(s)\widehat{f}^*_{0,\mathsf{DR}}(s)ds}, \quad \widehat{\mathcal{P}}^*_{a,\mathsf{DR}}(s) = \frac{\widehat{f}^*_{a,\mathsf{DR}}(s)}{\widehat{f}^*_{0,\mathsf{DR}}(s) + \widehat{f}^*_{1,\mathsf{DR}}(s)},$$

$$\widehat{\mu}^*_{a,\widehat{g}^*_{\mathsf{DR}},\mathsf{DR}} = n^{-1}\sum_{i=1}^{n}\bar{\mathcal{V}}_i\left\{\widehat{g}^*_{\mathsf{DR}}(S_i)\widehat{\omega}^*_{ai} - (\widehat{\omega}^*_{ai}-1)\hat{\zeta}^*_{a,\widehat{g}^*_{\mathsf{DR}}}(\mathbf{X}_i)\right\}, \quad \hat{\zeta}^*_{a,g}(\mathbf{x}) = \int g(s)\widehat{\psi}^*_{a,\mathsf{f}}(s,\mathbf{x})ds,$$

$$\widehat{\mu}^*_{a,\mathsf{DR}} = n^{-1}\sum_{i=1}^{n}\bar{\mathcal{V}}_i\left\{Y_i\widehat{\omega}^*_{ai} - (\widehat{\omega}^*_{ai}-1)\hat{\zeta}^*_{a}(\mathbf{X}_i)\right\}, \quad \hat{\zeta}^*_{a}(\mathbf{x}) = \int\widehat{\psi}^*_{a,\mathsf{m}}(s;\mathbf{x})\widehat{\psi}^*_{a,\mathsf{f}}(s;\mathbf{x})ds.$$

As described in the main text, we typically generate a large number, say $B = 500$, realizations for $\mathbf{V}$ and then obtain B realizations of the perturbed statistics of interest. Standard error estimates and confidence intervals can then be constructed based on empirical quantiles of these realizations.

A.6 Additional Figures

The additional figures referenced in the paper are provided at the end of the Supplementary Materials.

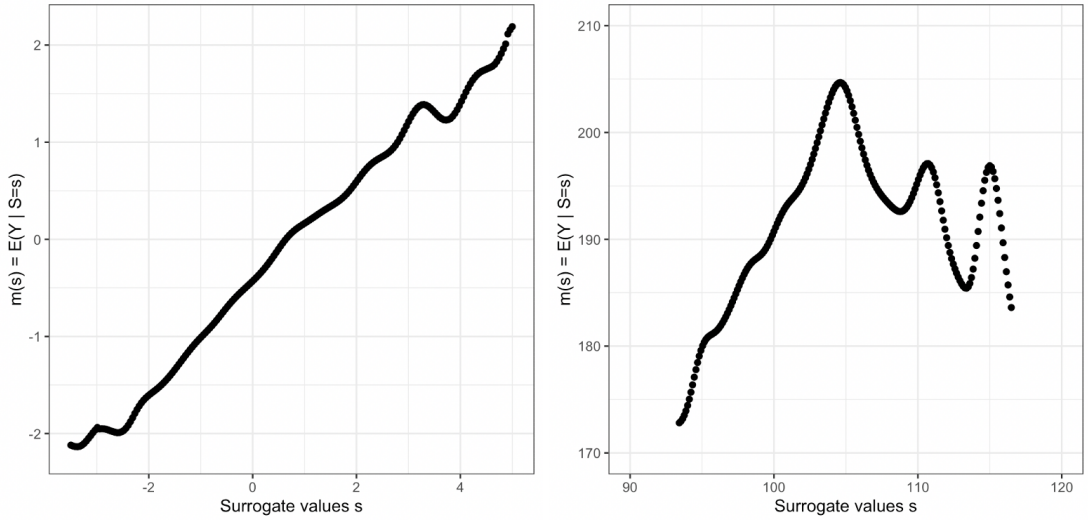


Figure A.1: Relationship between S and $E(Y \mid S = s)$ in setting I (left) and setting II (right)

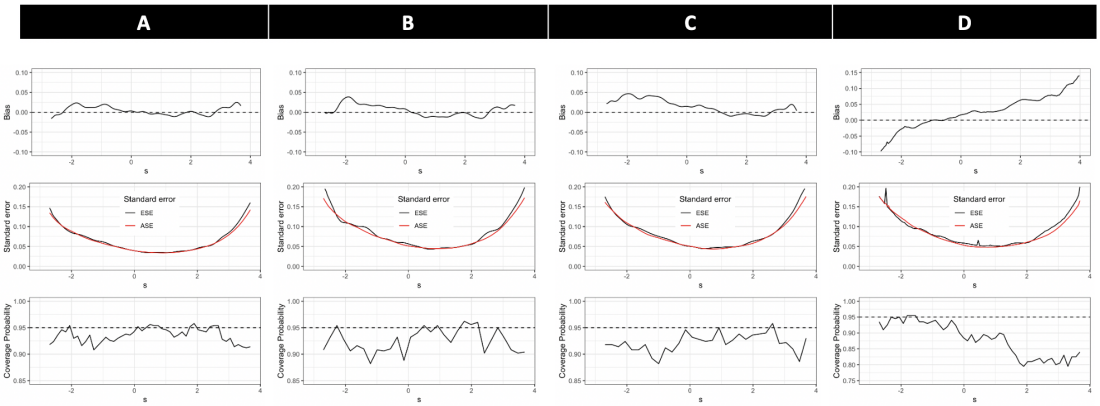


Figure A.2: Empirical bias, empirical standard error (ESE) versus average of the estimated standard error (ASE), and coverage probabilities of the 95% confidence intervals for $\widehat{g}_{\text{opt}}(s)$ when $n = 400$ and (A) both models are correctly specified, (B) PS model is misspecified, (C) OR model is misspecified, (D) both models are misspecified. Note the larger range in the y-axis for setting (D) due to the increased bias and undercoverage when both models are misspecified.

B

Proofs and Supplemental Materials for Federated Adaptive Causal Estimation (FACE) of Target Treatment Effects

Overview of Supplementary Materials

The Supplementary Materials are divided into four sections. In Section A, we illustrate the workflow of FACE to construct a global estimator in a federated data setting. In Section B, we provide a mild set of sufficient conditions for the necessary regularity conditions to hold in the special case with logistic regression models for the nuisance functions. In Section C, we provide proofs for the theoretical results in Section 4 of the main paper. In Section D, we provide supplementary tables corresponding to the real data analysis.

B.1 FACE Workflow

B.2 Special Case: Logistic Regression Models

For the special case with logistic regression models given in Section 2.3.5, we denote the asymptotic parameters as

$$\bar{\alpha}_k = \underset{\alpha \in \mathbb{R}^p}{\arg\min}\, \mathbb{E}\{\ell(A, \alpha^\top \mathbf{X}) \mid R = k\},$$

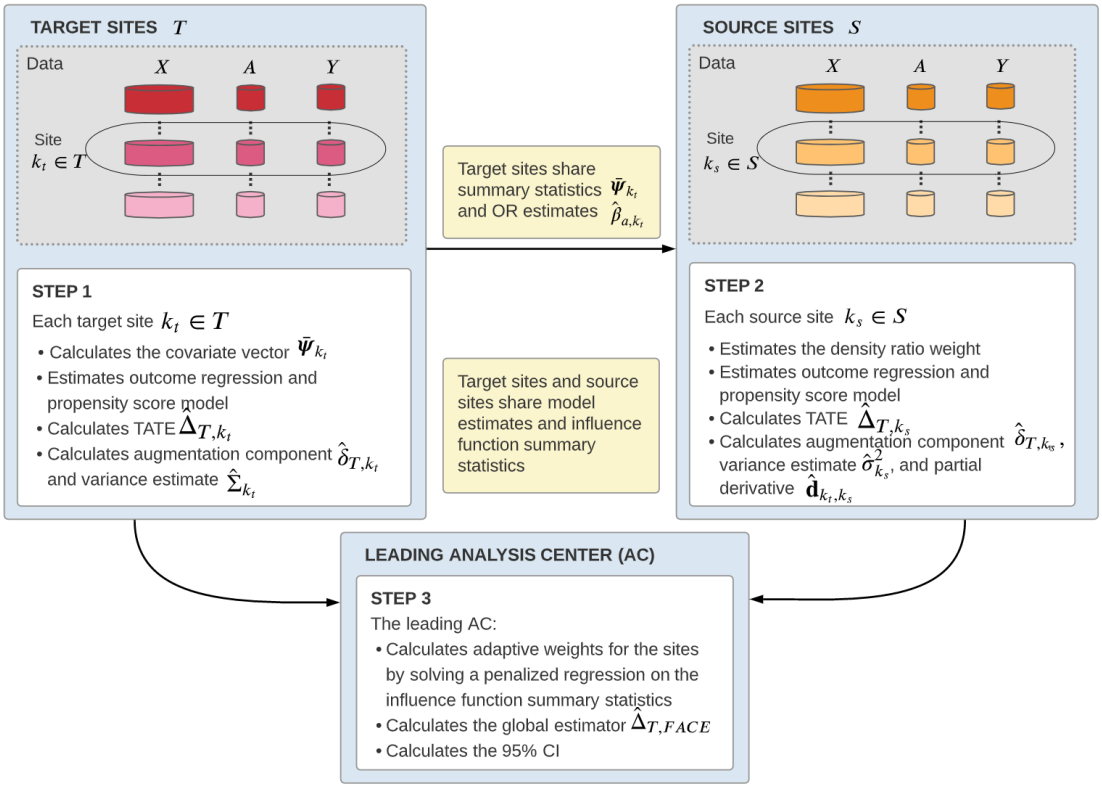


Figure S1: Workflow of FACE to construct a global estimator in a federated data setting

$$\bar{\beta}_{a,k} = \arg\min_{\alpha \in \mathbb{R}^p} \mathbb{E}\{\ell(Y, \alpha^\intercal \mathbf{X}) \mid A = a, R = k\},$$

$$\bar{\gamma}_{k_s} = \arg\min_{\gamma \in \mathbb{R}^q} \mathbb{E}\{\exp(\gamma^\intercal \mathbf{X}) - \gamma^\intercal \mathbb{E}(\mathbf{X} \mid R \in \mathcal{T}) \mid R = k_s\}.$$

We give a mild set of sufficient conditions for Assumption 2.

Assumption 5 *For absolute constants $M, \varepsilon > 0$,*

(a) (Design) $\|\mathbf{X}\|_\infty \leq M$ almost surely, and all eigenvalues of $\mathbb{E}(\mathbf{X}\mathbf{X}^\intercal)$ are in $[\varepsilon, M]$.

(b) (Overlap) For all $k = 1, \ldots, J+K$, $a = 0, 1$ and $i \in \mathcal{I}_k$, $g(\bar{\alpha}_k^\intercal \mathbf{X}_i)$, $g'(\bar{\beta}_{a,k}^\intercal \mathbf{X}_i)$ and $\exp\{\bar{\gamma}_{k_s}^\intercal \mathbf{X}_i\}$ are in $[\varepsilon, 1-\varepsilon]$ almost surely.

(c) (Double robustness) For each target site $k_t \in \mathcal{T}$, at least one of the two models is correctly specified:

-i the PS model is correct: $\mathbb{P}(A = 1 \mid \mathbf{X}, R = k_t) = g(\bar{\alpha}_{k_t}^\intercal \mathbf{X})$;

-ii the OR model is correct: $\mathbb{E}(Y \mid \mathbf{X}, A = a, R = k_t) = g(\bar{\beta}_{a,k_t}^\intercal \mathbf{X})$.

After verifying that Assumptions 1 and 5 imply the generic Assumption 2, we can apply Theorem 1 in that realization.

Corollary 2 *Under the setting of Section 2.3.5 and Assumptions 1 and 5, the FACE estimator is consistent and asymptotically normal with consistent variance estimation $\hat{\mathcal{V}}$,*

$$\sqrt{N/\hat{\mathcal{V}}}\left(\hat{\Delta}_{\mathcal{T},\mathsf{FACE}} - \Delta_{\mathcal{T}}\right) \rightsquigarrow \mathcal{N}(0,1).$$

B.3 Proofs

In this section, we provide proofs for the theoretical statements in the main text. In Sections B.3.1 and B.3.2, we declare and prove the key preliminary results. We then use these results to prove Theorem 1 and Corollary 1 in Section B.3.3, Corollary 2 in Section B.3.4, Proposition 1 in Section B.3.5 and Proposition 2 in Section B.3.6

B.3.1 Double Robustness of $\hat{\Delta}_{\mathcal{T},\mathcal{T}}$ and $\hat{\Delta}_{\mathcal{T},k_s}$

We first establish the consistency and asymptotic normality of the initial TATE estimator $\hat{\Delta}_{\mathcal{T},\mathcal{T}}$ and source site TATE estimator $\hat{\Delta}_{\mathcal{T},k_s}$.

Lemma 1 *Under Assumptions 1, 2(a)-2(c) and 2(e),*

$$\sqrt{N_{\mathcal{T}}}\left(\hat{\Delta}_{\mathcal{T},\mathcal{T}} - \Delta_{\mathcal{T}}\right) \rightsquigarrow \mathcal{N}(0, \sigma^2_{\mathcal{T},\mathcal{T}})$$

with asymptotic variance

$$\sigma^2_{\mathcal{T},\mathcal{T}} = \text{Var}\left(\zeta + \xi_{\mathcal{T}} \mid R \in \mathcal{T}\right).$$

Proof [Proof of Lemma 1]

From the influence function representation in Assumption 2(a)

$$\hat{\Delta}_{\mathcal{T},\mathcal{T}} - \bar{\Delta}_{\mathcal{T},\mathcal{T}} = \frac{1}{N_{\mathcal{T}}} \sum_{k_t \in \mathcal{T}} \sum_{i \in \mathcal{I}_{k_t}} \zeta_i + \xi_{i,\mathcal{T}} + o_p\left(N^{-1/2}\right),$$

where $\bar{\Delta}_{\mathcal{T},\mathcal{T}}$ is the asymptotic limit, and the stable variance in Assumption 2(c)

$$\text{Var}\left(\zeta + \xi_{\mathcal{T}} \mid R \in \mathcal{T}\right) \in [2\varepsilon, 2M],$$

we have the asymptotic normality of $\hat{\Delta}_{\mathcal{T},\mathcal{T}}$

$$\sqrt{N_{\mathcal{T}}}\left(\hat{\Delta}_{\mathcal{T},\mathcal{T}} - \bar{\Delta}_{\mathcal{T},\mathcal{T}}\right) \rightsquigarrow \mathcal{N}(0, \sigma^2_{\mathcal{T},\mathcal{T}}).$$

Under the typical Assumptions 1(a), 1(b), 1(d) and 2(e), the doubly robust estimator $\hat{\Delta}_{\mathcal{T},\mathcal{T}}$ converges to the TATE $\Delta_{\mathcal{T}}$[7]. Thus, we must have $\bar{\Delta}_{\mathcal{T},\mathcal{T}} = \Delta_{\mathcal{T}}$.

■

Lemma 2 *Under Assumptions 1 and 2(a)-2(c),*

$$\sqrt{n_{k_s}}\left(\hat{\Delta}_{\mathcal{T},k_s} - \bar{\Delta}_{\mathcal{T},k_s}\right) \rightsquigarrow \mathcal{N}(0, \sigma^2_{\mathcal{T},k_s})$$

with $\bar{\Delta}_{\mathcal{T},k_s} = \Delta_{\mathcal{T}} - \bar{\delta}_{\mathcal{T},\mathcal{T}} + \bar{\delta}_{\mathcal{T},k_s}$ *and*

$$\sigma^2_{\mathcal{T},k_s} = \mathrm{Var}\left(\xi_{k_s} \mid R = k_s\right) + n_{k_s} \sum_{k_t \in \mathcal{T}} n_{k_t}^{-1} \mathrm{Var}\left\{(\psi(\mathbf{X})^\intercal, v_1^\intercal, v_0^\intercal)\, \bar{\mathbf{d}}_{k_t,k_s} \mid R = k_t\right\}.$$

Additionally under Assumption 3(a), $\bar{\Delta}_{\mathcal{T},k_s} = \Delta_{\mathcal{T}}$ *for* $k_s \in \mathcal{S}'$.

Proof [Proof of Lemma 2] From the influence function representation in Assumption 2(a)

$$\begin{aligned}\hat{\Delta}_{\mathcal{T},k_s} - \bar{\Delta}_{\mathcal{T},k_s} = &\sum_{k_t \in \mathcal{T}} \frac{1}{n_{k_t}} \sum_{i \in \mathcal{I}_{k_t}} \left\{ \frac{n_{k_t}}{N_{\mathcal{T}}} \zeta_i + \left(\psi(\mathbf{X}_i)^\intercal - \mathbb{E}\{\psi(\mathbf{X}) \mid R = k_t\}^\intercal, v_{i,1}^\intercal, v_{i,0}^\intercal\right) \bar{\mathbf{d}}_{k_t,k_s} \right\} \\ &+ \frac{1}{n_{k_s}} \sum_{i \in \mathcal{I}_{k_s}} \xi_{i,k_s} + o_p\left(N^{-1/2}\right)\end{aligned}$$

and the stable variance in Assumption 2(c) $\mathrm{Var}\left(\xi_{i,k_s} \mid R = k_s\right) \in [\varepsilon, M]$ and

$$\mathrm{Var}\left\{ \frac{n_{k_t}}{N_{\mathcal{T}}} \zeta_i + \left(\psi(\mathbf{X}_i)^\intercal, v_{i,1}^\intercal, v_{i,0}^\intercal\right) \bar{\mathbf{d}}_{k_t,k_s} \mid R = k_t \right\} \leq M\left\{\mathbb{P}(R = k_t)^2 + \|\bar{\mathbf{d}}_{k_t,k_s}\|_2^2\right\},$$

we have the asymptotic normality of $\hat{\Delta}_{\mathcal{T},k_s}$

$$\sqrt{N_{\mathcal{T}}}\left(\hat{\Delta}_{\mathcal{T},k_s} - \bar{\Delta}_{\mathcal{T},k_s}\right) \rightsquigarrow \mathcal{N}(0, \sigma^2_{\mathcal{T},k_s}).$$

Similar to $\hat{\Delta}_{\mathcal{T},\mathcal{T}}$, the source site estimator $\hat{\Delta}_{\mathcal{T},k_s}$ is also doubly robust under Assumptions 1 and 3(a).

When the OR model is consistently estimated under Assumption 3(a)(i) (same as Assumption 2(e)-ii) but the density ratio model and PS model may be mis-specified, we have through classical asymptotic analysis

$$\hat{\Delta}_{\mathcal{T},\mathcal{T}} = \sum_{k_t \in \mathcal{T}} \frac{n_{k_t}}{N_{\mathcal{T}}} \left[\frac{1}{n_{k_t}} \sum_{i \in \mathcal{I}_{k_t}} \left\{ m(1, \mathbf{X}_i; \hat{\beta}_{1,k_t}) - m(0, \mathbf{X}_i; \hat{\beta}_{1,k_t}) \right\} \right.$$

$$
\begin{aligned}
&+ \frac{1}{n_{k_s}} \sum_{i \in \mathcal{I}_{k_s}} \omega_{k_t,k_s}(\mathbf{X}_i; \hat{\gamma}_{k_t,k_s}) \frac{(-1)^{1-A_i}}{\pi_{k_s}(A_i, \mathbf{X}_i; \hat{\alpha}_{k_s})} \{Y_i - m(A_i, \mathbf{X}_i; \hat{\beta}_{A_i,k_t})\} \Bigg] \\
&= O_p\left(N^{-1/2}\right) + \underbrace{\sum_{k_t \in \mathcal{T}} \frac{\mathbb{P}(R = k_t)}{\mathbb{P}(R \in \mathcal{T})} \mathbb{E}\{Y^{(1)} - Y^{(0)} \mid \mathbf{X}_i, R = k_t\}}_{= \Delta_{\mathcal{T}}} \\
&\quad + \underbrace{\sum_{k_t \in \mathcal{T}} \frac{\mathbb{P}(R = k_t)}{\mathbb{P}(R \in \mathcal{T})} \mathbb{E}\left[\omega_{k_t,k_s}(\mathbf{X}; \bar{\gamma}_{k_t,k_s}) \frac{(-1)^{1-A}}{\pi_{k_s}(A, \mathbf{X}; \bar{\alpha}_{k_s})} \{Y - \mathbb{E}(Y \mid A, \mathbf{X})\} \mid R = k_s\right]}_{= 0} \\
&= O_p\left(N^{-1/2}\right) + \Delta_{\mathcal{T}}.
\end{aligned}
$$

In the derivation, we utilized Assumption 1(d) to establish the "$= 0$" by the identity

$$
\mathbb{E}(Y \mid A, \mathbf{X}) = \mathbb{E}(Y \mid A, \mathbf{X}, R = k_s).
$$

Denote

$$
\omega^*_{k_t,k_s}(\mathbf{X}) = \frac{\mathbb{P}(R = k_t \mid \mathbf{X} = \mathbf{x}) \mathbb{P}(R = k_s)}{\mathbb{P}(R = k_s \mid \mathbf{X} = \mathbf{x}) \mathbb{P}(R = k_t)},
$$

which produces the identity

$$
\mathbb{E}\{\omega^*_{k_t,k_s}(\mathbf{X}) f(\mathbf{X}) \mid R = k_s\} = \mathbb{E}\{f(\mathbf{X}) \mid R = k_t\}.
$$

When the PS and density ratio models are consistently estimated under Assumption 3(a)(ii) but the OR model may be mis-specified, we have through classical asymptotic analysis

$$
\begin{aligned}
&\hat{\Delta}_{\mathcal{T},\mathcal{T}} \\
&= \sum_{k_t \in \mathcal{T}} \frac{n_{k_t}}{N_{\mathcal{T}}} \left[\frac{1}{n_{k_s}} \sum_{i \in \mathcal{I}_{k_s}} \omega_{k_t,k_s}(\mathbf{X}_i; \hat{\gamma}_{k_t,k_s}) \left\{ \frac{A_i}{\pi_{k_s}(1, \mathbf{X}_i; \hat{\alpha}_{k_s})} - \frac{1 - A_i}{\pi_{k_s}(0, \mathbf{X}_i; \hat{\alpha}_{k_s})} \right\} Y_i \right.
\end{aligned}
$$

$$
\begin{aligned}
&+ \frac{1}{n_{k_t}} \sum_{i \in \mathcal{I}_{k_t}} m(1, \mathbf{X}_i; \hat{\beta}_{1,k_t}) - \frac{1}{n_{k_s}} \sum_{i \in \mathcal{I}_{k_s}} \omega_{k_t,k_s}(\mathbf{X}_i; \hat{\gamma}_{k_t,k_s}) \frac{A_i}{\pi_{k_s}(1, \mathbf{X}_i; \hat{\alpha}_{k_s})} m(1, \mathbf{X}_i; \hat{\beta}_{1,k_t}) \\
&- \frac{1}{n_{k_t}} \sum_{i \in \mathcal{I}_{k_t}} m(0, \mathbf{X}_i; \hat{\beta}_{1,k_t}) + \frac{1}{n_{k_s}} \sum_{i \in \mathcal{I}_{k_s}} \omega_{k_t,k_s}(\mathbf{X}_i; \hat{\gamma}_{k_t,k_s}) \frac{1 - A_i}{\pi_{k_s}(0, \mathbf{X}_i; \hat{\alpha}_{k_s})} m(0, \mathbf{X}_i; \hat{\beta}_{0,k_t}) \Bigg] \\
&= O_p\left(N^{-1/2}\right) + \sum_{k_t \in \mathcal{T}} \frac{\mathbb{P}(R = k_t)}{\mathbb{P}(R \in \mathcal{T})} \Bigg(\mathbb{E}\left\{ \omega^*_{k_t,k_s}(\mathbf{X}) \frac{A}{\mathbb{P}(A = 1 \mid \mathbf{X}, R = k_t)} Y \mid R = k_t \right\} \\
&- \mathbb{E}\left\{ \omega^*_{k_t,k_s}(\mathbf{X}) \frac{1 - A}{\mathbb{P}(A = 0 \mid \mathbf{X}, R = k_t)} Y \mid R = k_t \right\} \\
&+ \mathbb{E}\{ m(1, \mathbf{X}; \bar{\beta}_{1,k_t}) - m(0, \mathbf{X}; \bar{\beta}_{0,k_t}) \mid R = k_t \} \\
&- \mathbb{E}[\omega^*_{k_t,k_s}(\mathbf{X}) \{ m(1, \mathbf{X}; \bar{\beta}_{1,k_t}) - m(0, \mathbf{X}; \bar{\beta}_{0,k_t}) \} \mid R = k_s] \Bigg) \\
&= \sum_{k_t \in \mathcal{T}} \frac{\mathbb{P}(R = k_t)}{\mathbb{P}(R \in \mathcal{T})} \mathbb{E}\left\{ \omega^*_{k_t,k_s}(\mathbf{X}) \mathbb{E}(Y^{(1)} \mid \mathbf{X}) \mid R = k_t \right\} - \mathbb{E}\left\{ \omega^*_{k_t,k_s}(\mathbf{X}) \mathbb{E}(Y^{(0)} \mid \mathbf{X}) \mid R = k_t \right\} \\
&+ O_p\left(N^{-1/2}\right) \\
&= \Delta_{\mathcal{T}} + O_p\left(N^{-1/2}\right).
\end{aligned}
$$

Therefore in either case $\bar{\Delta}_{\mathcal{T},k_s} = \Delta_{\mathcal{T}}$. ■

B.3.2 Optimal Aggregation

We next consider the aggregation of the initial $\hat{\Delta}_{\mathcal{T},\mathcal{T}}$ and the source site $\hat{\Delta}_{\mathcal{T},k_s}$. Denote

$$
\hat{L}(\eta) = N \left[\sum_{k_s \in \mathcal{S}} \eta_{k_s}^2 \frac{\hat{\sigma}^2_{k_s}}{n_{k_s}} + \sum_{k_t \in \mathcal{T}} \hat{\mathbf{h}}_{k_t}(\eta)^\intercal \frac{\hat{\Sigma}_{k_t}}{n_{k_t}} \hat{\mathbf{h}}_{k_t}(\eta) \right]. \tag{1}
$$

We define the oracle selection space for η as

$$
\mathcal{S}^* = \{ k_s \in \mathcal{S} : \bar{\Delta}_{\mathcal{T},k_s} = \Delta_{\mathcal{T}} \},\ \mathbb{R}^{\mathcal{S}^*} = \{ \eta \in \mathbb{R}^K : \eta_j = 0,\ \forall j \neq \mathcal{S}^* \}, \tag{2}
$$

and the asymptotic loss function

$$L^*(\eta) = \sum_{k_s \in \mathcal{S}^*} \eta_{k_s}^2 \text{Var}(\xi_{k_s} \mid R = k_s)/\mathbb{P}(R = k_s) + \sum_{k_t \in \mathcal{T}} \mathbf{h}_{k_t}^*(\eta)^\intercal \Sigma_{k_t} \mathbf{h}_{k_t}^*(\eta)/\mathbb{P}(R = k_t),$$

$$\mathbf{h}_{k_t}^*(\eta) = \left(\mathbb{P}(R = k_t \mid R \in \mathcal{T}), \mathbb{P}(R = k_t \mid R \in \mathcal{T}) \left(1 - \sum_{k_s \in \mathcal{S}^*} \eta_{k_s}\right), \sum_{k_s \in \mathcal{S}} \eta_{k_s} \bar{\mathbf{d}}_{k_t, k_s}^\intercal \right)^\intercal. \quad (3)$$

Any combination $\eta \in \mathbb{R}^{\mathcal{S}^*}$ results in a consistent aggregated estimator for the TATE. The asymptotically optimal combination is

$$\bar{\eta} = \underset{\eta \in \mathbb{R}^{\mathcal{S}^*}}{\arg\min}\, L^*(\eta). \quad (4)$$

In Lemma 3, we establish the asymptotic distribution of the aggregated estimator with fixed $\eta \in \mathbb{R}^{\mathcal{S}^*}$. In Lemma 4, we show that the estimator $\hat{\eta}$ recovers the optimal $\bar{\eta}$. In Lemma 5, we show that the uncertainty from $\hat{\eta}$ is negligible in estimating $\Delta_{\mathcal{T}}$ as $\hat{\Delta}_{\mathcal{T},\mathsf{FACE}}$.

Lemma 3 *Let* $\hat{\Delta}(\eta) = \hat{\Delta}_{\mathcal{T},\mathcal{T}} + \sum_{k_s \in \mathcal{S}'} \eta_{k_s} \left(\hat{\Delta}_{\mathcal{T},k_s} - \hat{\Delta}_{\mathcal{T},\mathcal{T}} \right)$ *be the aggregation with* $\eta \in \mathbb{R}^{\mathcal{S}'}$. *Under Assumptions 1 and 2, we have*

$$\sqrt{N} \left\{ \hat{\Delta}(\eta) - \Delta_{\mathcal{T}} \right\} \rightsquigarrow \mathcal{N}\left(0, L^*(\eta)\right).$$

Proof [Proof of Lemma 3] By Lemma 1, the initial estimator $\hat{\Delta}_{\mathcal{T},\mathcal{T}}$ is consistent for $\Delta_{\mathcal{T}}$. According to the definition of $\mathcal{S}^*$ (2), $\hat{\Delta}_{\mathcal{T},k_s}$ is consistent for $\Delta_{\mathcal{T}}$ for $k_s \in \mathcal{S}^*$. Thus, the weighted average $\hat{\Delta}(\eta)$ must also be consistent for $\Delta_{\mathcal{T}}$.

Next, we establish the asymptotic normality of $\hat{\Delta}(\eta)$. From Assumption 2(a), we have the influence function for $\hat{\Delta}(\eta)$

$$\hat{\Delta}(\eta) - \Delta_{\mathcal{T}}$$

$$
\begin{aligned}
&= o_p\left(N^{-1/2}\right) + \left(1 - \sum_{k_s \in \mathcal{S}^*} \eta_{k_s}\right) \frac{1}{N_{\mathcal{T}}} \sum_{k_t \in \mathcal{T}} \sum_{i \in \mathcal{I}_{k_t}} (\zeta_i + \xi_{i,\mathcal{T}}) \\
&+ \sum_{k_s \in \mathcal{S}^*} \eta_{k_s} \sum_{k_t \in \mathcal{T}} \frac{1}{n_{k_t}} \sum_{i \in \mathcal{I}_{k_t}} \left\{ \frac{n_{k_t}}{N_{\mathcal{T}}} \zeta_i + \left(\psi(\mathbf{X}_i)^\intercal - \mathbb{E}\{\psi(\mathbf{X}) \mid R = k_t\}^\intercal, v_{i,1}^\intercal, v_{i,0}^\intercal\right) \bar{\mathbf{d}}_{k_t,k_s} \right\} \\
&+ \sum_{k_s \in \mathcal{S}^*} \eta_{k_s} \frac{1}{n_{k_s}} \sum_{i \in \mathcal{I}_{k_s}} \xi_{i,k_s} \\
&= o_p\left(N^{-1/2}\right) + \frac{1}{N} \sum_{k_s \in \mathcal{S}^*} \sum_{i \in \mathcal{I}_{k_s}} \frac{\eta_{k_s} \xi_{i,k_s}}{\mathbb{P}(R = k_s)} \\
&+ \frac{1}{N} \sum_{k_t \in \mathcal{T}} \sum_{i \in \mathcal{I}_{k_t}} \left\{ \frac{\zeta_i + \left(1 - \sum_{k_s \in \mathcal{S}^*} \eta_{k_s}\right) \xi_{i,\mathcal{T}}}{\mathbb{P}(R \in \mathcal{T})} \right. \\
&\qquad \left. + \frac{\left(\psi(\mathbf{X}_i)^\intercal - \mathbb{E}\{\psi(\mathbf{X}) \mid R = k_t\}, v_{i,1}^\intercal, v_{i,0}^\intercal\right) \bar{\mathbf{d}}_{k_t,k_s}}{\mathbb{P}(R = k_t)} \right\}.
\end{aligned}
$$

We defined $L^*(\eta)$ to be precisely the variance of the influence function. To see this, we will show that $L^*(\eta)$ is the variance of $\left(1 - \sum_{k \in \mathcal{S}} \eta_k\right) \hat{\Delta}_{\mathcal{T},\mathcal{T}} + \sum_{k \in \mathcal{S}} \eta_k \hat{\Delta}_{\mathcal{T},k}$ and use the influence function representation from Assumption 2(a). Denote $\eta_{\mathcal{T}} = 1 - \sum_{k_s \in \mathcal{S}} \eta_{k_s}$ and define the asymptotic approximation of the aggregation under Assumption 2(a)

$$
\begin{aligned}
W(\eta) &= \frac{\eta_{\mathcal{T}}}{\sqrt{N}} \sum_{k_t \in \mathcal{T}} \sum_{i \in \mathcal{I}_{k_t}} \frac{N}{N_{\mathcal{T}}} \left(\zeta_i + \xi_{i,\mathcal{T}}\right) \\
&\quad + \sum_{k_s \in \mathcal{S}} \frac{\eta_{k_s}}{\sqrt{N}} \left\{ \sum_{k_t \in \mathcal{T}} \frac{N}{n_{k_t}} \sum_{i \in \mathcal{I}_{k_t}} \left\{ \frac{n_{k_t}}{N_{\mathcal{T}}} \zeta_i + \left(\psi(\mathbf{X}_i)^\intercal - \mathbb{E}\{\psi(\mathbf{X}) \mid R = k_t\}^\intercal, v_{i,1}^\intercal, v_{i,0}^\intercal\right) \bar{\mathbf{d}}_{k_t,k_s} \right\} \right. \\
&\qquad \left. + \frac{N}{n_{k_s}} \sum_{i \in \mathcal{I}_{k_s}} \xi_{i,k_s} \right\} \\
&= \eta_{\mathcal{T}} \sqrt{N} (\hat{\Delta}_{\mathcal{T},\mathcal{T}} - \bar{M}_{\mathcal{T},\mathcal{T}} - \bar{\delta}_{\mathcal{T},\mathcal{T}}) + \sum_{k_s \in \mathcal{S}} \eta_{k_s} \sqrt{N} (\hat{\Delta}_{\mathcal{T},k_s} - \bar{M}_{\mathcal{T},\mathcal{T}} - \bar{\delta}_{\mathcal{T},k_s}) + o_p(1).
\end{aligned}
$$

where we have merged by site and individual indices to obtain the last line. By this alternative representation of $W(\eta)$, it is clear that its variance equals $L^*(\eta)$. Under Assumption 1(c) and 2(c), $L^*(\eta)$

is stable

$$\frac{L^*(\eta)}{\|\eta\|_2^2 + \sum_{k_t \in \mathcal{T}} \|\mathbf{h}_{k_t}^*(\eta)\|_2^2} \in [\varepsilon, \mathcal{M}].$$

Further, under Asssumptions 1(c) and 2(a), we have

$$\varepsilon \leq \|\mathbf{h}_{k_t}^*(\eta)\|_2^2 \leq 2 + \|\eta\|_1 \left(1 + \max_{k_s \in \mathcal{S}} \|\bar{\mathbf{d}}_{k_t,k_s}\|_2\right) < \infty.$$

Hence for any bounded η, $L^*(\eta)$ is finite and nonzero, so we have

$$\sqrt{N}\left\{\hat{\Delta}(\eta) - \Delta_{\mathcal{T}}\right\} \rightsquigarrow \mathcal{N}\left(0, L^*(\eta)\right).$$

■

Lemma 4 *Under Assumptions 1 and 2, we have*

$$\lim_{N\to\infty} \mathbb{P}(\hat{\eta} \in \mathbb{R}^{\mathcal{S}^*}) = 1, \ \|\hat{\eta} - \bar{\eta}\| = O_p\left(N^{-1/2}\right).$$

Proof [Proof of Lemma 4] We define $\tilde{\eta}$ as the estimator under oracle selection

$$\tilde{\eta} = \underset{\eta \in \mathbb{R}^{\mathcal{S}^*}}{\arg\min}\, N\left[\sum_{k_s \in \mathcal{S}} \eta_{k_s}^2 \frac{\hat{\sigma}_{k_s}^2}{n_{k_s}} + \sum_{k_t \in \mathcal{T}} \hat{\mathbf{h}}_{k_t}(\eta)^\top \frac{\hat{\Sigma}_{k_t}}{n_{k_t}} \hat{\mathbf{h}}_{k_t}(\eta)\right] + \lambda \sum_{k_s \in \mathcal{S}} |\eta_{k_s}| \left(\hat{\delta}_{\mathcal{T},k_s} - \hat{\delta}_{\mathcal{T},\mathcal{T}}\right)^2. \quad (5)$$

We first show that $\|\tilde{\eta} - \bar{\eta}\| = O_p\left(N^{-1/2}\right)$. Then, we verify that $\tilde{\eta}$ satisfies the optimality condition, i.e., $\tilde{\eta} = \hat{\eta}$, with high probability. Note that $\hat{L}(\eta)$ and $L^*(\eta)$ are both quadratic functions of η, which can be expressed as

$$L(\eta) = \eta^\top \hat{H}\eta + \hat{\mathbf{g}}^\top \eta + \hat{c}, \ L^*(\eta) = \eta^\top H\eta + \mathbf{g}^\top \eta + c$$

Using Assumptions 2(d) and the Chebyshev inequality under Assumptions 2(a) and 2(c), it is clear that $\hat{H}$, $\hat{\mathbf{g}}$, and $\hat{c}$ are $\sqrt{N}$-consistent. Thus, $L(\eta) - L^*(\eta) \asymp (1 + \|\eta\|^2)/\sqrt{N}$, since H, $\mathbf{g}$ and c are bounded under Assumptions 2(a) and 2(c).

Under Assumptions 1(c) and 2(d), we have the uniform approximation of the loss in a compact neighborhood of $\bar{\eta}$ of S

$$\sup_{\|\eta - \bar{\eta}\| \leq M} |\hat{L}(\eta) - L^*(\eta)| = O_p\left(N^{-1/2}\right). \tag{6}$$

By Lemmata 1 and 2, we have for $k_s \in \mathcal{S}^*$

$$\hat{\delta}_{\mathcal{T},\mathcal{T}} - \hat{\delta}_{\mathcal{T},k_s} = \hat{\Delta}_{\mathcal{T},\mathcal{T}} - \hat{\Delta}_{\mathcal{T},k_s} = O_p\left(N^{-1/2}\right).$$

With $\lambda \lesssim N^{1/2}$, the penalty is small in the compact neighborhood of $\bar{\eta}$

$$\sup_{\|\eta - \bar{\eta}\| \leq M} \lambda \sum_{k_s \in \mathcal{S}} |\eta_{k_s}| \left(\hat{\delta}_{\mathcal{T},k_s} - \hat{\delta}_{\mathcal{T},\mathcal{T}}\right)^2 = O_p\left(N^{-1/2}\right). \tag{7}$$

Combining (6) and (7), we have the approximation of the penalized loss

$$\sup_{\|\eta - \bar{\eta}\| \leq M} \left| \hat{L}(\eta) + \lambda \sum_{k_s \in \mathcal{S}} |\eta_{k_s}| \left(\hat{\delta}_{\mathcal{T},k_s} - \hat{\delta}_{\mathcal{T},\mathcal{T}}\right)^2 - L^*(\eta) \right| = O_p\left(N^{-1/2}\right).$$

Following the convexity of $L^*(\eta)$ from Assumption 2(c), we have

$$\|\tilde{\eta} - \bar{\eta}\| = O_p\left(N^{-1/2}\right).$$

The optimality condition of the original problem (2.3.3) is

$$\frac{\partial}{\partial \eta_{k_s}} \hat{L} = -\text{sign}(\eta_{k_s}) \lambda \left(\hat{\delta}_{\mathcal{T},k_s} - \hat{\delta}_{\mathcal{T},\mathcal{T}} \right)^2, \ \eta_{k_s} \neq 0; \ \left| \frac{\partial}{\partial \eta_{k_s}} \hat{L} \right| \leq \lambda \left(\hat{\delta}_{\mathcal{T},k_s} - \hat{\delta}_{\mathcal{T},\mathcal{T}} \right)^2, \ \eta_{k_s} = 0.$$

For $j \in \mathcal{S}^*$, the conditions are shared with (5), so $\tilde{\eta}$ must satisfy them. To establish the optimality of $\tilde{\eta}$ for (2.3.3), it suffices to show

$$\left| \frac{\partial}{\partial \eta_{k_s}} \hat{L} \right| \leq \lambda \left(\hat{\delta}_{\mathcal{T},k_s} - \hat{\delta}_{\mathcal{T},\mathcal{T}} \right)^2, \ k_s \in \mathcal{S} \setminus \mathcal{S}^*. \tag{8}$$

By the definition of $\mathcal{S}^*$, we have for biased sites

$$\bar{\delta}_{\mathcal{T},k_s} - \bar{\delta}_{\mathcal{T},\mathcal{T}} = \bar{\Delta}_{\mathcal{T},k_s} - \bar{\Delta}_{\mathcal{T},\mathcal{T}} \neq 0.$$

By Lemmata 1 and 2, we have for $k_s \in \mathcal{S} \setminus \mathcal{S}^*$

$$\hat{\delta}_{\mathcal{T},k_s} - \hat{\delta}_{\mathcal{T},\mathcal{T}} = \bar{\Delta}_{\mathcal{T},k_s} - \bar{\Delta}_{\mathcal{T},\mathcal{T}} + O_p\left(N^{-1/2}\right)$$

bounded away from zero. With $\lambda \to \infty$, the penalty for biased sites diverges for $k_s \in \mathcal{S} \setminus \mathcal{S}^*$

$$\lambda \left(\hat{\delta}_{\mathcal{T},k_s} - \hat{\delta}_{\mathcal{T},\mathcal{T}} \right)^2 \to \infty. \tag{9}$$

Under Assumptions 1(c), 2(c) and 2(d), the derivative is tight

$$\frac{\partial}{\partial \eta_{k_s}} \hat{L} = \frac{\partial}{\partial \eta_{k_s}} L^* + O_p\left(N^{-1/2}\right) = O_p(1). \tag{10}$$

Combining (9) and (10), we must have (8) with high probability. This implies that $\hat{\eta}$ satisfies precisely the optimality condition with high probability. Therefore, we must have $\hat{\eta} = \tilde{\eta}$ according to the

convexity of the problem with high probability.

■

Lemma 5 *Under Assumptions 1 and 2,*

$$\sqrt{N}\left\{\hat{\Delta}(\bar{\eta}) - \hat{\Delta}_{\mathcal{T},\mathsf{FACE}}\right\} = o_p(1).$$

Proof [Proof of Lemma 5] We decompose the difference into informative source sites $k_s \in \mathcal{S}^*$ and biased source sites $k_s \in \mathcal{S} \setminus \mathcal{S}^*$

$$\begin{aligned}\sqrt{N}\left\{\hat{\Delta}(\bar{\eta}) - \hat{\Delta}_{\mathcal{T},\mathsf{FACE}}\right\} = &\sum_{k_s \in \mathcal{S}^*} (\bar{\eta}_{k_s} - \hat{\eta}_{k_s})\sqrt{N}\left(\hat{\Delta}_{\mathcal{T},k_s} - \hat{\Delta}_{\mathcal{T},\mathcal{T}}\right) \\ &+ \sum_{k_s \in \mathcal{S}\setminus\mathcal{S}^*} (\bar{\eta}_{k_s} - \hat{\eta}_{k_s})\sqrt{N}\left(\hat{\Delta}_{\mathcal{T},k_s} - \hat{\Delta}_{\mathcal{T},\mathcal{T}}\right).\end{aligned}$$

By the definition of $\mathcal{S}^*$ (2) and the conclusions of Lemmata 1 and 2, we have the tightness of terms for $k_s \in \mathcal{S}^*$

$$\sqrt{N}\left(\hat{\Delta}_{\mathcal{T},k_s} - \hat{\Delta}_{\mathcal{T},\mathcal{T}}\right) = O_p\left(N^{-1/2}\right).$$

Applying the conclusion of Lemma 4, we have for $k_s \in \mathcal{S}^*$

$$(\bar{\eta}_{k_s} - \hat{\eta}_{k_s})\sqrt{N}\left(\hat{\Delta}_{\mathcal{T},k_s} - \hat{\Delta}_{\mathcal{T},\mathcal{T}}\right) = O_p\left(N^{-1}\right) = o_p(1)$$

and for $k_s \in \mathcal{S} \setminus \mathcal{S}^*$

$$(\bar{\eta}_{k_s} - \hat{\eta}_{k_s})\sqrt{N}\left(\hat{\Delta}_{\mathcal{T},k_s} - \hat{\Delta}_{\mathcal{T},\mathcal{T}}\right) = 0$$

with large probability. Therefore, we have obtained

$$\sqrt{N}\left\{\hat{\Delta}(\bar{\eta}) - \hat{\Delta}_{\mathcal{T},\mathsf{FACE}}\right\} = o_p(1).$$

■

B.3.3 Proof of Theorem 1 and Corollary 1

Applying Lemmata 3 and 5, we have the asymptotic normality of $\hat{\Delta}_{\mathcal{T},\mathsf{FACE}}$,

$$\sqrt{N}\left(\hat{\Delta}_{\mathcal{T},\mathsf{FACE}} - \Delta_{\mathcal{T}}\right) \rightsquigarrow \mathcal{N}\left(0, L^*(\bar{\eta})\right).$$

Using the consistency of $\hat{\eta}$ for $\bar{\eta}$ and locally uniform convergence of $\hat{L}$ for L^* (see (1)-(4) for the definitions), we have the consistency of the variance estimator

$$\hat{\mathcal{V}} = \hat{L}(\hat{\eta}) = L^*(\bar{\eta}) + O_p\left(N^{-1/2}\right).$$

By the continuous mapping theorem, we have

$$\sqrt{N/\hat{\mathcal{V}}}\left(\hat{\Delta}_{\mathcal{T},\mathsf{FACE}} - \Delta_{\mathcal{T}}\right) \rightsquigarrow \mathcal{N}\left(0,1\right).$$

The coverage probability in Corollary 1 immediately follows.

B.3.4 Proof of Corollary 2

In the main text, we noted that the variance covariance matrix for the target site, $\hat{\Sigma}_1$ can be calculated as as $\hat{\Sigma}_1 = \frac{1}{n_{\mathcal{T}}^2}\sum_{i\in\mathcal{I}_1}\hat{\mathbf{U}}_i\hat{\mathbf{U}}_i^\intercal$ through the estimated influence functions, where $\hat{\mathbf{U}}_i = (\hat{\zeta}_i, \hat{\xi}_i, \psi(\mathbf{X}_i)^\intercal)^\intercal$.

Here, we provide the exact form for $\hat{\xi}_{i,1}$ and $\hat{\zeta}_i$.

$$
\begin{aligned}
\hat{\upsilon}_{i,1} &= \left\{ \frac{1}{n_{\mathcal{T}}} \sum_{i \in \mathcal{I}_1} g'\left(\hat{\alpha}_1^{\mathsf{T}} \mathbf{X}_j\right) \mathbf{X}_j \mathbf{X}_j^{\mathsf{T}} \right\}^{-1} \mathbf{X}_i \left\{ A_i - g\left(\hat{\alpha}_1^{\mathsf{T}} \mathbf{X}_i\right) \right\}, \\
\hat{\upsilon}_{i,0} &= \left\{ \frac{1}{n_{\mathcal{T}}} \sum_{j \in \mathcal{I}_1} (1 - A_j) g'\left(\hat{\beta}_{0,1}^{\mathsf{T}} \mathbf{X}_j\right) \mathbf{X}_j \mathbf{X}_j^{\mathsf{T}} \right\}^{-1} \mathbf{X}_i (1 - A_i) \left\{ Y_i - g\left(\hat{\beta}_{0,1}^{\mathsf{T}} \mathbf{X}_i\right) \right\}, \\
\hat{\xi}_{i,1} &= \frac{A_i}{g(\hat{\alpha}_1^{\mathsf{T}} \mathbf{X}_i)} \{ Y_i - g(\hat{\beta}_{1,i}^{\mathsf{T}} \mathbf{X}_i) \} - \frac{1 - A_i}{g(-\hat{\alpha}_1^{\mathsf{T}} \mathbf{X}_i)} \{ Y_i - g(\hat{\beta}_{0,i}^{\mathsf{T}} \mathbf{X}_i) \} \\
&\quad - \left[\frac{1}{n_{\mathcal{T}}} \sum_{j \in \mathcal{I}_1} e^{-(-1)^{A_j} \hat{\alpha}_1^{\mathsf{T}} \mathbf{X}_j} \left\{ Y_j - g\left(\hat{\beta}_{A_j,1}^{\mathsf{T}} \mathbf{X}_j\right) \right\} \mathbf{X}_j^{\mathsf{T}} \right] \left\{ \frac{1}{n_{\mathcal{T}}} \sum_{i \in \mathcal{I}_1} g'\left(\hat{\alpha}_1^{\mathsf{T}} \mathbf{X}_j\right) \mathbf{X}_j \mathbf{X}_j^{\mathsf{T}} \right\}^{-1} \\
&\qquad \mathbf{X}_i \left\{ A_i - g\left(\hat{\alpha}_1^{\mathsf{T}} \mathbf{X}_i\right) \right\} \\
&\quad - \left\{ \frac{1}{n_{\mathcal{T}}} \sum_{j \in \mathcal{I}_1} \frac{A_j}{g\left(\hat{\alpha}_1^{\mathsf{T}} \mathbf{X}_j\right)} g'\left(\hat{\beta}_{1,1}^{\mathsf{T}} \mathbf{X}_j\right) \mathbf{X}_j^{\mathsf{T}} \right\} \hat{\upsilon}_{i,1} \\
&\quad + \left\{ \frac{1}{n_{\mathcal{T}}} \sum_{j \in \mathcal{I}_1} \frac{1 - A_j}{g\left(-\hat{\alpha}_1^{\mathsf{T}} \mathbf{X}_j\right)} g'\left(\hat{\beta}_{0,1}^{\mathsf{T}} \mathbf{X}_j\right) \mathbf{X}_j^{\mathsf{T}} \right\} \hat{\upsilon}_{i,0}, \\
\hat{\zeta}_i &= g\left(\hat{\beta}_{1,1}^{\mathsf{T}} \mathbf{X}_i\right) - g\left(\hat{\beta}_{0,1}^{\mathsf{T}} \mathbf{X}_i\right) + \left\{ \frac{1}{n_{\mathcal{T}}} \sum_{j \in \mathcal{I}_1} g'\left(\hat{\beta}_{1,1}^{\mathsf{T}} \mathbf{X}_j\right) \mathbf{X}_j^{\mathsf{T}} \right\} \hat{\upsilon}_{i,1} \\
&\quad - \left\{ \frac{1}{n_{\mathcal{T}}} \sum_{j \in \mathcal{I}_1} g'\left(\hat{\beta}_{0,1}^{\mathsf{T}} \mathbf{X}_j\right) \mathbf{X}_j^{\mathsf{T}} \right\} \hat{\upsilon}_{i,0}, \\
\hat{\mathbf{U}}_i &= (\hat{\zeta}_i, \hat{\xi}_i, \psi(\mathbf{X}_i)^{\mathsf{T}}, \hat{\upsilon}_{i,1}^{\mathsf{T}}, \hat{\upsilon}_{i,0}^{\mathsf{T}})^{\mathsf{T}}.
\end{aligned}
$$

For source sites, the variance estimator $\hat{\sigma}_k^2$ can be calculated as $\hat{\sigma}_k^2 = \frac{1}{n_k} \sum_{i \in \mathcal{I}_k} \hat{\xi}_{i,k}^2$, where $\hat{\xi}_{i,k}$ is

$$
\begin{aligned}
\hat{\xi}_{i,k} &= e^{\hat{\gamma}_k^{\mathsf{T}} \mathbf{X}_i} \left[\frac{A_i}{g(\hat{\alpha}_k^{\mathsf{T}} \mathbf{X}_i)} \{ Y_i - g(\hat{\beta}_{1,i}^{\mathsf{T}} \mathbf{X}_i) \} - \frac{1 - A_i}{g(-\hat{\alpha}_k^{\mathsf{T}} \mathbf{X}_i)} \{ Y_i - g(\hat{\beta}_{0,i}^{\mathsf{T}} \mathbf{X}_i) \} \right] \\
&\quad - \left[\frac{1}{n_k} \sum_{j \in \mathcal{I}_k} e^{\left(\hat{\gamma}_k - (-1)^{A_j} \hat{\alpha}_k\right)^{\mathsf{T}} \mathbf{X}_j} \left\{ Y_j - g\left(\hat{\beta}_{A_j,k}^{\mathsf{T}} \mathbf{X}_j\right) \right\} \mathbf{X}_j^{\mathsf{T}} \right] \left\{ \frac{1}{n_k} \sum_{i \in \mathcal{I}_k} g'\left(\hat{\alpha}_k^{\mathsf{T}} \mathbf{X}_j\right) \mathbf{X}_j \mathbf{X}_j^{\mathsf{T}} \right\}^{-1}
\end{aligned}
$$

$$\mathbf{X}_i \left\{A_i - g\left(\hat{\alpha}_k^\intercal \mathbf{X}_i\right)\right\}$$
$$+ \hat{\mathbf{d}}_{k,\psi}^\intercal \left(e^{\hat{\gamma}_k^\intercal \mathbf{X}_i} \mathbf{X}_i - \bar{\psi}_{\mathcal{T}}\right).$$

As Assumption 2 is satisfied, the FACE estimator is consistent and asymptotically normal with consistent variance estimation $\hat{\mathcal{V}}$,

$$\sqrt{N/\hat{\mathcal{V}}}\left(\hat{\Delta}_{\mathcal{T},\mathsf{FACE}} - \Delta_{\mathcal{T}}\right) \rightsquigarrow \mathcal{N}(0,1).$$

B.3.5 Proof of Proposition 1

Since the initial estimator $\hat{\Delta}_{\mathcal{T},\mathcal{T}}$ corresponds to $\hat{\Delta}(0)$, the asymptotic variance of $\sqrt{N}(\hat{\Delta}_{\mathcal{T},\mathcal{T}} - \Delta_{\mathcal{T}})$ can be expressed as $L^*(0)$ by Lemma 3. By Lemmata 3 and 5, the asymptotic variance of $\sqrt{N}(\hat{\Delta}_{\mathcal{T},\mathsf{FACE}} - \Delta_{\mathcal{T}})$ is $L^*(\bar{\eta})$. By the definition of $\bar{\eta}$ as the minimum, we must have $L^*(\bar{\eta}) \leq L^*(0)$. Thus, we have shown the non-inferiority of $\hat{\Delta}_{\mathcal{T},\mathsf{FACE}}$.

To show that $L^*(\bar{\eta})$ is strictly smaller than $L^*(0)$, it suffices to find another $\check{\eta}$, an upper bound for $L^*(\bar{\eta})$ by the definition of $\bar{\eta}$, such that

$$L^*(\bar{\eta}) \leq L^*(\check{\eta}) < L^*(0). \tag{11}$$

Without loss of generality, we consider the simplified problem with one source site $k_* \in \mathcal{S}'$,

$$\check{\Delta}(\eta) = \hat{\Delta}_{\mathcal{T},\mathcal{T}} + \eta\left(\hat{\Delta}_{\mathcal{T},k_*} - \hat{\Delta}_{\mathcal{T},\mathcal{T}}\right).$$

Under Assumption 3(a), the TATE estimator of the site $\hat{\Delta}_{\mathcal{T},k_*}$ is consistent for $\Delta_{\mathcal{T}}$ and asymptotically normal by Lemma 2. Thus, $\check{\Delta}(\eta)$ is also consistent for $\Delta_{\mathcal{T}}$ and asymptotically normal with any η. The

optimal η is given by the projection

$$\eta_* = \frac{NCov\left(\hat{\Delta}_{\mathcal{T},\mathcal{T}}, \hat{\Delta}_{\mathcal{T},k_*} - \hat{\Delta}_{\mathcal{T},\mathcal{T}}\right)}{N\text{Var}\left(\hat{\Delta}_{\mathcal{T},k_*} - \hat{\Delta}_{\mathcal{T},\mathcal{T}}\right)}.$$

We can construct $\check{\eta}$ to be η_* for site-k_* and zero elsewhere such that $\hat{\Delta}(\check{\eta}) = \check{\Delta}(\eta_*)$. As long as $Cov\left(\hat{\Delta}_{\mathcal{T},\mathcal{T}}, \hat{\Delta}_{\mathcal{T},k_*} - \hat{\Delta}_{\mathcal{T},\mathcal{T}}\right) \neq 0$, the resulting estimator is different from the initial estimator $\check{\eta} \neq 0 \Rightarrow \hat{\Delta}(\check{\eta}) \neq \hat{\Delta}_{\mathcal{T},\mathcal{T}}$. Under Assumption 1(c) and 2(a), the asymptotic covariance between $\sqrt{N}\hat{\Delta}_{\mathcal{T},\mathcal{T}}$ and $\sqrt{N}\left(\hat{\Delta}_{\mathcal{T},k_*} - \hat{\Delta}_{\mathcal{T},\mathcal{T}}\right)$ takes the form

$$Cov\left(\frac{\zeta + \xi_{\mathcal{T}}}{\mathbb{P}(R \in \mathcal{T})}, -\frac{\xi_{\mathcal{T}}}{\mathbb{P}(R \in \mathcal{T})} + \sum_{k_t \in \mathcal{T}} \frac{\mathrm{I}(R = k_t)}{\mathbb{P}(R = k_t)} \left(\psi(\mathbf{X})^\intercal, v_1^\intercal, v_0^\intercal\right) \bar{\mathbf{d}}_{k_t,k_*} \mid R \in \mathcal{T}\right).$$

which is bounded away from zero by Assumption 3(b). Thus, we have found the suitable $\check{\eta}$ that separates the asymptotic variance of $\hat{\Delta}_{\mathcal{T},\mathsf{FACE}}$ and $\hat{\Delta}_{\mathcal{T},\mathcal{T}}$ through (11).

B.3.6 Proof of Proposition 2

Under the ideal setting of Assumption 4, the influence functions of the doubly robust $\hat{\Delta}_{\mathcal{T},\mathcal{T}}$ and $\hat{\Delta}_{\mathcal{T},2}$ admit much simpler forms[98] as a result of Neyman Orthogonality[21],

$$\begin{aligned}
\hat{\Delta}_{\mathcal{T},\mathcal{T}} - \Delta_{\mathcal{T}} = o_p\left(N^{-1/2}\right) + \frac{1}{n_{\mathcal{T}}} \sum_{i \in \mathcal{I}_1} \Bigg[& m(1, X_i; \bar{\beta}_1) - m(0, X_i; \bar{\beta}_0) - \Delta_{\mathcal{T}} \\
& + \frac{A_i\{Y_i - m(1, X_i; \bar{\beta}_1)\}}{\pi(1, \mathbf{X}_i; \bar{\alpha}_1)} - \frac{(1 - A_i)\{Y_i - m(0, X_i; \bar{\beta}_0)\}}{\pi(0, \mathbf{X}_i; \bar{\alpha}_1)} \Bigg] \\
\hat{\Delta}_{\mathcal{T},2} - \Delta_{\mathcal{T}} = o_p\left(N^{-1/2}\right) + \frac{1}{n_{\mathcal{T}}} \sum_{i \in \mathcal{I}_1} & \left[m(1, X_i; \bar{\beta}_1) - m(0, X_i; \bar{\beta}_0) - \Delta_{\mathcal{T}}\right] \\
& + \frac{1}{n_{\mathcal{S}}} \sum_{i \in \mathcal{I}_2} \omega_{1,2}(\mathbf{X}_i; \bar{\gamma}_{1,2}) \left[\frac{A_i\{Y_i - m(1, X_i; \bar{\beta}_1)\}}{\pi(1, \mathbf{X}_i; \bar{\alpha}_2)} - \frac{(1 - A_i)\{Y_i - m(0, X_i; \bar{\beta}_0)\}}{\pi(0, \mathbf{X}_i; \bar{\alpha}_2)}\right].
\end{aligned}$$

The asymptotic variance of the aggregation $\sqrt{N}\left\{(1-\eta)\hat{\Delta}_{\mathcal{T},\mathcal{T}} + \eta\hat{\Delta}_{\mathcal{T},2} - \Delta_{\mathcal{T}}\right\}$ takes the form

$$L^*(\eta) = \frac{N}{n_{\mathcal{T}}}\mathcal{V}_m^2 + \frac{N}{n_{\mathcal{T}}}(1-\eta)^2\mathcal{V}_{\mathcal{T}}^2 + \eta^2\frac{N}{n_{\mathcal{S}}}\mathcal{V}_{\mathcal{S}}^2.$$

Minimizing the quadratic function of η give the optimal solution

$$\bar{\eta} = \frac{n_{\mathcal{S}}\mathcal{V}_{\mathcal{T}}^2}{n_{\mathcal{S}}\mathcal{V}_{\mathcal{T}}^2 + n_{\mathcal{T}}\mathcal{V}_{\mathcal{S}}^2}.$$

We obtain the relative efficiency through

$$\frac{L^*(0)}{L^*(\bar{\eta})} = \frac{\mathcal{V}_m^2/n_{\mathcal{T}} + \mathcal{V}_{\mathcal{T}}^2/n_{\mathcal{T}}}{\mathcal{V}_m^2/n_{\mathcal{T}} + \mathcal{V}_{\mathcal{T}}^2\mathcal{V}_{\mathcal{S}}^2/(n_{\mathcal{T}}\mathcal{V}_{\mathcal{S}}^2 + n_{\mathcal{S}}\mathcal{V}_{\mathcal{T}}^2)} = 1 + \frac{\mathcal{V}_{\mathcal{T}}^4}{\mathcal{V}_m^2\mathcal{V}_{\mathcal{T}}^2 + n_{\mathcal{T}}\left(\mathcal{V}_m^2 + \mathcal{V}_{\mathcal{T}}^2\right)\mathcal{V}_{\mathcal{S}}^2/n_{\mathcal{S}}}.$$

B.4 Supplementary Tables

Table B.1: Baseline characteristics of veterans in each of five VA sites

	Site				
	1 North Atlantic ($n_1 = 143,076$)	2 Southwest ($n_2 = 128,792$)	3 Midwest ($n_3 = 123,228$)	4 Continental ($n_4 = 93,822$)	5 Pacific ($n_5 = 119,441$)
Age (years)					
18-49	12,264 (8.6%)	10,064 (7.8%)	9,753 (7.9%)	9,807 (10.5%)	12,936 (10.8%)
50-59	16,862 (11.8%)	16,906 (13.1%)	13,299 (10.8%)	13,146 (14.0%)	13,348 (11.2%)
60-69	35,709 (25.0%)	35,092 (27.2%)	29,943 (24.3%)	24,670 (26.3%)	27,906 (23.4%)
70-79	59,765 (41.8%)	50,839 (39.5%)	54,588 (44.3%)	36,230 (38.6%)	49,522 (41.5%)
80 or older	18,476 (12.9%)	15,891 (12.3%)	15,645 (12.7%)	9,969 (10.6%)	15,729 (13.2%)
Sex					
Female	11,752 (8.2%)	11,821 (9.2%)	8,829 (7.2%)	9,314 (9.9%)	9,897 (8.3%)
Male	131,324 (91.8%)	116,971 (90.8%)	114,399 (92.8%)	84,508 (90.1%)	109,544 (91.7%)
Race					
Asian	745 (0.5%)	391 (0.3%)	388 (0.3%)	535 (0.6%)	5,062 (4.2%)
Black	38,146 (26.7%)	34,064 (26.4%)	20,720 (16.8%)	24,182 (25.8%)	15,016 (12.6%)
White	96,890 (67.7%)	86,404 (67.1%)	94,769 (76.9%)	61,471 (65.5%)	82,750 (69.3%)
Other	7,295 (5.1%)	7,933 (6.2%)	7,351 (6.0%)	7,634 (8.1%)	16,613 (13.9%)
Ethnicity					
Hispanic	5,862 (4.1%)	16,768 (13.0%)	2,661 (2.2%)	9,127 (9.7%)	13,938 (11.7%)
Not Hispanic	137,214 (95.9%)	112,024 (87.0%)	120,567 (97.8%)	84,695 (90.3%)	105,503 (88.3%)
Urbanicity					
Rural	31,216 (21.8%)	25,223 (19.6%)	36,551 (29.7%)	21,932 (23.4%)	20,133 (16.9%)
Urban	111,860 (78.2%)	103,569 (80.4%)	86,677 (70.3%)	71,890 (76.6%)	99,308 (83.1%)
Comorbidities					
CLD*	43,186 (30.2%)	39,267 (30.5%)	41,912 (34.0%)	27,124 (28.9%)	30,780 (25.8%)
CVD**	40,565 (28.4%)	36,167 (28.1%)	38,512 (31.3%)	25,097 (26.7%)	28,999 (24.3%)
Hypertension	104,775 (73.2%)	97,584 (75.8%)	92,355 (74.9%)	68,454 (73.0%)	79,986 (67.0%)
T2D	56,641 (39.6%)	52,356 (40.7%)	49,660 (40.3%)	38,585 (41.1%)	42,170 (35.3%)
CKD	25,631 (17.9%)	24,029 (18.7%)	25,261 (20.5%)	17,396 (18.5%)	20,169 (16.9%)
Autoimmune†	49,135 (34.3%)	46,313 (36.0%)	45,952 (37.3%)	30,392 (32.4%)	38,870 (32.5%)
Obesity‡	39,626 (27.7%)	37,438 (29.1%)	36,465 (29.6%)	26,526 (28.3%)	31,330 (26.2%)

* Chronic lung diseases (CLD) included asthma, bronchitis, and chronic obstructive pulmonary disease.
** Cardiovascular disease (CVD) included acute myocardial infarction, cardiomyopathy, coronary heart disease, heart failure, and peripheral vascular disease.
† Autoimmune diseases included HIV infection, rheumatoid arthritis, etc. The full list of ICD-10 codes are given in the Supplement.
‡ Obesity was defined as a body-mass index of 30 or greater.

Table B.2: Baseline characteristics for veterans in each of the five sites in each vaccine group

	Site 1:	North Atlantic	Site 2:	Southwest	Site 3:	Midwest	Site 4:	Continental	Site 5:	Pacific
	Pfizer ($n = 69,903$)	Moderna ($n = 73,173$)	Pfizer ($n = 60,492$)	Moderna ($n = 68,300$)	Pfizer ($n = 57,853$)	Moderna ($n = 65,375$)	Pfizer ($n = 47,391$)	Moderna ($n = 46,431$)	Pfizer ($n = 57,498$)	Moderna ($n = 61,943$)
Age (years)										
18-49	6,920 (9.9%)	5,344 (7.3%)	5,381 (8.9%)	4,683 (6.9%)	5,082 (8.8%)	4,671 (7.1%)	5,449 (11.5%)	4,358 (9.4%)	7,070 (12.3%)	5,866 (9.5%)
50-59	9,180 (13.1%)	7,682 (10.5%)	8,407 (13.9%)	8,499 (12.4%)	6,131 (10.6%)	7,168 (11.0%)	7,207 (15.2%)	5,939 (12.8%)	6,968 (12.1%)	6,380 (10.3%)
60-69	18,442 (26.4%)	17,267 (23.6%)	16,371 (27.1%)	18,721 (27.4%)	13,716 (23.7%)	16,227 (24.8%)	12,513 (26.4%)	12,157 (26.2%)	13,427 (23.4%)	14,479 (23.4%)
70-79	27,601 (39.5%)	32,164 (44.0%)	23,196 (38.3%)	27,643 (40.5%)	25,967 (44.9%)	28,621 (43.8%)	17,919 (37.8%)	18,311 (39.4%)	22,990 (40.0%)	26,532 (42.8%)
80 or older	7,760 (11.1%)	10,716 (14.6%)	7,137 (11.8%)	8,754 (12.8%)	6,957 (12.0%)	8,688 (13.3%)	4,303 (9.1%)	5,666 (12.2%)	7,043 (12.2%)	8,686 (14.0%)
Sex										
Female	6,379 (9.1%)	5,373 (7.3%)	6,120 (10.1%)	5,701 (8.3%)	4,193 (7.2%)	4,636 (7.1%)	5,155 (10.9%)	4,159 (9.0%)	5,154 (9.0%)	4,743 (7.7%)
Male	63,524 (90.9%)	67,800 (92.7%)	54,372 (89.9%)	62,599 (91.7%)	53,660 (92.8%)	60,739 (92.9%)	42,236 (89.1%)	42,272 (91.0%)	52,344 (91.0%)	57,200 (92.3%)
Race										
Asian	479 (0.7%)	266 (0.4%)	224 (0.4%)	167 (0.2%)	196 (0.3%)	192 (0.3%)	323 (0.7%)	212 (0.5%)	2,270 (3.9%)	2,792 (4.5%)
Black	23,632 (33.8%)	14,514 (19.8%)	16,304 (27.0%)	17,760 (26.0%)	11,511 (19.9%)	9,209 (14.1%)	14,866 (31.4%)	9,316 (20.1%)	8,172 (14.2%)	6,844 (11.0%)
White	42,228 (60.4%)	54,662 (74.7%)	40,040 (66.2%)	46,364 (67.9%)	42,516 (73.5%)	52,253 (79.9%)	28,221 (59.5%)	33,250 (71.6%)	39,163 (68.1%)	43,587 (70.4%)
Other	3,564 (5.1%)	3,731 (5.1%)	3,924 (6.5%)	4,009 (5.9%)	3,630 (6.3%)	3,721 (5.7%)	3,981 (8.4%)	3,653 (7.9%)	7,893 (13.7%)	8,720 (14.1%)
Ethnicity										
Hispanic	2,929 (4.2%)	2,933 (4.0%)	5,951 (9.8%)	10,817 (15.8%)	1,531 (2.6%)	1,130 (1.7%)	5,062 (10.7%)	4,065 (8.8%)	6,615 (11.5%)	7,323 (11.8%)
Not Hispanic	66,974 (95.8%)	70,240 (96.0%)	54,541 (90.2%)	57,483 (84.2%)	56,322 (97.4%)	64,245 (98.3%)	42,329 (89.3%)	42,366 (91.2%)	50,883 (88.5%)	54,620 (88.2%)
Urbanicity										
Rural	11,546 (16.5%)	19,670 (26.9%)	11,701 (19.3%)	13,522 (19.8%)	12,442 (21.5%)	24,109 (36.9%)	8,598 (18.1%)	13,334 (28.7%)	8,538 (14.8%)	11,595 (18.7%)
Urban	58,357 (83.5%)	53,503 (73.1%)	48,791 (80.7%)	54,778 (80.2%)	45,411 (78.5%)	41,266 (63.1%)	38,793 (81.9%)	33,097 (71.3%)	48,960 (85.2%)	50,348 (81.3%)
Comorbidities										
CLD	19,423 (27.8%)	23,763 (32.5%)	18,356 (30.3%)	20,911 (30.6%)	18,253 (31.6%)	23,659 (36.2%)	13,031 (27.5%)	14,093 (30.4%)	14,598 (25.4%)	16,182 (26.1%)
CVD	18,573 (26.6%)	21,992 (30.1%)	16,902 (27.9%)	19,265 (28.2%)	17,335 (30.0%)	21,177 (32.4%)	12,546 (26.5%)	12,551 (27.0%)	13,742 (23.9%)	15,257 (24.6%)
Hypertension	49,985 (71.5%)	54,790 (74.9%)	45,094 (74.5%)	52,490 (76.9%)	42,622 (73.7%)	49,733 (76.1%)	34,362 (72.5%)	34,092 (73.4%)	37,453 (65.1%)	42,533 (68.7%)
T2D	26,872 (38.4%)	29,769 (40.7%)	23,884 (39.5%)	28,472 (41.7%)	22,770 (39.4%)	26,890 (41.1%)	19,549 (41.3%)	19,036 (41.0%)	19,841 (34.5%)	22,329 (36.0%)
CKD	12,241 (17.5%)	13,390 (18.3%)	11,287 (18.7%)	12,742 (18.7%)	11,197 (19.4%)	14,064 (21.5%)	8,665 (18.3%)	8,731 (18.8%)	9,542 (16.6%)	10,627 (17.2%)
Autoimmune	22,431 (32.1%)	26,704 (36.5%)	21,898 (36.2%)	24,415 (35.7%)	21,260 (36.7%)	24,692 (37.8%)	14,912 (31.5%)	15,480 (33.3%)	18,228 (31.7%)	20,642 (33.3%)
Obesity	18,799 (26.9%)	20,827 (28.5%)	18,406 (30.4%)	19,032 (27.9%)	16,731 (28.9%)	19,734 (30.2%)	13,168 (27.8%)	13,358 (28.8%)	15,190 (26.4%)	16,140 (26.1%)

C

Supplemental Materials for Robust and Optimal Sensitivity Analysis (ROSA) of Clinical Trial Designs

Overview of Supplementary Materials

The supplementary material includes a table of notation used in the paper.

C.1 Notation

Table C.1: Notation used in ROSA

Notation		Description
Θ	$\triangleq$	Unknown parameter space in $\mathbb{R}^d$
Θ'	$\triangleq$	Restricted unknown parameter subspace by prior knowledge in $\mathbb{R}^d$
Θ'_{re}	$\triangleq$	Restricted unknown parameter subspace by prior knowledge and fixing certain dimensions in $\mathbb{R}^d$
Θ^F	$\triangleq$	Diffuse and finite unknown parameter subspace in $\mathbb{R}^d$
$\theta = (\theta_1, ..., \theta_d)$	$\triangleq$	d-dimensional vector of unknown parameters
$\theta^t = (\theta^t_1, ..., \theta^t_d)$	$\triangleq$	d-dimensional training vector of unknown parameters
$\theta^v = (\theta^v_1, ..., \theta^v_d)$	$\triangleq$	d-dimensional validation vector of unknown parameters
$\{\theta_1, ..., \theta_K\}$	$\triangleq$	A set of K sensitivity scenarios
$\mathcal{S} = \{\theta^*_1, ..., \theta^*_K\}$	$\triangleq$	The ROSA set of K sensitivity scenarios optimizing loss $\mathcal{L}$
$\mathcal{S}_r = \{\theta^*_{1,r}, ..., \theta^*_{K,r}\}$	$\triangleq$	The ROSA set of K sensitivity scenarios optimizing marginal loss $\mathcal{L}_r$
$\mathbf{f}(\theta)$	$\triangleq$	R-vector of operating characteristics for unknown parameters θ
$\hat{\mathbf{f}}(\theta)$	$\triangleq$	Estimated R-vector of operating characteristics for unknown parameters θ
$\bar{\mathbf{f}}(\theta)$	$\triangleq$	Average across M simulations of the R-vector of operating characteristics for unknown parameters θ
$\phi(Z_{j,m}, \theta_j)$	$\triangleq$	Generic function to capture if a null hypothesis has been rejected, where $Z_{j,m}$ is the m^{th} trial under the j^{th} scenario, θ_j
$\mathcal{L}(\theta_1, ..., \theta_K)$	$\triangleq$	Loss function
$\mathcal{U}(\theta_1, ..., \theta_K)$	$\triangleq$	Utility criterion
$w_1, ..., w_r$	$\triangleq$	Fixed non-negative weights for operating characteristics $f_1, ..., f_R$
ω_1, ω_2	$\triangleq$	Weights for stage 1 and 2 p-values
$D[\cdot, \cdot]$	$\triangleq$	Pre-specified distance metric
$z^i_1, ..., z^i_K$	$\triangleq$	Gaussian noise in iteration i of simulated annealing
ρ_i	$\triangleq$	Acceptance probability in iteration i of simulated annealing
$T_0, T_1, ..., T_I$	$\triangleq$	Decreasing sequence of positive numbers (cooling schedule of simulated annealing)
r	$\triangleq$	Multiplicative reduction factor for simulated annealing in $(0, 1)$
U_i	$\triangleq$	Random variable distributed Uniform(0,1) for simulated annealing
e	$\triangleq$	Enrollment rate in $(0, \infty)$
N_a	$\triangleq$	Planned number of patients on arm $a = 0, 1$ at the final analysis
n_a	$\triangleq$	Planned number of patients on arm $a = 0, 1$ at the interim analysis
S	$\triangleq$	Binary auxiliary outcome
Y	$\triangleq$	Primary outcome
p_a	$\triangleq$	Response probability $P(Y = 1 \mid A = a)$
$\Delta = p_1 - p_0$	$\triangleq$	Treatment effect on Y
q_a	$\triangleq$	Response probability $P(S = 1 \mid A = a)$
ρ_a	$\triangleq$	Correlation between Y and S in $A = a$

References

[1] Agniel, D., Parast, L., & Hejblum, B. (2020). Doubly-robust evaluation of high-dimensional surrogate markers. *arXiv preprint arXiv:2012.01236*.

[2] Ananthakrishnan, A. N., Cagan, A., Cai, T., Gainer, V. S., Shaw, S. Y., Savova, G., Churchill, S., Karlson, E. W., Kohane, I., Liao, K. P., et al. (2016). Comparative effectiveness of infliximab and adalimumab in crohn's disease and ulcerative colitis. *Inflammatory bowel diseases*, 22(4), 880–885.

[3] Ananthakrishnan, A. N., Cai, T., Savova, G., Cheng, S.-C., Chen, P., Perez, R. G., Gainer, V. S., Murphy, S. N., Szolovits, P., Xia, Z., et al. (2013). Improving case definition of crohn's disease and ulcerative colitis in electronic medical records using natural language processing: a novel informatics approach. *Inflammatory bowel diseases*, 19(7), 1411–1420.

[4] Andrews, I. & Oster, E. (2017). Weighting for external validity.

[5] Athey, S., Chetty, R., & Imbens, G. (2020). Combining experimental and observational data to estimate treatment effects on long term outcomes. *arXiv preprint arXiv:2006.09676*.

[6] Athey, S., Chetty, R., Imbens, G. W., & Kang, H. (2019). *The surrogate index: Combining short-term proxies to estimate long-term treatment effects more rapidly and precisely*. Technical report, National Bureau of Economic Research.

[7] Bang, H. & Robins, J. M. (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61(4), 962–973.

[8] Battocchi, K., Dillon, E., Hei, M., Lewis, G., Oprescu, M., & Syrgkanis, V. (2021). Estimating the long-term effects of novel treatments. *arXiv preprint arXiv:2103.08390*.

[9] Beaver, J. A., Howie, L. J., Pelosof, L., Kim, T., Liu, J., Goldberg, K. B., Sridhara, R., Blumenthal, G. M., Farrell, A. T., Keegan, P., et al. (2018). A 25-year experience of us food and drug administration accelerated approval of malignant hematology and oncology drugs and biologics: a review. *JAMA oncology*, 4(6), 849–856.

[10] Bélisle, C. J. (1992). Convergence theorems for a class of simulated annealing algorithms on rd. *Journal of Applied Probability*, (pp. 885–895).

[11] Bentley, C., Cressman, S., van der Hoek, K., Arts, K., Dancey, J., & Peacock, S. (2019). Conducting clinical trials—costs, impacts, and the value of clinical trials networks: A scoping review. *Clinical Trials*, 16(2), 183–193.

[12] Berry, S. M., Carlin, B. P., Lee, J. J., & Muller, P. (2010). *Bayesian adaptive methods for clinical trials*. CRC press.

[13] Bonamici, S. (2016). 34—21st century cures act. In *114th Congress (2015–2016)* (pp. 114–255).

[14] Bookstein, F. L. (1989). Principal warps: Thin-plate splines and the decomposition of deformations. *IEEE Transactions on pattern analysis and machine intelligence*, 11(6), 567–585.

[15] Brat, G. A., Weber, G. M., Gehlenborg, N., Avillach, P., Palmer, N. P., Chiovato, L., Cimino, J., Waitman, L. R., Omenn, G. S., Malovini, A., et al. (2020). International electronic health record-derived covid-19 clinical course profiles: the 4ce consortium. *medRxiv*.

[16] Bruce, C. S., Brhlikova, P., Heath, J., & McGettigan, P. (2019). The use of validated and non-validated surrogate endpoints in two european medicines agency expedited approval pathways: a cross-sectional study of products authorised 2011–2018. *PLoS medicine*, 16(9).

[17] Carnell, R. & Carnell, M. R. (2016). Package 'lhs'. *CRAN. https://cran. rproject. org/web/packages/lhs/lhs. pdf*.

[18] Chen, X. & Xie, M.-g. (2014). A split-and-conquer approach for analysis of extraordinarily large data. *Statistica Sinica*, (pp. 1655–1684).

[19] Chen, Y., Dong, G., Han, J., Pei, J., Wah, B. W., & Wang, J. (2006). Regression cubes with lossless compression and aggregation. *IEEE Transactions on Knowledge and Data Engineering*, 18(12), 1585–1599.

[20] Cheng, D. & Cai, T. (2021). Adaptive combination of randomized and observational data. *arXiv preprint arXiv:2111.15012*.

[21] Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., & Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters. 21, C1–C68.

[22] Ciani, O., Buyse, M., Drummond, M., Rasi, G., Saad, E. D., & Taylor, R. S. (2017). Time to review the role of surrogate end points in health policy: state of the art and the way forward. *Value in Health*, 20(3), 487–495.

[23] Colombel, J. F., Rutgeerts, P., Reinisch, W., Esser, D., Wang, Y., Lang, Y., Marano, C. W., Strauss, R., Oddens, B. J., Feagan, B. G., et al. (2011). Early mucosal healing with infliximab is associated with improved long-term clinical outcomes in ulcerative colitis. *Gastroenterology*, 141(4), 1194–1201.

[24] Concato, J., Shah, N., & Horwitz, R. I. (2000). Randomized, controlled trials, observational studies, and the hierarchy of research designs. *New England journal of medicine*, 342(25), 1887–1892.

[25] Corrigan-Curay, J., Sacks, L., & Woodcock, J. (2018). Real-world evidence and real-world data for evaluating drug safety and effectiveness. *Jama*, 320(9), 867–868.

[26] Dahabreh, I. J., Robertson, S. E., Steingrimsson, J. A., Stuart, E. A., & Hernan, M. A. (2020). Extending inferences from a randomized trial to a new target population. *Statistics in medicine*, 39(14), 1999–2014.

[27] Degtiar, I. & Rose, S. (2021). A review of generalizability and transportability. *arXiv preprint arXiv:2102.11904*.

[28] Dickerman, B. A., Gerlovin, H., Madenci, A. L., Kurgansky, K. E., Ferolito, B. R., Figueroa Muñiz, M. J., Gagnon, D. R., Gaziano, J. M., Cho, K., Casas, J. P., & Hernán, M. A. (2021). Comparative effectiveness of bnt162b2 and mrna-1273 vaccines in u.s. veterans. *New England Journal of Medicine*.

[29] Dong, L., Yang, S., Wang, X., Zeng, D., & Cai, J. (2020). Integrative analysis of randomized clinical trials with real world evidence studies. *arXiv preprint arXiv:2003.01242*.

[30] Duan, R., Boland, M. R., Liu, Z., Liu, Y., Chang, H. H., Xu, H., Chu, H., Schmid, C. H., Forrest, C. B., Holmes, J. H., Schuemie, M. J., Berlin, J. A., Moore, J. H., & Chen, Y. (2019). Learning from electronic health records across multiple sites: A communication-efficient and privacy-preserving distributed algorithm. *Journal of the American Medical Informatics Association*, 27(3), 376–385.

[31] Duan, R., Boland, M. R., Moore, J. H., & Chen, Y. (2020a). ODAL: A one-shot distributed algorithm to perform logistic regressions on electronic health records data from multiple clinical sites. *Pacific Symposium on Biocomputing*, (pp. 30–41).

[32] Duan, R., Ning, Y., Wang, S., Lindsay, B., Carroll, R., & Chen, Y. (2020b). A fast score test for generalized mixture models. *Biometrics*, 76, 811–820.

[33] D'Amour, A., Ding, P., Feller, A., Lei, L., & Sekhon, J. (2021). Overlap in observational studies with high-dimensional covariates. *Journal of Econometrics*, 221(2), 644–654.

[34] Estellat, C. & Ravaud, P. (2012). Lack of head-to-head trials and fair control arms: randomized controlled trials of biologic treatment for rheumatoid arthritis. *Archives of internal medicine*, 172(3), 237–244.

[35] Fan, J. & Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456), 1348–1360.

[36] Fleming, T. R. & Powers, J. H. (2012). Biomarkers and surrogate endpoints in clinical trials. *Statistics in medicine*, 31(25), 2973–2984.

[37] Food, Administration, D., et al. (2020). Interacting with the fda on complex innovative trial designs for drugs and biological products. *Updated December*.

[38] Franklin, J. M., Glynn, R. J., Martin, D., & Schneeweiss, S. (2019). Evaluating the use of nonrandomized real-world data analyses for regulatory decision making. *Clinical Pharmacology & Therapeutics*, 105(4), 867–877.

[39] Freedman, L. S. & Schatzkin, A. (1992). Sample size for studying intermediate endpoints within intervention trials or observational studies. *American Journal of Epidemiology*, 136(9), 1148–1159.

[40] Gamerman, V., Cai, T., & Elsäßer, A. (2019). Pragmatic randomized clinical trials: best practices and statistical guidance. *Health Services and Outcomes Research Methodology*, 19, 23–35.

[41] Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT press.

[42] Green, S., Liu, P.-Y., & O'Sullivan, J. (2002). Factorial design considerations. *Journal of Clinical Oncology*, 20(16), 3424–3430.

[43] Gyawali, B., Hey, S. P., & Kesselheim, A. S. (2020). Evaluating the evidence behind the surrogate measures included in the fda's table of surrogate endpoints as supporting approval of cancer drugs. *EClinicalMedicine*, (pp. 100332).

[44] Hall, P., Racine, J., & Li, Q. (2004). Cross-validation and the estimation of conditional probability densities. *Journal of the American Statistical Association*, 99(468), 1015–1026.

[45] Han, A. K. (1987). Non-parametric analysis of a generalized regression model: the maximum rank correlation estimator. *Journal of Econometrics*, 35(2-3), 303–316.

[46] Han, K., Ren, M., Wick, W., Abrey, L., Das, A., Jin, J., & Reardon, D. A. (2014). Progression-free survival as a surrogate endpoint for overall survival in glioblastoma: a literature-based meta-analysis from 91 trials. *Neuro-oncology*, 16(5), 696–706.

[47] Hastie, T. & Tibshirani, R. (1993). Varying-coefficient models. *Journal of the Royal Statistical Society: Series B (Methodological)*, 55(4), 757–779.

[48] Hernán, M. A. & Robins, J. M. (2016). Using big data to emulate a target trial when a randomized trial is not available. *American journal of epidemiology*, 183(8), 758–764.

[49] Hernán, M. A. & Robins, J. M. (2020). Causal inference: What if?

[50] Hirshberg, D. A., Maleki, A., & Zubizarreta, J. R. (2019). Minimax linear estimation of the retargeted mean. *arXiv preprint arXiv:1901.10296*.

[51] Hirshberg, D. A. & Wager, S. (2021). Augmented minimax linear estimation. *The Annals of Statistics*, 49(6), 3206–3227.

[52] Hobbs, B. P., Barata, P. C., Kanjanapan, Y., Paller, C. J., Perlmutter, J., Pond, G. R., Prowell, T. M., Rubin, E. H., Seymour, L. K., Wages, N. A., et al. (2019). Seamless designs: current practice and considerations for early-phase drug development in oncology. *JNCI: Journal of the National Cancer Institute*, 111(2), 118–128.

[53] Hoffman, K. M., Trawalter, S., Axt, J. R., & Oliver, M. N. (2016). Racial bias in pain assessment and treatment recommendations, and false beliefs about biological differences between blacks and whites. *Proceedings of the National Academy of Sciences*, 113(16), 4296–4301.

[54] Hong, H., Leung, M. P., & Li, J. (2020). Inference on finite-population treatment effects under limited overlap. *The Econometrics Journal*, 23(1), 32–47.

[55] Hornik, K. (1991). Approximation capabilities of multilayer feedforward networks. *Neural networks*, 4(2), 251–257.

[56] Hripcsak, G., Ryan, P. B., Duke, J. D., Shah, N. H., Park, R. W., Huser, V., Suchard, M. A., Schuemie, M. J., DeFalco, F. J., Perotte, A., et al. (2016). Characterizing treatment pathways at scale using the ohdsi network. *Proceedings of the National Academy of Sciences*, 113(27), 7329–7336.

[57] Huang, C. & Huo, X. (2019). A distributed one-step estimator. *Mathematical Programming*, 174(1), 41–76.

[58] Husmann, K., Lange, A., & Spiegel, E. (2017). The r package optimization: Flexible global optimization with simulated-annealing.

[59] Iasonos, A., Gönen, M., & Bosl, G. J. (2015). Scientific review of phase i protocols with novel dose-escalation designs: how much information is needed? *Journal of Clinical Oncology*, 33(19), 2221.

[60] Imbens, G. W. & Rubin, D. B. (2015). *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press.

[61] Jarow, J. P., LaVange, L., & Woodcock, J. (2017). Multidimensional evidence generation and fda regulatory decision making: defining and using "real-world" data. *Jama*, 318(8), 703–704.

[62] Jenkins, M., Stone, A., & Jennison, C. (2011). An adaptive seamless phase ii/iii design for oncology trials with subpopulation selection using correlated survival endpoints. *Pharmaceutical statistics*, 10(4), 347–356.

[63] Jin, Z., Ying, Z., & Wei, L. (2001). A simple resampling method by perturbing the minimand. *Biometrika*, 88(2), 381–390.

[64] Jones, R. L., Attia, S., Mehta, C. R., Liu, L., Sankhala, K. K., Robinson, S. I., Ravi, V., Penel, N., Stacchiotti, S., Tap, W. D., et al. (2017). Tappas: An adaptive enrichment phase 3 trial of trc105 and pazopanib versus pazopanib alone in patients with advanced angiosarcoma (aas). *J. Clin. Oncol.*, 35, TPS11081.

[65] Josey, K. P., Yang, F., Ghosh, D., & Raghavan, S. (2020). A calibration approach to transportability with observational data. *arXiv preprint arXiv:2008.06615*.

[66] Khan, S. & Tamer, E. (2010). Irregular identification, support conditions, and inverse weight estimation. *Econometrica*, 78(6), 2021–2042.

[67] Kirkpatrick, S., Gelatt, C. D., & Vecchi, M. P. (1983). Optimization by simulated annealing. *Science*, 220(4598), 671–680.

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

[69] Leshno, M., Lin, V. Y., Pinkus, A., & Schocken, S. (1993). Multilayer feedforward networks with a nonpolynomial activation function can approximate any function. *Neural networks*, 6(6), 861–867.

[70] Lewis, J. D., Chuai, S., Nessel, L., Lichtenstein, G. R., Aberra, F. N., & Ellenberg, J. H. (2008). Use of the noninvasive components of the mayo score to assess clinical response in ulcerative colitis. *Inflammatory bowel diseases*, 14(12), 1660–1666.

[71] Li, R., Lin, D. K., & Li, B. (2013). Statistical inference in massive data sets. *Applied Stochastic Models in Business and Industry*, 29(5), 399–409.

[72] Lian, H. & Fan, Z. (2017). Divide-and-conquer for debiased l 1-norm support vector machine in ultra-high dimensions. *The Journal of Machine Learning Research*, 18(1), 6691–6716.

[73] Lin, D., Fleming, T., & De Gruttola, V. (1997). Estimating the proportion of treatment effect explained by a surrogate marker. *Statistics in medicine*, 16(13), 1515–1527.

[74] Lin, D.-Y., Gu, Y., Wheeler, B., Young, H., Holloway, S., Sunny, S.-K., Moore, Z., & Zeng, D. (2022). Effectiveness of covid-19 vaccines over a 9-month period in north carolina. *New England Journal of Medicine*.

[75] Masry, E. (1996). Multivariate local polynomial regression for time series: uniform strong consistency and rates. *Journal of Time Series Analysis*, 17(6), 571–599.

[76] Mayer, C., Perevozskaya, I., Leonov, S., Dragalin, V., Pritchett, Y., Bedding, A., Hartford, A., Fardipour, P., & Cicconetti, G. (2019). Simulation practices for adaptive trial designs in drug and device development. *Statistics in Biopharmaceutical Research*, 11(4), 325–335.

[77] McKay, M. D., Beckman, R. J., & Conover, W. J. (2000). A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. *Technometrics*, 42(1), 55–61.

[78] Mehta, C., Liu, L., & Theuer, C. (2019). An adaptive population enrichment phase iii trial of trc105 and pazopanib versus pazopanib alone in patients with advanced angiosarcoma (tappas trial). *Annals of Oncology*, 30(1), 103–108.

[79] Michael, J. & Schucany, W. (2002). The mixture approach for simulating new families of bivariate distributions with specified correlations. *The American Statistician*, 56(1), 48–54.

[80] Mitka, M. (2010). Us government kicks off program for comparative effectiveness research. *JAMA*, 304(20), 2230–2231.

[81] Neyman, J. (1923). On the application of probability theory to agricultural experiments. *Statistical Science*, 5(5), 463–480.

[82] Nguyen, T. Q., Ebnesajjad, C., Cole, S. R., & Stuart, E. A. (2017). Sensitivity analysis for an unobserved moderator in rct-to-target-population generalization of treatment effects. *The Annals of Applied Statistics*, (pp. 225–247).

[83] Nie, X., Imbens, G., & Wager, S. (2021). Covariate balancing sensitivity analysis for extrapolating randomized trials across locations. *arXiv preprint arXiv:2112.04723*.

[84] Niewczas, J., Kunz, C. U., & König, F. (2019). Interim analysis incorporating short-and long-term binary endpoints. *Biometrical Journal*, 61(3), 665–687.

[85] Pagan, A. & Ullah, A. (1999). *Nonparametric econometrics*. Cambridge university press.

[86] Pallmann, P., Bedding, A. W., Choodari-Oskooei, B., Dimairo, M., Flight, L., Hampson, L. V., Holmes, J., Mander, A. P., Odondi, L., Sydes, M. R., et al. (2018). Adaptive designs in clinical trials: why use them, and how to run and report them. *BMC medicine*, 16(1), 1–15.

[87] Pan, S. J. & Yang, Q. (2009). A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10), 1345–1359.

[88] Parast, L., McDermott, M. M., & Tian, L. (2016). Robust estimation of the proportion of treatment effect explained by surrogate marker information. *Statistics in Medicine*, 35(10), 1637–1653.

[89] Pathak, J., Kho, A. N., & Denny, J. C. (2013). Electronic health records-driven phenotyping: challenges, recent advances, and perspectives.

[90] Prentice, R. L. (1989). Surrogate endpoints in clinical trials: definition and operational criteria. *Statistics in Medicine*, 8(4), 431–440.

[91] Price, B. L., Gilbert, P. B., & van der Laan, M. J. (2018). Estimation of the optimal surrogate based on a randomized trial. *Biometrics*.

[92] Qin, J. (1998). Inferences for case-control and semiparametric two-sample density ratio models. *Biometrika*, 85(3), 619–630.

[93] Qin, J. & Liang, K.-Y. (2011). Hypothesis testing in a mixture case-control model. *Biometrics*, 67, 182–193.

[94] Rasmussen, C. E. (2003). Gaussian processes in machine learning. In *Summer school on machine learning* (pp. 63–71).: Springer.

[95] Ray, W. A., Stein, C. M., Daugherty, J. R., Hall, K., Arbogast, P. G., & Griffin, M. R. (2002). Cox-2 selective non-steroidal anti-inflammatory drugs and risk of serious coronary heart disease. *The Lancet*, 360(9339), 1071–1073.

[96] Razavi, S., Jakeman, A., Saltelli, A., Prieur, C., Iooss, B., Borgonovo, E., Plischke, E., Piano, S. L., Iwanaga, T., Becker, W., et al. (2021). The future of sensitivity analysis: an essential discipline for systems modeling and policy support. *Environmental Modelling & Software*, 137, 104954.

[97] Robins, J. M. & Richardson, T. S. (2010). Alternative graphical causal models and the identification of direct effects. *Causality and psychopathology: Finding the determinants of disorders and their cures*, (pp. 103–158).

[98] Robins, J. M., Rotnitzky, A., & Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American statistical Association*, 89(427), 846–866.

[99] Ross, J. S., Waldstreicher, J., Bamford, S., Berlin, J. A., Childers, K., Desai, N. R., Gamble, G., Gross, C. P., Kuntz, R., Lehman, R., et al. (2018). Overview and experience of the yoda project with clinical trial data sharing after 5 years. *Scientific data*, 5(1), 1–14.

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

[101] Rubin, D. B. (2005). Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association*, 100(469), 322–331.

[102] Rubin, D. B. (2009). Should observational studies be designed to allow lack of balance in covariate distributions across treatment groups? *Statistics in Medicine*, 28(9), 1420–1423.

[103] Rubin, D. T., Ananthakrishnan, A. N., Siegel, C. A., Sauer, B. G., & Long, M. D. (2019). Acg clinical guideline: ulcerative colitis in adults. *American Journal of Gastroenterology*, 114(3), 384–413.

[104] Sands, B. E., Peyrin-Biroulet, L., Loftus Jr, E. V., Danese, S., Colombel, J.-F., Törüner, M., Jonaitis, L., Abhyankar, B., Chen, J., Rogers, R., et al. (2019). Vedolizumab versus adalimumab for moderate-to-severe ulcerative colitis. *New England Journal of Medicine*, 381(13), 1215–1226.

[105] Schneeweiss, S. (2019). Real-world evidence of treatment effects: the useful and the misleading. *Clin Pharmacol Ther*, 106(1), 43–44.

[106] Scott, D. (1992). *Multivariate density estimation*. John Wiley & Sons.

[107] Seeger, J. D., Bykov, K., Bartels, D. B., Huybrechts, K., Zint, K., & Schneeweiss, S. (2015). Safety and effectiveness of dabigatran and warfarin in routine care of patients with atrial fibrillation. *Thrombosis and haemostasis*, 114(12), 1277–1289.

[108] Sherman, R. E., Anderson, S. A., Dal Pan, G. J., Gray, G. W., Gross, T., Hunter, N. L., LaVange, L., Marinac-Dabic, D., Marks, P. W., Robb, M. A., et al. (2016). Real-world evidence—what is it and what can it tell us. *N Engl J Med*, 375(23), 2293–2297.

[109] Sherman, R. P. (1993). The limiting distribution of the maximum rank correlation estimator. *Econometrica: Journal of the Econometric Society*, (pp. 123–137).

[110] Spall, J. C. (2005). *Introduction to stochastic search and optimization: estimation, simulation, and control*. John Wiley & Sons.

[111] Stuart, E. A., Ackerman, B., & Westreich, D. (2018). Generalizability of randomized trial results to target populations: design and analysis possibilities. *Research on social work practice*, 28(5), 532–537.

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

[113] Stuart, E. A., Cole, S. R., Bradshaw, C. P., & Leaf, P. J. (2011). The use of propensity scores to assess the generalizability of results from randomized trials. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 174(2), 369–386.

[114] Tan, Z. et al. (2020). Model-assisted inference for treatment effects using regularized calibrated estimation with high-dimensional data. *Annals of Statistics*, 48(2), 811–837.

[115] Taylor, P. C., Keystone, E. C., van der Heijde, D., Weinblatt, M. E., del Carmen Morales, L., Reyes Gonzaga, J., Yakushin, S., Ishii, T., Emoto, K., Beattie, S., et al. (2017). Baricitinib versus placebo or adalimumab in rheumatoid arthritis. *New England Journal of Medicine*, 376(7), 652–662.

[116] Thorlund, K., Haggstrom, J., Park, J. J., & Mills, E. J. (2018). Key design considerations for adaptive clinical trials: a primer for clinicians. *BMJ*, 360, k698.

[117] Tian, L., Zucker, D., & Wei, L. (2005). On the cox model with time-varying regression coefficients. *Journal of the American statistical Association*, 100(469), 172–183.

[118] Ungaro, R. & Colombel, J.-F. (2017). biologics in inflammatory bowel disease—time for direct comparisons. *Alimentary pharmacology & therapeutics*, 46(1), 68.

[119] Van der Vaart, A. W. (2000). *Asymptotic statistics*, volume 3. Cambridge university press.

[120] VanderWeele, T. J. (2013). Surrogate measures and consistent surrogates. *Biometrics*, 69(3), 561–565.

[121] Vo, T. V., Hoang, T. N., Lee, Y., & Leong, T.-Y. (2021). Federated estimation of causal effects from observational data. *arXiv preprint arXiv:2106.00456*.

[122] Wang, S., McCormick, T. H., & Leek, J. T. (2020a). Methods for correcting inference based on outcomes predicted by machine learning. *Proceedings of the National Academy of Sciences*, 117(48), 30266–30275.

[123] Wang, X., Parast, L., Han, L., Tian, L., & Cai, T. (2022). Robust approach to combining multiple markers to improve surrogacy. *Biometrics*.

[124] Wang, X., Parast, L., Tian, L., & Cai, T. (2020b). Model-free approach to quantifying the proportion of treatment effect explained by a surrogate marker. *Biometrika*, 107(1), 107–122.

[125] Wang, X., Yang, Z., Chen, X., & Liu, W. (2019). Distributed inference for linear support vector machine. *Journal of Machine Learning Research*, 20(113), 1–41.

[126] Wang, Y. & Taylor, J. M. (2002). A measure of the proportion of treatment effect explained by a surrogate marker. *Biometrics*, 58(4), 803–812.

[127] Weiss, K., Khoshgoftaar, T. M., & Wang, D. (2016). A survey of transfer learning. *Journal of Big data*, 3(1), 1–40.

[128] Wickström, K. & Moseley, J. (2017). Biomarkers and surrogate endpoints in drug development: a european regulatory view. *Investigative Ophthalmology & Visual Science*, 58(6), BIO27–BIO33.

[129] Xie, F., Chakraborty, B., Ong, M. E. H., Goldstein, B. A., & Liu, N. (2020). Autoscore: A machine learning–based automatic clinical score generator and its application to mortality prediction using electronic health records. *JMIR medical informatics*, 8(10), e21798.

[130] Xiong, R., Koenecke, A., Powell, M., Shen, Z., Vogelstein, J. T., & Athey, S. (2021). Federated causal inference in heterogeneous observational data. *arXiv preprint arXiv:2107.11732*.